Sequence-based Drug-Target Complex Pre-training Enhances Protein-Ligand Binding Process Predictions Tackling Crypticity

Shuo Zhang[1,2*], Li Xie[1], Daniel Tiourine[1], Lei Xie[1,2,3*]

[1]Department of Computer Science, Hunter College, The City University of New York, New York, 10065, NY, USA.
[2]Helen & Robert Appel Alzheimer's Disease Research Institute, Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, Cornell University, New York, 10065, NY, USA.
[3]Ph.D. Program in Computer Science, The Graduate Center, The City University of New York, New York, 10016, NY, USA.

*Corresponding author(s). E-mail(s): sz780@hunter.cuny.edu; lei.xie@hunter.cuny.edu;
Contributing authors: lixie9508@gmail.com; daniel.tiourine71@myhunter.cuny.edu;

Abstract

Predicting protein-ligand binding processes, such as affinity and kinetics, is critical for accelerating drug discovery. However, many existing computational methods face key limitations, including insufficient integration of comprehensive databases, inadequate representation of protein structural dynamics, and incomplete modeling of microscale protein-ligand interactions. To address these challenges, we introduce ProMoNet, a sequence-based pre-training and fine-tuning framework to enhance protein-ligand binding process prediction. ProMoNet connects protein and molecular foundation models to expand data coverage and enhance diversity, and it integrates large-scale binding site pre-training with efficient fine-tuning for affinity and kinetics prediction. During pre-training, it effectively models microscale protein-ligand interactions and captures the dynamic nature of proteins, including binding site crypticity, without relying on 3-dimensional structural inputs. Notably, ProMoNet's pre-training module surpasses or matches state-of-the-art structure-based methods in identifying exposed and cryptic binding sites. In the fine-tuning stage, it transfers pre-trained knowledge, achieving superior performance in affinity and kinetics prediction tasks with

high computational efficiency. The combination of ProMoNet's powerful modeling capabilities and demonstrated success across multiple tasks highlight its potential for broad applications in drug discovery.

Keywords: Deep learning, machine learning, transfer learning, drug-target interaction, protein-ligand interaction, binding affinity, binding kinetics

1 Introduction

Characterizing the binding process between proteins and small-molecule ligands is a critical step in drug discovery. These binding processes encompass structural properties like binding sites and poses [1, 2], thermodynamic properties including equilibrium constants K_{i}, K_{d} [3], and kinetic properties such as rate constants k_{on}, k_{off} [4]. Together, these properties provide essential insights into protein-ligand interactions, facilitating more effective compound screening [5]. To accelerate the characterization of protein-ligand binding processes and reduce associated costs, various computational approaches have been developed, demonstrating promising performance [6].

Despite significant progress in predicting characteristics of protein-ligand binding processes, existing methods still face several critical limitations. First, *the datasets used for training models are often limited and biased.* When experimental data for a specific binding characteristic is scarce, the performance of the trained models tends to suffer [7]. Additionally, while binding characteristics are closely interrelated [5], existing models often focus exclusively on the dataset for the characteristic being predicted and fail to leverage relevant information from datasets corresponding to other characteristics. This neglect of cross-characteristic information further limits their ability to generalize and accurately predict binding processes. Second, *the dynamic nature of protein structures is often overlooked or inefficiently captured.* Many structure-based methods rely solely on rigid 3-dimensional (3D) protein structures, failing to account for the dynamic interactions during binding [8–10]. Although some approaches address dynamics, they often rely on resource-intensive molecular dynamics (MD) simulations [11–15] or time-consuming diffusion-based generative modeling [16, 17], making them impractical for high-throughput compound screening. Third, *many models rely solely on protein- and molecule-level representations, neglecting their microscale interactions* [18, 19]. These interactions are essential for capturing the fundamental mechanisms of binding and are critical for accurately predicting characteristics of protein-ligand binding processes [20, 21].

To address the aforementioned limitations, we present **ProMoNet**, a sequence-based pre-training and fine-tuning framework designed to enhance protein-ligand binding process prediction. ProMoNet bridges independent protein and molecular foundation models, leveraging recent advancements in foundation models [22–24] and their pre-training and fine-tuning strategies [25, 26]. ProMoNet consists of two key modules. The **pre-training module**, ProMoSite, learns microscale protein-ligand interactions and captures dynamic structural information, including binding site crypticity by training on binding site annotations in protein-ligand complexes. The

fine-tuning module, ProMoBind, integrates ProMoSite's pre-trained capabilities to be fine-tuned for predicting characteristics of protein-ligand binding processes, including binding affinity and kinetics.

ProMoSite takes advantage of the transferability of foundation models, which are pre-trained on extensive databases to increase data size and diversity [22]. Independent protein and molecular foundation models [27, 28] are employed as encoders to extract embeddings from protein sequences and ligand SMILES strings [29]. Protein contact maps and ligand distance maps are derived from input data, providing structural information. Using the aforementioned information, ProMoSite models and updates the protein-ligand microscale interaction embeddings by satisfying the triangle inequality constraints [9, 30] in protein contact maps and ligand distance maps, enabling the capture of protein dynamics by modeling interactions in multistate protein conformations [31, 32]. Furthermore, ProMoSite employs a self-attention mechanism to model the physical and chemical effects [33] underlying protein-ligand interactions. To pre-train ProMoSite, we leverage protein-ligand binding site annotations from large-scale, open-source databases [34, 35] as signals of microscale interactions, enabling ProMoSite to effectively learn and predict them. Building on the pre-trained, frozen ProMoSite and independent foundation models, ProMoBind integrates their predictions and embeddings using a simple yet effective architecture that can be efficiently fine-tuned for downstream tasks, such as predicting protein-ligand binding affinity and kinetics. Specifically, ProMoSite's predicted residue scores act as a gating mechanism in ProMoBind, modulating the contribution of individual residues to the prediction of protein-ligand binding characteristics.

To comprehensively assess ProMoNet's performance, we evaluate its two modules, ProMoSite and ProMoBind, on diverse tasks related to protein-ligand binding process prediction. ProMoSite is evaluated on protein-ligand binding site prediction during pre-training. We evaluate its ability to predict both *exposed* binding sites, which are readily accessible on the protein surface, and *cryptic* binding sites, which emerge due to dynamic structural changes in the presence of ligands [14, 36]. Using only sequence data as input, ProMoSite surpasses or matches state-of-the-art methods, including those that rely on 3D protein structures, across multiple benchmarks. Notably, for cryptic binding site identification, ProMoSite is pre-trained solely on general binding site databases without using MD simulation data or protein conformations related to cryptic binding sites, yet achieves the best performance. These results highlight ProMoSite's ability to tackle binding site crypticity and capture the dynamics of multistate protein structures, even without explicit structural inputs. ProMoBind is evaluated on two downstream tasks: protein-ligand binding affinity prediction and protein-ligand binding kinetics prediction. Binding affinity prediction is a critical task and well-studied in drug discovery [6], while binding kinetics prediction has received relatively less attention, despite growing evidence that drug efficacy depends not only on the strength of drug-target interactions but also on binding speed and duration [37–39]. Existing computational methods for binding kinetics prediction often rely on MD simulations, enhanced sampling techniques, or 3D protein structures. However, MD simulations are computationally expensive, and reliable protein structures are often unavailable for many targets of interest. ProMoBind addresses these limitations by

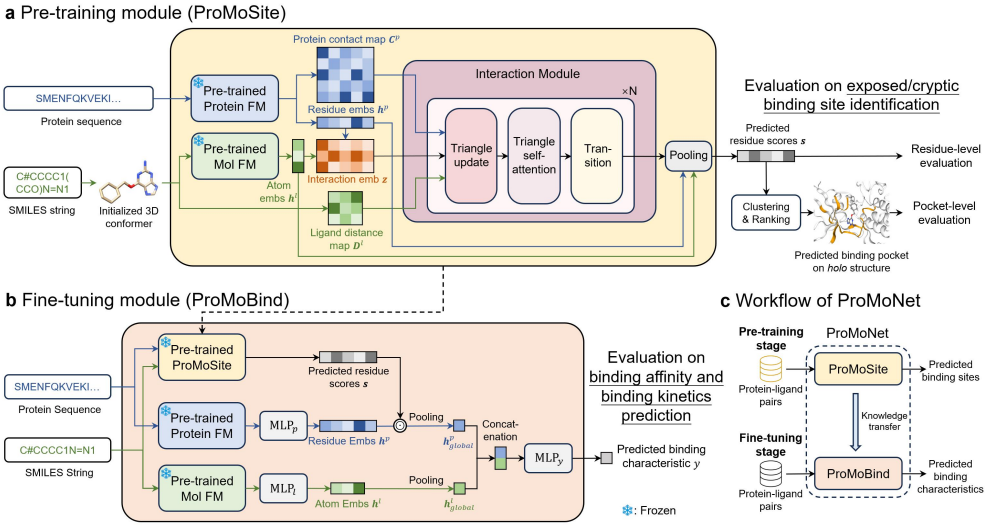

Fig. 1 Overview of ProMoNet. ProMoNet consists of two key modules: ProMoSite, the pre-training module, and ProMoBind, the fine-tuning module. (**a**) ProMoSite employs ESM-2 and Uni-Mol as foundation models to process protein sequences and ligand SMILES strings. Protein-ligand interaction embeddings are computed and iteratively updated through an interaction module constrained by protein contact maps and ligand distance maps to capture protein dynamics. These embeddings are then used to compute residue-level scores, which are pooled for exposed/cryptic binding site identification and evaluated using both residue-level and pocket-level metrics. (**b**) ProMoBind builds on the pre-trained ProMoSite to predict protein-ligand binding characteristics by integrating residue scores from ProMoSite into residue embeddings. These embeddings, along with pooled global representations of proteins and ligands, are concatenated and passed through a trainable MLP-based head for prediction. ProMoBind's efficient fine-tuning and inference design make it scalable and suitable for high-throughput compound screening. (**c**) Pre-training and fine-tuning workflow of ProMoNet. FM denotes foundation model throughout the figure.

requiring only sequence data and leveraging the pre-trained ProMoSite to improve performance. Across both tasks, ProMoBind outperforms state-of-the-art baselines while achieving high efficiency. These results demonstrate ProMoNet's potential as a robust and efficient sequence-based framework for high-throughput compound screening and broader applications in drug discovery.

2 Results and discussion

2.1 Overview of ProMoNet

The overall architecture of ProMoNet, including its two key modules, ProMoSite and ProMoBind, is illustrated in Figure 1. ProMoSite (Fig. 1a) serves as the pre-training module that connects independent protein and molecular foundation models via pre-training on binding site databases [34, 35]. It predicts whether an amino acid residue physically interacts with a ligand (i.e., binding site residue) and is evaluated for both exposed and cryptic binding site identification. ProMoBind (Fig. 1b),

the fine-tuning module, builds on the pre-trained ProMoSite to be efficiently fine-tuned on downstream protein-ligand binding characteristics prediction tasks, such as binding affinity and kinetics prediction. The pre-training and fine-tuning workflow of ProMoNet is illustrated in Fig. 1c. The details of ProMoNet are described in the following paragraphs.

In ProMoSite (Fig. 1a), we adopt ESM-2 [27] as the protein foundation model and use Uni-Mol [28] as the molecular foundation model. Both ESM-2 and Uni-Mol are pre-trained and the parameters are kept frozen during the pre-training of ProMoSite. To represent protein-ligand pairs, ProMoSite takes the protein sequences and the ligand SMILES strings [29] as input. The protein sequences are fed into ESM-2 to get residue embeddings and predict protein contact maps, while the SMILES strings are used to initialize 3D conformers via RDKit [40] to compute distance maps and as the input of Uni-Mol to get atom embeddings. Based on each pair of residue embedding and atom embedding, ProMoSite forms a protein-ligand interaction embedding to model the microscale interactions between protein residues and ligand atoms. Furthermore, the interaction embedding is updated via ProMoSite's interaction module which utilizes protein contact maps and ligand distance maps as triangle inequality constraints. Unlike previous works [9, 30] that aim to satisfy the constraints solely on distances, ProMoSite avoids using rigid 3D structure-based distance maps for proteins while using contact maps instead. By doing so, the dynamic structural information of multistate protein conformations that are contained in a predicted contact map [31, 32] can be captured when modeling microscale protein-ligand interactions. Besides, ProMoSite's interaction module uses the self-attention mechanism to capture the physical and chemical effects [33] between protein-ligand interactions for more comprehensive modeling. The updated interaction embedding from the interaction module as well as the initial residue and atom embeddings are fed into the pooling module to quantify a score between each residue and the entire ligand, where a higher score indicates a greater likelihood that the residue belongs to a binding site of the given ligand. Such scores can be directly used for residue-level metrics (e.g. ROC-AUC, PR-AUC) when evaluating ProMoSite for binding site identification. To additionally evaluate the identified binding site in 3D space, we post-process the scored residues with clustering and ranking. The filtered residues are mapped onto known *holo* structures to compute pocket-level metrics like DCA (distance from the center of the predicted binding pocket to the closest ligand heavy atom).

In ProMoBind (Fig. 1b), to predict the binding characteristic of a given protein-ligand pair, the same input formats are adopted as those used in ProMoSite, while ESM-2 and Uni-Mol are still utilized to extract protein residue embeddings and ligand atom embeddings. To achieve efficiency in fine-tuning, we freeze all parameters in these pre-trained foundation models and adapt trainable multi-layer perceptions (MLP) after the extracted embeddings to improve generalization and transferring ability on downstream tasks [26, 41]. Then ProMoBind incorporates the predicted residue scores from pre-trained ProMoSite by taking an element-wise product with the residue embeddings, which can apply weighting to these embeddings according to the protein-ligand interaction likelihoods that exist in predicted scores. Moreover, the resulting residue embeddings and atom embeddings are pooled to get corresponding

global embeddings of protein and ligand. Finally, ProMoBind concatenates the pair of global embeddings and uses a trainable MLP-based prediction head to predict the desired protein-ligand binding characteristic. In ProMoBind, only the parameters of three MLP modules are fine-tuned, and the foundation models only require a single forward pass during inference. Such efficiency of ProMoBind in training and inference is invaluable for resource-intensive tasks like high-throughput compound screening.

2.2 ProMoSite excels in predicting general binding sites

To evaluate ProMoSite's performance for identifying *general binding sites*, we used the scPDB v2017 database [34] for training and cross-validation, and benchmark datasets including COACH420 and HOLO4K [8] for testing. To avoid data leakage, we adopted the same strategy and data split as in [42], ensuring that the proteins in the training and test sets had neither similar global structures nor similar binding sites. For detailed dataset statistics and processing steps, please refer to the Dataset subsection in the Methods section.

For a comprehensive evaluation, we compared ProMoSite against two types of state-of-the-art methods, including 3D structure-based (Fpocket [43], Deepsite [44], P2Rank [8], Kalasanty [45], PUResNet [46], DeepPocket [42]) and 1D sequence-based methods (HoTS [47], DeepProSite [48]). The 3D structure-based methods require 3D protein coordinate information as input. Their predicted binding sites are evaluated in 3D space using metrics such as DCA, where predictions with DCAs < 4Å are considered successful [8, 42]. Based on the DCA criterion, we evaluated the models' ranking capability by measuring the success rates for the top n-ranked predictions, following [8, 42], where n is the number of annotated binding sites for a given protein. For 1D sequence-based methods, only protein sequences are required as input. Since these methods directly predict the probabilities of protein residues belonging to binding sites, binary classification metrics are typically used [48]. Specifically, we used the Area under the receiver operating characteristic curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC) for comparison.

As shown in Fig. 2a, ProMoSite outperforms all 3D structure-based baselines, achieving an averaged Top-n success rate of 68.30% $\pm$ 0.99 on COACH420 and 73.46% $\pm$ 0.65 on HOLO4K, respectively. These results demonstrate that ProMoSite, which requires only protein sequences as input, effectively incorporates 3D structural information that emerged from the pre-trained ESM-2 model to identify ligand binding sites in 3D space, including on both single-chain and multi-chain proteins. The ablation study results shown in Supplementary Table 1 further validate the effectiveness of ProMoSite's interaction module and pooling module, which play critical roles in modeling protein-ligand interactions.

As shown in Fig. 2b and Supplementary Fig. 1, when compared to 1D sequence-based methods, ProMoSite significantly outperforms them on COACH420 (ROC-AUC $= 0.933 \pm 0.001$; PR-AUC $= 0.530 \pm 0.003$) and HOLO4K (ROC-AUC $= 0.927 \pm 0.001$; PR-AUC $= 0.516 \pm 0.009$). Compared with HoTS, which uses 1D-CNN [49] and Morgan fingerprints [50] to extract protein and molecular representations, respectively, ProMoSite leverages pre-trained protein and molecular foundation models to

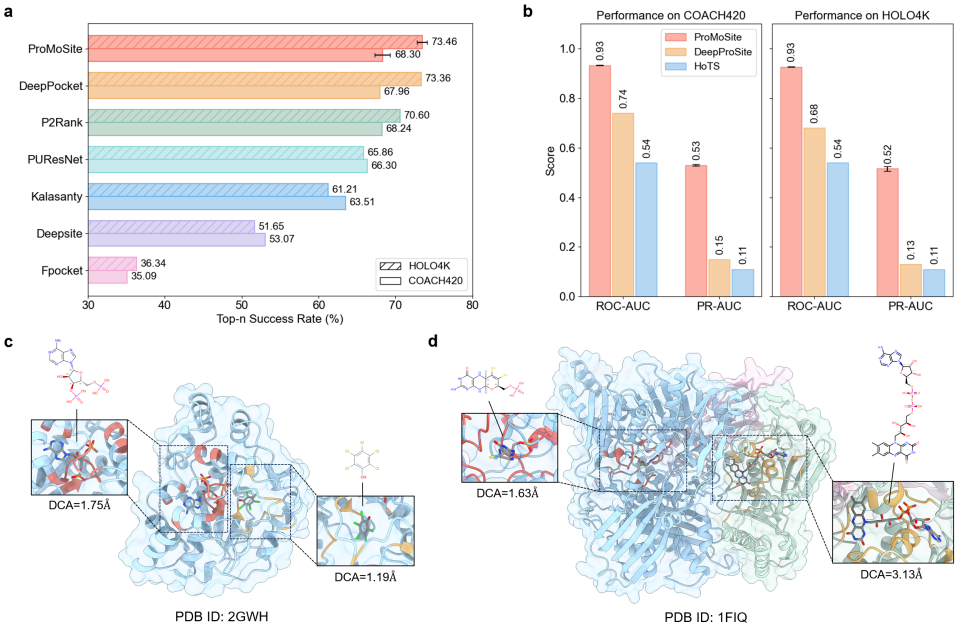

Fig. 2 Comparison of model performance in predicting general binding sites. (**a**) ProMoSite outperforms all 3D-based baselines on COACH420 and HOLO4K, achieving high Top-n success rates for identifying ligand binding sites in 3D space using only sequence data as input. (**b**) ProMoSite significantly outperforms sequence-based baselines on COACH420 and HOLO4K by leveraging pre-trained protein and molecular foundation models and incorporating geometric constraints via contact maps and distance maps. (**c**, **d**) Representative examples of ligand-specific binding site predictions by ProMoSite on proteins with multiple ligands, demonstrating its capability to accurately predict specific binding sites for each ligand.

improve model generalization. Although HoTS explicitly models ligand-protein interactions using transformer blocks [51], it lacks geometric constraints in protein-ligand binding. In contrast, ProMoSite incorporates protein contact maps and ligand distance maps as triangle inequality constraints to model microscale ligand-protein interactions, achieving significantly better performance. Compared with DeepProSite, which uses ESMFold-predicted 3D protein structures and incorporates a pre-trained protein language model for sequence embeddings, ProMoSite eliminates the need for 3D protein structures. Instead, it uses predicted protein contact maps to capture the dynamic structural information of multistate protein conformations. The superior performance of ProMoSite demonstrates the effectiveness of our approach.

Furthermore, it is important to note that ProMoSite is designed to predict *ligand-specific* binding sites when a given protein binds to multiple different ligands at different sites. While most methods only identify unspecific binding sites, a reliable prediction method should be able to identify the specific binding site for a query ligand [52]. Among the baselines in our experiments, only HoTS and DeepProSite can predict ligand-specific binding sites. However, DeepProSite requires retraining for each

ligand type, limiting its practical usability. In contrast, both HoTS and ProMoSite directly take ligand information as input, enabling the same model to generalize across any ligand structure. To demonstrate ProMoSite's capability of identifying ligand-specific binding sites, we provide representative visualizations (Fig. 2c, d) of its predictions in scenarios where a protein binds to multiple ligands at different sites. Fig. 2c shows a prediction example from COACH420 on human sulfotransferase SULT1C2 in complex with PAP and pentachlorophenol (PDB ID: 2GWH). ProMoSite accurately identifies the two specific binding sites corresponding to their ligands. Similarly, Fig. 2d presents a prediction example from HOLO4K on xanthine oxidase in complex with FAD and molybdopterin cofactor (PDB ID: 1FIQ). In this multi-chain protein example, ProMoSite accurately identifies ligand-specific binding sites across different chains, further demonstrating its effectiveness.

2.3 ProMoSite achieves state-of-the-art performance in detecting cryptic binding sites

Following our analysis of general binding site prediction, we also evaluate ProMoSite's performance in identifying *cryptic binding sites*, as accurately detecting these sites is crucial for understanding allosteric regulation and screening potential drugs for proteins currently considered undruggable [36]. For this task, we trained ProMoSite using the scPDB v2017 database [34] and evaluated it on the PocketMiner validation and test sets [14], which are specifically designed for cryptic binding site prediction. Details of these datasets are provided in the Methods section.

We compared ProMoSite with several state-of-the-art models. CryptoSite [53] and PocketMiner [14] require 3D protein structures obtained from MD simulations and directly output residue-level probabilities for cryptic binding sites. NeuralPLexer [16] and DynamicBind [17] are diffusion-based generative models that can reveal cryptic binding sites by generating 3D ligand-protein binding complexes. We evaluated two variants of NeuralPLexer based on the type of protein-related input: NeuralPLexer-fasta, which uses protein sequences, and NeuralPLexer-AF2, which utilizes AlphaFold2 [30]-predicted 3D protein structures. Additionally, we included P2Rank, which requires 3D protein structures as input but focuses on general binding site prediction. To use P2Rank for cryptic binding site prediction, we provide 3D *apo* protein structures as input, which haven't formed the potential cryptic pockets already.

For evaluation, we adopted different metrics depending on the model outputs. For CryptoSite and PocketMiner, we used ROC-AUC and PR-AUC following [14], as they output residue-level probabilities. For NeuralPLexer, DynamicBind, and P2Rank, whose predictions are either generated 3D complexes or 3D binding sites, we used the DCA metric. Additionally, for NeuralPLexer and DynamicBind, residues within 7Å of any ligand heavy atom in the generated binding complex were considered as cryptic binding site residues. To ensure consistency, all predicted residues were mapped onto experimental *holo* protein structures when calculating DCA. More details of the experimental settings are provided in the Methods section.

Fig. 3a presents a comparison of the ROC and PR curves for ProMoSite, CryptoSite, and PocketMiner on the test set. The ROC curves demonstrate that ProMoSite

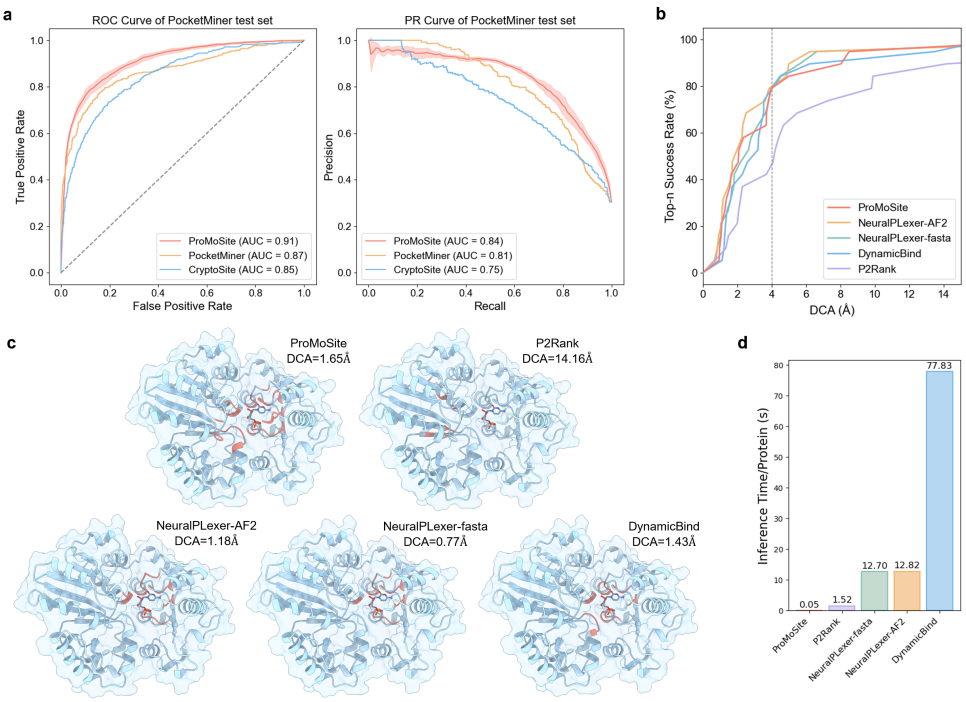

Fig. 3 Comparison of model performance in predicting cryptic binding sites. (**a**) ROC and PR curves comparing ProMoSite with CryptoSite and PocketMiner on the test set. ProMoSite achieves the highest ROC-AUC and PR-AUC, demonstrating superior performance even without requiring MD-derived data or 3D protein structures. (**b**) Top-n success rates as a function of DCA values for ProMoSite and baselines. At a 4.0Å threshold, ProMoSite matches NeuralPLexer-AF2 and NeuralPLexer-fasta while outperforming DynamicBind and P2Rank. (**c**) Visualization of cryptic binding site predictions for α-2,3-sialyltransferase using the *apo* structure (PDB ID: 4V38), mapped onto the experimental *holo* structure (PDB ID: 4V3B). ProMoSite, NeuralPLexer-AF2, NeuralPLexer-fasta, and DynamicBind accurately identify the cryptic binding site, whereas P2Rank fails due to its reliance on rigid protein structures. (**d**) Inference speed comparison among models. ProMoSite achieves the fastest inference time (0.05s/protein) while maintaining cryptic binding site identification performance comparable to generative models.

outperforms CryptoSite and PocketMiner, achieving a significantly higher ROC-AUC of 0.91 $\pm$ 0.01. Similarly, the PR curves show that ProMoSite maintains superior performance across a broader range of recall levels, achieving a higher PR-AUC of 0.84 $\pm$ 0.01. These results highlight ProMoSite's effectiveness. Even when trained solely on a general binding site dataset and without requiring MD-derived data or 3D protein structures, it surpasses state-of-the-art models that depend on 3D structures obtained from MD simulations.

In Fig. 3b, we compare the Top-n success rate as a function of DCA values for ProMoSite and relevant baselines. The detailed DCA values are listed in Supplementary Table 2. ProMoSite achieves performance comparable to the generative models NeuralPLexer-AF2, NeuralPLexer-fasta, and DynamicBind, while significantly

outperforming P2Rank. Specifically, at the commonly used threshold of 4.0Å, ProMoSite, NeuralPLexer-AF2, and NeuralPLexer-fasta all achieve a success rate of 78.9%, while DynamicBind achieves 73.7%. However, P2Rank, designed for general binding site prediction, only achieves a success rate of 42.1%. As a representative example, Fig. 3c visualizes the cryptic binding site predictions of these models when provided with the *apo* structure or sequence of α-2,3-sialyltransferase (PDB ID: 4V38), mapped onto the corresponding experimental *holo* structure (PDB ID: 4V3B). The visualizations show that all models except P2Rank accurately identify the cryptic binding site, which emerges through secondary structure motion. P2Rank fails to detect this site that is absent in the *apo* structure, as it treats protein structures as rigid and cannot capture protein dynamics. In contrast, when modeling microscale protein-ligand interactions, ProMoSite uses predicted protein contact map as geometric constraints, which can capture the dynamic structural information of multistate protein conformations [31, 32].

Further analysis in Fig. 3d compares the inference speed of these models. While the generative models NeuralPLexer-AF2, NeuralPLexer-fasta, and DynamicBind exhibit strong cryptic binding site identification performance, their inference times per protein are considerably slower due to their reliance on diffusion models [54] (NeuralPLexer-AF2: 12.82s, NeuralPLexer-fasta: 12.70s, DynamicBind: 77.83s). P2Rank, with an inference time of 1.52s per protein, is significantly faster but delivers much worse performance in cryptic binding site identification. In contrast, ProMoSite achieves cryptic binding site identification performance on par with the generative models while offering the fastest inference speed among all tested models at just 0.05s per protein. This combination of speed and accuracy demonstrates ProMoSite's potential for large-scale virtual screening and drug discovery applications.

2.4 ProMoBind outperforms baselines in binding affinity prediction

Following the evaluation of ProMoSite's superior performance in predicting general and cryptic binding sites, we now assess its downstream application in predicting binding characteristics using ProMoBind, which leverages the outputs from pre-trained ProMoSite. Specifically, we focus on binding affinity, one of the most critical and representative binding characteristics. To evaluate ProMoBind's performance, we conducted experiments on the commonly used PDBBind v2020 [67] benchmark. We used four metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Pearson correlation coefficient, and Spearman correlation coefficient. ProMoSite was pre-trained on the scPDB v2017 database [34] to predict residue scores, which were then integrated into ProMoBind. To assess the impact of these scores, we developed a variant of ProMoBind for comparison, named ProMoBind-Base, which excludes ProMoSite predicted scores for comparison.

For the experiments on PDBBind v2020, we applied a temporal split following [9, 21], where data points published before 2019 were used for training and validation, and those published during or after 2019 formed the test set. Each sample in PDBBind v2020 includes an experimentally determined 3D protein-ligand complex and its binding affinity, expressed in negative log-transformed dissociation ($\mathrm{p}K_\mathrm{d}$) or inhibition

Table 1 Comparison of model performance in predicting binding affinity on PDBBind v2020. The protein-related input type for each model is indicated. '1D Seq → 2D Map' denotes models that use both 1D protein sequences and their predicted 2D contact maps as input. The best results are highlighted in bold, and the second-best results are underlined.

Methods	Protein-related Input	RMSE ↓	MAE ↓	Pearson ↑	Spearman ↑
Pafnucy [55]	3D Complex	1.435	1.144	0.635	0.587
OnionNet [56]	3D Complex	1.403	1.103	0.648	0.602
IGN [57]	3D Complex	1.404	1.116	0.662	0.638
SIGN [20]	3D Complex	1.373	1.086	0.685	0.656
SMINA [58]	3D Struct	1.466	1.161	0.665	0.663
GNINA [59]	3D Struct	1.740	1.413	0.495	0.494
dMaSIF [60]	3D Struct	1.450	1.136	0.629	0.588
TankBind [9]	3D Struct	1.345	1.060	0.718	0.689
TransformerCPI [61]	1D Seq	1.493	1.201	0.604	0.551
MolTrans [62]	1D Seq	1.599	1.271	0.539	0.474
GraphDTA [19]	1D Seq	1.564	1.223	0.612	0.570
DrugBAN [63]	1D Seq	1.480	1.159	0.657	0.612
DGraphDTA [64]	1D Seq → 2D Map	1.493	1.201	0.604	0.551
WGNN-DTA [65]	1D Seq → 2D Map	1.501	1.196	0.605	0.562
STAMP-DPI [66]	1D Seq → 2D Map	1.503	1.176	0.653	0.601
PSICHIC [21]	1D Seq → 2D Map	1.314	**1.015**	0.710	0.686
ProMoBind-Base	1D Seq	<u>1.290</u>	1.041	<u>0.725</u>	<u>0.694</u>
ProMoBind	1D Seq → 2D Map	**1.268**	<u>1.020</u>	**0.738**	**0.707**

(pK_i) constant values. Accordingly, we compared ProMoBind against state-of-the-art baselines that utilize different types of protein-related inputs, including 3D protein-ligand complex structures (Pafnucy [55], OnionNet [56], IGN [57], and SIGN [20]), 3D *holo* protein structures (SMINA [58], GNINA [59], dMaSIF [60], and TankBind [9]), protein sequences (TransformerCPI [61], MolTrans [62], GraphDTA [19], and DrugBAN [63]), and protein sequences combined with predicted protein contact maps (DGraphDTA [64], WGNN-DTA [65], STAMP-DPI [66], and PSICHIC [21]). The results of all models are summarized in Table 1, where ProMoBind demonstrates the best overall performance, surpassing all baselines that use various types of protein-related inputs. Even ProMoBind-Base, which excludes the ProMoSite predicted scores, achieves the second-best scores on all evaluation metrics excluding MAE. This highlights the superiority and robustness of the protein and molecular foundation models pre-trained on diverse datasets, as well as the effectiveness of our approach in leveraging these models. Compared to ProMoBind-Base, ProMoBind achieves overall significant performance improvements (Table 1 and Supplementary Table 3), indicating that the incorporated ProMoSite predicted scores effectively capture microscale protein-ligand interaction information in our pre-training, thereby enhancing the model's ability to predict binding affinity more accurately.

To evaluate the efficiency and practicality of ProMoBind in large-scale virtual screening, we measured both its training and inference speeds. ProMoBind achieves an average training time of 39.8 minutes, which is overall efficient compared to the baseline training speeds listed in Supplementary Table 4. This highlights the effectiveness of ProMoBind's architecture, enabling it to be fine-tuned efficiently even when

incorporating the pre-trained ProMoSite and the foundation models. For inference, ProMoBind processes input sequentially with a batch size of 1 and achieves an average inference time of 0.0012 seconds per protein-ligand pair, making it the fastest model among all baselines, as shown in Supplementary Table 4. This exceptional efficiency underscores ProMoBind's scalability and practicality for high-throughput virtual screening in real-world drug discovery scenarios.

2.5 ProMoBind enables out-of-distribution binding kinetics prediction

Table 2 Comparison of model performance in predicting $k_{\mathbf{off}}$. The source of input for each model is indicated. The best-performing results are highlighted in bold, while the second-best results are underlined.

Methods	Source of Input	RMSE ↓	MAE ↓	Pearson ↑	Spearman ↑
Random Forest	ESM-2 + ECFP4	<u>0.994</u>	<u>0.720</u>	0.623	0.654
	ESM-2 + Uni-Mol	1.050	0.789	0.637	0.656
ProMoBind-Base	ESM-2 + Uni-Mol	1.024	0.780	<u>0.667</u>	<u>0.668</u>
ProMoBind	ESM-2 + Uni-Mol + ProMoSite	**0.927**	**0.707**	**0.723**	**0.720**

To further evaluate ProMoBind's performance in downstream tasks of predicted binding characteristics, we focus on binding kinetics prediction, an essential aspect of drug discovery and protein-ligand interaction studies [37–39]. Binding kinetics rates provide critical insights into the residence time of a ligand on its target, influencing drug efficacy and safety profiles. Accurate prediction of binding kinetics is therefore crucial for optimizing lead compounds in drug development. In our experiment, we collected dissociation rate constant (k_{off}) data from PDBbind-koff-2020 [68] and BindingDB [69]. Each sample in the dataset consists of a ligand structure, a protein sequence, and an experimentally determined k_{off} value. To validate the model's generalization capability in an out-of-distribution (OOD) scenario, we employed a random scaffold split with an 8:1:1 ratio for the training, validation, and testing sets, ensuring that the testing set contains ligands with distinct scaffolds not present in the training or validation sets.

For comparison, we adopted the random forest (RF) model as the baseline, which is a typical method for binding kinetics prediction [39, 68]. Specifically, the protein inputs were derived from ESM-2 embeddings, and the ligand inputs were represented using either ECFP4 (Extended-Connectivity Fingerprints) [50] and Uni-Mol embeddings. The comparison results, summarized in Table 2 and Supplementary Table 5, show that ProMoBind achieves the best performance across all metrics. Notably, even when only using the embeddings from foundation models (ESM-2 + Uni-Mol), ProMoBind-Base outperforms the RF baseline, demonstrating the advantages of our architecture in leveraging pre-trained embeddings for binding kinetics prediction. These findings highlight ProMoBind's robust generalization capability and its potential for accurately predicting binding kinetics in an OOD setting, further extending its applicability in real-world drug discovery scenarios.

3 Conclusion

In this study, we have presented ProMoNet, a sequence-based pre-training and fine-tuning framework, addressing key challenges in protein-ligand binding process prediction. By combining the strengths of protein and molecular foundation models, this framework delivers significant advances in predicting binding sites, affinity, and kinetics using only sequence-based inputs.

ProMoSite, the pre-training module of ProMoNet, excels in identifying both exposed and cryptic binding sites. Its design leverages protein contact maps to model dynamic structural information, avoiding the need for resource-intensive 3D structural modeling or MD simulations. This capability enables ProMoSite to achieve comparable or superior performance to state-of-the-art methods while maintaining scalability and efficiency. Particularly in cryptic binding site prediction, ProMoSite demonstrates its ability to capture dynamic protein-ligand interactions, highlighting the utility of integrating sequence-based features with geometric constraints.

ProMoBind, the fine-tuning module of ProMoNet, builds on the pre-trained ProMoSite and is optimized for downstream tasks such as binding affinity and kinetics prediction. Its strong performance across diverse benchmarks, including out-of-distribution scenarios, underscores the robustness and adaptability of this approach. By incorporating ProMoSite's residue-level binding site predictions, ProMoBind enhances its ability to capture microscale protein-ligand interactions and binding site crypticity, providing an efficient solution for high-throughput compound screening in drug discovery pipelines.

While the framework achieves promising results, certain limitations warrant further exploration. The reliance on sequence-based inputs, though advantageous for scalability, may limit its ability to capture high-resolution structural details required for certain applications. Future work could investigate the integration of computationally predicted 3D protein structures to complement the current approach. Additionally, extending the framework to the protein-protein binding scenario could further broaden its impact.

This work also underscores the potential of integrating pre-training and fine-tuning strategies in drug discovery. By connecting protein and molecular foundation models, ProMoNet exemplifies how large-scale datasets can be leveraged to address complex biological tasks. As more high-quality data become available, this approach could be extended to other aspects of protein-ligand interactions, accelerating innovation in computational drug discovery.

In conclusion, ProMoNet provides a pre-training and fine-tuning framework that combines scalability, efficiency, and strong performance. This unified framework holds significant promise for bridging computational models and practical applications in drug development.

4 Methods

4.1 ProMoSite architecture

Featurization. For protein, ProMoSite takes protein sequences as input. Leveraging a pre-trained ESM-2 model (ESM-2-650M) [27] as the protein encoder Encoder_p, the representation of the i-th amino acid residue in each provided protein sequence $\mathcal{A}^p = \{a_i\}$ can be computed: $\boldsymbol{h}_i^p = \text{Encoder}_p(a_i) \in \mathbb{R}^{d_p}$, where $\boldsymbol{h}_i^p$ is the i-th residue representation, a_i is the amino acid type of the i-th residue and d_p is the hidden dimension size. Additionally, we make use of the unsupervised contact prediction results obtained from ESM-2 to predict residue-residue contact map of each protein $\boldsymbol{C}^p = \text{Encoder}_p(\mathcal{A}^p) \in \mathbb{R}^{n_p \times n_p}$, where n_p is the number of residues in $\mathcal{A}^p$.

For ligand, ProMoSite takes SMILES strings [29] as input and uses RDKit [40] to initialize their 3D conformers $\mathcal{G}^l$. Then we use the pre-trained Uni-Mol [28] as the ligand encoder Encoder_l to compute the representation of all atoms in each ligand conformer: $\boldsymbol{h}^l = \text{Encoder}_l(\mathcal{G}^l) \in \mathbb{R}^{n_l \times d_l}$, where d_l is the hidden dimension size and n_l is the number of atoms in $\mathcal{G}^l$. We use $\boldsymbol{h}_i^l$ to denote the representation of the i-th atom. Besides, the distance map $\boldsymbol{D}^l \in \mathbb{R}^{n_l \times n_l}$ of each $\mathcal{G}^l$ is computed based on the 3D coordinates, where $\boldsymbol{D}_{ij}^l$ denotes the distance between the i-th and j-th atom of $\mathcal{G}^l$ in 3D Euclidean space.

We further encode the contact probabilities in $\boldsymbol{C}^p$ and distances in $\boldsymbol{D}^l$ using the radial basis function (RBF) in [70] and followed with a multi-layer perception (MLP):

$$\boldsymbol{c}^p = \text{MLP}_c(\text{RBF}(\boldsymbol{C}^p)) \in \mathbb{R}^{n_p \times n_p \times d}, \tag{1}$$

$$\boldsymbol{d}^l = \text{MLP}_d(\text{RBF}(\boldsymbol{D}^l)) \in \mathbb{R}^{n_l \times n_l \times d}. \tag{2}$$

Based on the protein residue representations $\boldsymbol{h}^p$ and ligand atom representations $\boldsymbol{h}^l$, we compute the interaction embedding $\boldsymbol{z} \in \mathbb{R}^{n_p \times n_l \times d}$ for each protein-ligand pair, which models the interactions between protein residues and ligand atoms. For $\boldsymbol{z}_{ij}$ between the i-th protein residue and j-th ligand atom, we have:

$$\boldsymbol{z}_{ij} = \text{MLP}_{p_0}(\boldsymbol{h}_i^p) \odot \text{MLP}_{l_0}(\boldsymbol{h}_j^l) \in \mathbb{R}^d, \tag{3}$$

where $\odot$ is the element-wise multiplication.

Then $\boldsymbol{z}$, $\boldsymbol{c}^p$, and $\boldsymbol{d}^l$ are passed through layer normalization [71] for further updates.

Interaction module. Our interaction module (Fig. 1a) contains three steps of updates including *triangle update*, *triangle self-attention*, and *transition* inspired by [9, 30] to update $\boldsymbol{z}$ by satisfying the triangle inequality constraints in protein contact maps $\boldsymbol{C}^p$ and ligand distance maps $\boldsymbol{D}^l$.

In the *triangle update* step, we update the interaction embedding $\boldsymbol{z}_{ij}$ by combining information within each triangle of edges ij, ik, and jk, where i is the i-th protein residue, j is the j-th ligand atom, and k can be either protein residue or ligand atom. In each hidden layer of the interaction module, we have:

$$\boldsymbol{z}_{ij} = \boldsymbol{z}_{ij} + g(\boldsymbol{z}_{ij}) \odot \psi \left(\sum\nolimits_{k=1}^{n_p} \phi(\boldsymbol{c}_{ik}^p)\phi(\boldsymbol{z}_{kj}) + \sum\nolimits_{k=1}^{n_l} \phi(\boldsymbol{z}_{ik})\phi(\boldsymbol{d}_{kj}^l) \right), \tag{4}$$

where $g(\cdot) = \text{sigmoid}(\text{Linear}(\cdot))$, $\psi(\cdot) = \text{Linear}(\text{LayerNorm}(\cdot))$, and $\phi(\cdot) = g(\cdot) \odot \text{Linear}(\cdot)$. In Equation 4, the first summation term corresponds to the case when i, k are protein residues and j is a ligand atom, while the second summation term corresponds to the case when i is protein residue and j, k are ligand atoms.

In the *triangle self-attention* step, we focus on updating the interaction embedding $\boldsymbol{z}_{ij}$ between the i-th protein residue and the j-th ligand atom by considering the effect of all interaction embeddings between the i-th protein residue and each ligand atoms. By doing so, we can capture the essential physical and chemical principles governing protein-ligand interactions such as excluded-volume, saturation, electrostatic, and hydrophobic effects [9, 33]. These rules collectively determine the specificity, driving forces, thermodynamics, and stability of binding, providing a comprehensive framework for modeling microscale protein-ligand interactions. In each hidden layer, we update $\boldsymbol{z}_{ij}$ as follows:

$$\alpha_{ijk}^{h} = \text{softmax}_k \left(\psi_W^h(\boldsymbol{z}_{ij})^\top \psi_W^h(\boldsymbol{z}_{ik}) \right), \tag{5}$$

$$\boldsymbol{z}_{ij} = \boldsymbol{z}_{ij} + \text{Linear}\left(\text{concat}_h \big(g^h(\boldsymbol{z}_{ij}) \odot \sum\nolimits_{k=1}^{n_l} \alpha_{ijk}^h \psi_W^h(\boldsymbol{z}_{ik}) \big) \right), \tag{6}$$

where h is the number of attention heads, $\psi_W(\cdot) = \text{LinearNoBias}(\text{LayerNorm}(\cdot))$, concat denotes concatenation.

In the *transition* step, we further refine the interaction embedding $\boldsymbol{z}_{ij}$ by normalization and nonlinear transformation:

$$\boldsymbol{z}_{ij} = \text{MLP}_z \big(\text{LayerNorm}(\boldsymbol{z}_{ij}) \big). \tag{7}$$

Pooling module. After having the updated interaction embedding $\boldsymbol{z} \in \mathbb{R}^{n_p \times n_l \times d}$, we design a pooling module to compute the score $s \in \mathbb{R}$ for each protein residue, representing the likelihood that the residue belongs to a binding site for the given ligand. We first use a linear transformation to reduce the last dimension size of $\boldsymbol{z}$ to 1 and then apply sum pooling over the dimension related to ligand atoms to compute a score s_1 for each protein residue. Additionally, another residue-level score, s_2, is computed directly using the protein residue representations $\boldsymbol{h}^p$ and ligand atom representations $\boldsymbol{h}^l$. The final score s for each residue is obtained by combining s_1 and s_2. For the i-th residue, we have:

$$s_{i_1} = \sum\nolimits_{j=1}^{n_l} \text{Linear}(\boldsymbol{z}_{ij}), \tag{8}$$

$$s_{i_2} = \text{MLP}_s \left(\text{MLP}_{p_0}(\boldsymbol{h}_i^p) \odot \text{MLP}_{l_1} \big(\sum\nolimits_{j=1}^{n_l} (\boldsymbol{h}_j^l) \big) \right) + \lambda \cdot \text{MLP}_{p_1}(\boldsymbol{h}_i^p), \tag{9}$$

$$s_i = \sigma \left(\frac{s_{i_1} + s_{i_2}}{2} \right), \tag{10}$$

where MLP_{p_0} shares the same parameters as the MLP_{p_0} used in the interaction module, λ is a scalar value, and $\sigma(\cdot)$ denotes the sigmoid function.

In cases where no ligand is provided as input, such as when predicting cryptic pockets on the PocketMiner dataset [14], which includes some rigid proteins without bound ligands, the score is computed as $s_i = \text{MLP}_{p_1}(\boldsymbol{h}_i^p)$.

Clustering and ranking. Although the scores s from the pooling module can be directly used for residue-level evaluation (e.g. ROC-AUC and PR-AUC) of ProMoSite, it is also necessary to detect the pockets in 3D space and have pocket-level evaluation (e.g. DCA) to be compared with 3D structure-based methods. To detect 3D pockets, we first normalize all scores $\{s_i\}_{i=1}^{n_p}$ of each protein to the range of 0 to 1. Following this, we apply a threshold t to filter out residues that have smaller s. With the remaining residues, we utilize the ESM-2 predicted contact map $\boldsymbol{C}^p$ to define the similarities between each pair of residues and construct a dendrogram that contains the hierarchical clustering information of the residues using the single linkage algorithm [72]. By cutting at a specific linkage distance cutoff c, we define the clusters as pocket candidates and then rank them based on the squared sum of the associated residue scores in each pocket candidate. The top-ranked ones are the detected pockets for pocket-level evaluation. For any protein with multiple chains, we process each single chain to find its pocket candidates and then rank all pocket candidates that belong to the same protein. An illustration of these steps can be found in Supplementary Fig. 2.

4.2 ProMoBind architecture

Featurization. As shown in Fig. 1b, ProMoBind uses the same way as in ProMoSite to get protein residue representations $\boldsymbol{h}^p$ and ligand atom representations $\boldsymbol{h}^l$ from the pre-trained ESM-2 and the pre-trained Uni-Mol, correspondingly. Moreover, ProMoBind takes the predicted scores $\{s_i\}_{i=1}^{n_p}$ of each protein from the pre-trained ProMoSite as input.

Fusion. To predict the binding characteristic $y \in \mathbb{R}$ between each protein-ligand pair, we first compute the global representation of the protein and ligand, $\boldsymbol{h}_{global}^p$ and $\boldsymbol{h}_{global}^l$, respectively, and then fuse them in the following way:

$$\boldsymbol{h}_{global}^p = P\big(\{\text{MLP}_p(\boldsymbol{h}_i^p) \odot s_i\}_{i=1}^{n_p}\big) \in \mathbb{R}^{d'}, \tag{11}$$

$$\boldsymbol{h}_{global}^l = P\big(\{\text{MLP}_l(\boldsymbol{h}_j^l)\}_{j=1}^{n_l}\big) \in \mathbb{R}^{d'}, \tag{12}$$

$$y = \text{MLP}_y\big(\text{concat}(\boldsymbol{h}_{global}^p, \boldsymbol{h}_{global}^l)\big), \tag{13}$$

where $P(\{\boldsymbol{x}_i\}_{i=1}^n)$ denotes pooling over $\boldsymbol{x}_i$ for all i from 1 to n, which can be sum or mean pooling.

Specifically in Equation 11, when pooling all residue representations $\{\boldsymbol{h}_i^p\}_{i=1}^{n_p}$ in a protein, we utilize the ProMoSite predicted scores $\{s_i\}_{i=1}^{n_p}$ as a gating mechanism to modulate the contribution of each $\boldsymbol{h}^p$ in $\boldsymbol{h}_{global}^p$ through element-wise multiplication. This approach allows the s to selectively emphasize or suppress each $\boldsymbol{h}^p$ depending on its predicted likelihood to be in a binding site.

4.3 Dataset

General binding site prediction. Following [42, 45], we use the scPDB v.2017 database [34] and adopt the same data splits for training and cross-validation of ProMoSite. Specifically, the dataset is cleaned and split into 10 cross-validation sets based on their Uniprot IDs to avoid data leakage. We train 10 separate models, each using 9 folds for training and 1 fold for validation. The best set of hyperparameters is selected based on the highest averaged F1 score achieved across the 10 models. This optimal set of hyperparameters is then used to train a final model on the full dataset for evaluation on the test sets. For each protein-ligand pair in scPDB v.2017, the protein is split into single chains, forming pairs between each single chain and each ligand. Irrelevant pairs that are not present in BioLiP [35], a curated database of biologically relevant protein-ligand binding interactions, are removed.

For evaluation, we use COACH420 and HOLO4K [8]. Consistent with [8, 42], we evaluate ProMoSite on the Mlig subsets of these datasets, which contain the relevant ligands for binding sites. The final test sets consist of 291 proteins and 359 ligands for COACH420, and 3413 proteins and 4288 ligands for HOLO4K. To avoid data leakage, we preprocess the training set following the approach in [42] that removes all proteins from the training set that have either sequence identity greater than 50% or ligand similarity greater than 0.9 and sequence identity greater than 30% to any structure in the test set. Additionally, we exclude training pairs where protein sequences exceed 1280 amino acids. After preprocessing, the training set consists of 17101 pairs for COACH420 and 10626 pairs for HOLO4K.

Cryptic binding site prediction. To train ProMoSite, we use the scPDB v.2017 database and preprocess the data to single chain-ligand pairs filtered by BioLiP similar to our preprocessing for general binding site prediction. For validation and testing, we adopt the PocketMiner validation and test sets collected by [14] for cryptic binding site prediction. These sets include proteins with ligand-binding cryptic pockets or rigid structures, comprising 28 and 38 protein-ligand binding pairs or rigid proteins in the validation and test sets, respectively. The best model is selected based on the lowest validation loss. To avoid data leakage, we remove the proteins in the training set that have sequence identity greater than 30% to any of the proteins in the validation and test sets. Additionally, training pairs with protein sequences exceeding 1280 amino acids are removed, resulting in a training set of 19254 single chain-ligand pairs.

For comparison with baselines that predict residue-level probabilities, we use the full PocketMiner test set, which includes both protein-ligand pairs and rigid proteins. For baselines that predict 3D complexes or 3D binding sites, we use a subset of the PocketMiner test set filtered by NeuralPLexer, consisting solely of proteins with ligand-binding cryptic pockets. This subset comprises 19 protein-ligand pairs. Only *apo* protein structures or sequences are used as input.

Binding affinity prediction. In this task, we pre-train ProMoSite on the BioLiP-filtered scPDB v.2017 database following a similar process as described in previous tasks. The processed training set contains 20403 single chain-ligand pairs. During pre-training, we use the PocketMiner validation set to select the best ProMoSite model

based on the lowest validation loss, which is then used to generate scores as inputs for ProMoBind.

For the training and evaluation of ProMoBind, we adopt PDBBind v2020 dataset [67], a widely used benchmark for binding affinity prediction. We follow the same time-based split in [9, 21], using data published before 2019 for training and validation, and data published during or after 2019 for testing. For multimers, we concatenate single chain sequences into a single sequence, following the process described in [21]. Sequences exceeding 3,000 amino acids are excluded. The final datasets have 17717, 961, and 363 protein-ligand pairs in the training, validation, and test sets, respectively.

Binding kinetics prediction. We use the same pre-trained ProMoSite model from the binding affinity prediction task for binding kinetics prediction. For ProMoBind, k_{off} data is collected from PDBbind-koff-2020 [68] and BindingDB [69]. Inconsistent entries, such as cases where $k_{\text{off}}/k_{\text{on}}$ values are not consistent with K_{d}, are removed. After filtering, 632 items from BindingDB and 669 items from PDBbind-koff-2020 are retained. To eliminate redundancy between the two datasets, Tanimoto coefficients (TC) are calculated for the ligands. Redundant pairs, defined as complexes involving the same protein and chemically similar ligands (TC$\geq$0.99), are excluded. Additional filtering removes protein-ligand pairs with mutated protein sequences or multiple k_{off} values, resulting in 911 unique pairs. Proteins with sequences longer than 700 amino acids are excluded. After all preprocessing steps, a final dataset of 683 protein-ligand pairs is used for k_{off} prediction.

4.4 Experimental setting

Implementation Details. In ProMoSite, the hyperparameters of radial basis function (RBF) used in Equation 1 and 2 are identical to those in [70]. The multi-layer perceptrons (MLPs) that encode the results of the RBF function in these equations consist of two linear layers and utilize the self-gated Swish activation function [73]. All other MLPs in ProMoSite and ProMoBind employ ReLU as the activation function and include layer normalization preceding the linear layers. During training, dropout is applied to the triangle update and triangle self-attention steps in ProMoSite to enhance generalization. Further details regarding hyperparameter configurations are provided in Supplementary Table 6.

ProMoSite is trained by minimizing the binary cross-entropy loss between the predicted residue scores and the ground-truth labels. Residues labeled as "unknown" or "unclassified" are excluded from the loss calculation on the PocketMiner dataset. ProMoBind is trained to minimize the mean squared error (MSE) loss between the predicted values and the ground-truth binding characteristics. Both models are optimized using the Adam optimizer.

To enhance memory efficiency while maintaining accuracy, we use PyTorch's Automatic Mixed Precision (AMP) package for mixed-precision training in ProMoNet. All experiments are conducted on an NVIDIA L40S GPU (48 GB). UCSF ChimeraX [74] is used to generate the visualizations of protein-ligand complexes shown in Fig. 2 and 3.

Baselines. For general binding site prediction, we compare ProMoSite with baseline results reported in [42] for fair comparison. Additionally, we evaluate PUResNet, DeepProSite, and HoTS using their publicly available trained models. PUResNet is accessible at https://github.com/jivankandel/PUResNet, DeepProSite at https://github.com/WeiLab-Biology/DeepProSite, and HoTS at https://github.com/GIST-CSBL/HoTS.

For cryptic binding site prediction, we use the results of CryptoSite and PocketMiner reported in [14]. Other baseline models, including NeuralPLexer-AF2, NeuralPLexer-fasta, DynamicBind, and P2Rank, are evaluated using their publicly available trained models. NeuralPLexer variants are available at https://github.com/zrqiao/NeuralPLexer, DynamicBind at https://github.com/luwei0917/DynamicBind, and P2Rank (version 2.4.2) at https://github.com/rdk/p2rank. When measuring inference speed, input data is processed sequentially for each model to ensure fair comparison, and preprocessing time is excluded.

For binding affinity prediction, we adopt the baseline results reported in [21].

For binding kinetics prediction, we implement the random forest baseline using the RandomForestRegressor from the scikit-learn library. ECFP4 fingerprints are computed with RDKit using a 1024-bit setting.

References

[1] Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J.: Docking and scoring in virtual screening for drug discovery: methods and applications. Nature reviews Drug discovery **3**(11), 935–949 (2004)

[2] Śledź, P., Caflisch, A.: Protein structure-based drug design: from docking to molecular dynamics. Current opinion in structural biology **48**, 93–102 (2018)

[3] Perozzo, R., Folkers, G., Scapozza, L.: Thermodynamics of protein–ligand interactions: history, presence, and future aspects. Journal of Receptors and Signal Transduction **24**(1-2), 1–52 (2004)

[4] Copeland, R.A., Pompliano, D.L., Meek, T.D.: Drug–target residence time and its implications for lead optimization. Nature reviews Drug discovery **5**(9), 730–739 (2006)

[5] 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)

[6] Dhakal, A., McKay, C., Tanner, J.J., Cheng, J.: Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. Briefings in Bioinformatics **23**(1), 476 (2022)

[7] Cai, C., Wang, S., Xu, Y., Zhang, W., Tang, K., Ouyang, Q., Lai, L., Pei, J.: Transfer learning for drug discovery. Journal of Medicinal Chemistry **63**(16),

8683–8694 (2020)

[8] Krivák, R., Hoksza, D.: P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. Journal of cheminformatics **10**, 1–12 (2018)

[9] Lu, W., Wu, Q., Zhang, J., Rao, J., Li, C., Zheng, S.: Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. Advances in neural information processing systems **35**, 7236–7249 (2022)

[10] Scardino, V., Di Filippo, J.I., Cavasotto, C.N.: How good are alphafold models for docking-based virtual screening? Iscience **26**(1) (2023)

[11] Wang, D.D., Ou-Yang, L., Xie, H., Zhu, M., Yan, H.: Predicting the impacts of mutations on protein-ligand binding affinity based on molecular dynamics simulations and machine learning methods. Computational and structural biotechnology journal **18**, 439–454 (2020)

[12] Decherchi, S., Cavalli, A.: Thermodynamics and kinetics of drug-target binding by molecular simulation. Chemical Reviews **120**(23), 12788–12833 (2020)

[13] Kuzmanic, A., Bowman, G.R., Juarez-Jimenez, J., Michel, J., Gervasio, F.L.: Investigating cryptic binding sites by molecular dynamics simulations. Accounts of chemical research **53**(3), 654–661 (2020)

[14] Meller, A., Ward, M., Borowsky, J., Kshirsagar, M., Lotthammer, J.M., Oviedo, F., Ferres, J.L., Bowman, G.R.: Predicting locations of cryptic pockets from single protein structures using the pocketminer graph neural network. Nature Communications **14**(1), 1177 (2023)

[15] Wang, J., Do, H.N., Koirala, K., Miao, Y.: Predicting biomolecular binding kinetics: A review. Journal of Chemical Theory and Computation **19**(8), 2135–2148 (2023)

[16] Qiao, Z., Nie, W., Vahdat, A., al.: State-specific protein–ligand complex structure prediction with a multiscale deep generative model. Nature Machine Intelligence, 1–14 (2024)

[17] Lu, W., Zhang, J., Huang, W., al.: Dynamicbind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. Nature Communications **15**(1), 1071 (2024)

[18] Öztürk, H., Özgür, A., Ozkirimli, E.: Deepdta: deep drug–target binding affinity prediction. Bioinformatics **34**(17), 821–829 (2018)

[19] Nguyen, T., Le, H., Quinn, T.P., Nguyen, T., Le, T.D., Venkatesh, S.: Graphdta:

predicting drug–target binding affinity with graph neural networks. Bioinformatics **37**(8), 1140–1147 (2021)

[20] Li, S., Zhou, J., Xu, T., Huang, L., Wang, F., Xiong, H., Huang, W., Dou, D., Xiong, H.: Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 975–985 (2021)

[21] Koh, H.Y., Nguyen, A.T., Pan, S., May, L.T., Webb, G.I.: Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. Nature Machine Intelligence, 1–15 (2024)

[22] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)

[23] Zhang, Q., Ding, K., Lyv, T., Wang, X., Yin, Q., Zhang, Y., Yu, J., Wang, Y., Li, X., Xiang, Z., et al.: Scientific large language models: A survey on biological & chemical domains. arXiv preprint arXiv:2401.14656 (2024)

[24] Li, Q., Hu, Z., Wang, Y., Li, L., Fan, Y., King, I., Song, L., Li, Y.: Progress and opportunities of foundation models in bioinformatics. arXiv preprint arXiv:2402.04286 (2024)

[25] Wang, X., Chen, G., Qian, G., Gao, P., Wei, X.-Y., Wang, Y., Tian, Y., Gao, W.: Large-scale multi-modal pre-trained models: A comprehensive survey. Machine Intelligence Research **20**(4), 447–482 (2023)

[26] Sledzieski, S., Kshirsagar, M., Baek, M., Dodhia, R., Lavista Ferres, J., Berger, B.: Democratizing protein language models with parameter-efficient fine-tuning. Proceedings of the National Academy of Sciences **121**(26), 2405840121 (2024)

[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., al.: Uni-mol: A universal 3d molecular representation learning framework. In: The Eleventh International Conference on Learning Representations (2022)

[29] Weininger, D.: Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of chemical information and computer sciences **28**(1), 31–36 (1988)

[30] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.*: Highly accurate protein structure prediction with alphafold. Nature **596**(7873), 583–589 (2021)

[31] Vassura, M., Margara, L., Di Lena, P., Medri, F., Fariselli, P., Casadio, R.: Reconstruction of 3d structures from protein contact maps. IEEE/ACM Transactions on Computational Biology and Bioinformatics **5**(3), 357–367 (2008)

[32] Li, J., Wang, L., Zhu, Z., Song, C.: Exploring the alternative conformation of a known protein structure based on contact map prediction. Journal of Chemical Information and Modeling **64**(1), 301–315 (2023)

[33] Li, J., Fu, A., Zhang, L.: An overview of scoring functions used for protein–ligand interactions in molecular docking. Interdisciplinary Sciences: Computational Life Sciences **11**, 320–328 (2019)

[34] Desaphy, J., Bret, G., Rognan, D., Kellenberger, E.: sc-pdb: a 3d-database of ligandable binding sites—10 years on. Nucleic acids research **43**(D1), 399–404 (2015)

[35] Yang, J., Roy, A., Zhang, Y.: Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic acids research **41**(D1), 1096–1103 (2012)

[36] Vajda, S., Beglov, D., Wakefield, A.E., Egbert, M., Whitty, A.: Cryptic binding sites on proteins: definition, detection, and druggability. Current opinion in chemical biology **44**, 1–8 (2018)

[37] Copeland, R.A.: Conformational adaptation in drug–target interactions and residence time. Future medicinal chemistry **3**(12), 1491–1501 (2011)

[38] Pan, A.C., Borhani, D.W., Dror, R.O., Shaw, D.E.: Molecular determinants of drug–receptor binding kinetics. Drug discovery today **18**(13-14), 667–673 (2013)

[39] Amangeldiuly, N., Karlov, D., Fedorov, M.V.: Baseline model for predicting protein–ligand unbinding kinetics through machine learning. Journal of Chemical Information and Modeling **60**(12), 5946–5956 (2020)

[40] Landrum, G.: RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. Academic Press (2013)

[41] Su, D., Xu, Y., Winata, G.I., Xu, P., Kim, H., Liu, Z., Fung, P.: Generalizing question answering system with pre-trained language model fine-tuning. In: Proceedings of the 2nd Workshop on Machine Reading for Question Answering, pp. 203–211 (2019)

[42] Aggarwal, R., Gupta, A., Chelur, V., Jawahar, C., Priyakumar, U.D.: Deeppocket: ligand binding site detection and segmentation using 3d convolutional neural networks. Journal of Chemical Information and Modeling **62**(21), 5069–5079 (2021)

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

[44] Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A.S., De Fabritiis, G.: Deepsite: protein-binding site predictor using 3d-convolutional neural networks. Bioinformatics **33**(19), 3036–3042 (2017)

[45] Stepniewska-Dziubinska, M.M., Zielenkiewicz, P., Siedlecki, P.: Improving detection of protein-ligand binding sites with 3d segmentation. Scientific reports **10**(1), 5035 (2020)

[46] Kandel, J., Tayara, H., Chong, K.T.: Puresnet: prediction of protein-ligand binding sites using deep residual neural network. Journal of cheminformatics **13**(1), 1–14 (2021)

[47] Lee, I., Nam, H.: Sequence-based prediction of protein binding regions and drug–target interactions. Journal of cheminformatics **14**(1), 5 (2022)

[48] Fang, Y., Jiang, Y., Wei, L., Ma, Q., Ren, Z., Yuan, Q., Wei, D.-Q.: Deepprosite: structure-aware protein binding site prediction using esmfold and pretrained language model. Bioinformatics **39**(12), 718 (2023)

[49] Redmon, J.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)

[50] Rogers, D., Hahn, M.: Extended-connectivity fingerprints. Journal of chemical information and modeling **50**(5), 742–754 (2010)

[51] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

[52] He, X., Zhao, L., Tian, Y., Li, R., Chu, Q., Gu, Z., Zheng, M., Wang, Y., Li, S., Jiang, H., *et al.*: Highly accurate carbohydrate-binding site prediction with deepglycansite. Nature Communications **15**(1), 5163 (2024)

[53] Cimermancic, P., Weinkam, P., Rettenmaier, T.J., al: Cryptosite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. Journal of molecular biology **428**(4), 709–719 (2016)

[54] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)

[55] Stepniewska-Dziubinska, M.M., Zielenkiewicz, P., Siedlecki, P.: Development and evaluation of a deep learning model for protein–ligand binding affinity prediction.

Bioinformatics **34**(21), 3666–3674 (2018)

[56] 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**(14), 15956–15965 (2019)

[57] Jiang, D., Hsieh, C.-Y., Wu, Z., Kang, Y., Wang, J., Wang, E., Liao, B., Shen, C., Xu, L., Wu, J., *et al.*: Interactiongraphnet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. Journal of medicinal chemistry **64**(24), 18209–18232 (2021)

[58] Koes, D.R., Baumgartner, M.P., Camacho, C.J.: Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. Journal of chemical information and modeling **53**(8), 1893–1904 (2013)

[59] McNutt, A.T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., Sunseri, J., Koes, D.R.: Gnina 1.0: molecular docking with deep learning. Journal of cheminformatics **13**(1), 43 (2021)

[60] Sverrisson, F., Feydy, J., Correia, B.E., Bronstein, M.M.: Fast end-to-end learning on protein surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15272–15281 (2021)

[61] Chen, L., Tan, X., Wang, D., Zhong, F., Liu, X., Yang, T., Luo, X., Chen, K., Jiang, H., Zheng, M.: Transformercpi: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics **36**(16), 4406–4414 (2020)

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

[63] Bai, P., Miljković, F., John, B., Lu, H.: Interpretable bilinear attention network with domain adaptation improves drug–target prediction. Nature Machine Intelligence **5**(2), 126–136 (2023)

[64] Jiang, M., Li, Z., Zhang, S., Wang, S., Wang, X., Yuan, Q., Wei, Z.: Drug–target affinity prediction using graph neural network and contact maps. RSC advances **10**(35), 20701–20712 (2020)

[65] Jiang, M., Wang, S., Zhang, S., Zhou, W., Zhang, Y., Li, Z.: Sequence-based drug-target affinity prediction using weighted graph neural networks. BMC genomics **23**(1), 449 (2022)

[66] Wang, P., Zheng, S., Jiang, Y., Li, C., Liu, J., Wen, C., Patronov, A., Qian, D., Chen, H., Yang, Y.: Structure-aware multimodal deep learning for drug–protein

interaction prediction. Journal of chemical information and modeling **62**(5), 1308–1317 (2022)

[67] Wang, R., Fang, X., Lu, Y., Yang, C.-Y., Wang, S.: The pdbbind database: methodologies and updates. Journal of medicinal chemistry **48**(12), 4111–4119 (2005)

[68] Liu, H., Su, M., Lin, H.-X., Wang, R., Li, Y.: Public data set of protein–ligand dissociation kinetic constants for quantitative structure–kinetics relationship studies. ACS omega **7**(22), 18985–18996 (2022)

[69] Liu, T., Hwang, L., Burley, S.K., Nitsche, C.I., Southan, C., Walters, W.P., Gilson, M.K.: Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. Nucleic Acids Research, 1075 (2024)

[70] Gasteiger, J., Groß, J., Günnemann, S.: Directional message passing for molecular graphs. In: International Conference on Learning Representations (ICLR) (2020)

[71] Lei Ba, J., Kiros, J.R., Hinton, G.E.: Layer normalization. ArXiv e-prints, 1607 (2016)

[72] Müllner, D.: Modern hierarchical, agglomerative clustering algorithms. arXiv preprint arXiv:1109.2378 (2011)

[73] Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. arXiv preprint arXiv:1710.05941 (2017)

[74] Meng, E.C., Goddard, T.D., Pettersen, E.F., Couch, G.S., Pearson, Z.J., Morris, J.H., Ferrin, T.E.: Ucsf chimerax: Tools for structure building and analysis. Protein Science **32**(11), 4792 (2023)