Multimodal Protein-Ligand Contrastive Pretraining for Effective and Efficient Drug Discovery

Zhen Wang[1†], Zhanfeng Wang[1†], Maohua Yang[1†], Long Pang[1], Fangyuan Nie[1], Siyuan Liu[1], Zhifeng Gao[1], Guojiang Zhao[1], Xiaohong Ji[1], Dandan Huang[1], Zhengdan Zhu[1], Dongdong Li[1], Yannan Yuan[1], Hang Zheng[1], Linfeng Zhang[1], Guolin Ke[1*], Dongdong Wang[1*], Feng Yu[1*]

[1]DP Technology, Beijing 100089, China.

*Corresponding author(s). E-mail(s): kegl@dp.tech; wangdd@dp.tech; yufeng@dp.tech;

[†]These authors contributed equally to this work.

Abstract

Accurate modeling of protein-ligand interactions (PLIs) is critical for drug discovery. Despite advancements, most existing PLIs modeling methods rely on single-modal data, restricting their effectiveness and applicability. In this study, we introduce Uni-Clip, a contrastive learning framework that incorporates multi-modalities, specifically structure and residue features of proteins, along with conformation and graph features of ligands. Through optimization with specifically designed CF-InfoNCE loss, Uni-Clip achieves comprehensive representations for PLIs. Uni-Clip demonstrates superior performance in benchmark evaluations on widely acknowledged datasets, LIT-PCBA and DUD-E, achieving a 147% and 218% improvements in enrichment factors at 1% compared to baselines. Furthermore, Uni-Clip serves as a practical tool for various applications in drug discovery, as demonstrated through virtual screening for a flat and challenging protein target GPX4, where it identified potent inhibitors with an IC_{50} of 4.17 µM, and through target fishing for benzbromarone, which highlights the potential for repurposing benzbromarone in cancer therapy.

1 Introduction

Protein-ligand interactions are crucial for essential biological processes such as signal transduction, molecular recognition, and drug action. The precise computational prediction and ranking of PLIs contribute substantially to tasks such as virtual screening, target fishing, and lead compound optimization [1–3]. With scalable computational capabilities and the ability to handle nonlinearities, various deep learning models have been developed to model PLIs [4–11]. The performance of deep learning models for PLIs heavily depends on the quality and quantity of training data [12]. Despite advancements in structure determination technologies such as high-throughput X-ray crystallography, cryo-electron microscopy (cryo-EM), and artificial intelligence exemplified by AlphaFold, the dataset of highly reliable protein-ligand complex structures remains limited [13–15]. Additionally, the experimentally measured affinity values exhibit considerable uncertainties and noise due to differences in experimental setups and measurement techniques [16].

Contrastive learning, recognized as a self-supervised learning paradigm, has achieved notable success in the domains of computer vision and natural language processing [17, 18]. Inspired by the success of models like SimCLR [19] and CLIP [20], recent studies have increasingly applied contrastive learning to model PLIs [10, 11, 21]. Unlike regression or classification methods that directly analyze protein-ligand complexes to predict PLIs, contrastive learning encodes proteins and ligands separately and uses a contrastive loss function to enhance the alignment of positive pairs and the differentiation of negative pairs. This approach improves the efficiency in utilizing PLI affinity data. Furthermore, by judiciously selecting positive sample pairs and adjusting the temperature parameter, contrastive learning demonstrates potential robustness against label noise [22].

Considering the data modalities, the deep learning models for PLIs can be broadly divided into two categories: one based on the 1D sequence and 2D graph data of proteins and ligands [4–7, 10], and the other on the 3D structure data of these entities [8, 9, 11]. Despite achieving promising results on benchmark assessments, most of these models face inherent limitations due to their reliance on single-modal data of proteins and ligands. These models lack adaptability, especially for PLI tasks that require multimodal data. For example, specific structural conformations and inter-domain interactions, which are crucial for some protein-ligand interactions, present important effects that are challenging to predict accurately from sequence data alone. Models based on the 3D structures of protein-ligand complexes tend to exhibit reduced accuracy when processing highly dynamic protein-ligand interactions [23]. Many "undruggable" proteins, such as transcription factors with flat interaction surfaces and intrinsically disordered proteins with highly dynamic structures, are found to lack pharmacologically targetable pockets on their surfaces [24, 25].

In this paper, we propose a contrastive protein-ligand pretraining (Uni-Clip) model, a multimodal contrastive learning framework developed to model PLIs. Uni-Clip concurrently processes protein structure and residue data, as well as ligand conformation and graph data, via the integration of diverse pretrained encoders, thereby encompassing the data features handled by other baselines, as shown in Fig. 1c. Subsequently,

we have developed the collision-free information noise-contrastive estimation (CF-InfoNCE) loss function, enabling Uni-Clip to more adeptly align the representations of protein-ligand pairs within a unified feature space, substantially enhancing the prediction of their interactions. Furthermore, we constructed the largest Multimodal Protein Ligand Binding (MMBind) dataset to our knowledge, specifically designed for the training of Uni-Clip. Consequently, we have shown that Uni-Clip substantially enhances performance compared to previous methods across two benchmark datasets in the field of virtual screening, while also maintaining optimal speed. We validated the efficacy of Uni-Clip through a practical drug discovery case targeting Glutathione Peroxidase 4 (GPX4), identifying DP021 as an effective non-covalent inhibitor with confirmed binding affinity. Uni-Clip also demonstrated its capability in target fishing using benzbromarone, identifying c-Myc as a potential target and showcasing its anti-cancer potential. Our explainable analyses of Uni-Clip effectively identified binding sites consistent with molecular dynamics and experimental results, even for challenging targets with low druggability or intrinsically disordered proteins like c-Myc. In summary, this research highlights Uni-Clip as a promising tool not only for achieving strong performance in standard benchmarks, but also for accelerating various drug design scenarios.

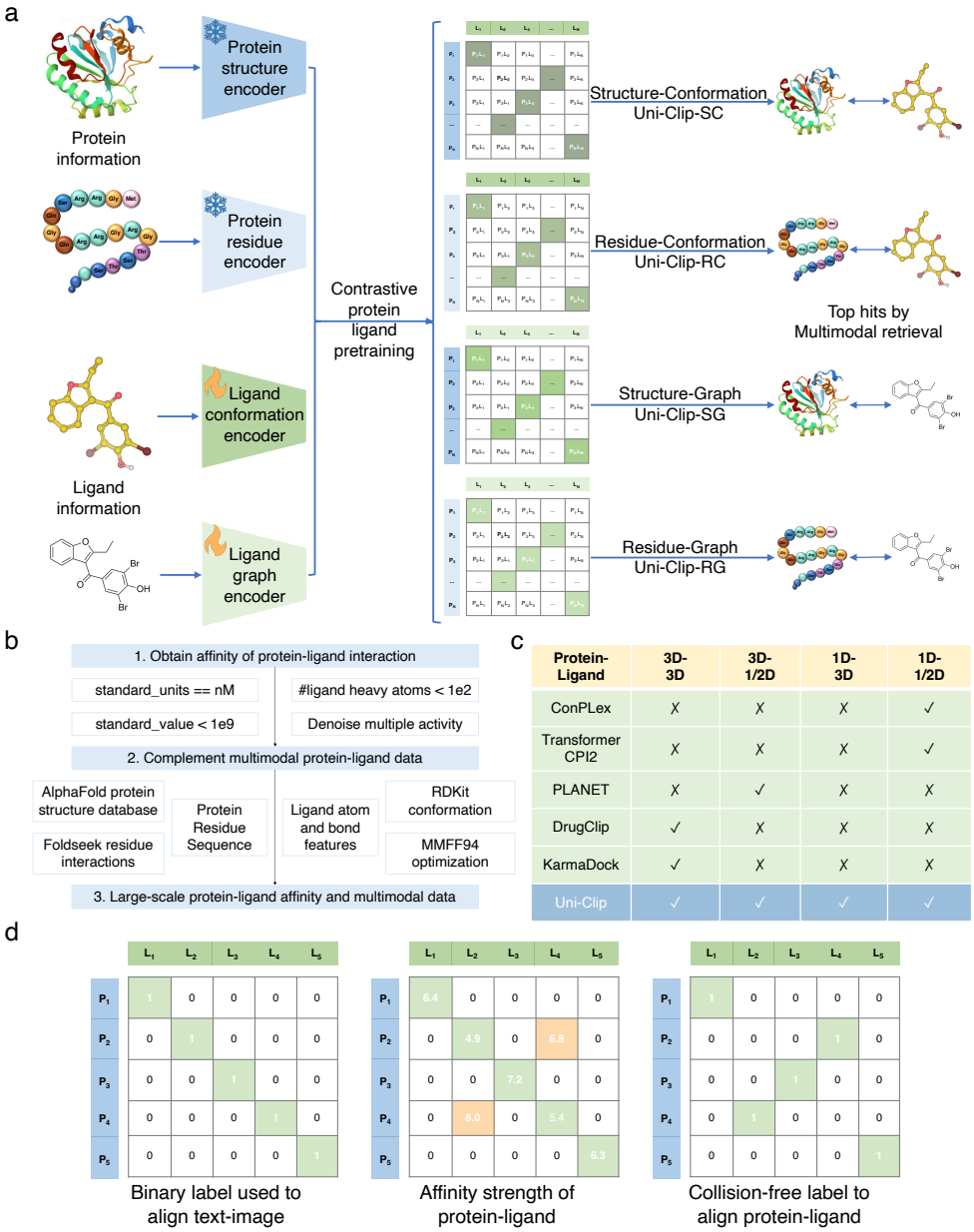
2 Results

2.1 Overview of Uni-Clip framework

We developed Uni-Clip, comprising approximately 100 million parameters, trained on a dataset of over 850,000 PLI data points. Uni-Clip represents a large-scale model with extensive protein-ligand coverage in the field of PLIs. The effective training of such a model was facilitated by three key components: model design, data curation and training methodology (detail ablation studies of these components are in Supplementary Notes).

The overall framework of Uni-Clip is illustrated in Fig. 1a. There are four distinct encoders: frozen SaProt [26], ESM2 [27], pretrained Uni-Mol [28] and pretrained Uni-Mol-Graph (detailed in Methods 4.2), which respectively encode protein *structure*, protein *residue*, ligand *conformation* and ligand *graph* multimodalities data, which are abbreviated as S, R, C and G. Subsequently, we apply contrastive learning to align the protein structure and residue encoders with the ligand conformation and graph encoders. This alignment leads to four sub-modules: Uni-Clip-SC, Uni-Clip-RC, Uni-Clip-SG and Uni-Clip-RG. By jointly optimizing these four sub-modules, we achieve the final Uni-Clip model, which is adept at handling multi-modal features across protein and ligand domains.

We constructed the MMBind dataset from publicly available resources, including the ChEMBL bioactivity database [29] and the AlphaFold protein structure database [30] as depicted in Fig. 1b. The ChEMBL database provides binding affinities of pairs of protein residue sequences and ligand SMILES. We then aligned multimodal data to each protein-ligand pair, including the protein structure data from the AlphaFold protein structure database and ligand conformation and graph data via RDKit. After data



Protein-Ligand	3D-3D	3D-1/2D	1D-3D	1D-1/2D
ConPLex	✗	✗	✗	✓
Transformer CPI2	✗	✗	✗	✓
PLANET	✗	✓	✗	✗
DrugClip	✓	✗	✗	✗
KarmaDock	✓	✗	✗	✗
Uni-Clip	✓	✓	✓	✓

Fig. 1: Uni-Clip predicts protein-ligand interactions using multimodal protein-ligand data. **a** The computational pipeline of Uni-Clip. **b** The data curation pipeline of large-scale multimodal protein-ligand data and affinity label. **c** Comparative analysis of input modalities in baselines and Uni-Clip for PLIs. **d** Our proposed collision-free label matrix aligning protein-ligand vs. vanilla binary label used to align text-image.

quality control, we obtained over 850,000 protein-ligand affinity data points (detailed in Methods 4.1).

We designed a training methodology for protein-ligand contrastive learning with CF-InfoNCE loss, an extension of InfoNCE loss [31], considering that 15% labels on the off-diagonal dominate over corresponding labels on the diagonal on average during training on the MMBind dataset. In contrast, InfoNCE hypothesizes that labels on the diagonal dominate over others in the corresponding row and column, as illustrated on the left side of Fig. 1d. Among the over 850k protein-ligand pairs in the MMBind dataset, there are over 529k ligands but only 5.2k proteins. Consequently a protein can form strong interactions with multiple ligands and vice versa when aligning protein-ligand, as depict in the middle of Fig. 1d. It is important to emphasize that there are two scenarios of collisions in a batch, one is that there may be identical proteins or ligands, and the other is that there may be multiple different ligands (or proteins) binding to the same protein (or ligand). The observed collisions (colored in golden in Fig. 1d) inspired us to develop CF-InfoNCE loss, enabling Uni-Clip to better align a batch of protein-ligand data. To enhance robustness and efficacy from contrastive learning, we only keep the maximal one as shown in the right of Fig. 1d, ensuring that there is one prominent label and significantly smaller remaining values in each row or column. The collision-free label matrix results in row-wise and column-wise normalizations, which separately guide protein-ligand and ligand-protein alignments (detailed in Methods 4.3).

2.2 The performance of Uni-Clip on benchmarks

To evaluate Uni-Clip's performance in virtual screening, we conducted a comparative analysis against several baseline methods: Vina [32], Glide-SP [33], TransformerCPI2 (TCPI2) [7], PLANET [8], and DrugCLIP [11]. Vina and Glide-SP are among the most commonly used molecular docking tools for virtual screening. TCPI2, PLANET, and DrugCLIP are deep learning models designed to utilize different modal representations: 1D sequence-based, 2D graph-based, and 3D structure features, respectively. The efficacy of these tools was benchmarked against two widely acknowledged datasets: DUD-E [34] and LIT-PCBA [35]. To assess screening performance, the Area Under the Receiver Operating Characteristic Curve (AUROC) and enrichment factors ($EF_{0.5\%}$, $EF_{1\%}$, and $EF_{5\%}$) were applied (detailed in Methods Section 4.5). The enrichment factors were calculated as the ratio of true actives in the selected subset versus the entire dataset at sampling rates of 0.5%, 1%, and 5%.

To assess the diversity in retrieval results across different sub-modules of Uni-Clip, we selected the LIT-PCBA dataset for analyzing variances in ligands retrieved by Uni-Clip-SC, Uni-Clip-RC, Uni-Clip-SG and Uni-Clip-RG. We identified the top 1% of ligands by rank from each sub-module for each protein target and computed diversity between the positive ligands. The results revealed that the diversity among the positive ligands retrieved by different retrieval methods exceeded 45% on average (Fig. 2a). This finding suggests that the different retrieval strategies implemented by Uni-Clip sub-modules are highly complementary, effectively enhancing the discovery of a wide range of candidate molecules for the target protein. Despite variations in network architecture and input data types among the encoders in Uni-Clip, we did not

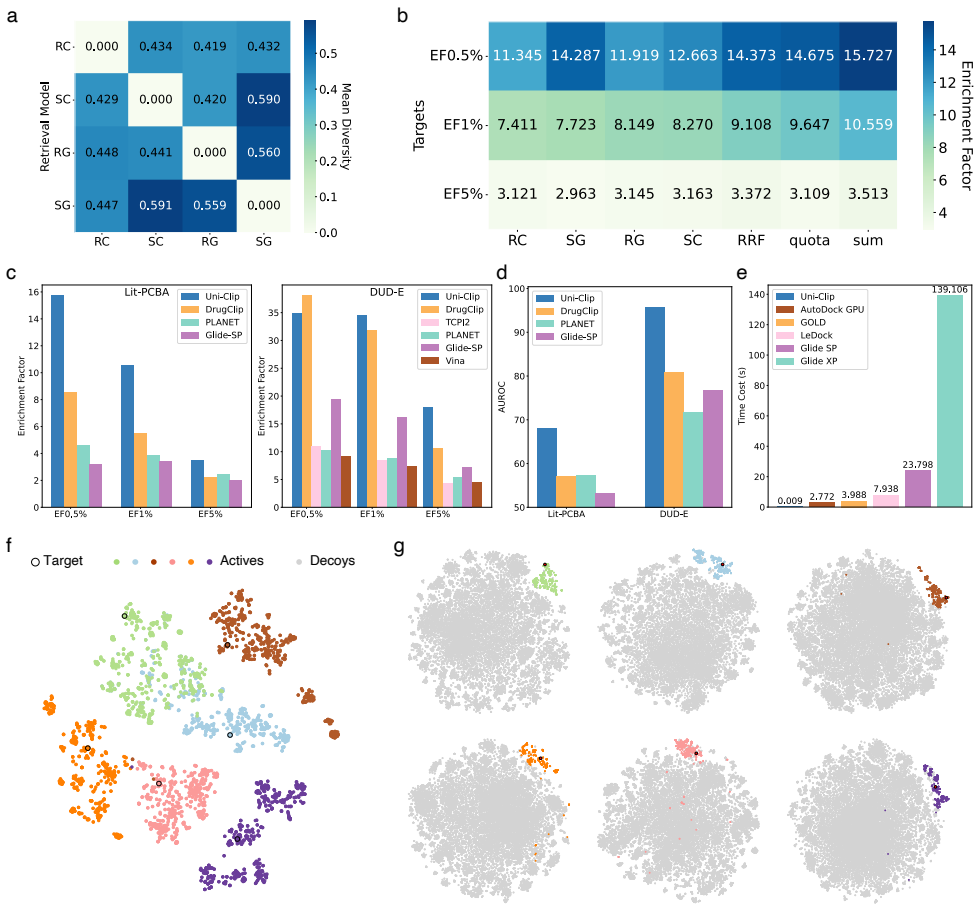


Fig. 2: Uni-Clip predicts protein ligand interactions with multimoldal data. **a** Diversity among different multimoldal retrival branches of Uni-Clip. **b** Improvements of Uni-Clip over different multimoldal retrival branches. **c-d** Enrichment factor and AUROC results of Uni-Clip and baseline models on the LIT-PCBA and DUD-E dataset. **e** Average time costs of the various models. **f-g** TSNE visualization of embeddings of DUD-E protein, actives (multicolored) and inactives (gray), produced by Uni-Clip.

impose explicit constraints during the optimization process to manage the divergence between Uni-Clip-SC, Uni-Clip-RC, Uni-Clip-SG and Uni-Clip-RG.

Given the variety of retrieval results produced by Uni-Clip sub-modules, we assessed three ensemble strategies to improve the efficiency of Uni-Clip's virtual screening process: Uni-Clip-sum, Uni-Clip-quota and Uni-Clip-RRF (detailed in the Methods 4.4). As shown in Fig. 2b, on the LIT-PCBA dataset, three ensemble methods outperformed single Uni-Clip sub-module at $EF_{0.5\%}$, $EF_{1\%}$, and $EF_{5\%}$. Particularly, Uni-Clip-sum demonstrated superiority, with an average improvement of 26%, 34%,

and 13% compared to the single retrieval method at $EF_{0.5\%}$, $EF_{1\%}$, and $EF_{5\%}$. More detailed results per target are depict in Supplementary Figure A1. In light of the promising performance manifested by Uni-Clip-sum, we employed its score as a representative for Uni-Clip in comparative analyses with other baselines.

As indicated in Fig. 2c, Uni-Clip shows superior performance across most evaluation metrics on the LIT-PCBA and DUD-E datasets. On the LIT-PCBA dataset, the Uni-Clip model outperforms the average performance of all baseline models by 240%, 147%, and 70% at $EF_{0.5\%}$, $EF_{1\%}$, and $EF_{5\%}$, respectively. On the DUD-E dataset, Uni-Clip surpasses the average performance of all baseline models by 162%, 218%, and 216% at $EF_{0.5\%}$, $EF_{1\%}$, and $EF_{5\%}$. Despite not achieving the peak $EF_{0.5\%}$ about DUD-E, the result of Uni-Clip remains competitive. More notably, Uni-Clip achieves the highest values for AUROC as shown in Fig. 2d. These comparative improvements highlight the consistency of Uni-Clip's predictive capabilities across different selection fractions of top-scoring molecules and its potential to substantially increase the effectiveness of virtual screening. Furthermore, to validate the efficiency of Uni-Clip in virtual screening, we illustrate the average time cost of Uni-Clip compared to various docking methods, including Glide-XP [36], Glide-SP [33], LeDock [37], GOLD [38], and AutoDock GPU [39], as shown in Fig. 2e. A lower bar indicates a faster method. In terms of screening power, Uni-Clip demonstrates a considerable advantage over established methods, notably achieving at least a 300-fold acceleration over AutoDock GPU, an already GPU-accelerated docking tool. This further underscores the potential application of Uni-Clip as an efficient tool for ultra-high-throughput ligand-based virtual screening.

For a more intuitive evaluation of the representations produced by Uni-Clip, we utilized t-SNE [40] to visualize the embeddings of Uni-Clip-RC for proteins and ligands from the DUD-E dataset. As shown in Fig. 2f, within the representation space, each target protein is clearly distinguishable, and the corresponding positive ligands display a noticeable clustering pattern. Furthermore, in Fig. 2g, we observed that the representations of positive ligands tend to cluster closely with the target proteins in the embedding space, while the representations of negative ligands are more dispersed and located at a greater distance from the target proteins. This suggests that through contrastive learning, Uni-Clip effectively creates clusters in the representation space, aligning with the protein's positive and negative samples, thereby enhancing the screening and ranking power for a given protein.

2.3 Drug discovery targeting Glutathione Peroxidase 4

To validate the efficacy of Uni-Clip in the context of practical drug discovery, we have chosen GPX4 as the target, which is known for its challenging druggability profile. GPX4, a crucial antioxidant enzyme in the selenoprotein family [41], is vital for maintaining cellular redox homeostasis by reducing lipid hydroperoxides and protecting cellular components from oxidative damage [42–45]. Its overexpression in various cancers has been linked to tumor progression and therapy resistance, making GPX4 inhibitors a promising cancer treatment strategy through inducing ferroptosis and reducing tumor growth [46, 47]. Currently, the majority of existing GPX4 inhibitors are covalent compounds, which utilize electrophilic warheads such as chloroacetamides

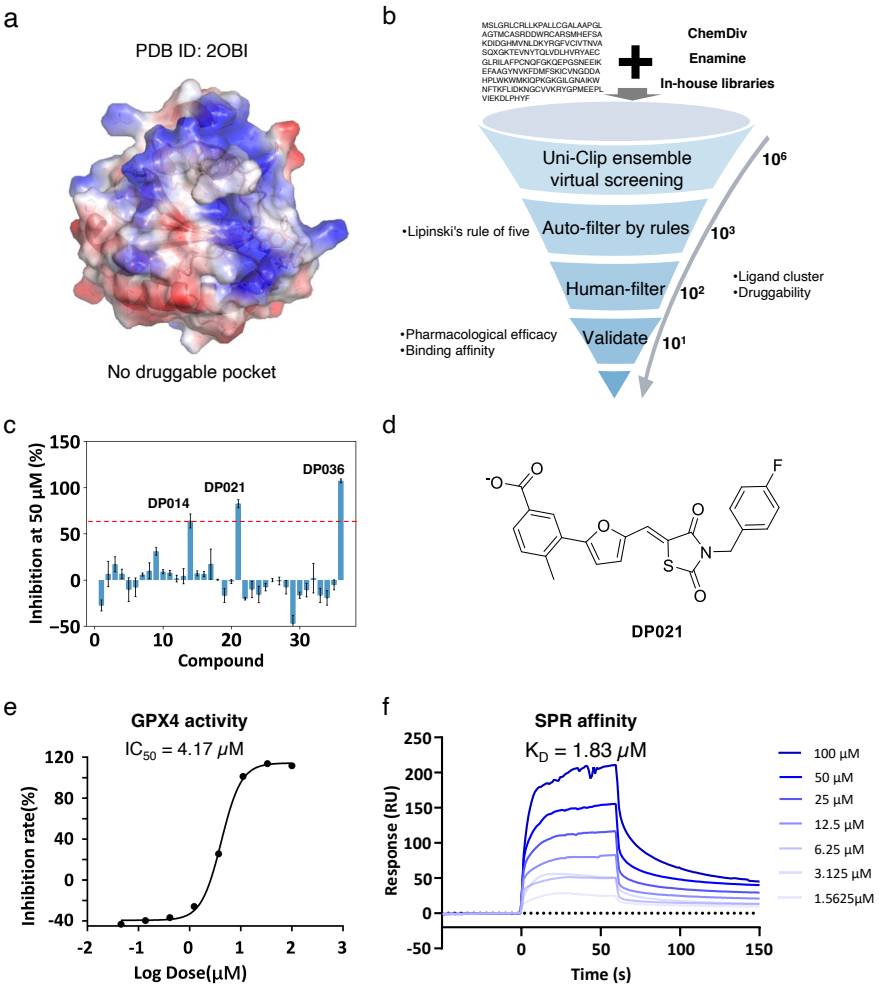


Fig. 3: Hit discovering of GPX4. **a** Crystal structure of GPX4 (PDB ID: 2OBI) is depicted, with surface representation. Notably, no druggable pockets are observed in this structure. **b** Schematic representation of the Uni-Clip virtual screening workflow, illustrating the process used to identify potential GPX4 inhibitors. **c** Biological activity inhibition of the 36 candidates at the concentration of 50 μM. **d** Chemical structure of compound DP021, presented in a 2D format. **e** The half-maximal inhibitory concentration (IC_{50}) for DP021 against GPX4 enzyme activity is determined to be 4.17 μM. **f** The binding affinity constant (K_D) of DP021 to GPX4, as measured by SPR, is found to be 1.83 μM.

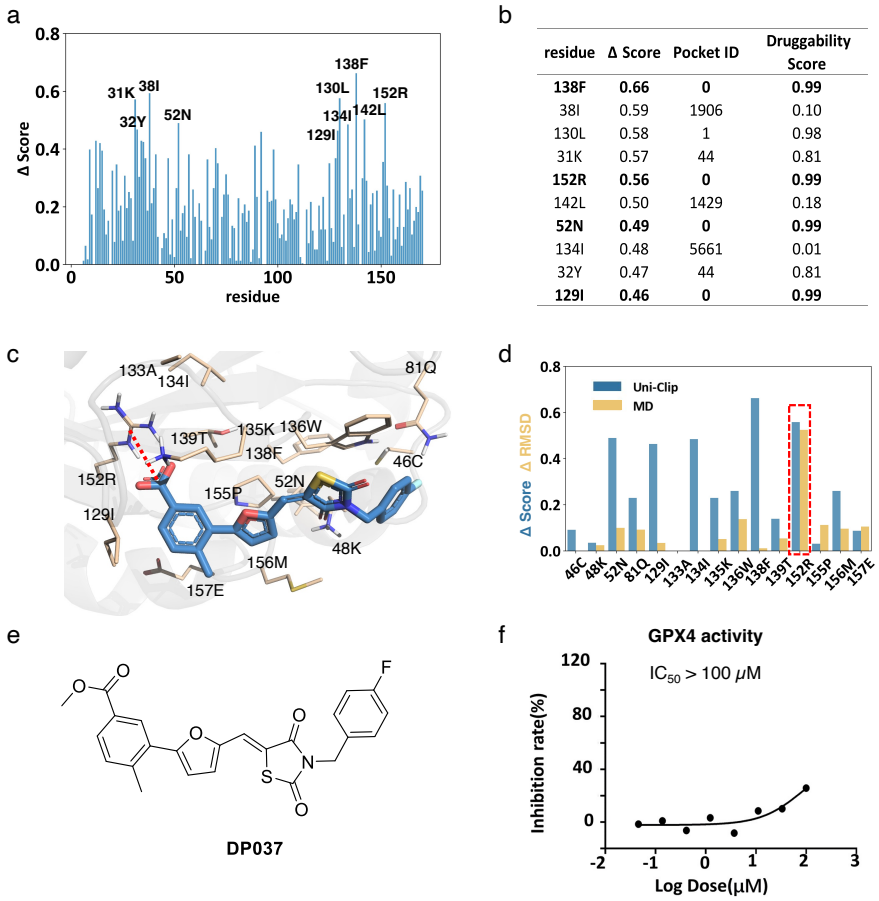
[48–52], leading to potential issues of selectivity and toxicity. Consequently, there is an urgent demand for potent, drug-like non-covalent small molecule inhibitors of GPX4. However, X-ray crystallography has revealed that the surface of GPX4 is relatively

flat and lacks druggable pockets (Fig. 3a and Supplementary Table A3). The highest druggability score observed from the crystal structure of GPX4 is 0.196 on a scale from 0 to 1, where higher values indicate better druggability. This extremely low score indicates that it poses a great challenge for traditional structure-based drug discovery (SBDD).

Hence, we employed a multimodal method, Uni-Clip, for the virtual screening of GPX4. Given the protein's flat surface and the lack of identifiable druggable pockets, it was not clear which of the Uni-Clip sub-modules (Uni-Clip-SC, Uni-Clip-RC, Uni-Clip-SG, Uni-Clip-RG, or Uni-Clip-sum) would be most appropriate. To address this challenge, we devised an ensemble virtual screening protocol that harnesses the five retrieval sub-modules of Uni-Clip, termed "Uni-Clip ensemble virtual screening" (Fig. 3b). We implemented these five approaches to screen and evaluate compounds from two extensive commercial compound libraries, ChemDiv and Enamine, each containing millions of compounds, as well as our in-house compound library comprising approximately 1,900 compounds. Utilizing the five retrieval approaches of Uni-Clip, we selected the top 900 compounds from the commercial libraries and the top 100 from our in-house library, totaling $(900 + 100) \times 5 = 5,000$ compounds for subsequent analysis. Compounds violating Lipinski's rule of five or having a druglikeness score [53] below -6 were excluded. The remaining compounds were clustered based on their extended-connectivity fingerprints (ECFP), resulting in 127 distinct chemotype clusters. Representative compounds were selected to ensure structure diversity, and 36 candidates were ultimately chosen for further experimental evaluation.

These compounds underwent an initial assessment of their biological activity through enzyme inhibition assays. Two distinct concentrations, 10 μM and 50 μM, were employed for this purpose. The inhibition rate at 50 μM is depicted in Fig. 3c, while a comprehensive presentation of the results can be found in Supplementary Tables A4. Certain compounds function as inhibitors, exerting a positive effect, while some others seemingly operate as agonists, resulting in a negative impact. Remarkably, the enzyme inhibition rate for three of the compounds exceeded 60% (Fig. 3c). The two compounds exhibiting the highest inhibition rates, DP021 and DP036, were further subjected to multi-concentration fitting analysis. The half-maximal inhibitory concentration (IC_{50}) values for DP021 and DP036 were determined to be 4.17 μM (Fig. 3d and 3e) and 26.25 μM (data not shown), respectively. Consequently, DP021 was chosen for more in-depth analysis and experimental testing. Subsequent Surface Plasmon Resonance (SPR) experiments were performed to verify the direct binding interaction between DP021 and the GPX4 protein. The binding affinity constant (K_D) was found to be 1.83 μM (Fig. 3f). To the best of our knowledge, this represents the most potent non-covalent GPX4 inhibitor reported in the literature thus far.

To determine the precise binding site of DP021 on GPX4, we employed alanine scanning mutagenesis within the Uni-Clip framework (refer to Methods) to evaluate the individual contribution of each residue (Δscore) to the binding affinity of DP021 for GPX4 (Fig. 4a). A higher score denotes a greater contribution to the interaction. For alanine scanning analysis, we used the Uni-Clip-RC sub-module, as DP021 was initially identified through this approach. The top 10 residues, identified as potential



residue	Δ Score	Pocket ID	Druggability Score
138F	**0.66**	**0**	**0.99**
38I	0.59	1906	0.10
130L	0.58	1	0.98
31K	0.57	44	0.81
152R	**0.56**	**0**	**0.99**
142L	0.50	1429	0.18
52N	**0.49**	**0**	**0.99**
134I	0.48	5661	0.01
32Y	0.47	44	0.81
129I	**0.46**	**0**	**0.99**

Fig. 4: The binding site of DP021 on GPX4. **a** Alanine scanning mutagenesis results from Uni-Clip, illustrating the individual contributions of each residue to the binding affinity of DP021 for GPX4. **b** Pockets of the top 10 residues, as derived from the MixMD. **c** The optimal docking pose of DP021 obtained via induced fit docking. The residues in Pocket 0 are visualized as lines, while DP021 is represented as sticks (PyMol). **d** The specific contributions of each residue in Pocket 0 to the binding of DP021 were evaluated using Uni-Clip alanine scanning (Δ score) and molecular dynamics simulation (Δ RMSD). **e** The chemical structure of the modified compound DP037 is presented in a 2D format. **f** The half-maximal inhibitory concentration (IC_{50}) of DP037 on GPX4 enzyme activity.

interactors with DP021, are listed in Fig. 4a. To identify potential binding pockets on the planar surface of GPX4, we performed mixed-solvent molecular dynamics (MixMD) simulations [54] using various small molecules as solvents. A 20 ns production run yielded 1,000 conformations for analysis. Fpocket [55] was subsequently utilized to identify 11,599 pockets. The pockets that contain one or more of the top

10 residues identified by Uni-Clip and have the highest druggability score are listed in Fig. 4b and Supplementary Table A5. Residues 138F, 152R, 52N and 129I were found within the same pocket (Pocket 0), which had the highest druggability score, suggesting this pocket may be the most probable binding site for DP021.

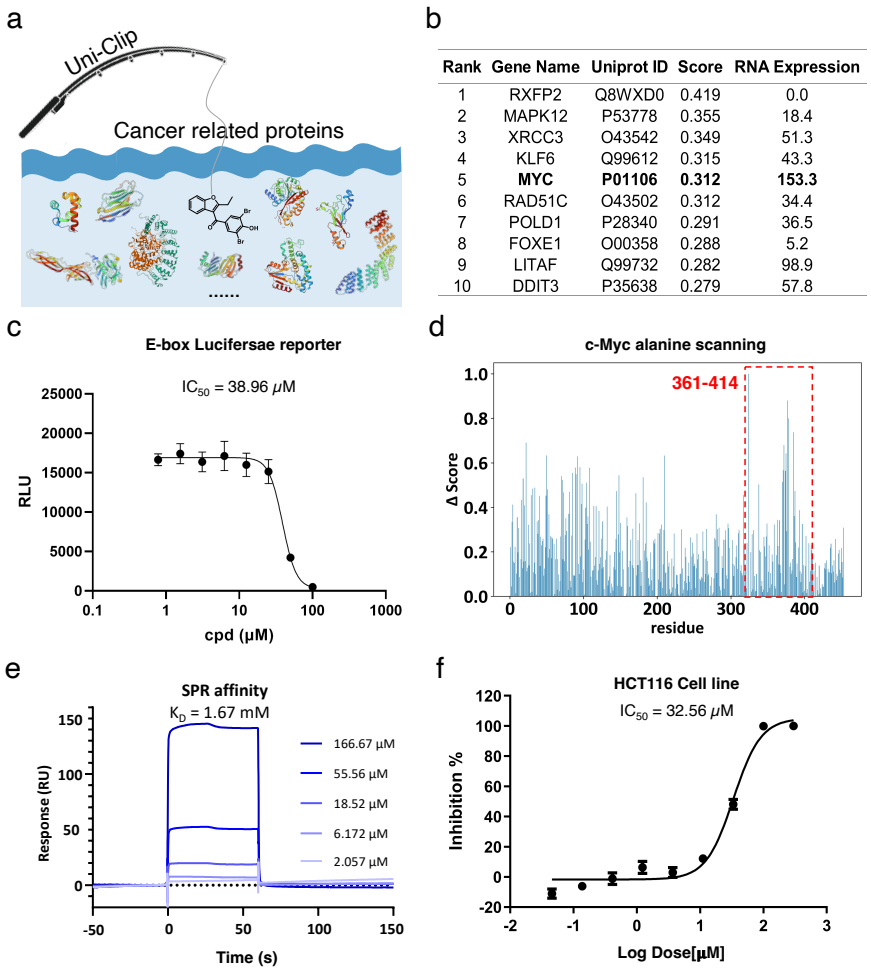
To validate this hypothesis, we docked DP021 into this pocket using the induced fit docking (IFD) method and selected the docking pose with the highest docking score for further analysis, as shown in Fig. 4c. We mutated each pocket residues of DP021 to alanine and performed molecular dynamics simulations to confirm the key residues in this pocket. The RMSD of the "wild-type pose" indicates considerable stability (Supplementary Figure A2). For the "mutated pose," the RMSD differences (ΔRMSD) from the wild-type pose are used to assess the importance of each residue (Fig. 4d). Notably, the "mutated pose" of residue R152 is the most unstable (Supplementary Figure A2), with ΔRMSD exceeding 0.5 nm (Fig. 4d), underscoring the critical role of 152R in the interaction between DP021 and GPX4. The docking pose also indicates that the primary interaction arises from a salt bridge formed by residue 152R with the carboxyl functional group of DP021. Therefore, we modified the carboxyl group of DP021 by capping it with a methyl group, creating a novel molecule, DP037 (Fig. 4e). Using Uni-Clip to evaluate DP037, the score decreases by 61% from DP021 to DP037. Subsequent synthesis and IC_{50} testing of DP037 against GPX4 via enzyme activity assays showed that the inhibitory effects of DP037 had considerably decreased with IC_{50} greater than 100uM (Fig. 4f). This result suggests that residue R152 and the formed salt bridge are crucial for DP021 binding, highlighting Pocket 0 as the potential binding site for DP021. This finding also indicates that Uni-Clip can accurately identify the key residues for ligand binding.

In summary, our application of the "Uni-Clip ensemble virtual screening" successfully led to the identification of DP021 as a GPX4 inhibitor. Additionally, Uni-Clip facilitated the precise identification of the binding pocket and key residues for DP021. Our research demonstrates how Uni-Clip can aid in discovering new inhibitors for challenging targets like GPX4.

2.4 Target fishing of benzbromarone

Here, we also assess the target fishing capability of Uni-Clip, an emerging approach in drug discovery used to identify potential protein targets for a query molecule [56]. We chose benzbromarone (Fig. 5a), a well-established uricosuric agent and a non-purine selective inhibitor of xanthine oxidase [57], to demonstrate the target fishing capability of Uni-Clip. Benzbromarone is mainly prescribed for hyperuricemia and gout treatment [58], targeting urate transporter 1 (URAT1) to increase renal excretion of uric acid. Despite being approved in several countries for gout management in patients with inadequate response to standard therapy or contraindications to other urate-lowering drugs, benzbromarone's potential in other therapeutic areas remains largely unexplored.

Our investigation was prompted by a recent study that utilized machine learning with graph attention networks (GATs) to predict benzbromarone as a potential therapeutic agent for lung adenocarcinoma (LUAD) [59]. Building on this insight, we also



Rank	Gene Name	Uniprot ID	Score	RNA Expression
1	RXFP2	Q8WXD0	0.419	0.0
2	MAPK12	P53778	0.355	18.4
3	XRCC3	O43542	0.349	51.3
4	KLF6	Q99612	0.315	43.3
5	**MYC**	**P01106**	**0.312**	**153.3**
6	RAD51C	O43502	0.312	34.4
7	POLD1	P28340	0.291	36.5
8	FOXE1	O00358	0.288	5.2
9	LITAF	Q99732	0.282	98.9
10	DDIT3	P35638	0.279	57.8

Fig. 5: Target Fishing Analysis of benzbromarone. **a** Schematic representation of the Target Fishing strategy utilizing the Uni-Clip. **b** Ranking of the top 10 genes selected by the Uni-Clip-RG approach. RNA expression levels (in nTPM units) were sourced from the Human Protein Atlas database. **c** The c-Myc E-box reporter assay of benzbromarone shows an IC_{50} value of 38.96 μM. **d** Alanine scanning results from the Uni-Clip. The red rectangular dot highlights the binding site of benzbromarone located within the C-terminal domain (residues 361-414) of c-Myc. **e** Surface Plasmon Resonance (SPR) analysis confirming the interaction between benzbromarone and the c-Myc protein, yielding a dissociation constant (K_D) of 1.67 mM. **f** Cellular activity of benzbromarone in HCT116 cells, with an IC_{50} value of 32.56 μM.

used the five retrieval approaches of Uni-Clip (Uni-Clip-SC, Uni-Clip-RC, Uni-Clip-SG, Uni-Clip-RG and Uni-Clip-sum) to evaluate the interaction of benzbromarone

with 330 human proteins associated with tumor, as catalogued in the UniProt database (Fig. 5a). The resulting scores were ranked according to the predicted interaction probability. From this analysis, we shortlisted the top 10 proteins from each Uni-Clip method as candidate targets for benzbromarone. To further refine our target selection, we retrieved RNA expression data for lung cancer cell lines from the Human Protein Atlas database [60–63]. Guided by the RNA expression profiles, we identified c-Myc as the primary research target for benzbromarone, which was highlighted by the Uni-Clip-RG method (Fig. 5b). Additional findings of other four retrieval approaches are detailed in Supplementary Table A9.

c-Myc, an intrinsically disordered protein (IDP) and a transcription factor, plays a pivotal role in regulating gene expression and orchestrating a myriad of biological processes [64, 65]. The aberrant expression of c-Myc protein is observed in approximately 70% of tumors, resulting from gene amplification, chromosomal translocation, upregulated mRNA, and abnormal protein stability [64, 66–68]. To investigate the potential of benzbromarone as a c-Myc inhibitor, we employed a systematic approach: first, we verified the benzbromarone's ability to inhibit c-Myc function; second, we confirmed the benzbromarone's binding region to c-Myc; and finally, we evaluated the benzbromarone's effect on related disease cell phenotypes of c-Myc.

To ascertain whether benzbromarone could inhibit c-Myc function, we initially employed a luciferase reporter assay utilizing an E-box element (Fig. 5c). Our findings indicate that benzbromarone exerts an inhibitory effect on c-Myc function, with a half-maximal inhibitory concentration (IC_{50}) of 38.96 μM.

To elucidate the direct binding of benzbromarone to c-Myc and identify the binding region, we utilized the alanine scanning approach of Uni-Clip (Fig. 5d). This analysis localized the binding sites to the C-terminal domain (residues 361-414) of c-Myc. We then synthesized a corresponding c-Myc peptide spanning this region and conducted a binding assay using SPR to evaluate the direct interaction with benzbromarone. The SPR data demonstrates a dose-dependent binding of benzbromarone to the c-Myc protein on the biosensor chip, with a dissociation constant (K_D) of 1.67 mM, confirming the direct binding of benzbromarone to the c-Myc C-terminal domain (residues 361-414).

Further investigations were conducted to evaluate the impact of benzbromarone on the cancer cells, specifically examining its effects on HCT116 cells. Benzbromarone demonstrates inhibitory effects on the proliferation of HCT116 cells, with an IC_{50} value of 32.56 μM (Fig. 5f). These findings collectively suggest that benzbromarone may exert anti-cancer effects through its direct interaction with c-Myc, thereby inhibiting cancer cell growth.

This study underscores the potential of benzbromarone as a therapeutic agent targeting the c-Myc protein, offering a promising avenue for cancer treatment through drug repurposing. The ability of benzbromarone to directly bind to and inhibit the transcriptional activity of c-Myc highlights its potential utility in combating cancers characterized by dysregulated c-Myc expression.

3 Discussion

In this study, we developed Uni-Clip, a multimodal contrastive learning model designed to improve the modeling of PLIs. By combining for pretrained encoders for both proteins and ligands, Uni-Clip effectively integrates protein structure and residue data with ligand conformation and graph data through the CF-InfoNCE loss, resulting in a comprehensive representation of protein-ligand interactions. Consequently, Uni-Clip consists of four submodules: Uni-Clip-SC, Uni-Clip-RC, Uni-Clip-SG, and Uni-Clip-RG. The multimodal design of Uni-Clip, in conjunction with the multimodal PLI dataset we constructed, MMBind, enhances its versatile applicability in diverse drug discovery scenarios, such as virtual screening and target fishing.

Compared to traditional single-modal approaches, Uni-Clip offers significant advantages by leveraging multimodal data to improve predictive accuracy and computational efficiency on widely acknowledged datasets LIT-PCBA and DUD-E. In the actual drug design process, Uni-Clip has efficaciously identified DP021 as a prospective inhibitor targeting GPX4, with its inhibitory efficacy corroborated by experimental assays. Furthermore, Uni-Clip delineated the binding interaction between benzbromarone, an established uricosuric agent, and the oncogenic protein c-Myc, thereby underscoring benzbromarone's potential as an anticancer therapeutic. Moreover, utilizing the alanine scanning method, Uni-Clip demonstrates potential in identifying binding sites within real-world applications.

While Uni-Clip has shown promising results, it still has potential for further refinement. First, as a data-driven neural network model, Uni-Clip's generalizability can be greatly improved by expanding the PLI dataset to cover a broader range of protein and ligand distributions. Additionally, collecting more high-quality experimental data and leveraging data derived from first-principles calculations are crucial steps for enhancing its performance. Second, although Uni-Clip has been effective in utilizing multimodal data of PLIs, incorporating more detailed information, such as protein binding pockets and ligand fragments, could enhance both the model's effectiveness and flexibility. In addition, Uni-Clip offers multimodal encoders for proteins and ligands. In practical applications, when addressing diverse protein-ligand pairs, incorporating a Mixture of Experts module would facilitate the dynamic selection and weighting of outputs from different encoders based on the input features of proteins and ligands. This adaptive mechanism would enable Uni-Clip to flexibly adjust to various protein-ligand combinations and their inherent complexities. Despite these challenges, Uni-Clip represents a substantial advancement in PLI modeling, offering enhanced accuracy, efficiency, and wide applications in virtual screening and target fishing. Uni-Clip could be a valuable tool for medicinal chemists, with great potential to accelerate the drug discovery process.

4 Methods

This section offers a comprehensive summary of each component in the proposed method and provides details of the experiments. We elucidate the key characteristics of our workflow and configurations, encompassing data curation, model architecture

design, evaluation metrics, and procedures for experiments conducted both in silico and in the wet lab.

4.1 MMBind Dataset

The construction of MMBind dataset includes the collection of protein-ligand affinity values and preparation of multi-modal ligand and protein data. We achieve the first goal by cleaning and refining ChEMBL33, a large database of bioactive assays collected from published experimental data, providing ligand canonical SMILES, protein UniProt IDs and their assay results. Initially, the original ChEMBL dataset was first filtered by assay type of 'B', target type of 'SINGLE PROTEIN' and molecule type of 'Small molecule'. All samples with concentration unit in nM molar concentration, which occupies over 99.6% of all samples, were included for further processing while others were dropped. Moreover, data with abnormal concentration values were further removed, such as negative and extremely large ones. To make full use of the dataset, a threshold of 562 μM was set to define positive samples. Additionally, we filtered out molecules with more than 100 heavy atoms. To achieve a zero-shot test, all targets in the DUD-E and LIT-PCBA datasets were excluded. Given the UniProt IDs of the targets, we accessed protein sequences using API provided by UniProt. The UniProt IDs was also used to access the AlphaFold database to get 3D fold structure. Although some of the proteins had experimental crystalline structure, we chose to include only AlphaFold structures for consistency. 3D conformations of molecules for Uni-Mol were randomly generated and optimized using the MMFF94 force field by RDKit. Molecule topology was also obtained in this procedure. After these steps, we constructed one of the largest datasets in field of PLIs, containing 853932 pairs composed of 533656 different ligands and 5245 proteins, as shown in Supplementary Table A6. In this research, the MMBind dataset is randomly split into training and validation subsets at a ratio of 10:1.

4.2 Encoders for Uni-Clip

Uni-Clip encodes ligands and proteins using four distinct encoders for each modality. For ligands, which were essentially small molecules, we used the Uni-Mol as the structure encoder and Uni-Mol-Graph as the molecular graph encoder. For proteins in 1D sequence and 3D structure, we chose ESM2 and SaProt, respectively. Uni-Mol is a robust 3D representation model for molecules. It has been pretrained on 209 million molecule conformations using unsupervised learning, achieving state-of-the-art results in various molecular representation learning tasks [28]. Uni-Mol integrates atomic and coordinate features to capture the three-dimensional structure of ligands by a two-track transformer architecture.

To develop an encoder that efficiently handles molecular graph information such as chemical bonds, we developed Uni-Mol-Graph, a molecule encoder leveraging graph features like shortest path within the molecular graph. Uni-Mol and Uni-Mol-Graph shared the same model backbone. However, unlike Uni-Mol, Uni-Mol-Graph excludes inter-atomic distance expansions from its attention mechanism and requires only a 2D molecular graph as input. Detailed input features were specified in Supplementary

Table A7 and Table A8. Similar to Uni-Mol, Uni-Mol-Graph was pretrained on the Uni-Mol dataset using a masked atom tokens recovery strategy. After the pretraining phase, we employed Uni-Mol-Graph as the graph encoder for ligands. Both Uni-Mol and Uni-Mol-Graph were finetuned during the PLI training process.

Evolutionary scale modeling 2 (ESM2) is chosen as protein residue encoder. ESM2 is a BERT-like protein language model proposed by the Meta Fundamental AI Research(FAIR) Protein Team. It is pretrained on 250 million protein sequence data in a masked language model manner. Among sequence-only methods, ESM2 stands out as one of the most formidable competitors. The 650M version of ESM2 is used in this research. Since fine-tuning a model at this scale is computationally expensive, we detach ESM2 from the training process and save ESM2 representations for all sequences in advance.

Apart from experimentally characterized proteins, which account for only a small fraction of known proteins, the structures of the majority of proteins remain to be explored. The success of computational folding methods, such as AlphaFold, enables the introduction of structure information for these proteins. Trained on 40 million protein structures generated by AlphaFold, SaProt effectively captures protein structure information at the residue level by integrating both structure tokens and residue tokens.

Atom-level representations from molecule encoders and residue-level representations from protein encoders are mean-pooled to form molecule- and protein-level representations. Given the multimodal representations, a projection head with sparse activation function s [69] is applied to each of them to align the semantic space as:

$$y'_m = \mathrm{s}(W_m(y_m) + b_m) \ , \tag{1}$$

$$\mathrm{s}(x) = ReLU(x).detach() + GELU(x) - GELU(x).detach() \ , \tag{2}$$

m refers to a certain modality. Specifically, the sparse activation function s gives a neuron ReLU outputs but GELU [70] gradient to avoid the dead neuron problem. Activated by ReLU, the representations become rather sparse, which greatly facilitates the calculation of cosine similarity between modalities.

4.3 Collision-free InfoNCE loss

The collision-free InfoNCE loss is designed to solve the multiple-positive-samples issue in context of PLI. It incorporates an dynamic labeling strategy into the conventional InfoNCE loss, which writes:

$$\mathcal{L}_{InfoNCE} = -\log \frac{exp(f(x_i, y_i^+)/\tau)}{\sum_j exp(f(x_i, y_j)/\tau)} \ , \tag{3}$$

in which x_i is the representation of sample i, y_i^+ is the corresponding positive sample's representation, and τ is a temperature hyperparameter. f is a scoring function measuring the similarity between representations. Since the available experimental data contains majorly positive samples of binding proteins and ligands, construction

of negative pairs is of necessity. Similar to CLIP, we adopt a dynamic in-batch sampling strategy. Given a batch of N pairs $\{(P_i, L_i)\}_{i=1}^{N}$, the combination of $\{P_i\}_{i=1}^{N}$ and $\{L_j\}_{j=1}^{N}$ results in N^2 pairs (P_i, L_j).

Usually, only pairs with $i = j$ are treated as positives samples while others are negative ones. However, as shown in Fig. 1d, it is possible that the dynamic constructed off-diagonal ($i \neq j$) pairs turn to be positive samples. We refer this issue as "collision of positive samples" and try to solve it with the following dynamic labeling strategy.

Specifically, each pairs (P_i, L_j) is first assigned with a label $l_{i,j}$ related to its assay value:

$$l_{i,j} = \begin{cases} \max(0,\ (4.01 - \log_{10}\ v_{i,j})) & (P_i, L_j)\,\text{is positive}\ , \\ 0 & \text{otherwise}\ . \end{cases} \tag{4}$$

where v_{ij} is the assay value of pair (P_i, L_j), For each row in the resulting label matrix, the pair with largest label, i.e., the pair with strongest binding affinity, is selected as the positive sample, while other pairs are regarded as negative samples. After that, the label matrix is normalized row-wise. In this way, the label is made sure to be one-hot and the dominant pair is preserved, avoiding the collision problem. The collision-free InfoNCE loss $\mathcal{L}_{\mathcal{CF}}$ can be written as:

$$\mathcal{L}_i^{CF} = -\log \frac{exp(f(x_i, y_k)/\tau)}{\sum_j exp(f(x_i, y_j)/\tau)}\ , \tag{5}$$

where $k = \arg\max_j l_{ij}$. Considering the protein and ligand parts, the overall loss function of a batch data for a retrieval branch $B \in (SC, RC, SG, RG)$ is:

$$\mathcal{L}_B = \frac{1}{2}\sum_{i=1}^{N}(\mathcal{L}_i^P + \mathcal{L}_i^L)\ , \tag{6}$$

where $\mathcal{L}_{\mathcal{B}}$ is the InfoNCE loss calculated with collision-free label. In Uni-Clip, we align 4 different representations: the SaProt and ESM2representations for proteins with the Uni-Mol and Uni-Mol-Graph representations for ligands. Their combination, i.e. 4 contrastive pairs L_{SC},L_{RC},L_{SG} and L_{RG}, forms the final loss as:

$$\mathcal{L} = \mathcal{L}_{SC} + \mathcal{L}_{RC} + \mathcal{L}_{SG} + \mathcal{L}_{RG}\ . \tag{7}$$

4.4 Ensemble strategy of Uni-Clip

To synthetically consider the four retrieval branches formed by the combination of different encoders, three ensemble stategies are tried in this research, including sum strategy, quota fusion strategy and reciprocal rank fusion (RRF) strategy. In the sum ensemble strategy, all candidates are re-ranked according to the summed score from Uni-Clip-SC, Uni-Clip-RC, Uni-Clip-SG and Uni-Clip-RG results. Since equal weights is assigned to each part of the loss function corresponding to the combination above during training, we keep their weights equal when conduct the ensemble:

$$\text{score}_{\text{sum}} = S_{SC} + S_{RC} + S_{SG} + S_{RG}\ , \tag{8}$$

$\text{score}_{\text{sum}}$ is the score of certain pair in sum ensemble, S_{SC},S_{RC},S_{SG},S_{RG} represents the score given by Uni-Clip-SC, Uni-Clip-RC, Uni-Clip-SG, Uni-Clip-RG, respectively. The quota fusion strategy and RRF strategy is based on rank of pairs rather than their scores. In quota fusion, the candidates with highest scores from each combination is took out to form a retrieval set. This operation is repeated until the set achieves given capacity.

$$R_{\text{q}} = \bigcup_{i \in \{SC,SG,RC,RG\}} (R_i)_{\frac{q}{4}} \ , \tag{9}$$

in which R_{q} stands for the retrieved set at a given quota q of the candidate set. The ensemble retrieval results are then sorted according to their Uni-Clip-RC scores.

RRF strategy ensembles different ranking results base on a scoring funcion of rank:

$$\text{score}_{\text{RRF}} = \sum_{r \in R} \frac{1}{k + r(d)} \ , \tag{10}$$

in which $r(d)$ is the rank of d in a ranking result R, k is a hyper-parameter and is set to 60 in this reasearch. High $\text{score}_{\text{RRF}}$ means the candidate is favored in RRF ensembling.

4.5 Evaluation metrics and datasets

Evaluation metrics includes AUROC and enrichment factor (EF) on 0.5%, 1% and 5%. While AUROC is widely used in classification tasks, enrichment factor is a metric for virtual screening, measuring how good a screening method performs compare to random guess:

$$EF_{\alpha} = \frac{N_{active(\alpha)}}{N_{active(total)} \times \alpha} \ , \tag{11}$$

where α is a proportion, $N_{active(\alpha)}$ is the number of true active compounds in the top α candidates sorted by model. $N_{active(total)}$ is the total number of active compounds among all candidates.

Diversity between candidates sets is obtained from different retrieval branches. This metric is calculated from the fingerprints' similarity between the molecules they contain. The diversity between set A against set B is:

$$D(A,B) = 1 - \sum_{a \in A} \frac{max_{b \in B}(\langle FP_a, FP_b \rangle)}{N_A} \ , \tag{12}$$

D is diversity value, FP is the Morgan fingerprint of a molecule, N_A is the number of elements of set A.

DUD-E and LIT-PCBA have been chosen to evaluate Uni-Clip's screening power. DUD-E is one of the most popular virtual screening benchmark datasets [71–73]. It contains 102 proteins. Each protein is associated with a known ligand defining the binding pocket and relevant binders as well as non-binders in a ratio of 1:50.

LIT-PCBA is another virtual screening benchmark dataset proposed to address biased data problem in previous benchmarks [35]. It is built upon dose-response PubChem bio-assays, consisting of 15 targets with 7,844 experimentally active and 407,381 inactive compounds.

4.6 Analysis tools for interpretation

As shown in results, Uni-Clip can not only handle scoring and ranking tasks, but also help determine the key interactions in the protein-ligand pairs. This is achieved through two analysis tools: substitution effect analysis and alanine scanning. Below, we briefly introduce these methods.

In substitution effect analysis, the protein is fixed and the effect of substituting functional groups in its active compounds is examined. Specifically, potential key functional groups in these compounds are substituted with less chemically active ones and the modified molecules' affinity scores with the protein are subsquently evaluated. The model is expected to give lower scores and recognize the activity cliffs in these molecules.

Alanine scanning is a commonly accepted method to identify important residues in a protein sequence. For proteins, we anchor an active compound and mutate each residue into alanine, which usually does not interact with small molecules due to its lack of a functional group but has similar contributions to protein structure as other amino acids. We check whether the scores predicted by Uni-Clip for the original protein and mutated protein change substantially as $\Delta score = |score_{mutation} - score_{wt}|$.

4.7 Virtual screening of GPX4

We leveraged five retrieval approaches of Uni-Clip (Uni-Clip-SC, Uni-Clip-RC, Uni-Clip-SG, Uni-Clip-RG and Uni-Clip-sum) to assess and score potential compounds from three comprehensive chemical libraries. These included ChemDiv, with an inventory of 1,566,730 compounds, Enamine, housing 2,056,498 compounds, and our proprietary library, which contains approximately 1,900 compounds. From the commercial libraries, we selected the top 900 compounds from each method, complemented by the top 100 from our in-house collection, amassing a total of 5,000 compounds for further analysis.

Our objective was achieved through a meticulous two-tier filtering process. Initially, compounds contravening Lipinski's Rule of Five (a criterion for assessing the drug-likeness of compounds) were eliminated. Simultaneously, we applied DataWarrior software [53] to exclude molecules with a computed drug-likeness score below -6.

Subsequently, the remaining compounds underwent automated clustering based on their ECFP. This molecular descriptor encapsulates the structure attributes of the compounds, leading to the identification of approximately 127 unique clusters, each representing a distinct chemotype.

From these clusters, we meticulously selected representative compounds based on their intra-cluster ranking and their structure diversity relative to other highly ranked compounds. This strategy was designed to ensure a broad spectrum of structure

diversity within the final candidate pool. Ultimately, 36 candidates were procured for subsequent experimental validation.

4.8 GPX4 protein expression and purification

The GPX4 (residues 28-197) gene was cloned into the pCDNA3.4 vector and co-expressed with SECISBP2 in HEK293 cells, as previously described [74, 75]. The cells were then harvested and resuspended in lysis buffer, containing 50 mM Tris-HCl (pH 7.5), 500 mM NaCl, 10% glycerol, 1 mM dithiothreitol (DTT), and 0.1 mM phenylmethylsulfonyl fluoride (PMSF). The cells were subsequently lysed through sonication, and the cell debris was removed by centrifugation to obtain the supernatant. Next, the supernatant was applied to a Ni-NTA resin column to isolate the protein of interest via affinity chromatography. The column was washed, and the target protein was eluted. To further purify the protein, size-exclusion chromatography was performed using a Superdex 200 column, pre-equilibrated with SEC buffer composed of 50 mM Tris-HCl (pH 7.5), 150 mM NaCl, and 10% glycerol. Finally, the purified protein was aliquoted, flash-frozen in liquid nitrogen, and stored at -80 °C for future use.

4.9 GPX4 activity assay

An enzymatic activity inhibition assay was conducted to evaluate the inhibitory potential of selected compounds on GPX4. This assay was performed using the commercially available Cayman GPX4 Inhibitor Screening Assay Kit (catalog number 701880) as per the manufacturer's instructions. The assay is based on the principle of measuring GPX4 activity indirectly through a coupled reaction with glutathione reductase (GR), where the oxidation of NADPH to NADP+ and the subsequent decrease in absorbance at 340 nm is proportional to GPX4 activity. This colorimetric method allowed us to assess the inhibitory effects of various compounds on GPX4 by monitoring the changes in absorbance. We conducted a preliminary screening of our 36 candidate molecules to evaluate their inhibitory effects, using two different concentrations: 10 μM and 50 μM. Following the identification of the most promising candidates, we determined their half-maximal inhibitory concentration (IC_{50}) values. To obtain these values, we performed dose-response experiments by treating the target with increasing concentrations of the selected inhibitors and monitoring the inhibitory response. This comprehensive evaluation of the candidate molecules enabled us to select the most effective inhibitors for further in-depth studies.

4.10 GPX4 SPR

The GPX4 SPR biosensing experiments were conducted using a Biacore T200 instrument (GE Healthcare). The running buffer for immobilization was prepared using phosphate-buffered saline (PBS, pH 7.4) supplemented with 5% DMSO, 750 mM NaCl, and 0.05% TWEEN80, which is referred to as PBST. The GPX4 (residues 28-197) protein was immobilized onto a CM5 sensor chip (GE Healthcare), with immobilization levels reaching approximately 3900 response units (RUs). Interaction studies were carried out in PBST at a constant temperature of 25°C. The test molecules were injected for 60 seconds at a flow rate of 50 μL/min, and the dissociation phase was monitored

for up to 240 seconds. Data processing and analysis were performed using the Biacore T200 evaluation software (GE Healthcare). The data were fitted to a single-site binding model, enabling the calculation of dissociation constants (K_D values).

4.11 Induced fit docking and molecular dynamics simulations

4.11.1 Mixed-solvent molecular dynamics simulations

Due to the planar surface of GPX4, we employed a mixed-solvent molecular dynamics (MixMD) [54] simulation approach to investigate the potential formation of pockets within the protein. Five small fragments were selected as solvents: acetate ion, ethyl-dimethyl-ammonium ion, 9-methyl-9H-purin-6-amine, N-methylformamide, and 2-methylpyridine. Each solvent was mixed with water at a 1:100 ratio. The AMBER14SB force field [76] was used for the protein, while the GAFF2 [77, 78] force field was applied for the solvent molecules. The TIP3P water model was used for water molecules. A cubic box with a side length of 7.10 nm was constructed, containing a total of 25,394 atoms. Molecular dynamics simulations were performed using the OpenMM [79] software package. The system was first energy minimized, followed by equilibration in the NVT ensemble for 100 ps and then in the NPT ensemble for 100 ps to ensure system stability. Finally, a 20 ns production run was conducted in the NPT ensemble, with conformations saved every 10 ps, resulting in a total of 1,000 frames for subsequent analysis. Fpocket [55] software was used to analyze the pockets in the obtained 1,000 conformations, yielding a total of 11,599 pockets. We identified the pocket with the highest druggability score and contains one of the top 10 residues as the pocket of that residue.

4.11.2 Induced fit docking

We employed the IFD method to investigate the binding interactions between small molecules and the GPX4 protein. Due to the relatively flat surface of GPX4, traditional rigid docking approaches may not accurately capture the conformational changes induced by ligand binding. In contrast, the IFD method allows for flexibility in both the ligand and the protein receptor, thereby providing a more realistic representation of the binding process. The IFD protocol was carried out using the Hermite web platform at https://hermite.dp.tech. For each small molecule, multiple docking poses were generated and evaluated based on their IFD scores, which take into account both the binding affinity and the conformational energy of the protein-ligand complex. The top-scoring pose was selected as the representative binding conformation for further analysis.

4.11.3 Conventional molecular dynamics simulations

To validate the importance of the pocket residues identified for the DP0221 molecule, we performed mutagenesis and conventional molecular dynamics (MD) simulations on the GPX4-DP021 complex obtained from the IFD analysis. Protein mutations were introduced using the protein preparation tool in the Hermite web platform at https://hermite.dp.tech. The pocket residues were individually mutated to alanine

(Ala). The force field parameters used for the MD simulations were consistent with those employed in the mixMD approach. The simulations were carried out using the uni-GBSA automated simulation tool [80], with each mutation being simulated for 300 ns. To account for the stochastic nature of MD simulations, three parallel replicas were run for each mutation. For trajectory analysis, the 50-100 ns segment of each trajectory was considered. The root-mean-square deviation (RMSD) values of the small molecule relative to its initial conformation were calculated for all conformations sampled after 50 ns in each of the three replicas. The average RMSD value for each mutation was then computed and subtracted from the corresponding value for the wild-type (WT) complex. This difference in RMSD was used as a stability score for each mutation site, reflecting the impact of the mutation on the binding stability of DP021.

4.12 Cell lines

HCT116 cells (ATCC, CCL-247) were cultured in McCoy's 5A medium (16600-082, Gibco), while HEK293T cells (catalog no. CBP60439, Cobioer) were maintained in Dulbecco's Modified Eagle Medium (DMEM; catalog no. C0162-811S, CellWorld). Both culture media were supplemented with 10% (v/v) heat-inactivated fetal bovine serum (FBS; catalog no. 10099141C, Gibco) and 1% (v/v) penicillin-streptomycin solution (P/S; catalog no. C0160-611, CellWorld) to support optimal cell growth and viability.

4.13 c-Myc E-box reporter assay

HEK293T cells were seeded onto a 6-well plate at a density of 1×10^6 cells per well, following harvesting and counting. After a 6-hour incubation period, which allowed the cells to adhere to the plate, a transfection mixture was prepared. This mixture contained 16 μL of FuGENE® HD Transfection Reagent (E2311, Promega) and 4 μg of Ebox-Luciferase plasmid (11544ES03, YEASEN). The components were mixed thoroughly and incubated for 15 minutes before being added to each well. The cells were then cultured in a 37°C incubator with 5% CO_2 for an additional 18 hours.

Subsequently, transfected cells were seeded onto a 96-well plate at a density of 3×10^4 cells per well. Compounds were added simultaneously to each well at various concentrations (50 μM maximum concentration, double dilution series, 10 concentration gradients), with DMSO serving as a positive control. Three replicates were performed for each well. After a 24-hour incubation period, the cell supernatant was removed, and luminescence was measured using the Luciferase Reporter Gene Assay Kit (11401ES80, YEASEN), following the manufacturer's instructions.

Specifically, 100 μL of lysis buffer was added to the cells, and samples were incubated for 5 minutes on ice. Next, 20 μL of cell lysate was transferred to a white 96-well plate (410096M, in vitro scientific) and combined with 100 μL of Firefly luciferase test reagent. Luminescence was then measured to assess luciferase activity. Additionally, 20 μL of the cell lysate was transferred to a black 96-well plate (J09602, JingAn biological), where 20 μL of CellTiter-Glo (G9242, Promega) was added to determine cell viability through luminescence measurements. Luminescence values were normalized

by the corresponding relative cell viability (CTG value of each well/mean CTG value of each compound).

4.14 c-Myc SPR

The c-Myc SPR assay was performed utilizing a c-Myc peptide (residues 361-414) synthesized by Hefei Guotai Biology Science and Technology. The peptide was immobilized onto a CM5 sensor chip, which had been pre-treated with a 1:1 mixture of 0.4 M 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) and 0.1 M N-hydroxysuccinimide (NHS). The immobilization level achieved was 6192 response units (RU). Subsequently, the surface of the CM5 chip was blocked with an ethanolamine solution for 10 minutes to eliminate non-specific binding. The compound of interest was prepared at the desired concentration in a buffer containing 5% dimethyl sulfoxide (DMSO) and phosphate-buffered saline with Tween 20 (PBST). The Biacore T200 instrument (GE Healthcare) was employed to determine the association (K(a)) and dissociation (K(d)) constants of the compound, enabling the calculation of its binding affinity. The equilibrium dissociation constant (K_D) was computed using the steady-state model and a 1:1 binding kinetic model.

4.15 Proliferation assay

HCT116 cells were seeded in a 96-well plate at a density of 500 cells per well. Concurrently, compounds were added to each well at varying concentrations, ranging from 300 μM as the maximum concentration, with a triple dilution series consisting of 9 gradients. DMSO-treated wells were included as positive controls, and each condition was replicated twice. The cells were then incubated for 6 days at 37°C in a 5% CO_2 atmosphere.

Following incubation, 50 μL of CellCounting-Lite 2.0 Luminescent Cell Viability Assay reagent (G7573, Promega) was added to each well. The plates were gently shaken for 2 minutes, after which the relative luminescence units (RLU) were measured to determine cell viability. The inhibition rate was calculated using the following formula:

$$\text{Inhibition}(\%) = \frac{\text{RLU}_{\text{control}} - \text{RLU}_{\text{compound}}}{\text{RLU}_{\text{control}} - \text{RLU}_{\text{blank}}} \times 100\% \ , \tag{13}$$

where $\text{RLU}_{\text{compound}}$ represents the RLU values for compound-treated wells, $\text{RLU}_{\text{blank}}$ denotes the RLU values for DMSO wells without cells, and $\text{RLU}_{\text{control}}$ signifies the RLU values for positive control wells.

5 Data availability

ChEMBL database is available at https://www.ebi.ac.uk/chembl/. The commercial Enamine Library is available at https://enamine.net/, and ChemDiv Library is available at https://www.chemdiv.com/. All data are available from the corresponding author upon request. The source data can be accessed through figshare at https://figshare.com/s/cb245e12c48a2298d610.

6 Code availability

The source code will be available after paper acceptance. Several servers, such as virtual screening, targt fishing and binding site prediction, are available as a non-commercial usage at https://bohrium.dp.tech/apps/uni-clip/job?type=app.

7 Acknowledgements

This study was supported by the National Key Research and Development Program of China (2022YFA1004302).

References

[1] Hughes, J.P., Rees, S., Kalindjian, S.B., Philpott, K.L.: Principles of early drug discovery. British journal of pharmacology **162**(6), 1239–1249 (2011)

[2] Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., Ji, X.-L., Liu, S.-Q.: Insights into protein–ligand interactions: mechanisms, models, and methods. International journal of molecular sciences **17**(2), 144 (2016)

[3] Zhao, L., Zhu, Y., Wang, J., Wen, N., Wang, C., Cheng, L.: A brief review of protein–ligand interaction prediction. Computational and Structural Biotechnology Journal **20**, 2831–2838 (2022)

[4] Tsubaki, M., Tomii, K., Sese, J.: Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics **35**(2), 309–318 (2019)

[5] Lee, I., Keum, J., Nam, H.: Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS computational biology **15**(6), 1007129 (2019)

[6] Huang, K., Xiao, C., Glass, L.M., Sun, J.: Moltrans: molecular interaction transformer for drug–target interaction prediction. Bioinformatics **37**(6), 830–836 (2021)

[7] Chen, L., Fan, Z., Chang, J., Yang, R., Hou, H., Guo, H., Zhang, Y., Yang, T., Zhou, C., Sui, Q., *et al.*: Sequence-based drug design as a concept in computational drug design. Nature Communications **14**(1), 4217 (2023)

[8] Zhang, X., Gao, H., Wang, H., Chen, Z., Zhang, Z., Chen, X., Li, Y., Qi, Y., Wang, R.: Planet: a multi-objective graph neural network model for protein–ligand binding affinity prediction. Journal of Chemical Information and Modeling (2023)

[9] Zhang, X., Zhang, O., Shen, C., Qu, W., Chen, S., Cao, H., Kang, Y., Wang, Z., Wang, E., Zhang, J., *et al.*: Efficient and accurate large library ligand docking with karmadock. Nature Computational Science **3**(9), 789–804 (2023)

[10] Singh, R., Sledzieski, S., Bryson, B., Cowen, L., Berger, B.: Contrastive learning in protein language space predicts interactions between drugs and protein targets. Proceedings of the National Academy of Sciences **120**(24), 2220778120 (2023)

[11] Gao, B., Qiang, B., Tan, H., Jia, Y., Ren, M., Lu, M., Liu, J., Ma, W.-Y., Lan, Y.: Drugclip: Contrasive protein-molecule representation learning for virtual screening. Advances in Neural Information Processing Systems **36** (2024)

[12] Fan, F.J., Shi, Y.: Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction. Bioorganic & Medicinal Chemistry **72**,

117003 (2022)

[13] Ruff, K.M., Pappu, R.V.: Alphafold and implications for intrinsically disordered proteins. Journal of molecular biology **433**(20), 167208 (2021)

[14] Fraser, J.S., Murcko, M.A.: Structure is beauty, but not always truth. Cell **187**(3), 517–520 (2024)

[15] Meli, R., Morris, G.M., Biggin, P.C.: Scoring functions for protein-ligand binding affinity prediction using structure-based deep learning: a review. Frontiers in bioinformatics **2**, 885983 (2022)

[16] Siebenmorgen, T., Menezes, F., Benassou, S., Merdivan, E., Didi, K., Mourão, A.S.D., Kitel, R., Liò, P., Kesselheim, S., Piraud, M., et al.: Misato: machine learning dataset of protein–ligand complexes for structure-based drug discovery. Nature Computational Science, 1–12 (2024)

[17] Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1), 2 (2020)

[18] Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learning: A framework and review. Ieee Access **8**, 193907–193934 (2020)

[19] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607 (2020). PMLR

[20] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR

[21] Xue, Y., Whitecross, K., Mirzasoleiman, B.: Investigating why contrastive learning benefits robustness against label noise. In: International Conference on Machine Learning, pp. 24851–24871 (2022). PMLR

[22] Wu, J., Chen, J., Wu, J., Shi, W., Wang, X., He, X.: Understanding contrastive learning via distributionally robust optimization. Advances in Neural Information Processing Systems **36** (2024)

[23] Zhang, Y., Li, S., Meng, K., Sun, S.: Machine learning for sequence and structure-based protein–ligand interaction prediction. Journal of Chemical Information and Modeling (2024)

[24] Zhang, G., Zhang, J., Gao, Y., Li, Y., Li, Y.: Strategies for targeting undruggable targets. Expert Opinion on Drug Discovery **17**(1), 55–69 (2022)

[25] Xie, X., Yu, T., Li, X., Zhang, N., Foster, L.J., Peng, C., Huang, W., He, G.:

Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials. Signal Transduction and Targeted Therapy **8**(1), 335 (2023)

[26] Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., Yuan, F.: Saprot: protein language modeling with structure-aware vocabulary. bioRxiv, 2023–10 (2023)

[27] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., *et al.*: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023)

[28] Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., Ke, G.: Uni-mol: A universal 3d molecular representation learning framework. In: The Eleventh International Conference on Learning Representations (2023)

[29] Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J.P.: ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research **40**(D1), 1100–1107 (2011) https://doi.org/10.1093/nar/gkr777

[30] Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., *et al.*: Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic acids research **52**(D1), 368–375 (2024)

[31] Oord, A., Li, Y., Vinyals, O.: Representation Learning with Contrastive Predictive Coding (2019)

[32] 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**(2), 455–461 (2010)

[33] Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K., *et al.*: Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. Journal of medicinal chemistry **47**(7), 1739–1749 (2004)

[34] 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**(14), 6582–6594 (2012). PMID: 22716043

[35] 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**(9), 4263–4273 (2020). PMID: 32282202

[36] Friesner, R.A., Murphy, R.B., Repasky, M.P., Frye, L.L., Greenwood, J.R., Halgren, T.A., Sanschagrin, P.C., Mainz, D.T.: Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein- ligand

complexes. Journal of medicinal chemistry **49**(21), 6177–6196 (2006)

[37] Zhao, H., Caflisch, A.: Discovery of zap70 inhibitors by high-throughput docking into a conformation of its kinase domain generated by molecular dynamics. Bioorganic & medicinal chemistry letters **23**(20), 5721–5726 (2013)

[38] Jones, G., Willett, P., Glen, R.C., Leach, A.R., Taylor, R.: Development and validation of a genetic algorithm for flexible docking. Journal of molecular biology **267**(3), 727–748 (1997)

[39] Santos-Martins, D., Solis-Vasquez, L., Tillack, A.F., Sanner, M.F., Koch, A., Forli, S.: Accelerating autodock4 with gpus and gradient-based local search. Journal of chemical theory and computation **17**(2), 1060–1073 (2021)

[40] Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

[41] Thomas, J.P., Maiorino, M., Ursini, F., Girotti, A.W.: Protective action of phospholipid hydroperoxide glutathione peroxidase against membrane-damaging lipid peroxidation. in situ reduction of phospholipid and cholesterol hydroperoxides. Journal of Biological Chemistry **265**(1), 454–461 (1990)

[42] Conrad, M., Angeli, J.P.F.: Glutathione peroxidase 4 (gpx4) and ferroptosis: what's so special about it? Molecular & Cellular Oncology **2**(3), 995047 (2015). PMID: 27308484

[43] Yant, L.J., Ran, Q., Rao, L., Van Remmen, H., Shibatani, T., Belter, J.G., Motta, L., Richardson, A., Prolla, T.A.: The selenoprotein gpx4 is essential for mouse development and protects from radiation and oxidative damage insults. Free Radical Biology and Medicine **34**(4), 496–502 (2003)

[44] Ursini, F., Maiorino, M.: Lipid peroxidation and ferroptosis: The role of gsh and gpx4. Free Radical Biology and Medicine **152**, 175–185 (2020)

[45] Kühn, H., Borchert, A.: Regulation of enzymatic lipid peroxidation: the interplay of peroxidizing and peroxide reducing enzymes1 1this article is part of a series of reviews on "regulatory and cytoprotective aspects of lipid hydroperoxide metabolism." the full list of papers may be found on the homepage of the journal. Free Radical Biology and Medicine **33**(2), 154–172 (2002)

[46] Sun, S., Shen, J., Jiang, J., Wang, F., Min, J.: Targeting ferroptosis opens new avenues for the development of novel therapeutics. Signal transduction and targeted therapy **8**(1), 372 (2023)

[47] Zhou, Q., Meng, Y., Li, D., Yao, L., Le, J., Liu, Y., Sun, Y., Zeng, F., Chen, X., Deng, G.: Ferroptosis in cancer: From molecular mechanisms to therapeutic strategies. Signal Transduction and Targeted Therapy **9**(1), 55 (2024)

[48] Eaton, J.K., Furst, L., Ruberto, R.A., Moosmayer, D., Hilpmann, A., Ryan, M.J., Zimmermann, K., Cai, L.L., Niehues, M., Badock, V., *et al.*: Selective covalent targeting of gpx4 using masked nitrile-oxide electrophiles. Nature chemical biology **16**(5), 497–506 (2020)

[49] Eaton, J.K., Furst, L., Cai, L.L., Viswanathan, V.S., Schreiber, S.L.: Structure–activity relationships of gpx4 inhibitor warheads. Bioorganic & medicinal chemistry letters **30**(23), 127538 (2020)

[50] Eaton, J.K., Ruberto, R.A., Kramm, A., Viswanathan, V.S., Schreiber, S.L.: Diacylfuroxans are masked nitrile oxides that inhibit gpx4 covalently. Journal of the American Chemical Society **141**(51), 20407–20415 (2019)

[51] Xu, C., Xiao, Z., Wang, J., Lai, H., Zhang, T., Guan, Z., Xia, M., Chen, M., Ren, L., He, Y., *et al.*: Discovery of a potent glutathione peroxidase 4 inhibitor as a selective ferroptosis inducer. Journal of Medicinal Chemistry **64**(18), 13312–13326 (2021)

[52] Liu, H., Forouhar, F., Lin, A.J., Wang, Q., Polychronidou, V., Soni, R.K., Xia, X., Stockwell, B.R.: Small-molecule allosteric inhibitors of gpx4. Cell chemical biology **29**(12), 1680–1693 (2022)

[53] Sander, T., Freyss, J., Von Korff, M., Rufener, C.: Datawarrior: an open-source program for chemistry aware data visualization and analysis. Journal of chemical information and modeling **55**(2), 460–473 (2015)

[54] Raman, E.P., Yu, W., Lakkaraju, S.K., MacKerell, A.D.J.: Inclusion of multiple fragment types in the site identification by ligand competitive saturation (silcs) approach. Journal of Chemical Information and Modeling **53**(12), 3384–3398 (2013). PMID: 24245913

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

[56] Galati, S., Di Stefano, M., Martinelli, E., Poli, G., Tuccinardi, T.: Recent advances in in silico target fishing. Molecules **26**(17) (2021)

[57] Sriranganathan, M.K.: Benzbromarone: A review. In: Reference Module in Biomedical Sciences. Elsevier, ??? (2018)

[58] Azevedo, V.F., Kos, I.A., Vargas-Santos, A.B., Pinheiro, G.d.R.C., Paiva, E.d.S.: Benzbromarone in the treatment of gout. Advances in Rheumatology **59**, 37 (2019)

[59] Huang, R.X., Siriwanna, D., Cho, W.C., Wan, T.K., Du, Y.R., Bennett, A.N., He, Q.E., Liu, J.D., Huang, X.T., Chan, K.H.K.: Lung adenocarcinoma-related target gene prediction and drug repositioning. Frontiers in Pharmacology **13**,

936758 (2022)

[60] Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., Sjöstedt, E., Butler, L., Odeberg, J., Dusart, P., *et al.*: A single–cell type transcriptomics map of human tissues. Science advances **7**(31), 2169 (2021)

[61] Thul, P.J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L.M., *et al.*: A subcellular map of the human proteome. Science **356**(6340), 3321 (2017)

[62] Uhlen, M., Karlsson, M.J., Zhong, W., Tebani, A., Pou, C., Mikes, J., Lakshmikanth, T., Forsström, B., Edfors, F., Odeberg, J., *et al.*: A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. Science **366**(6472), 9198 (2019)

[63] Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., *et al.*: Towards a knowledge-based human protein atlas. Nature biotechnology **28**(12), 1248–1250 (2010)

[64] Meyer, N., Penn, L.Z.: Reflecting on 25 years with myc. Nature Reviews Cancer **8**(12), 976–990 (2008)

[65] Duffy, M.J., O'Grady, S., Tang, M., Crown, J.: Myc as a target for cancer treatment. Cancer treatment reviews **94**, 102154 (2021)

[66] Dang, C.V.: Myc on the path to cancer. Cell **149**(1), 22–35 (2012)

[67] Gabay, M., Li, Y., Felsher, D.W.: Myc activation is a hallmark of cancer initiation and maintenance. Cold Spring Harbor perspectives in medicine **4**(6), 014241 (2014)

[68] Lancho, O., Herranz, D.: The myc enhancer-ome: long-range transcriptional regulation of myc in cancer. Trends in cancer **4**(12), 810–822 (2018)

[69] Wang, Y., Zhang, Q., Guo, Y., Wang, Y.: Non-negative contrastive learning. In: ICLR (2024)

[70] Hendrycks, D., Gimpel, K.: Gaussian Error Linear Units (GELUs) (2023)

[71] Chen, L., Fan, Z., Chang, J., Yang, R., Hou, H., Guo, H., Zhang, Y., Yang, T., Zhou, C., Sui, Q., Chen, Z., Zheng, C., Hao, X., Zhang, K., Cui, R., Zhang, Z., Ma, H., Ding, Y., Zhang, N., Lu, X., Luo, X., Jiang, H., Zhang, S., Zheng, M.: Sequence-based drug design as a concept in computational drug design. Nature communications **14**(1), 4217 (2023)

[72] Zhang, X., Gao, H., Wang, H., Chen, Z., Zhang, Z., Chen, X., Li, Y., Qi, Y., Wang, R.: Planet: A multi-objective graph neural network model for protein-ligand binding affinity prediction. Journal of chemical information and modeling

64(7), 2205–2220 (2024)

[73] Gao, B., Qiang, B., Tan, H., Jia, Y., Ren, M., Lu, M., Liu, J., Ma, W.-Y., Lan, Y.: DrugCLIP: Contrasive protein-molecule representation learning for virtual screening. In: Neural Information Processing Systems (2023)

[74] Sakamoto, K., Sogabe, S., Kamada, Y., Matsumoto, S.-i., Kadotani, A., Sakamoto, J.-i., Tani, A.: Discovery of gpx4 inhibitory peptides from random peptide t7 phage display and subsequent structural analysis. Biochemical and Biophysical Research Communications **482**(2), 195–201 (2017)

[75] Eaton, J.K., Furst, L., Ruberto, R.A., Moosmayer, D., Hilpmann, A., Ryan, M.J., Zimmermann, K., Cai, L.L., Niehues, M., Badock, V., *et al.*: Selective covalent targeting of gpx4 using masked nitrile-oxide electrophiles. Nature chemical biology **16**(5), 497–506 (2020)

[76] Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E., Simmerling, C.: ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. Journal of chemical theory and computation **11**(8), 3696–3713 (2015)

[77] Wang, J., Wang, W., Kollman, P.A., Case, D.A.: Automatic atom type and bond type perception in molecular mechanical calculations. Journal of molecular graphics and modelling **25**(2), 247–260 (2006)

[78] Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., Case, D.A.: Development and testing of a general amber force field. Journal of computational chemistry **25**(9), 1157–1174 (2004)

[79] Eastman, P., Swails, J., Chodera, J.D., McGibbon, R.T., Zhao, Y., Beauchamp, K.A., Wang, L.-P., Simmonett, A.C., Harrigan, M.P., Stern, C.D., *et al.*: Openmm 7: Rapid development of high performance algorithms for molecular dynamics. PLoS computational biology **13**(7), 1005659 (2017)

[80] Yang, M., Bo, Z., Xu, T., Xu, B., Wang, D., Zheng, H.: Uni-gbsa: an open-source and web-based automatic workflow to perform mm/gb (pb) sa calculations for virtual screening. Briefings in Bioinformatics **24**(4), 218 (2023)

[81] Hougardy, S.: The floyd–warshall algorithm on graphs with negative cycles. Information Processing Letters **110**(8-9), 279–281 (2010)

Appendix A Supplementary Notes, Tables and Figures

Here is the supplementary ablation study of Uni-Clip. To further substantiate the hypothesis that multimodal training can enhance PLI performance, we conduct experiments that training Uni-Clip on single-modal and multimodal PLI data. All experiments are trained on the MMBind dataset and evaluated on the LIT-PCBA dataset. The experiment results of multimodal training are summarized in Supplementary Table A1, multimodal training outmatches any single-modal training on the LIT-PCBA dataset, with an average improvement of 40%, 30%, and 8% at $EF_{0.5\%}$, $EF_{1\%}$, and $EF_{5\%}$.

As Supplementary Table A2 show, CF-InfoNCE surpasses vanilla InfoNCE on the LIT-PCBA dataset, with an improvement of 15% and 14% at $EF_{0.5\%}$ and $EF_{1\%}$. Before scoring PLIs, non-negative reparameterization like ReLU, can be operated to protein and ligand features, inducing good feature sparsity and robustness. But ReLU output features may suffer from the dead neuron problem during training. So we adopt straight-through Relu to resurrecte the dead neurons while ensuring sparsity (detailed in Methods 4.2). The Supplementary Table A2 reveals the sparse activation function can obtain an improvement of 4% and 13% at $EF_{1\%}$, and $EF_{5\%}$ on the LIT-PCBA dataset. We also conduct an ablation study to evaluate the pretrained models: pretrained Uni-Mol, pretrained Uni-Mol-Graph. Results are summarized in Supplementary Table A2. We can see that performance improves by adding each of these pretrained models, with an averaged improvement of 16% and 30% at $EF_{0.5\%}$ and $EF_{1\%}$.

Table A1: Multimodal training.

Multimodal training				EF ↑		
SC	SG	RC	RG	0.5%	1%	5%
✓	×	×	×	13.195	9.462	3.446
×	✓	×	×	12.602	8.095	3.890
×	×	✓	×	9.416	6.996	2.750
×	×	×	✓	10.484	8.279	3.107
✓	✓	✓	✓	15.727	10.559	3.513

Table A2: Ablation Settings.

Ablation Settings				EF ↑		
Pretrained Uni-Mol	Pretrained Uni-Mol-Graph	Collision-free loss	Sparse activation	0.5%	1%	5%
×	✓	✓	✓	13.012	7.876	3.572
✓	×	✓	✓	14.191	8.294	3.496
✓	✓	×	✓	13.650	9.250	3.763
✓	✓	✓	×	15.083	9.327	3.413
✓	✓	✓	✓	15.727	10.559	3.513

Table A3: Pockets found in the cystal structure (PDBID: 2OBI).

Pocket ID	Druggability Score	Residues
0	0.196	8,110,111,118,121,122,132
1	0.076	21,22,27,32,90,93,94,98,100,101,102,103
2	0.050	47,48,51,52,155,156,158
3	0.049	18,19,20,26,73,104,105,106,114
4	0.041	76,77,105,106,107
5	0.011	57,60,61,97,99
6	0.006	125,127,128,129,135,139,150,151,152,154,155,157
7	0.000	44,45,82,83,84,88,89,92

(a) DP001-DP018

Compound	Conc. (μM)	Inhibition % repeat01	repeat02	Mean	SD
DP001	50.00	-33.6	-21.7	-27.6	6.0
	10.00	1.5	39.5	20.5	19.0
DP002	50.00	-7.2	20.1	6.5	13.6
	10.00	10.0	11.4	10.7	0.7
DP003	50.00	25.2	8.9	17.0	8.1
	10.00	5.3	10.3	7.8	2.5
DP004	50.00	11.7	1.8	6.7	5.0
	10.00	4.5	8.7	6.6	2.1
DP005	50.00	-23.4	2.4	-10.5	12.9
	10.00	5.5	22.0	13.8	8.2
DP006	50.00	-18.1	2.1	-8.0	10.1
	10.00	7.9	10.3	9.1	1.2
DP007	50.00	7.6	4.1	5.9	1.7
	10.00	2.3	-7.9	-2.8	5.1
DP008	50.00	18.6	1.2	9.9	8.7
	10.00	2.7	-16.7	-7.0	9.7
DP009	50.00	27.0	34.9	30.9	3.9
	10.00	7.2	2.1	4.6	2.6
DP010	50.00	11.3	6.7	9.0	2.3
	10.00	3.9	-0.9	1.5	2.4
DP011	50.00	10.3	5.1	7.7	2.6
	10.00	8.0	-26.4	-9.2	17.2
DP012	50.00	5.1	-2.0	1.5	3.6
	10.00	2.4	10.9	6.7	4.3
DP013	50.00	-4.1	12.3	4.1	8.2
	10.00	-3.8	-10.0	-6.9	3.1
DP014	50.00	56.2	71.4	63.8	7.6
	10.00	8.0	38.2	23.1	15.1
DP015	50.00	4.5	9.9	7.2	2.7
	10.00	2.6	6.2	4.4	1.8
DP016	50.00	3.7	9.2	6.4	2.8
	10.00	-4.6	0.6	-2.0	2.6
DP017	50.00	0.9	33.4	17.1	16.2
	10.00	-0.8	11.0	5.1	5.9
DP018	50.00	-0.3	0.9	0.3	0.6
	10.00	-0.3	-7.2	-3.7	3.4

(b) DP019-DP036

Compound	Conc. (μM)	Inhibition % repeat01	repeat02	Mean	SD
DP019	50.00	-24.3	-9.4	-16.8	7.4
	10.00	-18.7	-7.2	-13.0	5.8
DP020	50.00	0.5	-3.9	-1.7	2.2
	10.00	-0.7	0.2	-0.3	0.4
DP021	50.00	78.0	87.0	82.5	4.5
	10.00	81.9	85.9	83.9	2.0
DP022	50.00	-19.0	-21.7	-20.4	1.3
	10.00	-6.5	-3.7	-5.1	1.4
DP023	50.00	-1.4	-19.1	-10.2	8.9
	10.00	-3.0	2.6	-0.2	2.8
DP024	50.00	-24.7	-6.8	-15.8	8.9
	10.00	-0.2	-3.4	-1.8	1.6
DP025	50.00	-11.7	-3.7	-7.7	4.0
	10.00	-7.8	-4.4	-6.1	1.7
DP026	50.00	-1.9	1.7	-0.1	1.8
	10.00	-8.7	-5.7	-7.2	1.5
DP027	50.00	-5.3	3.1	-1.1	4.2
	10.00	-12.1	-6.7	-9.4	2.7
DP028	50.00	-0.5	-15.0	-7.8	7.3
	10.00	-39.1	-28.6	-33.9	5.3
DP029	50.00	-38.4	-56.0	-47.2	8.8
	10.00	-40.6	-30.9	-35.8	4.8
DP030	50.00	-19.8	-13.2	-16.5	3.3
	10.00	-33.7	-2.8	-18.2	15.5
DP031	50.00	-18.3	-3.7	-11.0	7.3
	10.00	-30.9	-36.2	-33.5	2.7
DP032	50.00	-13.9	17.9	2.0	15.9
	10.00	-31.2	4.6	-13.3	17.9
DP033	50.00	-9.2	-24.3	-16.8	7.6
	10.00	-25.5	-20.4	-23.0	2.6
DP034	50.00	-27.3	-11.6	-19.5	7.9
	10.00	-36.2	-49.4	-42.8	6.6
DP035	50.00	-10.6	0.1	-5.2	5.4
	10.00	-45.6	-52.5	-49.0	3.5
DP036	50.00	109.3	105.1	107.2	2.1
	10.00	102.6	97.8	100.2	2.4

Table A4: Enzymatic activity inhibition assay results of 36 compounds (DP001-DP036).

Table A5: Pockets found in the mixMD structure which contains the top 10 key reidues.

Pocket ID	Druggability Score	Residues
0	0.988	46,47,48,**52**,81,125,127,128,**129**,135,136, 137,**138**,139,150,151,**152**,153,154,155,156
1	0.981	46,47,48,81,127,128,129,**130**,135,136,137 139,151,152,153,154,155,156,157,161,164,165
44	0.808	21,22,23,27,28,29,**31**,**32**,52,87,90, 91,93,94,98,100,101,102,103
1429	0.175	125,**142**,150,151,164,165,168,169
1906	0.100	**38**,56,59,68,70,162,166
5661	0.005	130,131,133,**134**,135

Table A6: Constuction of training data in several learning methods

	Database	Training Samples
Uni-Clip	ChEMBL33	853932
TCPI2	ChEMBL27	217,732
ConPlex	BindingDB/DUD-E[1]	12,668/415,204
PLANET	PDBbind 2020/ChEMBL29[2]	15,616/161,547
DrugCLIP	PDBbind 2019/ChEMBL	758,107[3]

[1]ConPlex is trained on BindingDB and fine-tune on DUD-E

[2]Decoys are collected from ChEMBL while binders are from PDBbind

[3]DrugCLIP proposes a data augmentation method, which greatly expands the training set

Table A7: Atom features

features	**size**	**description**
atom type	119	type of atoms including C, N, O, etc, by atomic number
chirality	6	type of chirality like Tetrahedral chirality
degree	11	the degree of an atom in molecule
formal charge	11	integer electronic charge assigned to atom
number of H	9	number of bonded hydrogen atoms
number of radical electrons	5	number of radical electrons
hybridization	5	SP, SP2, SP3, SP3D, SP3D2
aromaticity	1	whether an atom is part of an aromatic system
in ring	1	whether an atom is within a ring structure

Table A8: Bond features

features	size	description
bond type	4	SINGLE, DOUBLE, TRIPLE, AROMATIC
stereo	6	NONE, Z, E, CIS, TRANS, ANY,
conjugated	1	whether the bond is conjugated,
SPD	512	the shortest distances between each pair of connected atoms by the Floyd-Warshall algorithm[81]

(a) Uni-Clip-SG

Rank	Gene Name	Uniprot ID	Score	RNA Expression
1	MSR1	P21757	0.306	2.6
2	MAPK12	P53778	0.305	18.4
3	RAD51	Q06609	0.304	20.9
4	CD274	Q9NZQ7	0.292	7.4
5	SLC25A11	Q02978	0.288	64.1
6	HNF1B	P35680	0.275	2.8
7	XRCC3	O43542	0.266	51.3
8	TBXT	O15178	0.265	8.2
9	TRIP13	Q15645	0.262	62.3
10	CHCHD7	Q9BUK0	0.261	46.5

(b) Uni-Clip-SC

Rank	Gene Name	Uniprot ID	Score	RNA Expression
1	MSR1	P21757	0.380	2.6
2	CD274	Q9NZQ7	0.363	7.4
3	RAD51	Q06609	0.325	20.9
4	MAX	P61244	0.319	47.1
5	XRCC3	O43542	0.315	51.3
6	MSLN	Q13421	0.298	28.4
7	RAD51C	O43502	0.296	34.4
8	CRTC1	Q6UUV9	0.294	9.6
9	HNF1B	P35680	0.294	2.8
10	DDIT3	P35638	0.292	57.8

(c) Uni-Clip-sum

Rank	Gene Name	Uniprot ID	Score	RNA Expression
1	XRCC3	O43542	0.327	51.3
2	RXFP2	Q8WXD0	0.323	0.0
3	MAPK12	P53778	0.312	18.4
4	RAD51C	O43502	0.306	34.4
5	KLF6	Q99612	0.293	43.3
6	MSR1	P21757	0.292	2.6
7	RAD51	Q06609	0.288	20.9
8	SLC25A11	Q02978	0.287	64.1
9	CD274	Q9NZQ7	0.286	7.4
10	DDIT3	P35638	0.278	57.8

(d) Uni-Clip-RC

Rank	Gene Name	Uniprot ID	Score	RNA Expression
1	RXFP2	Q8WXD0	0.456	0.0
2	XRCC3	O43542	0.376	51.3
3	RAD51C	O43502	0.369	34.4
4	PARN	O95453	0.362	22.0
5	KLF6	Q99612	0.361	43.3
6	POLD1	P28340	0.355	36.5
7	**MYC**	**P01106**	**0.344**	**153.3**
8	CREB1	P16220	0.343	20.3
9	MAPK12	P53778	0.324	18.4
10	DDX6	P26196	0.320	44.4

Table A9: Top ten genes selected by (a) Uni-Clip-SG model, (b) Uni-Clip-SC model, (c) Uni-Clip-sum model and (d) Uni-Clip-RC model.

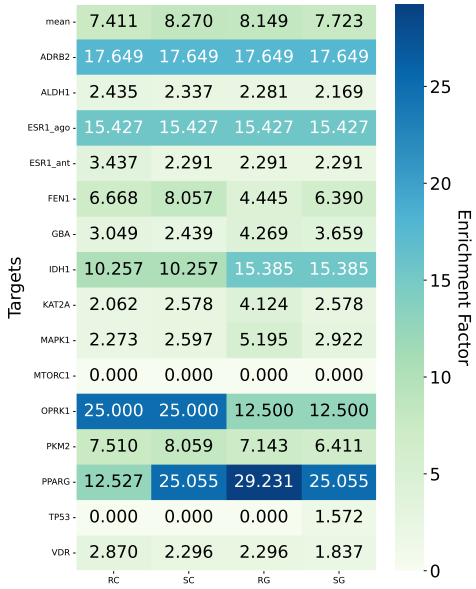


Fig. A1: Impact of multimoldal protein ligand data on the $EF_{1\%}$ performance of UniClip for different targets on the LIT-PCBA dataset.

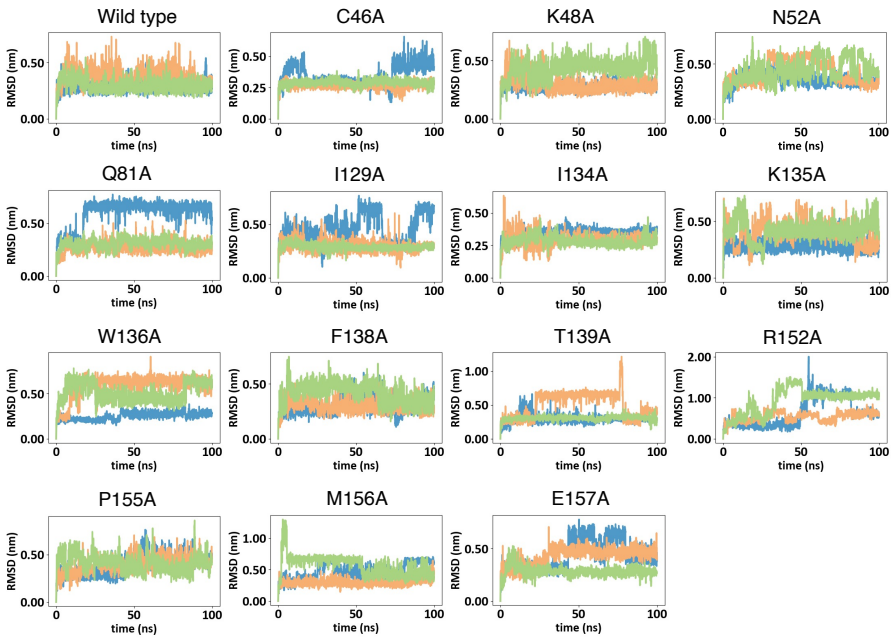


Fig. A2: Line plot of RMSD versus time for different mutant protein complexes with DP021, where different colors represent different replicates.