# A foundation model for protein-ligand affinity prediction through jointly optimizing virtual screening and hit-to-lead optimization

Bin Feng[1], Zijing Liu[1*], Mingjun Yang[2], Junjie Zou[2], He Cao[1], Yu Li[1*], Lei Zhang[1], Sheng Wang[3*]

[1] International Digital Economy Academy (IDEA), Shenzhen, China

[2] XtalPi (Shenzhen) Co., Ltd., Shenzhen, China

[3] Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA

[#]Corresponding authors. Emails: liuzijing@idea.edu.cn, liyu@idea.edu.cn, swang@cs.washington.edu

**Abstract**

Protein-ligand binding affinity plays an important role in drug discovery, especially during virtual screening and hit-to-lead optimization. Computational chemistry and machine learning methods have been developed to investigate these tasks. Despite the encouraging performance, virtual screening and hit-to-lead optimization are often studied separately by existing methods, partially because they are performed sequentially in the existing drug discovery pipeline, thereby overlooking their interdependency and complementarity. To address this problem, we propose LigUnity, a foundation model for protein-ligand binding prediction by jointly optimizing virtual screening and hit-to-lead optimization. In particular, LigUnity learns coarse-grained active/inactive distinction for virtual screening, and fine-grained pocket-specific ligand preference for hit-to-lead optimization. We demonstrate the effectiveness and versatility of LigUnity on eight benchmarks across both tasks. In virtual screening, LigUnity outperforms 24 competing methods with more than 50% improvement on the DUD-E and Dekois 2.0 benchmarks, and shows robust generalization to novel proteins. In hit-to-lead optimization, LigUnity achieves the best performance on split-by-time, split-by-scaffold, and split-by-unit settings, further demonstrating its potential as a cost-effective alternative to free energy perturbation (FEP) calculations. We further showcase how LigUnity can be employed in an active learning framework to efficiently identify active ligands for TYK2, a therapeutic target for autoimmune diseases, yielding over 40% improved prediction performance. Collectively, these comprehensive results establish LigUnity as a versatile foundation model for both virtual screening and hit-to-lead optimization, offering broad applicability across the drug discovery pipeline through accurate protein-ligand affinity predictions.

# Main

Breakthroughs in protein structure prediction, represented by AlphaFold[1] and RoseTTAFold,[2] have provided us with highly accurate three-dimensional (3D) structures of almost every protein. More recently, AlphaFold3 has demonstrated remarkable advancement in predicting protein-ligand complex structures, substantially accelerating the drug discovery pipeline by providing novel insights into the interactions between proteins and drug molecules.[3] The next challenge in the drug discovery pipeline after deciphering the structures is to predict protein-ligand binding affinity. In particular, protein-ligand binding affinity reflects how strongly a potential drug molecule interacts with its target protein, serving as a key determinant of drug efficacy and target selectivity.[4]

Protein-ligand binding affinity has played a key role in two sequential steps in the early stage drug discovery: high-throughput virtual screening, which aims to discover active compounds that can bind to an interested human protein from large-scale chemical libraries,[5–7] and hit-to-lead optimization, which aims to refine the structure of these active ligands to improve their binding affinity and drug-like properties.[8,9] For virtual screening, docking methods have been widely utilized with several successful applications.[10] Still, they are either poorly correlated with experimental binding affinities or require large computational resources to take protein flexibility into consideration.[11] As for hit-to-lead optimization, physics-based calculation methods developed on the foundation of molecular dynamics (MD), including free energy perturbation (FEP)[12] and molecular mechanics-generalized born surface area (MM/GBSA),[13] have shown encouraging performance, but suffer from the trade-off between the large chemical search space and computational cost.[14,15] To address the challenges faced by computational methods, in recent years, machine learning (ML) methods have been developed for both virtual screening (e.g., RTMScore,[16] DrugCLIP[17]) and hit-to-lead optimization (e.g., ActFound,[18] PBCNet[19]). These ML approaches achieve comparable performance to traditional computational methods while offering significant improvements in computational efficiency, thereby making them highly efficient alternatives for large-scale applications.

Despite the encouraging performance of existing ML methods, virtual screening and hit-to-lead optimization are often studied separately by these methods, partially because they are performed sequentially in the drug discovery pipeline.[20] However, these two tasks are interdependent and complementary to each other. A model that only focuses on hit-to-lead optimization might be limited within a restricted chemical space, making it ineffective to generalize to ligands with unseen chemical scaffolds. On the other hand, a model only developed for virtual screening might overlook crucial substructures that determine the protein-ligand interaction, thus falling short of distinguishing ligands with subtle structure differences. Based on these observations, we hypothesize that a unified foundation model that jointly studies virtual screening and hit-to-lead optimization can leverage the synergy between these two tasks to enhance performance. Specifically, virtual screening data provides diverse protein-ligand interactions that help establish global structure-activity relationships, while hit-to-lead optimization data offers detailed features into how subtle structural differences affect binding affinity.

To this end, we propose LigUnity (**Lig**and **U**nified affi**nity**), a foundation model for protein-ligand affinity prediction that jointly optimizes both virtual screening and hit-to-lead optimization in a unified manner. The key idea of LigUnity is to learn coarse-grained active/inactive distinction for virtual screening and fine-grained pocket-specific ligand preference for hit-to-lead optimization. In particular, to optimize the affinity for virtual screening, we exploit contrastive learning[17,21] to jointly embed ligand and protein pocket, allowing the model to differentiate between active and inactive

ligands. For hit-to-lead optimization, we employ a listwise ranking[22,23] approach that orders active ligands for each pocket, allowing the model to learn fine-grained ligand-pocket interaction features. Once the embeddings are computed, the designed framework enables rapid screening of large virtual libraries and efficient identification of active ligands within the large chemical search space, making LigUnity a versatile foundation model suited for both virtual screening and hit-to-lead optimization.

To validate the effectiveness of LigUnity as a foundation model for protein-ligand affinity, we conducted comprehensive evaluations on eight benchmarks across six settings. On three virtual screening benchmarks (DUD-E,[24] Dekois,[25] and LIT-PCBA[26]), LigUnity outperforms all 24 competing methods, demonstrating its effectiveness in identifying active ligands. When applied to unseen proteins, our model achieves consistent improvement, demonstrating its robust generalization ability to novel targets. For hit-to-lead optimization, we evaluated LigUnity on two FEP benchmarks[12,27] and observed superior performance in both zero-shot and few-shot settings, revealing that LigUnity can be potentially used as an efficient alternative to the costly FEP calculations.[12] We further assessed LigUnity's adaptability as a foundation model in various settings, including split-by-time, split-by-scaffold, and split-by-unit, on diverse assays from ChEMBL[28] and BindingDB.[29] LigUnity again outperforms eight competing methods. To further illustrate its versatility in the real-world drug discovery pipeline, we integrated LigUnity into an active learning framework to simulate the multi-iteration optimization in drug discovery. LigUnity successfully identifies ligands with optimal binding affinity in just a few iterations, illustrating its practical value in drug discovery. Collectively, LigUnity serves as a foundation model for both virtual screening and hit-to-lead optimization, offering broad applicability across the drug discovery pipeline through the accurate prediction of protein-ligand affinity.

# Results

## Overview of LigUnity

LigUnity is a foundation model for protein-ligand affinity prediction. The key idea of LigUnity is to leverage the interdependency between virtual screening and hit-to-lead optimization to pre-train the pocket and ligand encoders. In particular, for virtual screening, LigUnity employs a contrastive learning loss to effectively capture the coarse-grained distinction between active and inactive compounds. The model learns to bring the embeddings of positive pocket-ligand pairs closer and push the negative pairs apart. For hit-to-lead optimization, LigUnity employs a listwise ranking loss that alleviates the batch effect across assays to learn the fine-grained pocket-specific ligand preference. These two loss functions complement each other, thus being optimized simultaneously in the pre-training stage (**Fig. 1a**).

After pre-training, LigUnity is adapted separately for virtual screening and hit-to-lead optimization to leverage task-specific data. For virtual screening, LigUnity employs a novel graph-based approach to utilize large-scale protein-ligand interactions in existing databases (**Fig. 1b**). Specifically, we first built a large-scale knowledge pocket-ligand graph, containing 0.83 million pocket-ligand edges with known interaction and 16 million pocket-pocket edges based on structural similarity. We then applied a graph neural network to refine the query pocket embedding by aggregating neighboring information. For hit-to-lead optimization, experimentally measured affinities for a small set of interested ligands are often available, and we fine-tuned LigUnity on these measured ligands using a ranking loss (**Fig. 1c**). This enables the model to accurately rank unmeasured ligands for the target protein pocket and find highly active ligands among them.

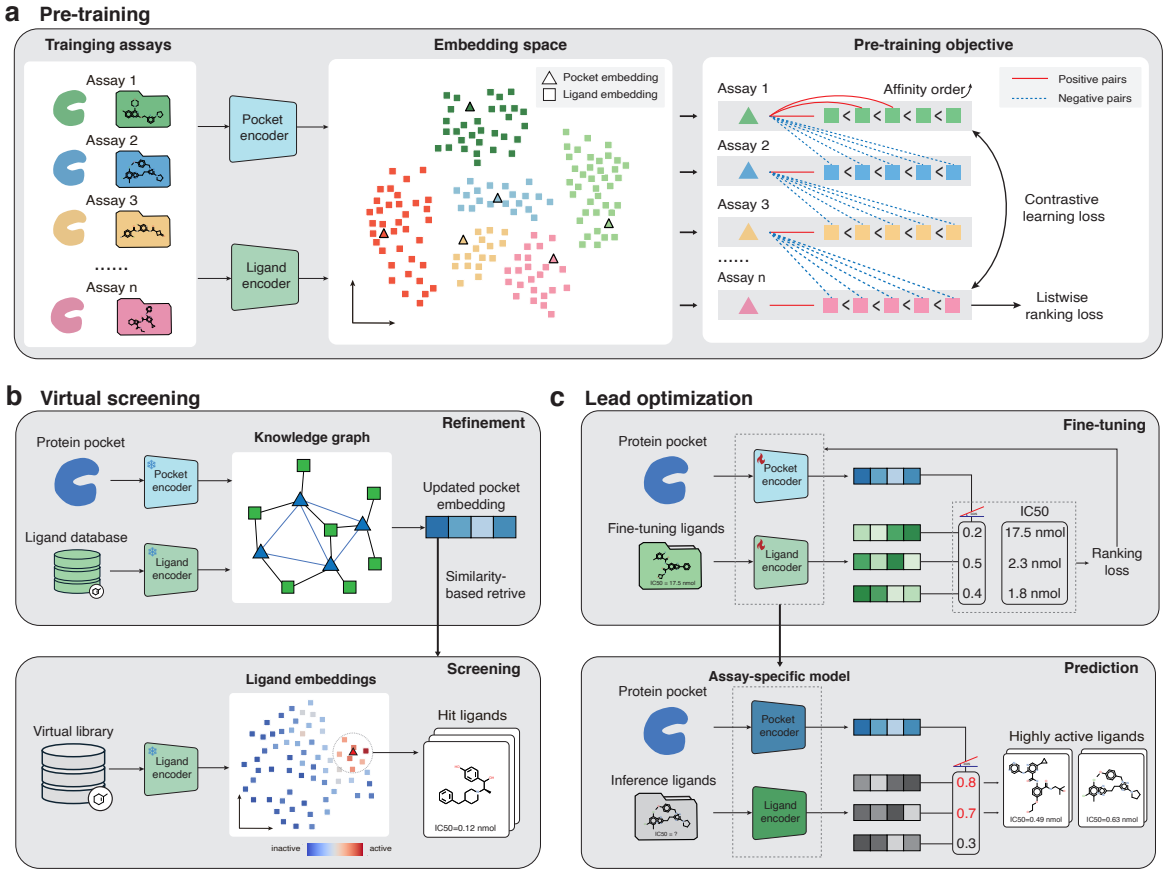



**Figure 1. Overview of LigUnity. a,** LigUnity first pre-trains the pocket and ligand encoders by jointly optimizing virtual screening and hit-to-lead optimization. It first projects pockets and ligands into the same embedding space. It then exploits contrastive learning to capture coarse-grained active/inactive distinction and a listwise ranking loss to learn fine-grained pocket-specific ligand preference. **b,** For virtual screening, LigUnity first refines the query pocket embedding by a graph neural network, and the updated pocket embedding is then utilized to retrieve hit ligands in the candidate ligand embedding space. **c,** For lead optimization, LigUnity is first fine-tuned according to the ranking loss computed on measured ligands. The fine-tuned model is then used to rank unmeasured candidate ligands and find highly active ligands among them.

During inference, LigUnity takes two inputs: 3D protein pockets and RDKit-generated ligand conformations. It processes ligand conformations through the ligand encoders and computes an affinity score based on cosine similarity with the pocket embedding. In addition to improved prediction performance, LigUnity substantially reduces the computational cost during inference, achieving six orders of magnitude speedup compared to traditional docking methods.

## LigUnity improves virtual screening

We first evaluated LigUnity for virtual screening on 102 protein targets from the DUD-E benchmark and 81 protein targets from the DEKOIS 2.0 benchmark.[24,25] These two benchmarks cover a diverse set of targets spanning enzymes, ion channels, GPCRs, and transcription factors (**Supplementary Figures. 1, 2**). We compared LigUnity with 24 different competing methods, including molecular docking methods,[30,31] structure-based methods,[16,32,33] and structure-free methods,[17,34] and observed that LigUnity outperformed all competing methods in both benchmarks (**Fig. 2a,b, Supplementary Figures. 3-6**). Specifically, when compared to the best competing structure-based method Denvis-

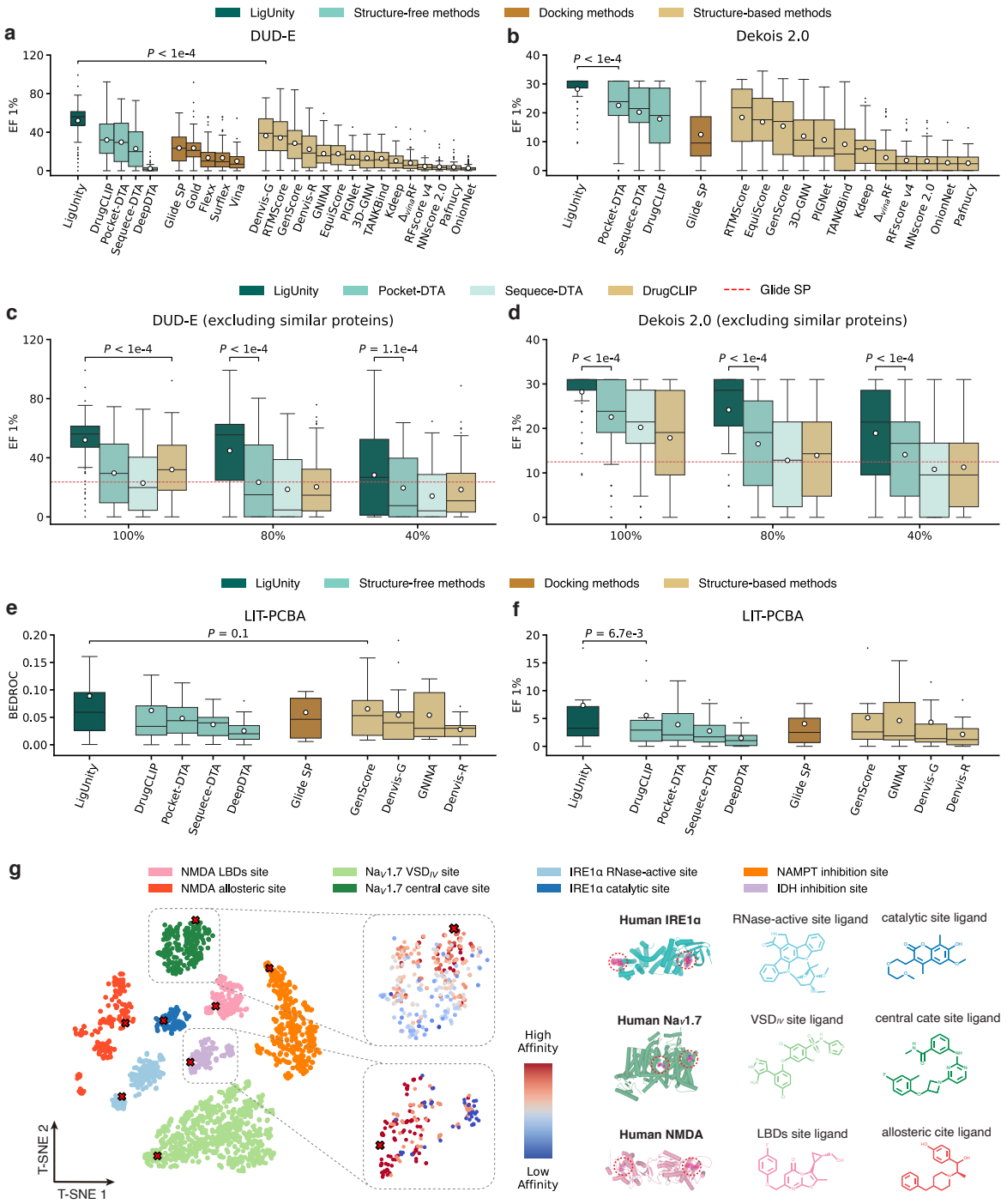



**Figure 2. Evaluation on virtual screening. a,b,** Box plots comparing LigUnity and competing methods in the DUD-E and Dekois-2.0 benchmarks in terms of enrichment factor (EF) 1%. The mean values (white dots) are calculated across n = 102 and 81 targets for the DUD-E and Dekois-2.0 benchmarks, respectively. **c,d,** Box plots comparing LigUnity and structure-free methods in the DUD-E and Dekois-2.0 benchmarks in terms of enrichment factor (EF) 1% using different protein training sets. The $x$-axis denotes the maximum sequence similarity between the training and test sets. The mean values (white dots) are calculated across n = 102 and 81 targets for the DUD-E and Dekois-2.0 benchmarks, respectively. **e,f,** Box plots comparing LigUnity and competing methods on the LIT-PCBA benchmark in terms of enrichment factor (EF) 1% and BEDROC score ($\alpha = 80.5$). The mean values (white dots) are calculated across n = 15 targets. **g,** t-distributed stochastic neighbor embedding (t-SNE) plot showing ligand embeddings (dots) and pocket embeddings (cross marks) derived by LigUnity (left panel); the binding sites in three proteins (circled in red) and their corresponding active ligands (right panel). All ligands and pockets are from the test set.

G[33] and RTMScore,[16] LigUnity still achieves over 50% improvements in the enrichment factor (EF) 1% ($P < 10^{-9}$). Meanwhile, once the embeddings are computed, LigUnity is $10^6$ times faster than RTMScore in the screening speed as LigUnity does not need docking poses. To assess the ability of LigUnity to generalize to unseen proteins, we excluded training proteins that are similar to any test protein in terms of sequence similarity (**Fig. 2c,d, Supplementary Figures. 7, 8**). When using a stringent threshold of 40% sequence similarity, LigUnity still significantly outperforms both DrugCLIP and the commercial docking software Glide(SP) ($P < 0.05$), demonstrating its generalizability to novel proteins.

Next, we evaluated LigUnity on the LIT-PCBA benchmark, which contains experimentally measured affinities for active and inactive ligands across 15 targets. Compared to DUD-E and DEKOIS, LIT-PCBA presents a more challenging setting for virtual screening, as it contains more inactive ligands that are structurally similar to active ligands. In contrast, DUD-E and DEKOIS utilize experimentally unverified decoys from the public compound database (ZINC),[35] while excluding decoys structurally similar to active ligands to avoid potential false negatives, which may make them relatively easy for ML models. Furthermore, LIT-PCBA maintains a 1:1000 active-to-inactive ratio, which better reflects the practical conditions encountered in real-world drug discovery. Nevertheless, LigUnity still achieves superior performance on the more challenging LIT-PCBA benchmark (**Fig. 2e,f**). The improvement is also achieved when excluding training proteins that have large sequence similarity to any test protein using varying thresholds (**Supplementary Figure. 9**), reassuring the effectiveness and generalizability of LigUnity for virtual screening.

Moreover, we conducted ablation studies to study the contribution of each module in LigUnity (**Supplementary Figures. 10-12**). First, we found that ablating the contrastive learning component leads to a performance drop of over 60% in terms of EF 1%, confirming its crucial role in distinguishing active ligands from decoys in virtual screening tasks. Second, the ablation of the heterogeneous GNN leads to consistent decreases in performance, demonstrating its capability to utilize information from the pocket-ligand heterogeneous graph to improve the predictive power. Furthermore, we found that the listwise ranking is particularly effective on LIT-PCBA (**Supplementary Figure. 12**), as it helps LigUnity distinguish subtle structural differences among similar ligands that are crucial to their affinity scores. Together, our ablation studies demonstrate the importance of all three key technical ideas proposed by LigUnity, necessitating the joint optimization of virtual screening and hit-to-lead optimization.

Finally, we examined the embedding space derived by LigUnity to gain deeper insights into the predictive capabilities of LigUnity. We visualized the generated embeddings with the t-distributed stochastic neighbor embedding (t-SNE)[36] (**Fig. 2g**). The embedding space demonstrates meaningful hierarchical patterns. At the pocket level, ligands binding to the same pocket form visible clusters, whereas ligands targeting different pockets of the same protein are well-separated, indicating that embeddings of LigUnity can capture coarse-grained binding site information through optimizing virtual screening. Furthermore, the presence of different binding pockets for the same protein highlights the necessity of utilizing pockets for affinity prediction, as pockets with different structures should have different active ligands. At a more fine-grained ligand level, the embedding space shows that ligands with higher binding affinity values are closer to their target pockets, indicating that LigUnity learns structure-level differences affecting binding affinity, showing potential for hit-to-lead optimization.

## LigUnity improves hit-to-lead optimization

Hit-to-lead optimization aims to optimize the structure of the ligand for improved affinity by ranking candidate ligands that are commonly structurally similar to initial hits identified to be active. After observing the substantial improvement of LigUnity in virtual screening, especially its ability to capture ligand structure differences, we next investigated whether LigUnity can be applied to hit-to-lead optimization. We evaluated LigUnity on two FEP benchmarks (JACS[12] and Merck[27]) to predict binding free energies, which are the most widely used indicators for hit-to-lead optimization. These two benchmarks contain 16 targets, and each target has an average of 29 experimentally measured ligands. Following previous work,[18] assays similar to those included in the FEP benchmarks are excluded in the pre-training stage to avoid potential data leakage (see **Methods**). Ligands in the FEP benchmarks are also removed from the pre-training data.

In the zero-shot setting, where the model does not use any measured ligands for test proteins during pre-training and fine-tuning, our model surpasses existing computational methods (e.g. Glide-SP,[30] MM/GBSA[37]), structure-based methods (e.g. GenScore[32]), and structure-free methods (e.g. DrugCLIP[17]) (**Fig. 3a**), on 8 targets of the Merck FEP benchmark, demonstrating the effectiveness of our method. To further examine the performance of LigUnity in hit-to-lead optimization, we studied three more challenging settings: (1) no similar ligands setting, where training ligands with more than 50% Tanimoto similarity to any test ligand are excluded; (2) no similar proteins setting, where training proteins with more than 40% sequence similarity to any test protein are excluded;[3] (3) no similar ligands and proteins setting, where the criteria of the first two settings are both applied to exclude similar proteins or ligands. LigUnity again consistently outperforms competing methods across all three settings (**Supplementary Figure. 13**). Notably, on 16 targets from Merck and JACS benchmarks, it achieves a significant improvement of 38.1% over the sequence-only variant LigUnity(seq) under the most difficult no similar ligands and proteins setting (15.4% under the no same ligands setting), demonstrating the importance of incorporating pocket information for better generalization to novel proteins.

In real-world drug discovery, binding data for a few ligands of interest are often experimentally measured. A successful hit-to-lead optimization model should be able to leverage these binding data to enhance its predictions. To examine the ability to employ these data, we fine-tuned LigUnity and competing methods with varying proportions of ligand binding data. We observed that LigUnity consistently outperforms competing methods across all four settings (**Fig. 3b-e, Supplementary Figure. 14**). Even under the most challenging setting where both similar ligands and proteins are excluded, LigUnity achieves an $r^2$ of 0.501 on the Merck benchmark after fine-tuning on 80% of ligands (23.2 ligands on average for each target), reaching performance comparable to FEP+(OPLS4) ($r^2 = 0.528$), a leading commercial computational software for calculating the binding free energy. These results demonstrate the practical value of LigUnity for affinity prediction, offering an accurate and efficient alternative to resource-intensive computational chemistry approaches.

Similar to our analyses in the virtual screening, we conducted ablation studies to investigate the contribution of each component in LigUnity (**Supplementary Figure. 15**). Our results demonstrate that excluding the contrastive learning or the listwise ranking component leads to substantial performance drops. Notably, ablating the listwise ranking leads to a performance drop of over 50% in both zero-shot and few-shot settings, highlighting the effectiveness of the ranking objective in capturing the relative activity differences for hit-to-lead optimization tasks.

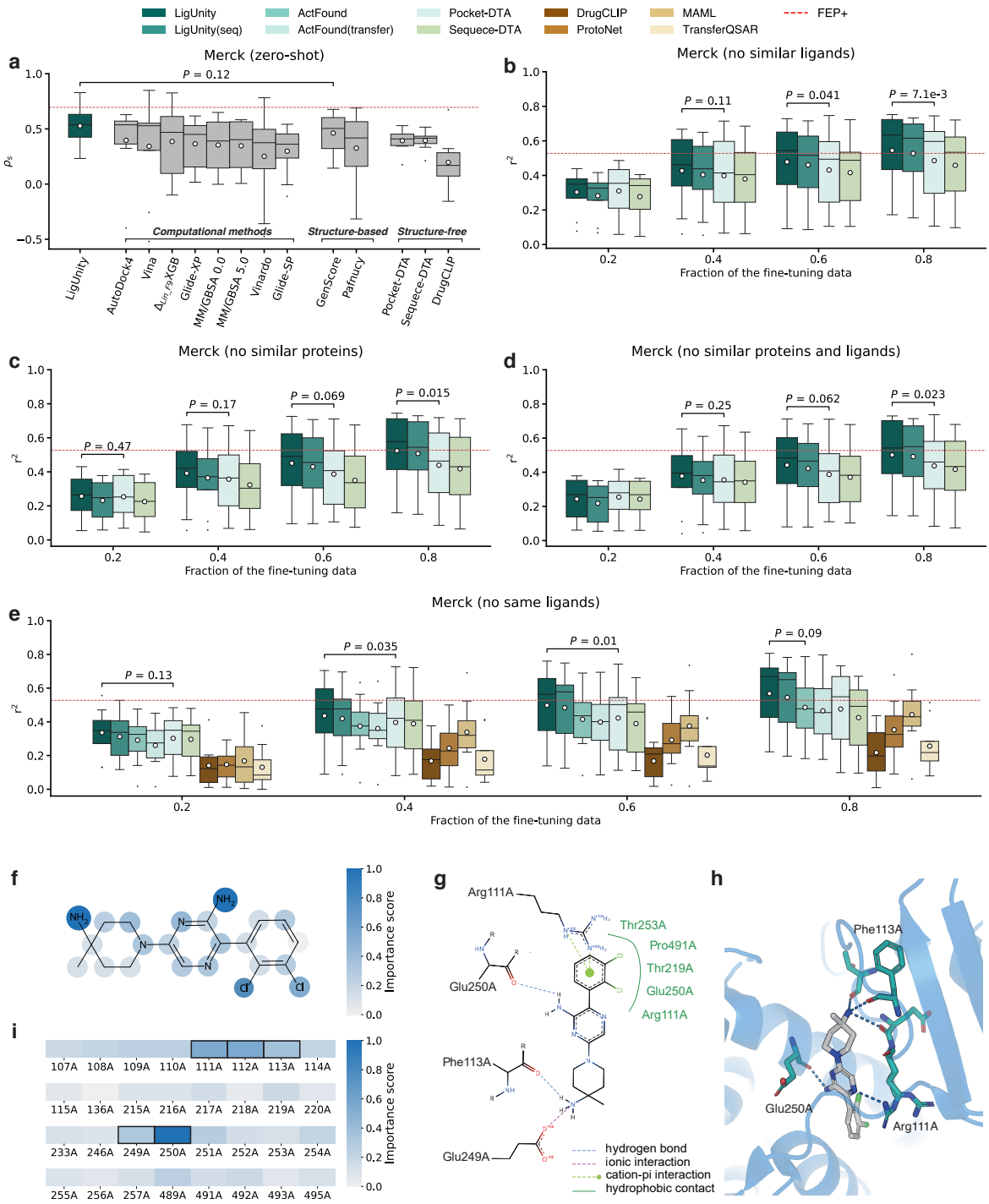



**Figure 3. Evaluation on hit-to-lead optimization. a,** Box plot comparing the binding affinity prediction on the Merck benchmark in terms of Spearman's rank correlation ($\rho_s$) in the zero-shot setting. The mean values (white dots) are calculated across n = 8 targets. **b-e,** Box plots comparing the binding affinity prediction across different settings on the Merck benchmark in terms of $r^2$ when 20%, 40%, 60%, 80% of the experimental binding affinities are used for fine-tuning. The mean values (white dots) are calculated across n = 8 targets. **f,** Case study on the SHP2 target showing the importance score of each ligand atom predicted by LigUnity. **g,h,** 2D **(g)** and 3D interaction graph **(h)** showing the non-covalent interaction between the ligand and pocket (PDB ID: 5EHR). **i,** Case study on the SHP2 target showing the importance score of each pocket residue predicted by LigUnity. The top five predicted key residues are highlighted with black boxes.

Finally, to help understand the structural significance of individual atoms and residues involved in the interaction, we calculated the importance score of each atom and protein pocket residue on the SHP2 target by masking them individually (**Fig. 3f-i**, see **Methods**). By examining the atom importance scores, LigUnity correctly identifies two amino groups (-NH2) as key contributing factors to the predicted binding affinity (**Fig. 3f**), aligning with the crystallographic observations of their hydrogen bonding and ionic interactions (**Fig. 3g,h**). Furthermore, using protein residue importance scores, the model correctly highlights the crucial role of Glu250E in hydrogen bonding, along with the polar interactions involving Phe113A, Arg249A, and Arg111A (**Fig. 3i**). Although the pose-free nature of our approach results in certain limitations, such as underestimating the contributions from specific hydrophobic cavity residues (Thr253A, Pro491A), LigUnity successfully identifies key interaction patterns. These results suggest that LigUnity can serve as an interpretable computational tool for studying structure-activity relationships and guiding hit-to-lead optimization.

## LigUnity as a versatile foundation model for different applications

To further explore the ability of LigUnity as a foundation model for affinity prediction, we next sought to evaluate its performance on assays from ChEMBL[28] and BindingDB,[29] two comprehensive public databases containing experimentally measured binding affinities covering various aspects of assay types, such as protein-based, cell-based, cell-membrane-based, and subcellular-based assays.[28,29] We first used a split-by-time setting to evaluate the performance, where the model is pre-trained using 18,552 assays released before March 2019 and tested on 161 assays released after March 2019 in ChEMBL and BindingDB. Each test assay has 48.6 ligands on average. To avoid potential data leakage, we further excluded test assays whose targets exist in the pre-training data.

LigUnity attains consistent improvement using varying numbers of fine-tuning data ranging from 4 to 16 (**Fig. 4a, Supplementary Figure. 16**). When compared to ActFound,[18] the best-competing method specifically designed for few-shot affinity prediction, LigUnity achieved an average improvement of 8.5% in terms of $r^2$. Since ActFound is a ligand-based model without considering pocket information, our substantial improvement reflects the effectiveness of using protein pocket information for affinity prediction. On a more challenging split-by-scaffold setting where test ligands have distinct scaffolds from training ligands, LigUnity again obtains the best performance (**Fig. 4b**), demonstrating its strong generalizability to unseen chemical scaffolds.

Next, we conducted a case study on phosphodiesterase 10A (PDE10A), a verified therapeutic target for antipsychotic therapy with no approved drugs, to assess the applicability of LigUnity in the real-world drug discovery pipeline. This dataset was released by Roche Ltd. in 2022, comprising 1,162 ligands with experimentally measured binding affinity and timestamp of each experiment.[38] We used a split-by-time setting for ligands in the PDE10A dataset, where the first measured 20% ligands are used for fine-tuning and the remaining 80% are used for testing (**Fig. 4c**). To ensure a rigorous split-by-time evaluation, we also restricted the pre-training data to the assays released before March 2019. To this end, our fine-tuned model achieves a Pearson correlation of 0.55 on the test set (**Fig. 4d**), substantially outperforming the 0.37 Pearson correlation of DrugCLIP (**Fig. 4e**). Building on this performance, our model demonstrates a strong ability to identify highly potent ligands, successfully identifying 4 highly active ligands (IC50 $<$ 1 nmol) among the top-10 predicted ligands (**Fig. 4f**) and 24 highly active ligands among the top-50 predicted ligands (compared to 11 for DrugCLIP). Collectively, these results demonstrate the superior capability of LigUnity in identifying highly potent ligands, holding

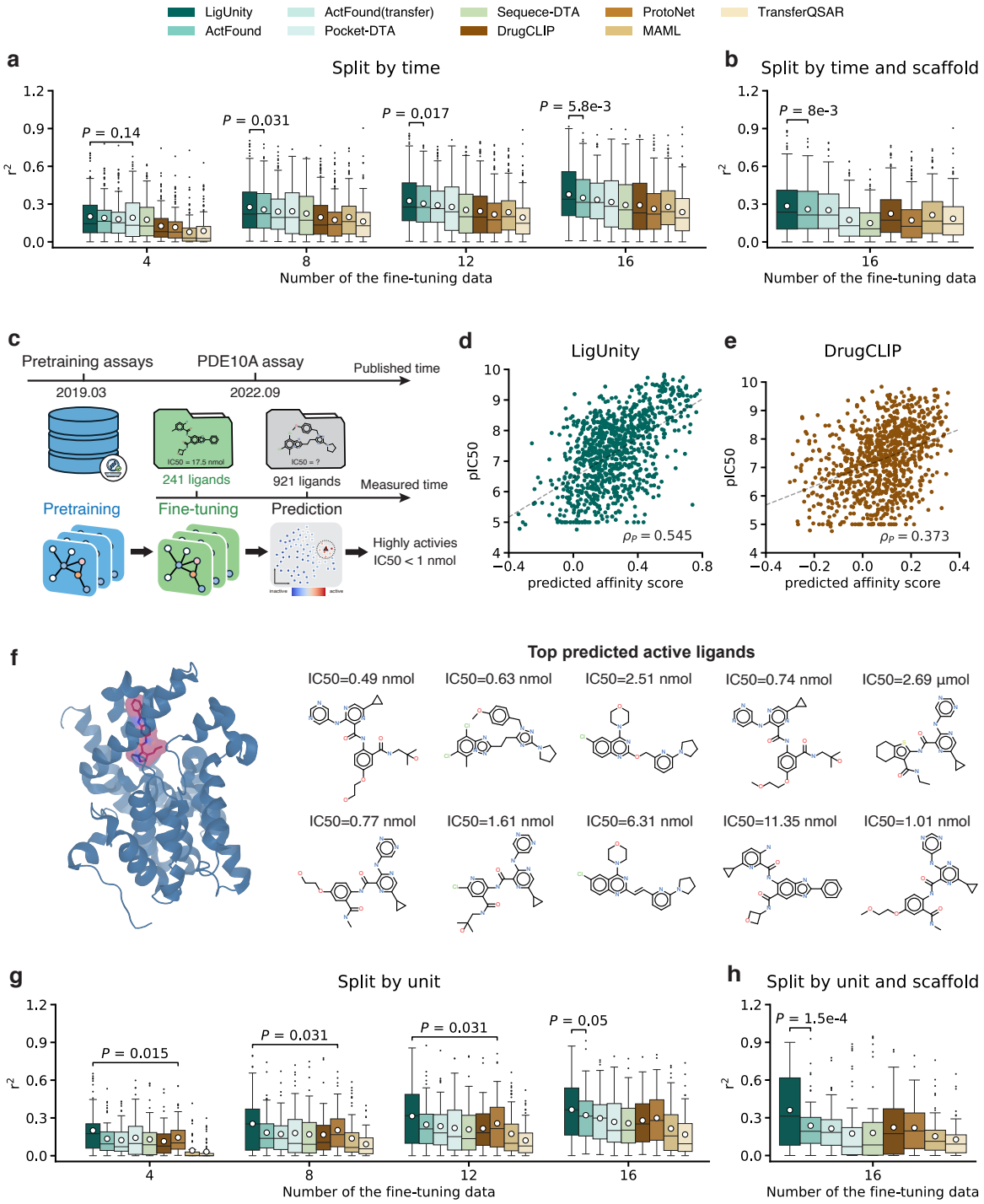



**Figure 4. Evaluation on diverse settings. a,b,** Box plots comparing the binding affinity prediction on the split-by-time setting in terms of $r^2$ when 4, 8, 12, 16 experimentally measured ligands are used for fine-tuning. The mean values (white dots) are calculated across n = 161 assays. **c,** Experimental setting for case study on the phosphodiesterase 10A (PDE10A) target. **d,e,** Scatter plots comparing the predicted affinity score and pIC50$=-\log_{10}$IC50 for LigUnity (**f**) and DrugCLIP (**g**) when finetuned with the first measured 20% ligands. **f,** 3D structure for the PDE10A target and its active ligand, and the top 10 ligands ranked by affinities predicted using LigUnity, along with their experimental IC50 values. **g,h,** Box plots comparing the binding affinity prediction on the split-by-unit setting in terms of $r^2$ when 4, 8, 12, 16 experimentally measured ligands are used for fine-tuning. The mean values (white dots) are calculated across n = 65 assays.

the potential to reduce the cost of hit-to-lead optimization.

Finally, we studied a split-by-unit setting where the test assays use percentage ("%") units, a widely used format (24.4% in ChEMBL) in practical drug design with limited exploration in ML methods. Assays with percentage units pose challenges for affinity prediction, as they distribute differently from pre-training assays with molar concentration (e.g., nmol) or density units (e.g., $\mu$g/ml). We hypothesize that our listwise ranking approach enables LigUnity to be robust against assays with different units, as it aims to learn the pocket-specific ligand ranking instead of the absolute affinity value. To this end, we examined a split-by-unit setting by using ChEMBL assays reporting activities in percentage (%) as test assays and assays with defined molar concentration or density units as pre-training assays. This test set contains 65 assays, and each assay has on average 30.5 ligands. When fine-tuned with different numbers of ligands (4 to 16), LigUnity consistently surpasses all competing methods (**Fig. 4g,h, Supplementary Figure. 16**). Notably, compared to the regression model Pocket-DTA, LigUnity shows a 40.2% improvement in this split-by-unit setting, substantially higher than the 13.8% improvement in the split-by-time setting, demonstrating the generalization ability of the listwise ranking approach across different measurement units, reassuring LigUnity as an effective foundation model for drug discovery.

## LigUnity boosts the active learning framework for drug discovery

In the real-world drug discovery pipeline, multiple iterations of molecular optimization are often required due to limited experimental resources and the difficulty in accurately predicting protein-ligand binding affinity. As a result, active learning is often exploited in this process. Building on this, we explored how LigUnity can be integrated with active learning to optimize ligands for Tyrosine Kinase 2 (TYK2), a therapeutic target for autoimmune diseases. We employed a dataset that contains binding free energy for 10,000 ligands calculated using FEP, consuming approximately 80,000 GPU hours (9.1 GPU years).[39] Following established active learning protocols,[40] we aimed to identify highly active ligands with limited computational resources in terms of the amount of FEP calculations.

To this end, we integrated LigUnity in an active learning framework, where we first trained the model using a small number of randomly selected ligands with known binding free energies. In each subsequent iteration, we selected a subset of unlabeled ligands and performed FEP calculations to determine their binding free energies. These newly labeled data points are then incorporated into the training set, and the model is retrained to refine its predictions (**Fig. 5a**).

We compared all methods under two selection strategies: a greedy strategy and an exploration-exploitation strategy (see **Methods**). We first found that LigUnity successfully identified ligands with the best binding free energy within a few iterations (**Fig. 5b**). The proportion of highly active ligands continually increased after each iteration (**Fig. 5c**), demonstrating the effectiveness of the active learning framework in identifying highly active ligands. Moreover, we found that LigUnity consistently outperforms other competing methods under both greedy and exploration-exploitation strategies, achieving over 40% improvement in terms of $r^2$ (**Fig. 5d**). By comparing two different selection strategies, we found that the greedy strategy has a higher recall in early iterations by focusing on the most promising candidates (**Fig. 5d, Supplementary Figure. 17**). Interestingly, the exploration-exploitation strategy attains superior performance in later iterations, primarily because of its more diverse ligand sampling in early iterations that helps improve model accuracy. These findings demonstrate the broad applicability of LigUnity in drug design, particularly its effectiveness in

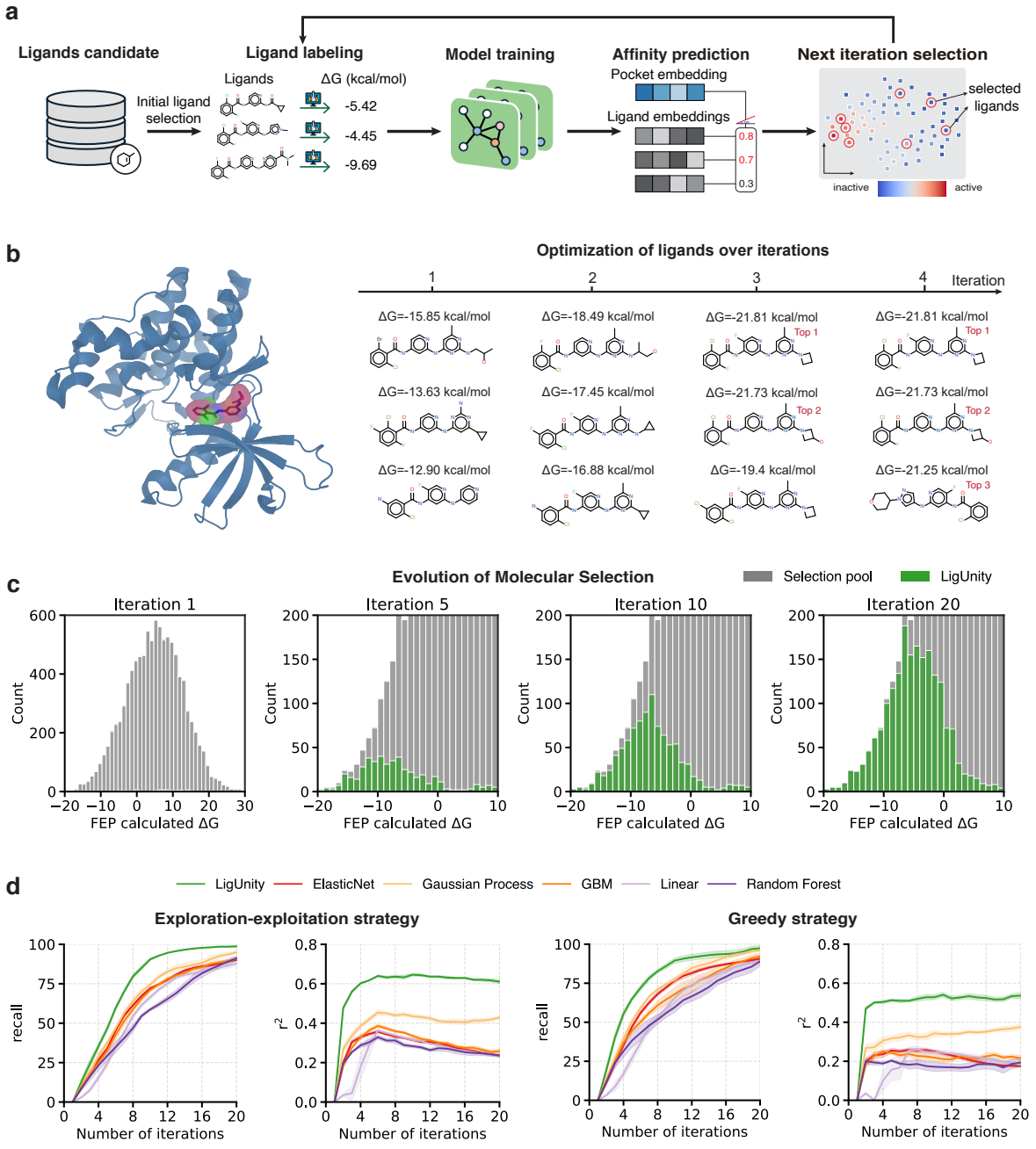



**Figure 5. Evaluation of LigUnity in an active learning framework on TYK2. a,** An active learning framework based on LigUnity. **b,** 3D structure for the TYK2 target (left panel) and the top ligands selected for iteration 1, 2, 3, and 4 (right panel) under the exploration-exploitation strategy, FEP-calculated binding free energies for each ligand are also shown. **c,** Histogram plots showing the distribution of ΔG values of selected ligand for each iteration based on a greedy selection strategy. **d,** Plots comparing LigUnity and competing methods on the TYK2 dataset in terms of top 2% recall and $r^2$ when using the exploration-exploitation strategy (left) and greedy selection strategy (right).

identifying highly active ligands using limited experimental and computational resources.

# Discussion

We have presented LigUnity, a protein-ligand affinity foundation model jointly optimized for both virtual screening and hit-to-lead optimization. Our model exploits the listwise ranking to learn pocket-specific ligand preference and contrastive learning to distinguish between active and inactive compounds. For virtual screening, our experiments have shown that LigUnity outperformed 24 competing methods across three benchmarks (DUD-E, LIT-PCBA, and Dekois 2.0). For hit-to-lead optimization, our model achieves the best performance in two FEP benchmarks and demonstrates promising accuracy in split-by-time, split-by-scaffold, and split-by-unit settings. Moreover, by developing an active learning framework to simulate real-world lead optimization, LigUnity efficiently identifies optimal binding compounds with limited computational resources. Collectively, these results demonstrate the strong generalization ability and practical usability of LigUnity throughout the early-stage drug discovery pipeline, establishing our confidence in LigUnity as a versatile foundation model for computer-aided drug discovery.

Two main lines of work are related to LigUnity, including structure-free Drug-Target Affinity (DTA) prediction methods and structure-based methods. Compared to DTA methods,[34,41–49] which directly predict binding affinity using protein and ligand information, LigUnity presents three key differences. First, DTA methods are challenged by the small number of proteins with measured affinity (2,773 in BindingDB), whereas LigUnity incorporates fine-grained pocket information and leverages a comprehensive training set covering 53,406 pockets for 4,847 proteins, enabling robust performance on under-studied proteins. Second, existing DTA methods perform direct regression of absolute affinity values, whereas LigUnity employs listwise ranking to alleviate batch effect and contrastive learning to capture protein-ligand interaction, reflected by the consistent performance improvement in our experiments. Third, DTA methods require intensive computation for each prediction, whereas LigUnity enables rapid screening by pre-computing and storing embeddings, leading to better efficiency for large-scale applications. Compared to structure-based methods[16,32,33,50–61] that rely on 3D binding poses, LigUnity presents two improvements: First, structure-based methods are constrained by the scarcity of ligands with co-crystal structures. For example, only 7.5% of 0.57 million compounds with measured activities in BindingDB have crystal structures.[29] In contrast, LigUnity is developed using a diverse training set of 0.45 million unique ligands, enabling effective generalization to novel chemical scaffolds; Second, structure-based methods require computationally expensive pose generation (approximately one minute per compound using Glide), whereas LigUnity achieves competitive performance without the need of pose information, making it applicable to large-scale screening.

Despite its promising performance in binding affinity prediction, LigUnity has two limitations that can be addressed in future works. First, LigUnity can only be applied to assays with known protein targets, limiting its applications to diverse assays without target information (e.g., phenotype assays). To expand the applicability of our model, we plan to incorporate additional modalities such as text descriptions and cell line information in future work. Second, although our current approach achieves strong results without structural information, integrating both binding pose data and activity data could potentially enhance prediction accuracy, particularly in scenarios where sufficient computational resources are available to obtain precise binding poses.

# Method

## Problem setting

LigUnity aims to predict binding affinity between given binding pockets and unmeasured ligands. Our training data are organized by assay, which is defined as the experimental procedure designed to evaluate the binding affinity of ligands against a specific protein target. Each assay has multiple measured ligands and only one target protein. A key consideration is that affinity values are only comparable within the same assay, and comparing affinity values from different assays targeting the same protein may not be reliable.[62] This incomparability arises from variations in experimental conditions (e.g., cofactor concentration, pH, temperature), assay formats (e.g., cell-based vs. target-based assays, different detection methods), and affinity measurements (e.g., IC50, Ki, Kd). Therefore, our method focuses on learning relative affinity rankings within each assay instead of absolute values.

Let $\mathcal{A}$ denote the set of assays. For each assay $A_i \in \mathcal{A}$, $L_i$ denotes the set of tested ligands, and $v_i(l)$ denotes the affinity value of ligand $l \in L_i$. Each assay corresponds to one target protein, which may have multiple PDB structures available. We define $P_i$ as the set of protein pocket structures for assay $A_i$ and randomly sample one pocket structure from $P_i$ at each training step.

## Pocket and ligand Encoder

In the pre-training stage, our pocket encoder and ligand encoder are initialized with Uni-Mol,[63] an SE(3)-invariant graph transformer for 3D moleculars, which is a structure-free method and does not rely on the binding pose as input. For the ligand branch, we first use RDKit to generate 3D conformations for each ligand which have been proven effective for various tasks,[64] and then input them into the Uni-Mol ligand encoder to generate ligand embeddings. Similarly, for the pocket branch, we input the protein pocket conformation into the Uni-Mol pocket encoder to generate the protein embedding. For a pocket-ligand pair $(p, l)$, their vector representations are obtained through:

$$(\mathbf{p}, \mathbf{l}) = (\mathrm{E_p}(p), \mathrm{E_l}(l)), \tag{1}$$

where $\mathrm{E_p}$ and $\mathrm{E_l}$ denote the pocket and ligand encoders respectively, $\mathbf{p}$ denotes the protein embedding, and $\mathbf{l}$ denotes the ligand embedding.

## Listwise Ranking and Contrastive Learning

In the pre-training stage, LigUnity integrates two fundamental strategies: contrastive learning and listwise ranking. The contrastive learning approach captures the coarse-grained distinction between active and inactive pocket-ligand pairs, which aims to bring the embeddings of positive pocket-ligand pairs closer and push the negative pairs apart. The listwise ranking approach learns the fine-grained affinity ranking of measured ligands for their binding pocket, which can alleviate the batch effect across assays. These two approaches complement each other, and we jointly optimize them to make the model suitable for both virtual screening and hit-to-lead optimization scenarios.

During inference, the model only needs to take 3D protein pockets and 3D ligand conformations (generated by RDKit) as input. It then calculates the pocket and ligand embeddings and outputs their cosine similarity as the predicted affinity score.[17] This differs from most previous machine learning

affinity prediction models, which typically model affinity prediction as a regression task and use either dual-tower architectures (structure-free methods)[34,45,47] or require binding poses (structure-based methods)[16,33] as input. Once the pocket and ligand embeddings are pre-computed, LigUnity enables fast screening by similarity-based comparison in joint embedding space, achieving six magnitudes of speedup compared to docking methods.[17]

### Listwise Ranking

Listwise ranking is a learning-to-rank approach that directly optimizes the order of an entire list of items. Unlike pointwise approaches that predict absolute scores or pairwise approaches that compare item pairs, listwise ranking considers the entire ranked list simultaneously, making it particularly suitable for problems where the relative ordering matters more than absolute values. We adopt the Plackett-Luce model,[22] a statistical model specifically designed for ranking data. The fundamental idea of the Plackett-Luce model is to assign each item a positive "strength" parameter, which determines the probability of the item being selected. In our context, the strength parameter $\theta_i$ for each ligand is computed as the exponential of its pocket-ligand similarity score: $\theta_i = \exp(\text{sim}(\mathbf{p}, \mathbf{l}_i))$. For a set of items with strength parameters $\{\theta_1, ..., \theta_m\}$, the probability of selecting item $i$ is defined as:

$$P(i \text{ is chosen}) = \frac{\theta_i}{\sum_{j=1}^{m} \theta_j}.$$

The Plackett-Luce model treats ranking as a sequential selection process. Given $n$ ligands to rank, the model first selects the top-ranked ligands from all candidates using the above probability formula. Then, it selects the second-ranked ligand from the remaining $n-1$ candidates, again using the same probability formula but only considering the remaining ligands. This process continues until all ligands are ranked. The probability of observing a complete ranking sequence $\pi$ is the product of these conditional probabilities at each selection step:

$$P(\pi) = \prod_{k=1}^{n} \frac{\theta_{\pi(k)}}{\sum_{j=k}^{n} \theta_{\pi(j)}},$$

where $\pi(k)$ denotes the index of the ligand at position $k$ in the ranking. The model parameters are optimized by minimizing the negative log-likelihood of observing the true ranking based on measured affinity values. This framework naturally handles varying numbers of ligands across different assays and provides a probabilistic interpretation of the ranking process.

Given this probabilistic framework, we can define the ranking loss for assay $A_i$ as the weighted negative log-likelihood of the observed ranking probability. For each assay $A_i \in \mathcal{A}$, we order ligands according to their measured affinity values. At the $k$-th step of the selection process, considering all ligands with affinity values no greater than that of the $k$-th ranked ligand, we define the selection probability for the $k$-th ranked ligand as:

$$\pi(l_k | L_i^{\leq k}, p) = \frac{\exp(\text{sim}(\mathbf{p}, \mathbf{l})/\tau)}{\sum\limits_{j \in L_i^{\leq k}} \exp(\text{sim}(\mathbf{p}, \mathbf{l}_j)/\tau)}, \tag{2}$$

$$\text{where } \mathbf{p} = \text{E}_p(p), \quad \mathbf{l} = \text{E}_l(l), \tag{3}$$

where $p \in P_i$ is the 3D structure of the pocket tested in the assay $A_i$; $l_k$ the $k$-th ranked ligand; $L_i^{\leq k}$ represents the set of ligands with affinity values no greater than $v_i(l_k)$; $\tau$ is a temperature hyper-parameter; and $\text{sim}(\mathbf{p}, \mathbf{l})$ denotes the cosine similarity between protein and ligand embedding vectors.

Based on these step-wise selection probabilities, we compute the ranking loss as the weighted sum of the negative log probabilities over all selection steps:

$$\mathcal{L}_i^r = -\sum_{k=1}^{|L_i|} \mu_k \log \pi(l_k | L_i^{\leq k}, p), \tag{4}$$

$$\text{where } \mu_k = \frac{1}{\sqrt{|L_i|}\log(k+1)}, \tag{5}$$

where $\mu_k$ is a decay factor with two components: $1/\log(k+1)$ prioritizes top-ranked molecules and $1/\sqrt{|L_i|}$ prevents assays with many ligands from dominating the training loss. We also experimented with $1/|L_i|$ and constant scaling terms and empirically found that $1/\sqrt{|L_i|}$ yields the best performance.

### Contrastive Learning

Contrastive learning aims to maximize the similarity between active pocket-ligand pairs while minimizing the similarity between inactive pairs. Following Gao et al.,[17] we formulate our contrastive loss using in-batch softmax, which consists of two components: the pocket-to-ligand loss $\mathcal{L}_i^{p\to l}$ and ligand-to-pocket loss $\mathcal{L}_i^{l\to p}$:

$$\mathcal{L}_i^{p\to l} = -\frac{1}{|L_i^a|} \sum_{k \in L_i} \log \frac{\exp(\text{sim}(\mathbf{p}, \mathbf{l}_k)/\tau)}{\sum_{m \in L_{batch}} \exp(\text{sim}(\mathbf{p}, \mathbf{l}_m)/\tau)}, \tag{6}$$

$$\mathcal{L}_i^{l\to p} = -\frac{1}{|L_i^a|} \sum_{k \in L_i} \log \frac{\exp(\text{sim}(\mathbf{p}, \mathbf{l}_k)/\tau)}{\sum_{n \in P_{batch}} \exp(\text{sim}(\mathbf{p}_n, \mathbf{l}_k)/\tau)}, \tag{7}$$

$$\text{where } \mathbf{p} = \text{E}_p(p), \quad \mathbf{l} = \text{E}_l(l), \tag{8}$$

where $p \in P_i$ is a 3D structure of the pocket tested in the assay $A_i$; $L_{batch}$ and $P_{batch}$ denote the set of ligands and pockets sampled in this training batch, respectively; $\tau$ is a temperature hyper-parameter, and $L_i^a$ denotes the set of active ligands tested in the assay $A_i$.

Building on this formulation, we define how positive and negative pairs are identified in our framework: pocket-ligand pairs with experimentally measured affinity ($\leq 10\mu$mol) are treated as positive pairs following Mayr et al.,[65] while pairs without measured affinity values serve as negative pairs. Although the boundary between active and inactive ligands varies across different protein targets, our listwise ranking framework naturally addresses this issue by learning the relative affinity among ligands rather than relying solely on binary active/inactive labels. However, this in-batch negative sampling may introduce potentially harmful false negative pocket-ligand pairs. To mitigate this, we devised a simple strategy to identify potentially active pocket-ligand pairs. Specifically, pocket-ligand pairs formed between assay A and assay B are excluded from negative sampling if the protein targets of the two assays share the same UniProt ID. This exclusion is based on the observation that protein pockets with the same UniProt ID are likely to share overlapping active ligands.

The total contrastive loss is defined as:

$$\mathcal{L}^c = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \gamma_i (\mathcal{L}_i^{p\to l} + \mathcal{L}_i^{l\to p}). \tag{9}$$

where $\gamma_i = 1/\sqrt{|L_i|}$ prevents assays with many ligands from dominating the training loss.

### LigUnity Loss

To optimize both virtual screening and hit-to-lead optimization, LigUnity is pre-trained with a combined ranking and contrastive loss:

$$\mathcal{L} = \mathcal{L}^c + \mathcal{L}^r. \tag{10}$$

Collectively, this approach enables LigUnity to predict affinity rankings of active ligands while distinguishing active from inactive ligands, supporting its effectiveness in both tasks. We demonstrated that jointly optimizing these two tasks can help each other, which is shown in our ablation studies (**Supplementary Figures. 12, 15**).

## Heterogeneous GNN

Based on the observation that structurally similar pockets tend to have similar interaction patterns with ligands, we hypothesize that integrating information from existing large-scale pocket-ligand databases should improve the performance of LigUnity. To this end, we propose a Heterogeneous Graph Neural Network model (H-GNN) to refine pocket embeddings by incorporating information from a large-scale pocket-ligand knowledge graph.

Our H-GNN model consists of two main components: a pocket-pocket aggregator and a pocket-ligand aggregator. This model receives a directed knowledge graph $G = (V_p, V_l, E_{l\to p}, E_{p\to p})$ of pockets and ligands as input, where $V_p = \{\mathbf{p}_i\}, V_l = \{\mathbf{l}_i\}$ are two sets of nodes representing the pockets and ligands respectively; $E_{l\to p}$ is the ligand-to-pocket edge set and $E_{p\to p}$ is the pocket-to-pocket edge set. For ligand-to-pocket edges, we add an edge from a ligand to a pocket if the ligand is active for the protein pocket with experimentally measured affinity. For pocket-to-pocket edges, we add edges between pockets based on sequence alignment scores, which are computed using scikit-bio with the BLOSUM50 substitution matrix. We only add edges between pockets with high alignment scores. Specifically, if the alignment score between pocket A and pocket B is larger than half of the self-alignment score of pocket A, we add an edge from pocket B to pocket A.

We constructed this large-scale pocket-ligand knowledge graph from our pre-training data. The graph contains 53,406 pocket nodes, 0.45 million ligand nodes, 0.83 million ligand-to-pocket edges, and 16 million pocket-to-pocket edges. During the virtual screening, the H-GNN model refines the embedding of the query protein pocket by incorporating information from the heterogeneous graph: $\mathbf{p}_k^* = \text{H-GNN}(\mathbf{p}_k)$. We then use this refined pocket embedding $\mathbf{p}_k^*$ to retrieve hit ligands in the candidate ligand embedding space.

### Pocket-Pocket Aggregator

We first use pocket-pocket aggregator to incorporate information from similar pockets, based on the observation that structure-similar pockets often exhibit similar interaction patterns with ligands. The aggregation function is defined as:

$$\mathbf{p}_k^p = \text{AGG}_p(\mathbf{p}_i | (i,k) \in E_{p\to p}), \tag{11}$$

where $\mathbf{p}_k^p$ is the refined embedding of the $k$-th pocket $\mathbf{p}_k$ that aggregates information from its neighboring pockets; $\text{AGG}_p$ is implemented using an attention mechanism, which is defined as follows:

$$\text{AGG}_p(\mathbf{p}_i|(i,k) \in E_{p\to p}) = \sum_{(i,k)\in E_{p\to p}} \alpha_i \mathbf{p}_i, \tag{12}$$

$$\text{where } \alpha_i = \frac{\exp(\tilde{\alpha}_i)}{\sum_{(n,k)\in E_{p\to p}} \exp(\tilde{\alpha}_n)}, \tag{13}$$

$$\text{where } \tilde{\alpha}_i = \mathbf{W}_2^T \cdot \text{ReLU}(\mathbf{W}_1 \cdot [\mathbf{p}_i|\mathbf{p}_k] + \mathbf{b}_1), \tag{14}$$

where $\alpha_i$ denotes the normalized attention weight for the neighboring pocket $i$; $[\mathbf{p}_i|\mathbf{p}_k]$ denotes the concatenation of vectors $\mathbf{p}_i$ and $\mathbf{p}_k$; $\mathbf{W}_1 \in \mathbb{R}^{d_h \times 2d}$, $\mathbf{b}_1 \in \mathbb{R}^{d_h}$, and $\mathbf{W}_2 \in \mathbb{R}^{d_h \times 1}$ are three learnable parameters, where $d$ is the hidden dimension of pocket embeddings and $d_h$ is the hidden dimension of the attention network.

### Pocket-Ligand Aggregator

After aggregating information from similar pockets, We next utilize the pocket-ligand aggregator to aggregate information from 2-hop ligand neighbors, based on the observation that structure-similar pockets should have structure-similar active ligands. Let $\mathcal{N}_k^2$ denote the set of 2-hop ligand neighbors of pocket $k$:

$$\mathcal{N}_k^2 = \{j|\exists i : (j,i) \in E_{l\to p} \wedge (i,k) \in E_{p\to p}\}, \tag{15}$$

the aggregation function is defined as:

$$\mathbf{p}_k^l = \text{AGG}_l(\Phi([\mathbf{l}_j|\mathbf{r}_{kj}])|j \in \mathcal{N}_k^2), \tag{16}$$

where $\mathbf{p}_k^l$ is the refined embedding of the $k$-th pocket that aggregates information from the active ligands of its neighbor pockets; $\mathbf{l}_j$ is the ligand embedding of the $j$-th ligand; $\mathbf{r}_{kj}$ is a relation embedding derived from the similarity score between the $k$-th pocket and its intermediate neighbor (the $i$-th pocket) where the $j$-th ligand is active for it; $\Phi$ is a 2-layer MLP with the hidden size of $2d$. Similar to the pocket-pocket aggregator, $\text{AGG}_l$ is also implemented using an attention mechanism with different learnable parameters.

### Refined Pocket Embedding

Finally, we combine the original pocket embedding with the refined embeddings from both aggregators to obtain the final refined pocket embedding:

$$\mathbf{p}_k^* = \gamma_1 \mathbf{p}_k^l + \gamma_2 \mathbf{p}_k^p + (1 - \gamma_1 - \gamma_2)\mathbf{p}_k, \tag{17}$$

where $\gamma_1, \gamma_2 \in [0,1]$ and $\gamma_1 + \gamma_2 \leq 1$ are parameters that control the contribution of each component; This weighted combination allows the model to adaptively balance information from different sources: similar pockets, their active ligands, and the pocket's intrinsic features. Collectively, our proposed H-GNN integrates information from both similar pockets and their active ligands, leveraging the observation that structure-similar binding pockets tend to have similar active ligands, to enhance the performance of LigUnity in finding active ligands for the query protein pocket. The effectiveness of H-GNN is shown in our ablation studies (**Supplementary Figures. 10- 12**).

## Implementation details

The details of our training strategy are listed as follows. In the pre-training stage, we trained LigUnity for 50 epochs, and we used the averaged parameter of the last 10 epochs for testing. We warmed up the learning rate for the first 6% step and adopted a polynomial decay learning rate scheduler afterward. We used the Adam optimizer with a maximum learning rate of 0.0001. For each training iteration, we sampled 24 assays for computing loss. For each assay, we randomly sampled one pocket structure and a maximum of 16 measured ligands coming from the assay, making sure that their experimental affinity values were comparable to each other. We used listwise ranking for measured ligands that belonged to the same assay and used contrastive loss between the whole sampled ligands and pockets in this batch.

For Heterogeneous-GNN (H-GNN), we trained it separately from the pocket and ligand encoder. We trained H-GNN for 10 epochs and used the last epoch for testing. We adopted the constant learning rate scheduler with a learning rate of 0.001. For each training iteration, we randomly sampled 128 assays, and we randomly selected one pocket structure and one active ligand for each assay. As the large-scale nature of the pocket-ligand knowledge graph makes it hard to fine-tune H-GNN on a few measured ligands, we adopted H-GNN only for virtual screening tasks and used the contrastive loss between the whole ligands and pockets sampled in this batch for training.

For LigUnity(seq), the protein-based version of LigUnity, we used ESM2[66] 35M for encoding the protein sequence in replacement of the 3D pocket encoder, and we updated the parameters of ESM2 during training. Furthermore, we also attempted to use the larger ESM2 3B model with a frozen gradient but observed a drop in test performance. To fully make use of the pocket and protein information, the results of LigUnity are reported as the ensemble of the pocket-based model and the protein-based model, except for the experimental setting where training proteins with >40% sequence similarity are removed. This exception is made because protein-based models typically show limited generalization ability to unseen protein family with low sequence similarity.

We also trained Pocket-DTA and Sequence-DTA, the regression-based version of LigUnity and LigUnity(seq) respectively, to compare different protein-ligand affinity prediction strategies. In the pre-training stage, we used mean square loss and margin loss for training. First, for pocket-ligand pairs with experimental affinities, we used the mean square loss for training which is defined as the mean square error between experimental affinity and predicted affinity $\left(L_{SE} = (y_i - \hat{y}i)^2\right)$. Second, to enable Pocket-DTA to distinguish between active and inactive ligands, we introduced in-batch margin loss. For unmeasured in-batch pocket-ligand pairs, we consider them to be inactive and employ margin loss to penalize them $\left(L_{Margin} = \max(0, \hat{y} - (y_{min} - m))^2\right)$, where $y_{min}$ is the minimum known activity known to the pocket and $m$ is a margin hyper-parameter. We set the margin to 2.0 after a grid search on the validation set.

## Data Curation

The purpose of our data curation is to collect high-quality pocket-ligand binding affinity data and organize them by assays. We developed a curation pipeline based on the BindingDB[29] and ChEMBL[28] datasets. ChEMBL is a large-scale bioactivity database that integrates experimental data from scientific publications, patents, and other datasets. It covers various aspects of drug discovery and development, containing 2.5 million unique compounds and 19 million experimentally measured activities. BindingDB is another large-scale database that focuses specifically on protein-ligand binding

affinities. It is curated mainly from scientific publications and US patents, containing 0.5 million compounds, 1.1 million measured affinities, and 2,773 protein targets.

Our data curation pipeline consists of three stages: organizing affinities into assays, finding pockets in PDB, and removing duplicate assays. In this study, we used the 34 version of ChEMBL and the 2024m5 version of BindingDB downloaded from their official websites. The details of our curation pipeline are as follows:

### Organizing affinities into assays

Considering that the affinities from different assays are incompatible due to varying experimental conditions (e.g., cofactor concentration, pH, temperature), assay formats (e.g., cell-based vs. target-based assays, and different detection methods), and affinity measurements (e.g., IC50, Ki, Kd), we first organized the experimental affinities into assays. For ChEMBL, we collected all assays with "Binding" type and known protein targets, then we grouped affinity data by ChEMBL assay ID. We only retained affinity data with molar concentration units (e.g., nmol) or density units (e.g., $\mu$g/ml) to ensure data quality, as the exact meaning of other units (e.g., % or no unit) was difficult to determine. This filtering step removed 33.1% of the activity data in ChEMBL. For BindingDB, we grouped the affinity data with the same protein target, affinity measurement, source document, and assay description into an assay following Actfound,[18] ensuring that affinity values within the same assay were comparable. To ensure data quality, we removed all affinity data whose standard relation is not equal to "=", as this indicates that their specific activity values are unknown. We also removed inorganic ligands and large molecular ligands (molecular weights $>$ 1,000). Finally, we retained assays with more than 5 affinity values for further curation. We collected 69,872 and 25,902 assays from ChEMBL and BindingDB respectively, with each assay having one corresponding protein target.

### Finding pockets in PDB

Based on the observation that most assays are designed for a specific binding site of interest, we hypothesize that ligands in an assay should bind to the same pocket of a protein. Therefore, once we identified a binding pocket for one ligand in an assay, we assumed that other ligands in the same assay would bind to the identical pocket. For each assay, we identified binding pockets in the PDB (Protein Data Bank)[67] using two criteria: protein sequence similarity and ligand Tanimoto similarity. First, we searched for protein-ligand complex structures whose protein sequence similarity with the protein studied in this assay is above 40%, as proteins with such similarity typically share high structural and functional homology.[3] Then, we calculated the Tanimoto similarity between the crystal ligands in these PDB structures and the assay ligands. If the ECFP4 Tanimoto similarity exceeded 70%, we collected the binding pocket of the crystal ligand as the candidate pockets for this assay, as ligands with such high similarity are likely to bind to the same pocket in a protein.[68]

To discard biologically irrelevant pockets that have weak binding interactions with the ligand (i.e., those having a small number of binding residues with ligands), we used the processed binding pocket database provided by BioLip.[69] After this procedure, we collected 25,883 and 16,863 assays from ChEMBL and BindingDB respectively. For each assay, there are 7.19 identified pockets on average. Although the identified pockets for the same assay have different PDB IDs, they mostly belong to the same binding site. To verify this, we aligned complex PDB structures retrieved for each assay and computed the maximum distance between the centers of all pockets. We identified 1,229 and

802 assays whose maximum distance between retrieved pockets exceeded 10 Å for ChEMBL and BindingDB respectively, accounting for only 4.8% of the total assays.

### Removing duplicate assays

As ChEMBL and BindingDB share common data sources including scientific literature and patents, there could be many overlapping assays between these two datasets. Additionally, there are some repeating assays within the ChEMBL or BindingDB database as well. To reduce the negative impact of data redundancy on pre-training, we detected and removed identical assays following pQSAR-ChEMBL.[70] Specifically, if the Pearson correlation ($\rho_p$) of affinities between the shared ligands of two assays exceeds a threshold of 0.95, we considered those two assays to be identical. After this stage, we detected and removed 15,664 repeating assays, ending up with 26,748 non-repeating assays curated from ChEMBL and BindingDB.

After the above curation procedures, we collected 26,748 assays with 428,767 unique ligands and 36,662 unique pockets from ChEMBL, BindingDB, and PDB, covering 2,196 unique proteins. Each assay has on average 30.29 measured ligands and 10.96 pockets from PDB. Our curation process has created a large-scale pocket-ligand affinity dataset while maintaining the quality of the curated protein binding pockets, which is supported by the small fraction of assays with distinct pockets.

However, due to the limited number of unique proteins in the original binding database (2,773 in BindingDB), models trained solely using these data might result in poor generalization on unseen proteins. To address this limitation, we further enlarged our curated dataset by utilizing the protein-ligand complex data from PDB,[71] which includes diverse proteins. For each protein-ligand complex with crystal structure, we considered the ligand to be active following previous methods[16,17] and created one assay that contained only one protein pocket and one active ligand. We did not merge ligands binding to the same protein into one assay, as we could not determine the relative affinity ranking between these ligands. To avoid the negative impact of the low-quality complex structure data in PDB (those with low resolution, inactive ligands, or biologically irrelevant ligands), we directly used the PDBBind v2020[72] dataset for training, following previous methods.[16,17] The PDBBind dataset contains 3,889 unique proteins, and the number of unique proteins and pockets in our curated pre-training data finally reached 4,847 and 53,406 respectively.

## Competing Methods

In this paper, we compared LigUnity with three types of competing methods: (1). docking score methods including Glide (SP and XP),[30] Vina,[31] AutoDock4,[73] Vinardo,[74] Gold,[75] Surflex,[76] and Flexx;[77] (2). structure-based ML methods that take protein-ligand complex structure as input including NNscore 2.0,[52] RFscore v4,[53] $\Delta_{\text{Vina}}$RF,[54] $\Delta_{\text{Lin_F9}}$XGB,[56] RTMScore,[16] Genscore,[32] Pafnucy,[57] OnionNet,[58] Planet,[60] PIGNet,[59] EquiScore,[51] Kdeep,[78] TANKBind,[79] 3D-GNN,[59] Denvis[33] and GNINA;[61] (3). structure-free methods DeepDTA,[34] DrugCLIP,[17] Pocket-DTA (the regression-based version of LigUnity) and its protein sequence-based version Sequence-DTA. For experiments on the Merck and JACS benchmarks, we additionally compared LigUnity with physics-based binding free energy calculation methods including Prime-MM/GBSA[37] and FEP+.[12] For active learning experiments on the TYK2 dataset, we compared five machine learning methods including ElasticNet,[80] Gaussian Process,[81,82] Gradient Boosting Machine (GBM),[83] Linear Regression, and Random Forest,[84] and these methods take 2048-dimensional ECFP4 fingerprint as input.

## Experiments on virtual screening benchmarks

For virtual screening, we evaluated our method on three benchmark datasets (DUD-E,[24] Dekois,[25] and LIT-PCBA[26]). For DUD-E and LIT-PCBA benchmarks, we used the processed version from DrugCLIP's repository.[17] For the Dekois 2.0 benchmark, we downloaded it from the official website and extracted residues within 6 Å distance from the crystal ligand to form the binding pocket. To prevent data leakage from overlapping protein targets between our curated pre-training data and test benchmarks, we excluded proteins that appear in the three benchmarks from our pre-training data and used the model trained on this processed dataset for testing.

The DUD-E benchmark contains 102 targets, with an average of 224.4 active ligands per target. For each active ligand, there are approximately 50 decoys with similar physicochemical properties, with over 99% of decoys not experimentally verified. The Dekois 2.0 benchmark follows a similar construction approach, containing 81 targets, each with 30 active ligands and 40 generated decoys per active ligand. The LIT-PCBA benchmark has 15 targets with an average of 533.3 active ligands per target. It contains 2.64 million inactive ligands with an active-to-inactive ratio of 1:1000, better reflecting real-world drug design scenarios. Additionally, all active and inactive ligands in LIT-PCBA have been experimentally verified. In contrast, DUD-E and Dekois 2.0 benchmarks exclude decoys with similar structures to avoid potential active ligands, which may make them relatively easy for ML models. For DUD-E, the results of Denvis-G, Denvis-R, DeepDTA, Gold, Surflex, Flexx, Vina, and GNINA are from Krasoulis et al.;[33] the results of other structure-based methods are from Cao et al.[51] For Dekois 2.0, the results of all docking methods and structure-based methods are from Cao et al.[51] For LIT-PCBA, the evaluation results of Denvis-G, Denvis-R, GNINA, and DeepDTA are from Krasoulis et al.;[33] the results of GenScore and Glide SP are from Shen et al.[32]

Following previous works,[16,17,33] we evaluated methods using three metrics: Enrichment Factor (EF), Area Under the Receiver Operating Characteristic curve (AU-ROC), and Boltzmann-Enhanced Discrimination of ROC (BEDROC). EF quantifies a model's ability to identify actives among top-ranked ligands, calculated as the proportion of actives in top-ranked ligands (e.g., top 1%) divided by the proportion of actives in all candidate ligands. We use EF 1% for evaluation, which ranges from 0 to 100. AU-ROC assesses a model's ability to correctly rank active ligands before inactive ligands and is defined as the area under the plot of true positive rate against false positive rate. However, since AU-ROC treats all ranked ligands equally, it may not be ideal for virtual screening tasks where only top-ranked ligands undergo costly experimental verification. To address this limitation, we employ BEDROC, which assigns higher weights to top-ranked ligands. Following previous works,[16,17,33] we set the $\alpha$ parameter to 80.5 to balance the evaluation of top-ranked ligands and overall ranking performance. AU-ROC and BEDROC both range from 0 to 1. For all metrics, higher values indicate better performance.

To further assess LigUnity's generalizability to unseen proteins, we conducted experiments on unseen protein families. We excluded training proteins with more than 80% and 40% sequence similarities to any test protein on the three benchmarks; then we used the model trained on this processed dataset for evaluation. We used cd-hit to compute protein sequence similarity, which is defined as the number of identical residues after alignment.

## Experiments on FEP benchmarks

To test the performance of LigUnity on hit-to-lead optimization scenarios, we evaluated all methods on two FEP calculation benchmarks (JACS[12] and Merck[27]). The JACS and Merck benchmark datasets, along with the results of FEP+(OPLS4)[12] calculated using the state-of-the-art OPLS4[85] force field, were obtained from the public repository https://github.com/schrodinger/public_binding_free_energy_benchmark.[86] These FEP benchmarks contain 16 targets (assays), with each target having an average of 29 experimentally measured ligands. We extracted residues within a 6 Å distance from the crystal ligand as the binding pocket. For experiments on these FEP calculation benchmarks, we used Spearman correlation ($\rho_s$) and $r^2 = \max(0, \rho_p)^2$ for evaluation.

For zero-shot evaluation on FEP benchmarks, we directly applied LigUnity for inference without fine-tuning. To mitigate potential data leakage in the pre-training data, we removed assays that were similar to assays in test benchmarks from the pre-training data. Following previous methods,[18,70] we defined two assays as identical if the affinities of overlapping ligands reached a Pearson's correlation of 0.95. We also removed all ligands from the pre-training data that existed in FEP benchmarks to ensure a zero-shot setting. To reduce variability in evaluation metrics due to the limited number of targets in the FEP benchmarks, we trained five models with different random seeds for all our implemented methods (LigUnity, LigUnity(seq), Pocket-DTA, and Sequence-DTA), and reported the average evaluation metrics. The evaluation results ($\rho_s$) of all computational methods and structure-based methods on the Merck benchmark are from Shen et al.[32]

For few-shot experiments, the fine-tuning ligands for each test assay were randomly sampled with a uniform distribution, and we evaluated LigUnity's performance on the remaining ligands. We evaluated all methods with varying proportions of fine-tuning ligands ranging from 20% to 80%. Following previous methods,[18] we repeated the test experiments 40 times. For all our implemented methods (LigUnity, LigUnity(seq), Pocket-DTA, and Sequence-DTA) and DrugCLIP, we performed 10 steps of fine-tuning on the fine-tuning ligands using the same learning rate as in pre-training, and the final evaluation metrics were obtained by averaging the model's predictions on test ligands from the last 5 steps (steps 6-10).

To study the generalization ability of LigUnity on unseen chemical scaffolds and proteins, we investigated three increasingly challenging settings: (1) no similar ligands setting, where we excluded training ligands with more than 50% ECFP4 Tanimoto similarity to any test ligand, the similarity is computed using 2048-dimensional ECFP4 fingerprints generated by RDKit; (2) no similar proteins setting, where we excluded training proteins with more than 40% sequence similarity to any test protein (threshold set following AlphaFold3[3]), and we used cd-hit to compute protein sequence similarity defined as the number of identical residues after alignment; (3) no similar ligands and proteins setting, where we applied both criteria from the first two settings.

For calculating the importance score of each ligand atom and protein residue, we replaced atomic features with padding tokens and measured the resulting decrease in affinity prediction, where larger decreases indicated higher importance of the masked atoms. In this study, we used the model trained on the "no same ligands" setting, and we used the PDB structure 5EHR and its crystal ligand 5OD for the case study.

## Experiments under diverse settings

After evaluation on FEP benchmarks, we next sought to evaluate LigUnity on assays from ChEMBL[28] and BindingDB,[29] which contain diverse assays covering various aspects of assay types. We evaluated all methods under two assay split settings: split-by-time and split-by-unit. Under each assay split setting, we also investigated two ligand split settings: uniformly split at random and split-by-scaffold.

For the split-by-time setting, we first pre-trained all models using assays released before March 2019, and then conducted testing on assays released after March 2019. To evaluate generalization capability on truly unseen proteins, we only evaluated on assays containing previously unseen proteins. This test set comprised 161 assays, with an average of 48.6 ligands per assay.

For the split-by-unit setting, we collected 65 assays with unit "%" and assay type "Activity" from ChEMBL for testing, differing from our pre-training assays that have molar concentration units (e.g., nmol) or density units (e.g., $\mu$g/ml). This test set contains 65 assays in total, with an average of 30.5 ligands per assay. In this setting, we used the model trained in the split-by-time setting for evaluation.

For few-shot evaluation on both split-by-time and split-by-unit settings, we evaluated all methods with varying proportions of fine-tuning ligands ranging from 4 to 16. We repeated the test experiments 10 times. For all our implemented methods (LigUnity, LigUnity(seq), Pocket-DTA, and Sequence-DTA) and DrugCLIP, we performed 10 steps of fine-tuning on the fine-tuning ligands using the same learning rate as in pre-training, and the final evaluation metrics were obtained by averaging the model's predictions on test ligands from the last 5 steps (steps 6-10). For the split-by-scaffold setting on each test assay, we first generated Bemis-Murcko chemical scaffolds for all ligands using RDKit; then we iteratively selected molecules sharing the same scaffold as fine-tuning ligands until reaching the preset number, while using the remaining molecules for testing.

For evaluation on the PDE10A target, we used the dataset released by Roche Ltd. in 2022.[38] This dataset comprises 1,162 ligands with experimentally measured binding affinities and corresponding experimental timestamps. To simulate the realistic drug development scenarios, we followed the split-by-time setting at both assay and ligand levels (**Fig. 4e**). At the assay splitting level, we used the model pre-trained on assays published before March 2019, while the PDE10A dataset was released in 2022. At the ligand splitting level, we used 20% of the ligands measured earlier for fine-tuning and the remaining 80% of ligands measured later for testing. As the number of fine-tuning ligands is more than 100, we performed 20 steps of fine-tuning, and the final evaluation metrics were obtained by averaging the model's predictions on test ligands from the last 5 steps (steps 16-20).

## Active learning experiment on the TYK2 target

For the active learning experiment on the TYK2 target, we used the TYK2 kinase inhibitor dataset released by Google Research.[39] This dataset contains 10,000 ligands with binding free energies calculated using FEP. The dataset was initially constructed from 573 measured ligands found in public patents and was then expanded to 10,000 ligands through R-group assembly and filtering. The binding free energies were calculated using the Relatively Absolute Binding Free Energy (RABFE) protocol,[87] which achieved an $R^2$ of 0.41 with experimental affinities in validation tests. The entire calculation consumed 80,000 GPU hours (equivalent to 9.1 GPU years), which would cost 160,000$.

We implemented an active learning framework to evaluate LigUnity and competing methods. Our goal was to identify ligands with optimal binding free energy within specific FEP calculation constraints. The procedure consisted of the following steps: (1) Sampling a pool of ligands using

a defined selection strategy; (2) Calculating binding free energies for selected ligands using the FEP method and incorporating them into the training set (in this study, we directly used existing values from the dataset); (3) Training a model using ligands with known binding free energies; (4) Using the trained model to predict binding affinities for unmeasured ligands; (5) Repeating steps 1-4 until reaching a predetermined number of accessed ligands. This procedure simulates real-world drug discovery scenarios, providing insights into the practical utility of different methods.

For each iteration, we explored 100 ligands, starting with a randomly sampled initial training set of 100 ligands, and we conducted 20 active learning iterations in total. For each iteration, we performed 20 fine-tuning steps on the fine-tuning ligands. The final predictions were obtained by averaging the model's predictions on test ligands from the last 5 steps (steps 16-20). We evaluated each method using two types of metrics: (1) top 1%, 2%, and 5% recall, defined as the ratio of correctly identified top ligands to the total number of top ligands, and (2) $r^2$, calculated based on the ligands in the test set.

# Data availability

The datasets for training LigUnity were collected from ChEMBL version 34[28] and BindingDB version 2024m5.[29] Our processed Dekois 2.0 benchmark dataset is available on figshare https://doi.org/10.6084/m9.figshare.27967422. The LIT-PCBA and DUD-E benchmark datasets were obtained from Gao et al. https://github.com/bowen-gao/DrugCLIP.[17] The FEP calculation benchmarks[12,27] and the calculated results using FEP+(OPLS4)[12] were obtained from https://github.com/schrodinger/public_binding_free_energy_benchmark.[86] The experimental affinity data for the PDE10A data released by Roche Ltd. was obtained from Tosstorff et al.[38] The binding free energy dataset on the TYK2 target released by Google Research was obtained from Thompson et al.[39]

# Code availability

The code is available in our GitHub project https://github.com/IDEA-XL/LigUnity, including the model weights of LigUnity and relevant source code. We have included details of our experiments in the Methods to enable independent replication.

# Acknowledgments

We'd like to express our gratitude to Dr. Zequn Liu, Dr. Kangjie Zheng, and Dr. Junwei Yang from Peking University for their valuable discussions and assistance in this work. We are also grateful to Dr. Michael K. Gilson and Dr. Tiqing Liu from the University of California, San Diego (UCSD) for their expert guidance on data curation from BindingDB.

This project was supported by Shenzhen Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone, under Grant No. HTHZQSWS-KCCYB-2023052.

# References


[1] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

[2] Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

[3] Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **630**, 493 (2024).

[4] Du, X. *et al.* Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences* **17**, 144 (2016).

[5] Gorgulla, C. *et al.* An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663–668 (2020).

[6] Gentile, F. *et al.* Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. *Nature Protocols* **17**, 672–697 (2022).

[7] Zhou, G. *et al.* An artificial intelligence accelerated virtual screening platform for drug discovery. *Nature Communications* **15**, 7761 (2024).

[8] Heifetz, A., Southey, M., Morao, I., Townsend-Nicholson, A. & Bodkin, M. J. Computational methods used in hit-to-lead and lead optimization stages of structure-based drug discovery. *Computational Methods for GPCR Drug Discovery* 375–394 (2018).

[9] Huang, L. *et al.* A dual diffusion model enables 3d molecule generation and lead optimization based on target pockets. *Nature Communications* **15**, 2657 (2024).

[10] Cerqueira, N. M. *et al.* Receptor-based virtual screening protocol for drug discovery. *Archives of biochemistry and biophysics* **582**, 56–67 (2015).

[11] Pagadala, N. S., Syed, K. & Tuszynski, J. A. Software for molecular docking: a review. *Biophysical Reviews* **9**, 91 – 102 (2017). URL https://api.semanticscholar.org/CorpusID:33687753.

[12] Wang, L. *et al.* Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society* **137**, 2695–2703 (2015).

[13] Wang, E. *et al.* End-point binding free energy calculation with mm/pbsa and mm/gbsa: strategies and applications in drug design. *Chemical reviews* **119**, 9478–9508 (2019).

[14] Cournia, Z., Allen, B. K. & Sherman, W. Relative binding free energy calculations in drug discovery: Recent advances and practical considerations. *Journal of chemical information and modeling* **57 12**, 2911–2937 (2017). URL https://api.semanticscholar.org/CorpusID:25855226.

[15] Wang, E. *et al.* End-point binding free energy calculation with mm/pbsa and mm/gbsa: Strategies and applications in drug design. *Chemical reviews* (2019). URL https://api.semanticscholar.org/CorpusID:195696542.

[16] Shen, C. *et al.* Boosting protein–ligand binding pose prediction and virtual screening based on residue–atom distance likelihood potential and graph transformer. *Journal of Medicinal Chemistry* **65**, 10691–10706 (2022).

[17] Gao, B. *et al.* DrugCLIP: Contrastive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems* (2023).

[18] Feng, B. *et al.* A bioactivity foundation model using pairwise meta-learning. *Nature Machine Intelligence* (2024). URL https://doi.org/10.1038/s42256-024-00876-w.

[19] Yu, J. *et al.* Computing the relative binding affinity of ligands based on a pairwise binding comparison network. *Nature Computational Science* **3**, 860–872 (2023).

[20] Hughes, J. P., Rees, S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *British journal of pharmacology* **162**, 1239–1249 (2011).

[21] Luo, D., Liu, D., Qu, X., Dong, L. & Wang, B. Enhancing generalizability in protein-ligand binding affinity prediction with multimodal contrastive learning. *Journal of chemical information and modeling* (2024). URL https://api.semanticscholar.org/CorpusID:268238440.

[22] Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F. & Li, H. *Learning to rank: from pairwise approach to listwise approach* (2007). URL https://api.semanticscholar.org/CorpusID:207163577.

[23] Ru, X., Ye, X., Sakurai, T. & Zou, Q. Nerltr-dta: drug-target binding affinity prediction based on neighbor relationship and learning to rank. *Bioinformatics* (2022). URL https://api.semanticscholar.org/CorpusID:246678858.

[24] Mysinger, M. M., Carchia, M., Irwin, J. J. & Shoichet, B. K. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry* **55**, 6582–6594 (2012).

[25] Vogel, S. M., Bauer, M. R. & Boeckler, F. M. Dekois: Demanding evaluation kits for objective in silico screening–a versatile tool for benchmarking docking programs and scoring functions. *Journal of chemical information and modeling* **51**, 2650–2665 (2011).

[26] Tran-Nguyen, V.-K., Jacquemard, C. & Rognan, D. Lit-pcba: an unbiased data set for machine learning and virtual screening. *Journal of chemical information and modeling* **60**, 4263–4273 (2020).

[27] Schindler, C. E. *et al.* Large-scale assessment of binding free energy calculations in active drug discovery projects. *Journal of Chemical Information and Modeling* **60**, 5457–5474 (2020).

[28] Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research* **52**, D1180–D1192 (2024).

[29] Gilson, M. K. *et al.* Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* **44**, D1045–D1053 (2016).

[30] Friesner, R. A. *et al.* Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry* **47**, 1739–1749 (2004).

[31] Trott, O. & Olson, A. J. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **31**, 455–461 (2010).

[32] Shen, C. *et al.* A generalized protein–ligand scoring framework with balanced scoring, docking, ranking and screening powers. *Chemical Science* **14**, 8129–8146 (2023).

[33] Krasoulis, A., Antonopoulos, N., Pitsikalis, V. & Theodorakis, S. Denvis: scalable and high-throughput virtual screening using graph neural networks with atomic and surface protein pocket features. *Journal of Chemical Information and Modeling* **62**, 4642–4659 (2022).

[34] Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).

[35] Irwin, J. J. *et al.* Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of chemical information and modeling* **60**, 6065–6073 (2020).

[36] Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9** (2008).

[37] Lyne, P. D., Lamb, M. L. & Saeh, J. C. Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and mm-gbsa scoring. *Journal of medicinal chemistry* **49**, 4805–4808 (2006).

[38] Tosstorff, A. *et al.* A high quality, industrial data set for binding affinity prediction: performance comparison in different early drug discovery scenarios. *Journal of Computer-Aided Molecular Design* **36**, 753–765 (2022).

[39] Thompson, J. *et al.* Optimizing active learning for free energy calculations. *Artificial Intelligence in the Life Sciences* **2**, 100050 (2022).

[40] Reker, D. & Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug discovery today* **20**, 458–465 (2015).

[41] Huang, K. *et al.* Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics* **36**, 5545–5547 (2020).

[42] Chatterjee, A. *et al.* Improving the generalizability of protein-ligand binding predictions with ai-bind. *Nature communications* **14**, 1989 (2023).

[43] Luo, Y., Liu, Y. & Peng, J. Calibrated geometric deep learning improves kinase–drug binding predictions. *Nature Machine Intelligence* **5**, 1390–1401 (2023).

[44] Pei, Q. *et al.* Breaking the barriers of data scarcity in drug–target affinity prediction. *Briefings in Bioinformatics* **24**, bbad386 (2023).

[45] Nguyen, T. *et al.* GraphDTA: predicting drug–target binding affinity with graph neural networks. *Bioinformatics* **37**, 1140–1147 (2021).

[46] Zhao, Q., Xiao, F., Yang, M., Li, Y. & Wang, J. *AttentionDTA: prediction of drug–target binding affinity using attention model*, 64–69 (IEEE, 2019).

[47] Yang, Z., Zhong, W., Zhao, L. & Chen, C. Y.-C. MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical science* **13**, 816–833 (2022).

[48] Lin, S., Shi, C. & Chen, J. Generalizeddta: combining pre-training and multi-task learning to predict drug-target binding affinity for unknown drug discovery. *BMC bioinformatics* **23**, 367 (2022).

[49] Yuan, W., Chen, G. & Chen, C. Y.-C. FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. *Briefings in Bioinformatics* **23**, bbab506 (2022).

[50] Zhang, X. *et al.* Efficient and accurate large library ligand docking with karmadock. *Nature Computational Science* **3**, 789–804 (2023).

[51] Cao, D. *et al.* Generic protein–ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. *Nature Machine Intelligence* 1–13 (2024).

[52] Durrant, J. D. & McCammon, J. A. Nnscore 2.0: a neural-network receptor–ligand scoring function. *Journal of chemical information and modeling* **51**, 2897–2903 (2011).

[53] Ballester, P. J. & Mitchell, J. B. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* **26**, 1169–1175 (2010).

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

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

[56] Yang, C. & Zhang, Y. Delta machine learning to improve scoring-ranking-screening performances of protein–ligand scoring functions. *Journal of Chemical Information and Modeling* **62**, 2696–2712 (2022).

[57] Stepniewska-Dziubinska, M. M., Zielenkiewicz, P. & Siedlecki, P. Pafnucy - a deep neural network for structure-based drug discovery. *ArXiv* **abs/1712.07042** (2017).

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

[59] Moon, S., Zhung, W., Yang, S., Lim, J. & Kim, W. Y. Pignet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science* **13**, 3661–3673 (2022).

[60] Zhang, X. *et al.* Planet: a multi-objective graph neural network model for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling* **64**, 2205–2220 (2023).

[61] McNutt, A. T. *et al.* Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics* **13**, 43 (2021).

[62] Kalliokoski, T., Kramer, C., Vulpetti, A. & Gedeck, P. Comparability of mixed ic50 data–a statistical analysis. *PloS one* **8**, e61007 (2013).

[63] Zhou, G. *et al.* *Uni-mol: A universal 3d molecular representation learning framework* (2023).

[64] Li, H. *et al.* Decoupled peak property learning for efficient and interpretable electronic circular dichroism spectrum prediction. *Nature Computational Science* 1–11 (2025).

[65] Mayr, A. *et al.* Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chemical Science* **9**, 5441 – 5451 (2018). URL https://api.semanticscholar.org/CorpusID:52110581.

[66] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

[67] Zardecki, C., Dutta, S., Goodsell, D. S., Voigt, M. & Burley, S. K. Rcsb protein data bank: A resource for chemical, biochemical, and structural explorations of large and small biomolecules. *Journal of Chemical Education* **93**, 569–575 (2016).

[68] Maggiora, G., Vogt, M., Stumpfe, D. & Bajorath, J. Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry* **57**, 3186–3204 (2014). URL https://doi.org/10.1021/jm401411z. PMID: 24151987.

[69] Yang, J., Roy, A. & Zhang, Y. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research* **41**, D1096–D1103 (2012).

[70] Martin, E. J. *et al.* All-assay-Max2 pQSAR: activity predictions as accurate as four-concentration ic50s for 8558 novartis assays. *Journal of chemical information and modeling* **59**, 4450–4459 (2019).

[71] Berman, H. M. *et al.* The Protein Data Bank. *Nucleic acids research* **28**, 235–242 (2000).

[72] Wang, R., Fang, X., Lu, Y. & Wang, S. The pdbbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry* **47**, 2977–2980 (2004).

[73] Huey, R., Morris, G. M., Olson, A. J. & Goodsell, D. S. A semiempirical free energy force field with charge-based desolvation. *Journal of computational chemistry* **28**, 1145–1152 (2007).

[74] Quiroga, R. & Villarreal, M. A. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PloS one* **11**, e0155183 (2016).

[75] Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein–ligand docking using gold. *Proteins: Structure, Function, and Bioinformatics* **52**, 609–623 (2003).

[76] Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of medicinal chemistry* **46**, 499–511 (2003).

[77] Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A fast flexible docking method using an incremental construction algorithm. *Journal of molecular biology* **261**, 470–489 (1996).

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

[79] Lu, W. *et al.* Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems* **35**, 7236–7249 (2022).

[80] Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**, 301–320 (2005).

[81] Williams, C. & Rasmussen, C. Gaussian processes for regression. *Advances in neural information processing systems* **8** (1995).

[82] DiFranzo, A., Sheridan, R. P., Liaw, A. & Tudor, M. Nearest neighbor gaussian process for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling* **60**, 4653–4663 (2020).

[83] Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J. & Gifford, E. M. Extreme gradient boosting as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling* **56**, 2353–2360 (2016).

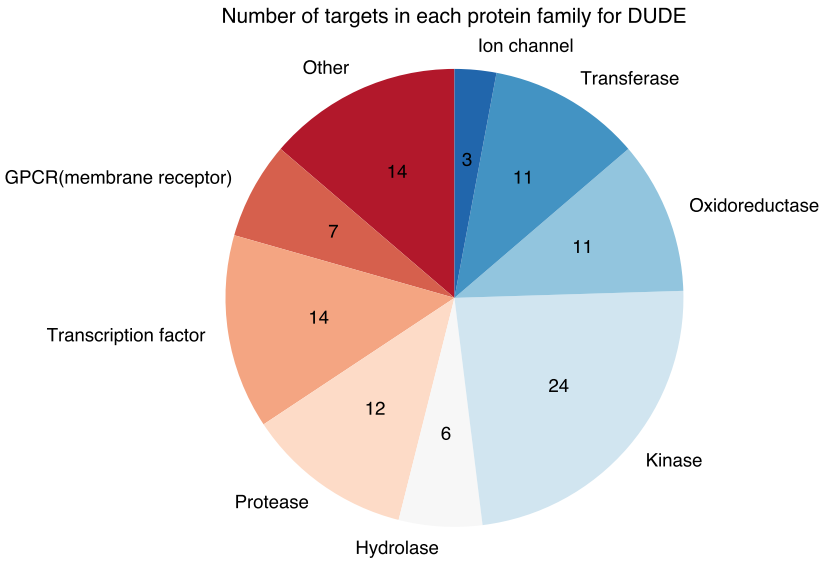
[84] Svetnik, V. *et al.* Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences* **43**, 1947–1958 (2003).

[85] Lu, C. *et al.* OPLS4: Improving force field accuracy on challenging regimes of chemical space. *Journal of chemical theory and computation* **17**, 4291–4300 (2021).

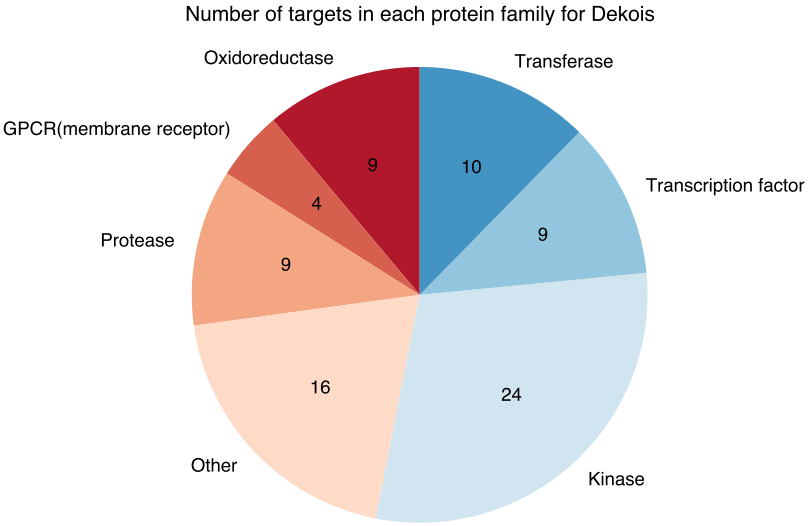
[86] Ross, G. A. *et al.* The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Communications Chemistry* **6**, 222 (2023).

[87] Rocklin, G. J., Mobley, D. L. & Dill, K. A. Separated topologies—a method for relative binding free energy calculations using orientational restraints. *The Journal of chemical physics* **138** (2013).

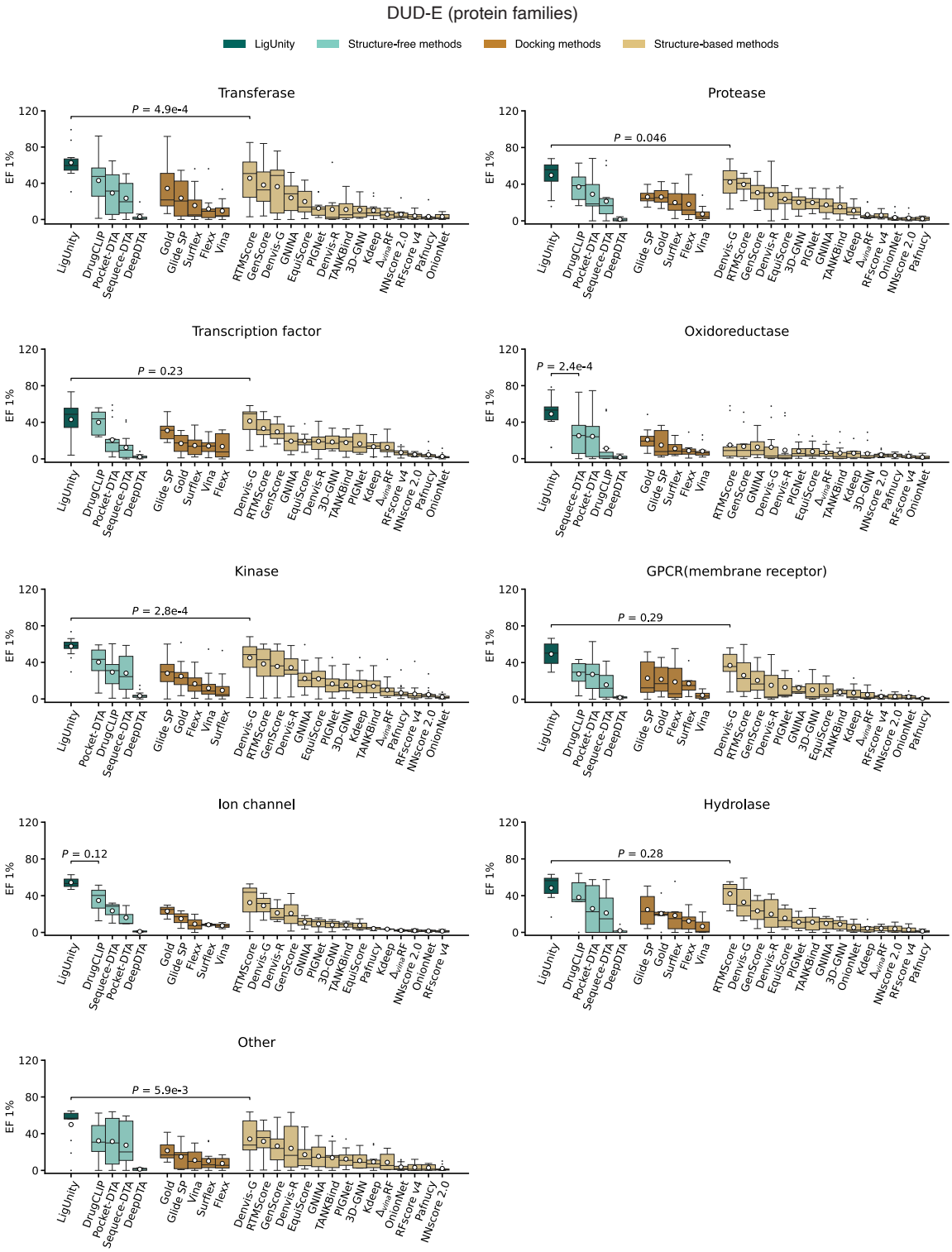
# Supplementary figures



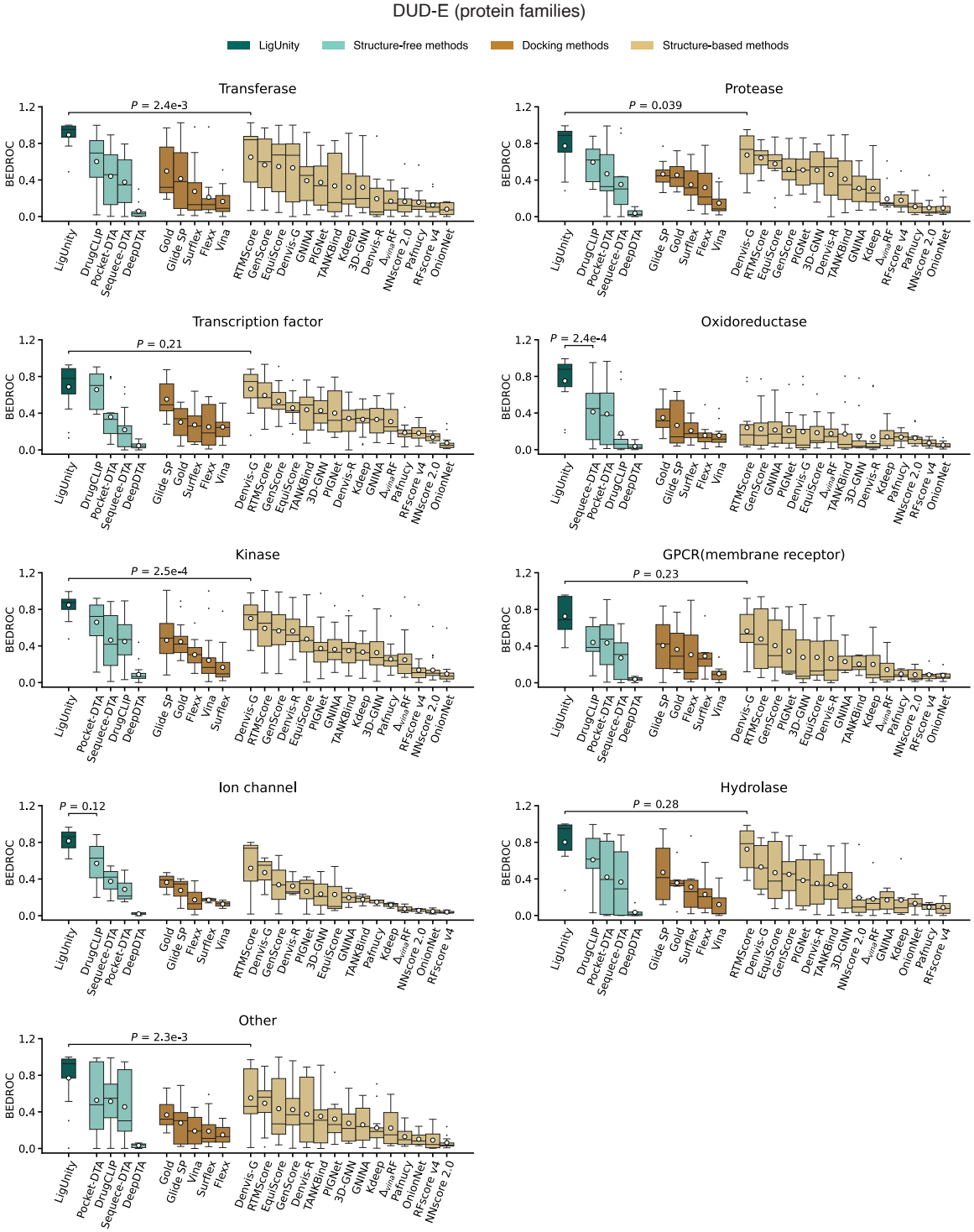

**Supplementary Figure 1:** Pie chart showing the number of targets for each protein family in the DUD-E benchmark. The protein families come from ChEMBL.[28]



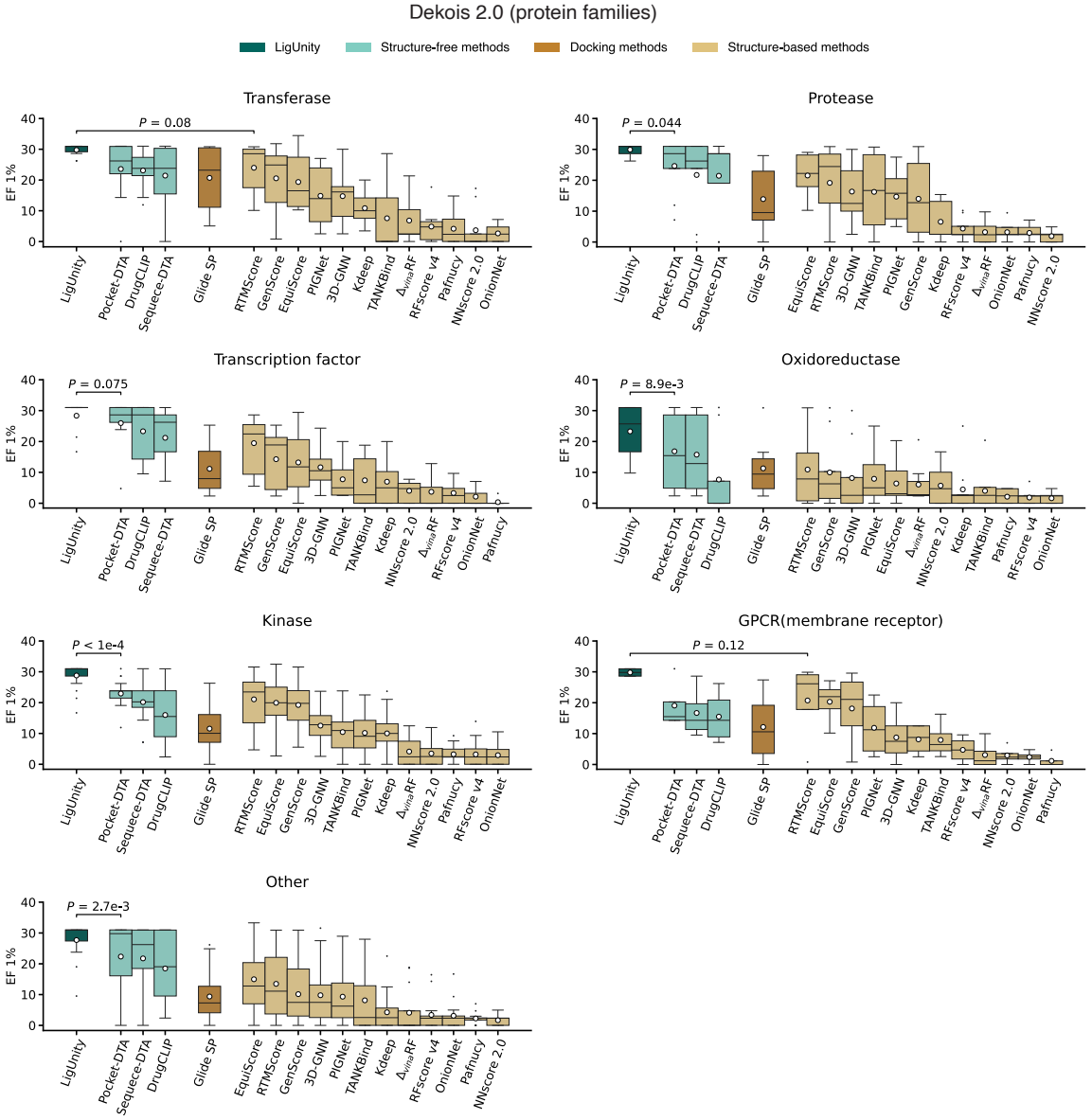

**Supplementary Figure 2:** Pie chart showing the number of targets for each protein family in the Dekois 2.0 benchmark. The protein families come from ChEMBL.[28]



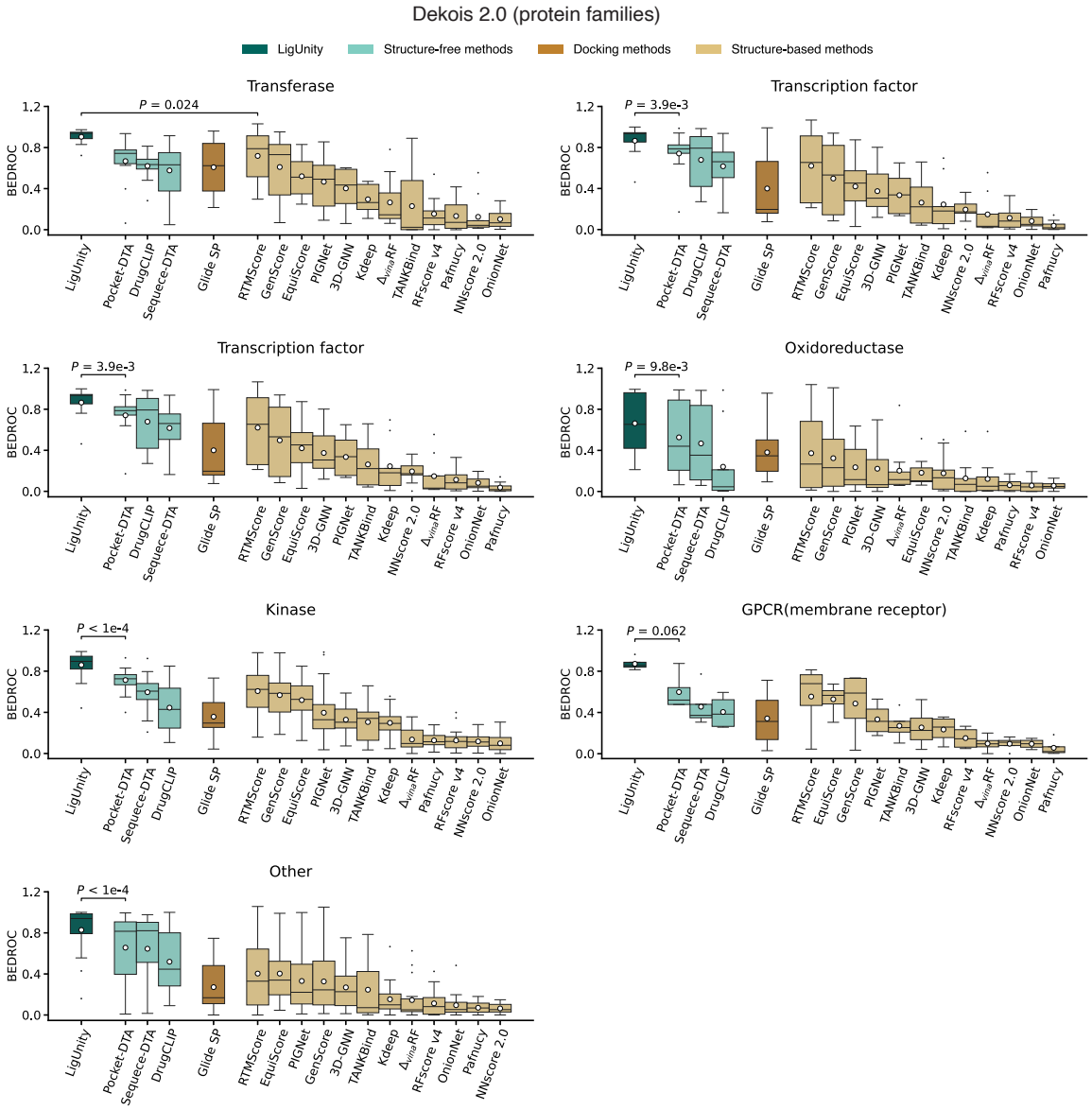

**Supplementary Figure 3:** Box plots comparing LigUnity and competing methods on the DUD-E benchmark in terms of enrichment factor (EF) 1%. Targets of different protein families are tested separately.



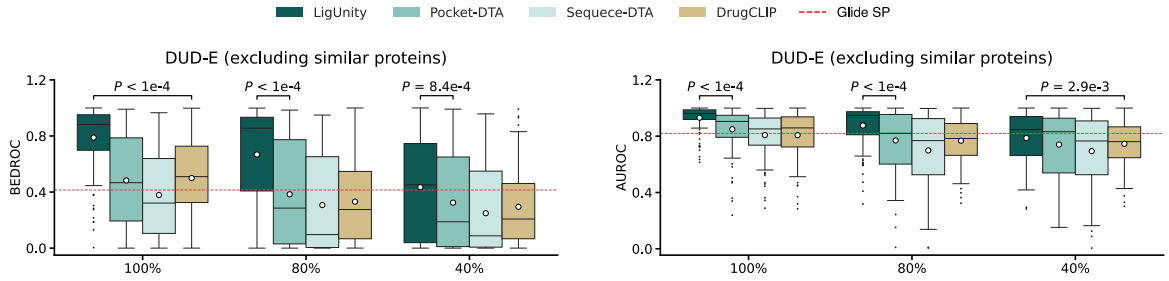

**Supplementary Figure 4:** Box plots comparing LigUnity and competing methods on the DUD-E benchmark in terms of BEDROC. Targets of different protein families are tested separately.



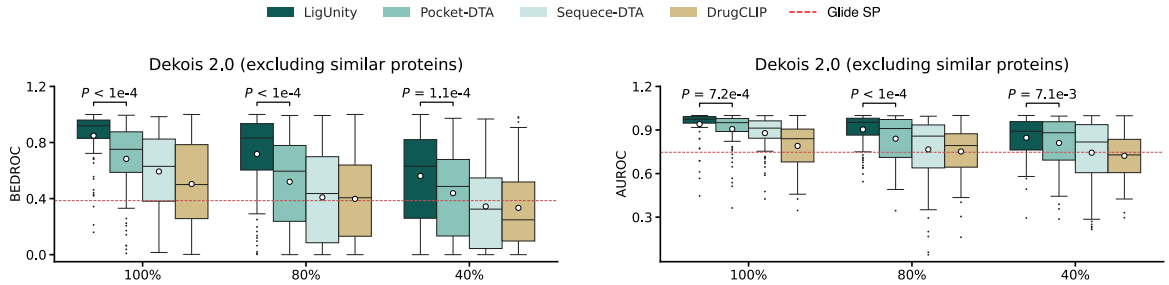

**Supplementary Figure 5:** Box plots comparing LigUnity and competing methods on the Dekois 2.0 benchmark in terms of enrichment factor (EF) 1%. Targets of different protein families are tested separately.



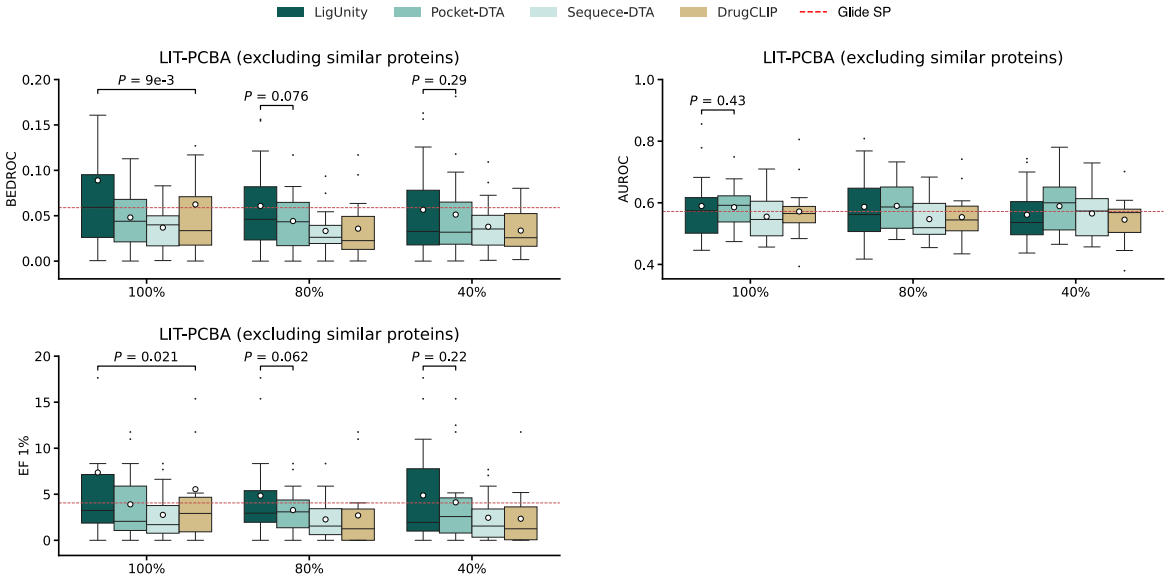

**Supplementary Figure 6:** Box plots comparing LigUnity and competing methods on the Dekois 2.0 benchmark in terms of BEDROC. Targets of different protein families are tested separately.



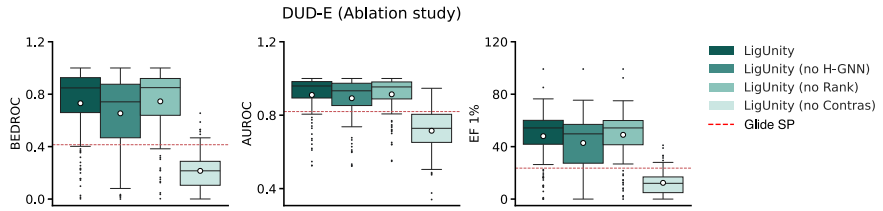

**Supplementary Figure 7:** Box plots comparing LigUnity and competing methods on the DUD-E benchmark in terms of BEDROC and AU-ROC using different protein training sets. The $x$-axis denotes the maximum sequence similarity between the training and test sets. The mean values (white dots) are calculated across n = 102 targets.



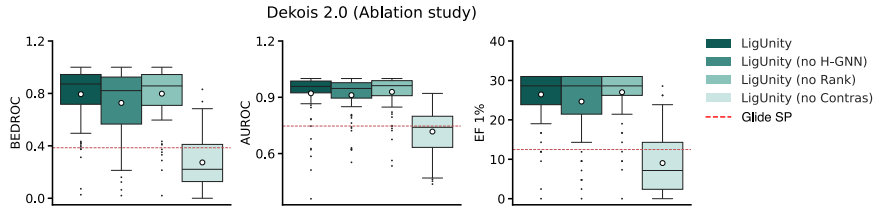

**Supplementary Figure 8:** Box plots comparing LigUnity and competing methods on the Dekois 2.0 benchmark in terms of BEDROC and AU-ROC using different protein training sets. The $x$-axis denotes the maximum sequence similarity between the training and test sets. The mean values (white dots) are calculated across n = 81 targets.



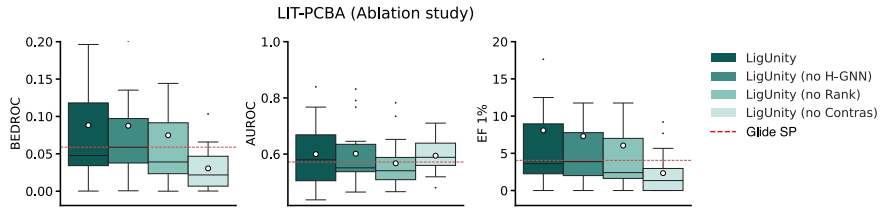

**Supplementary Figure 9:** Box plots comparing LigUnity and competing methods on the LIT-PCBA benchmark in terms of enrichment factor (EF) 1%, BEDROC, and AU-ROC using different protein training sets. The $x$-axis denotes the maximum sequence similarity between the training and test sets. The mean values (white dots) are calculated across n = 15 targets.



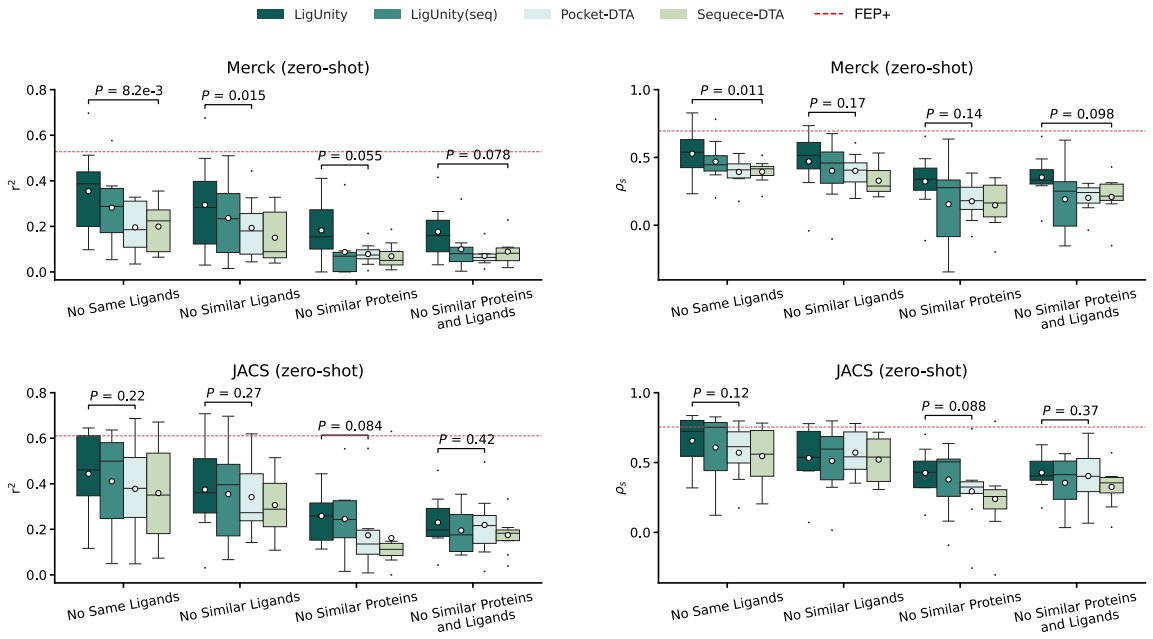

**Supplementary Figure 10:** Box plots of ablation study evaluating the contribution of different modules on the DUD-E benchmark in terms of enrichment factor (EF) 1%, BEDROC, and AU-ROC. The mean values (white dots) are calculated across n = 102 targets.



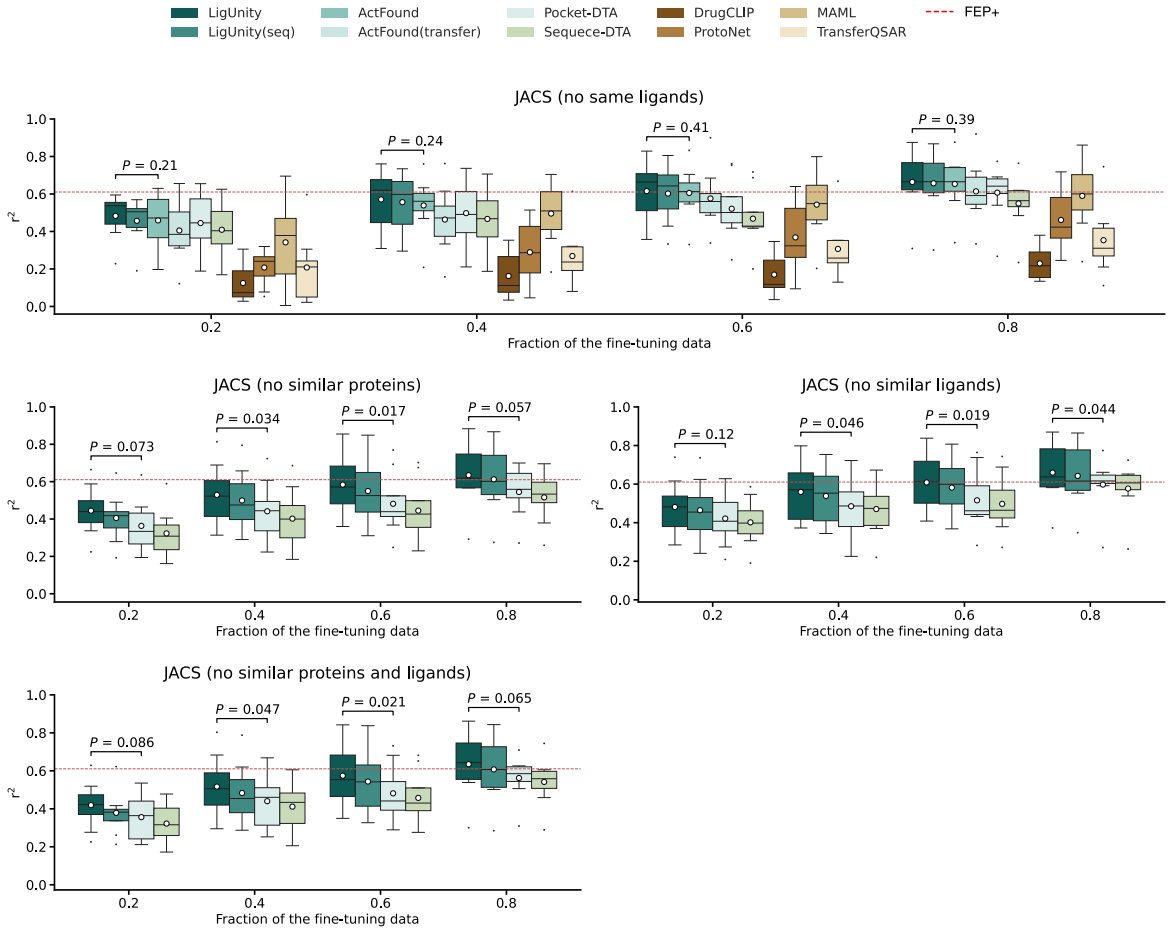

**Supplementary Figure 11:** Box plots of ablation study evaluating the contribution of different modules on the Dekois 2.0 benchmark in terms of enrichment factor (EF) 1%, BEDROC, and AU-ROC. The mean values (white dots) are calculated across n = 81 targets.



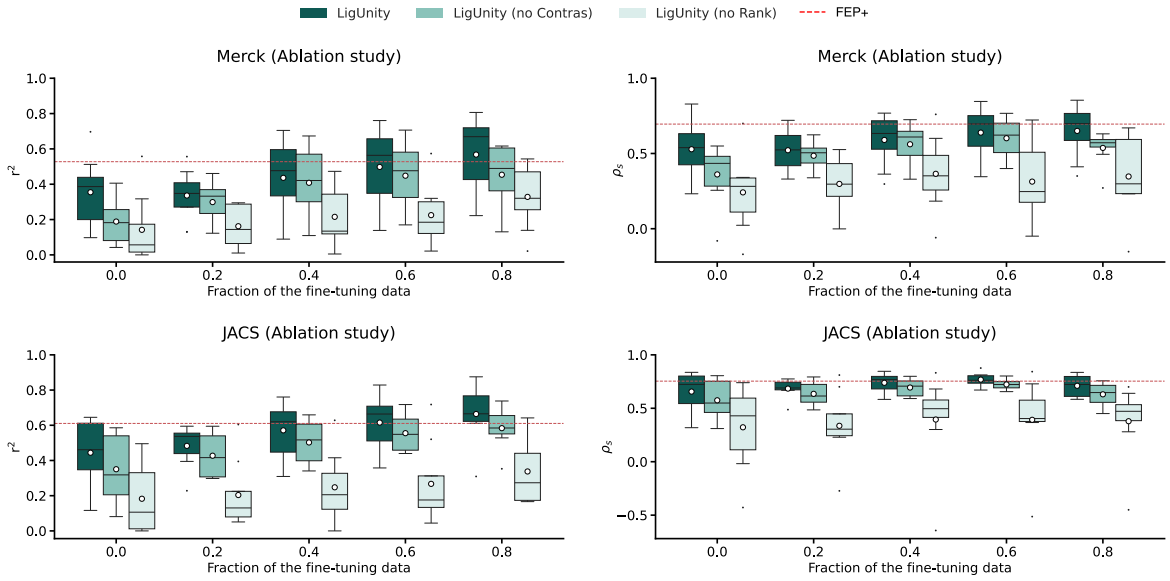

**Supplementary Figure 12:** Box plots of ablation study evaluating the contribution of different modules on the LIT-PCBA benchmark in terms of enrichment factor (EF) 1%, BEDROC, and AU-ROC. The mean values (white dots) are calculated across n = 15 targets.



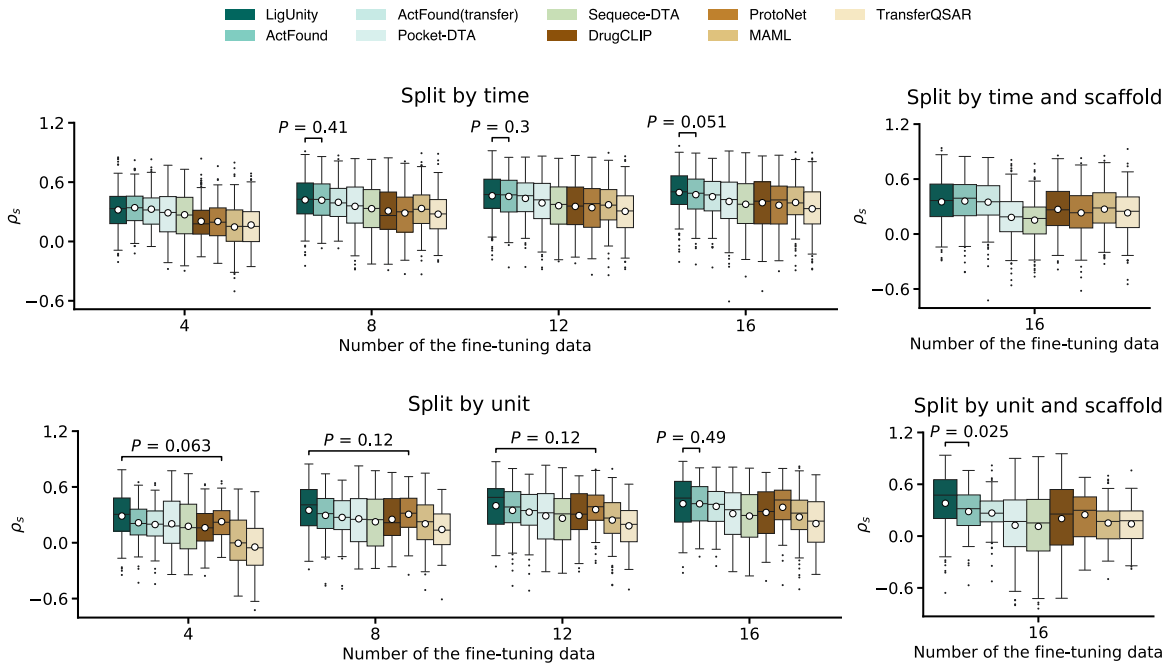

**Supplementary Figure 13:** Box plots comparing LigUnity and competing methods across different settings on Merck and JACS benchmarks in terms of $r^2$ and Spearman's rank correlation ($\rho_s$). The mean values (white dots) are calculated across n = 8 targets.



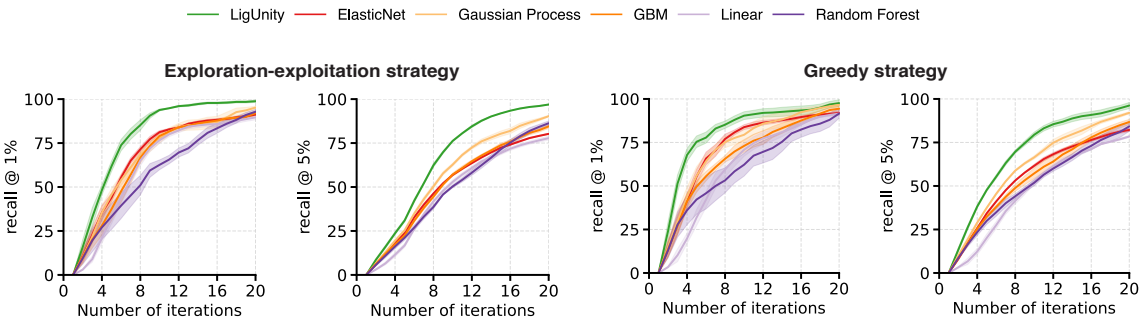

**Supplementary Figure 14:** Box plots comparing LigUnity and competing methods across different settings on the JACS benchmark in terms of $r^2$ and Spearman's rank correlation ($\rho_s$) when 20%, 40%, 60%, 80% of the experimental binding affinities are used for fine-tuning. The mean values (white dots) are calculated across n = 8 targets.




**Supplementary Figure 15:** Box plots of ablation study evaluating the contribution of different modules on Merck and JACS benchmarks in terms of $r^2$ and Spearman's rank correlation ($\rho_s$) when 0% (zero-shot), 20%, 40%, 60%, 80% of the experimental binding affinities are used for fine-tuning. The mean values (white dots) are calculated across n = 8 targets.




**Supplementary Figure 16:** Box plots comparing the binding affinity prediction on the split-by-time and split-by-unit setting in terms of $\rho_s$ when 4, 8, 12, and 16 of the experimental binding affinity is used for fine-tuning. The mean values (white dots) are calculated across n = 161 assays for split-by-time setting and n = 65 assays for split-by-time setting.




**Supplementary Figure 17:** Plots comparing LigUnity and competing methods on the TYK2 dataset in terms of top 1% recall and top 5% recall when using the exploration-exploitation strategy (left) and greedy selection strategy (right).