Protein-Ligand Binding Site Prediction and de Novo Ligand Generation from Cryo-EM Maps

Chunyang Lu[1], Kaustav Mitra[1], Kiran Mitra[1], Hanze Meng[2], Shane Thomas Rich-New[3], Fengbin Wang[3], Dong Si[1*]

[1]Division of Computing and Software Systems, University of Washington, 18115 Campus Way NE, Bothell, 98011, WA, United States.
[2]Department of Computer Science, Duke University, 308 Research Drive, Durham, 27708, NC, United States.
[3]Department of Biochemistry and Molecular Genetics, University of Alabama at Birmingham, 1720 2nd Ave S, Birmingham, 35233, AL, United States.

*Corresponding author(s). E-mail(s): dongsi@uw.edu;

Abstract

Identification of protein-ligand binding sites is one of the most challenging tasks in drug discovery and design. Recent advances in machine learning community, particularly deep learning, inspired considerable research into deep learning-based methods for protein ligand binding site prediction (PLBP) and have achieved promising results. However one limitation of these methods and models is that they are developed and evaluated using conventional databases consisting of relatively small protein structures determined from X-ray crystallography and NMR with high resolution. When current PLBP methods are directly applied to large protein complexes, they either fail or produce very poor results. Given the increasing popularity of protein structures determined from Cryo-EM maps, which usually capture large protein complexes at relatively low resolution, there is a strong need to apply PLBP methods to Cryo-EM maps. For this purpose, we created a novel database (EMD-Ligand Dataset) consisting of Cryo-EM maps and information about their ligand-binding partners. Our study of several state-of-the-art PLBP methods show that even though they perform reasonably well on conventional databases, they do poorly on the novel EMD-Ligand Dataset, which calls for more research and work to be done by the community. As current methods for determining protein structures experimentally from Cryo-EM maps are time-consuming and expensive, we integrated PLBP methods with DeepTracer, an automatic and fast, de novo Cryo-EM protein structure modeling method,

and established an end-to-end system capable of predicting ligand binding sites directly from Cryo-EM maps. Additionally, we leveraged an existing ligand generator to generate drug-like ligands from our predicted ligand-binding pockets and demonstrated its effectiveness.

Keywords: Protein-Ligand Binding Site Prediction, Machine Learning, Bioinformatics, Drug-Like Molecule Generation, Cryo-EM, Drug Design

1 Background

1.1 Motivation

Proteins perform crucial biological functions through their interactions with binding partners (ligands). Interactions that are critical for various pathologies have frequently been the subject of study in drug design, however, designing drugs that bind specifically to a protein's unique three-dimensional structure is often challenging and can be costly on several fronts including time, chemical resources, and financially. To mitigate these factors, computational prediction of protein-ligand interactions has recently been integrated into the modern drug discovery field. Most recently, due to an abundance of experimental structural data, machine learning and artificial intelligence have been applied to investigate protein-ligand interactions. Prediction of protein-ligand binding sites is essential for structure-based drug design, which may assist in predicting potential therapeutic side effects. Protein-ligand bindings are typically highly specific intermolecular interactions that occur at binding pockets on proteins.

There is a limitation in current popular deep-learning based protein-ligand binding site prediction (PLBP) methods such as Kalasanty [1], PUResNet [2] and DeepSurf [3]. These methods are developed and evaluated on public datasets like ScPDB [4] and PDBbind [5], which are mainly comprised of small protein complexes determined by X-ray crystallography and Nuclear Magnetic Resonance (NMR) with high resolution, typically <3Å. When directly applied to large protein complexes, these PLBP methods either fail or produce very poor results due to their limited experience with larger protein sizes and greater numbers of atoms. Given the increasing popularity of protein structures determined by Cryo-EM maps, which are usually images of large protein complexes at relatively low resolution, there is a strong need to study PLBP methods on proteins from Cryo-EM maps. To our knowledge, there is no published database explicitly designed for predicting ligand binding sites on proteins with Cryo-EM maps, which motivated us to create a novel database for this research: the EMD-Ligand Dataset. The Cryo-EM maps used in this database were retrieved from the Protein Data Bank (PDB) linked to EMDB, which is the database that has the Cryo-EM maps. This database enabled us to test the possibility of extending the functionality of DeepTracer [6], allowing for the prediction of protein-ligand binding sites directly from Cryo-EM maps without experimentally-determined protein 3D structures, as well as ligand generation using protein structures determined by DeepTracer.

1.2 Previous Protein-Ligand Interaction Prediction Methods

There has been considerable research in applying deep learning-based methods to predict ligand binding sites. For example, the SURFNET [6] architecture successfully predicted the ligand-binding site as the largest pocket in over 80% of all tested cases. Additionally, other methods such as MaSIF [7], DeepCSeqSite [8], DELIA [9], Kalasanty [1], DeepSurf [3], and PUResNet [2] have been developed to predict binding sites. MaSIF exploits geometric deep learning to learn interaction fingerprints in protein molecular surfaces. DeepCSeqSite is a sequence-based method using a deep convolutional neural network (dCNN) to predict protein-ligand binding residues. DELIA is a hybrid deep neural network integrating a CNN with a bidirectional long short-term memory network (BiLSTM) to mobilize 1D sequence feature vectors and 2D distance metrics. It improves the data imbalance between binding and nonbinding residues using the over-sampling in minibatch, random under-sampling and stacking ensemble strategies. Kalasanty is a 3D convolutional neural network based on U-Net architecture [10]. DeepSurf is based on deep learning architecture which mobilizes surface-based representation (implementation of 3D voxelized grids) along with state-of-the-art deep learning architectures to predict potential druggable sites on proteins. PUResNet implemented a deep ResNet as the backbone of the network, which is given as input a 3D protein structure and outputs the ligandability of each voxel. It also used a unique training data cleaning process.

Current PLBP methods rely on publicly available datasets like ScPDB [4] and PDBbind [5] to train and evaluate their models. However, these datasets mainly comprise smaller protein complexes determined by X-ray crystallography and NMR with high resolution. The "resolution revolution" of Cryo-EM that began in 2014 has enabled the imaging of an increasing number of structures using the technique, resulting in an increasing proportion of protein structures in the PDB Database[11] derived from Cryo-EM maps. This increase and the lack of PLBP methods utilizing Cryo-EM maps represents an unmet need in the study of protein-ligand interactions, drug discovery, and drug design. Further, Cryo-EM is an indispensable tool for studying the structures and functions of macromolecular assemblies and provides another basis for the design of novel therapeutic drugs.

In addition to the protein and ligand binding site prediction methods discussed earlier, there are several other related methods in the field. For instance, SSnet utilizes a 1D representation based on the curvature and torsion of the protein backbone. This method represents the protein using the Cα carbons of the backbone, which clearly defines a recognizable structure for peptides and proteins alike and is very useful for categorization [12]. Ligands are represented using SMILES strings, which represent various chemical bonds and orientations present in the molecule. Another approach, utilizing in silico chemogenomic informatics (which is based on mining the chemical space of the molecules in question) predicts the interactions between small molecules and proteins, which is crucial for understanding many biological processes and plays a critical role in drug discovery since this approach can identify and create a selectivity profile of similar proteins (homologues) in order to prevent side effects [13]. Docking

approaches and quantitative prediction of binding affinity using regression methods in the joint space are also commonly used when the target 3D structure is available. SIGN [14] is a structure-aware interactive graph neural network that comprises polar-inspired graph attention layers (PGAL) and pairwise interactive pooling (PiPool) and is used for prediction of binding affinities as well as per-atom structural and interaction information. DeepAffinity [15] is a deep learning technique based on a recurrent neural network that represents a protein with amino acid sequences and a molecule with SMILES. This method embeds 3D structural information of a protein-ligand complex in an adjacency matrix, devises a distance-aware attention algorithm to differentiate various types of intermolecular interactions, and introduces a variant of graph neural networks suitable for learning protein-ligand interactions. DeepDock [16] is a geometric deep learning-based method that extracts features from the input data to identify key ligand-target interactions. It involves two residual graph convolutional neural networks. GemSpot [17] is useful for the robust identification of ligand poses and drug discovery efforts through Cryo-EM. This tool aids in the correct interpretation and modeling of ligand densities and will greatly help drug discovery efforts based on Cryo-EM.

1.3 Introduction to Our Pipeline

To support our research, we introduce the EMD-Ligand Dataset, the first protein-ligand binding database created from EMDataResource and Protein Data Bank (PDB). We evaluate current PLBP methods/models (with modifications so that they can apply to large protein complexes) on this new database, as well as testing these methods against the conventionally used, publicly-accessible protein-ligand binding databases which will be introduced in Section 2.1.1. Furthermore, we demonstrate how integrating PLBP methods with DeepTracer [18], a fully-automated deep learning-based method for de novo multichain protein complex structure modeling based on Cryo-EM maps, can create a fast and automated pipeline to directly predict ligand binding sites from Cryo-EM maps. Additionally, we show how extending this pipeline with Pocket2Mol [19], a deep generative model for efficient molecular sampling based on 3D protein pockets, can generate high-quality drug-like molecules without any prior design input.

In Section 2, we begin with a comprehensive comparative study of three popular PLBP methods (PUResNet, Kalasanty, DeepSurf) across multiple public datasets. Through our independent study, we aim to fairly assess these methods and use the best methods in our PLBP study on Cryo-EM maps. We found that there is no single method that consistently outperforms the others, so we evaluated all three methods on the EMD-Ligand Dataset. Next, we present the DeepTracer Protein Ligand Binding Prediction pipeline (DeepTracer-PLBP) and demonstrate its effectiveness. We also show the extended DeepTracer Drug-Like Molecules Generator pipeline (DeepTracer-DLMGen). In Section 3, we present the detailed evaluation results of these PLBP methods, as well as the evaluation of DeepTracer-DLMGen generated drug-like molecules based on the metrics that are widely used to evaluate the quality of generated drug candidates. Finally, in Section 4, we conclude with a discussion and

our conclusions covering the case-proven successes and failures of utilizing DeepTracer in a pipeline to predict ligand binding sites and generate de novo drug-like ligands from Cryo-EM maps. The appendix contains the sample list of EMD maps whose ligand binding sites are correctly predicted by our DeepTracer-PLBP pipeline.

We showcase the designed high-level workflow of our method in Figure 1, which integrates the DeepTracer pipeline to predict ligand binding sites and generate drug-like ligands directly from Cryo-EM maps. Our multi-step approach starts with an input Cryo-EM map and utilizes the DeepTracer pipeline to generate the protein's backbone structure. This is then followed by the prediction of ligand binding sites using one of the PLBP methods. Finally, both the predicted protein structure and the predicted binding site are fed to Pocket2Mol to generate drug-like ligands.

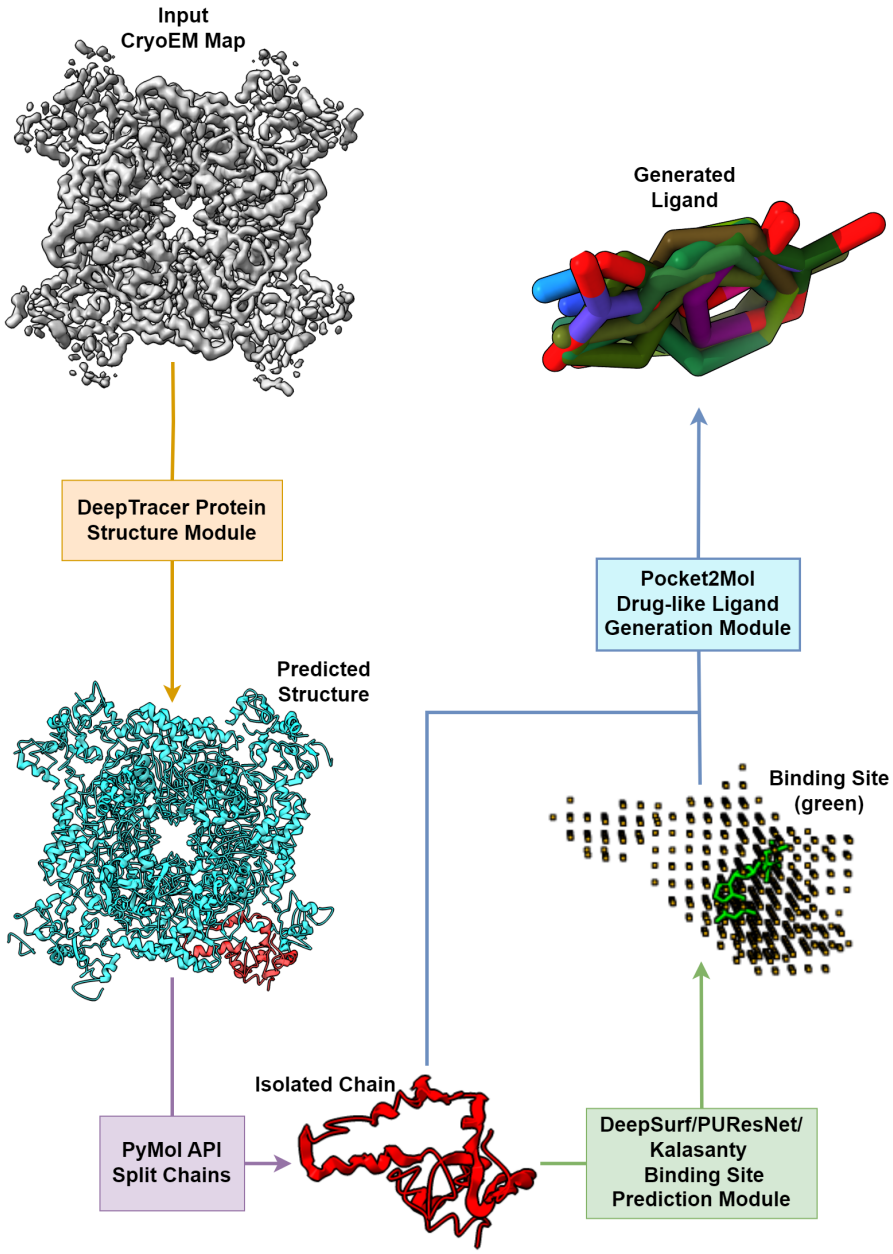

Fig. 1 DeepTracer Ligand Binding Site Prediction and the Drug-like Molecules Generator pipeline. (created in ChimeraX).

2 Our Pipeline

2.1 Comparative Studies of State-Of-The-Art PLBP Methods

To begin, we introduce the popular datasets used for protein-ligand binding studies, followed by a discussion of our evaluation metrics. We introduce the three state-of-the-art methods that we are evaluating, namely Kalasanty, PUResNet, and DeepSurf. After that, we introduce our newly created EMD-Ligand Dataset which provides the feasibility to do the benchmark experiment. Further, integrating with DeepTracer will be introduced. Last, it is the case proven introduction of the integration with Pocket2Mol drug-like ligand generator.

2.1.1 Existing Datasets for Protein-Ligand Binding Study

Examining the datasets used in training and evaluating PLBP algorithms is very helpful for this study. The PDB provides the 3D structures of many proteins and their ligands, and it is a widely used resource in structural biology. The BioLip [20], ScPDB [4], and BindingMOAD [21] databases provide additional curated data for studying ligand-protein interactions. The coach420 [22] and BU48 [23] datasets are independent test sets used to evaluate the performance of machine learning models in predicting protein-ligand binding sites. The core set within PDBBind [5] contains a diverse set of high-quality protein-ligand complexes, which serve as an essential benchmark dataset for different scoring functions.

The PDB database stores the primary information of protein structures, including coordinate files for all deposited biological molecules. These files list the atoms in each protein and their 3D location space. The basic syntax and format comprise category and attribute names. These entities describe the chemistry and identity of the molecules of a chemically distinct part of a structure.

PDBBind, the latest version updated in 2020 and a new 2021 version on the way, contains information for 23,496 protein-ligand binding interactions. A series of computer programs were implemented to screen the entire PDB database to identify four major types of molecular interactions: protein-small ligand, nucleic acid-small ligand, protein-nucleic acid, and protein-protein interactions [5]. It contains a special subset [24], called the "core set", with high-quality protein-ligand complexes with remarkable diversity in structure as well as binding data. This subset is used in the popular Comparative Assessment of Scoring Function (CASF) evaluation. It is composed of 285 protein-ligand complexes in 57 clusters [24].

BioLip [20], updated weekly, is a curated database of ligand-protein binding interactions that have biological relevance. After the automated process for determining a ligand's biological relevance is completed, a thorough manual review is performed to correct any errors. The structural data is primarily obtained from the PDB while biological insights are obtained from literature and other databases. Eventually, the manual check is performed and possible false-positive entries are verified by reading the original literature and consulting other databases that ensure the completeness

and high quality of BioLip. The latest version (2023) has a total of 833,327 entries.

The ScPDB [4] database is a comprehensive and up-to-date selection of druggable binding sites of the Protein Data Bank. Druggable sites are defined as complexes between a protein and a pharmacological ligand. The database provides the all-atom description of the protein, the associated ligand, the binding site, and the binding mode of the protein-ligand interaction. It is continuously updated. The ScPDB release 2017 version is the most current release and it archives 16,034 entries, 4,782 proteins and 6,326 ligands [4]. Its features include identifying binding sites, correcting structures, annotating protein function, noting ligand properties, and characterizing their binding mode. Each entry also contains a transformed 3D MOL2 file of the structure of each protein-ligand complex [4].

BindingMOAD [21] contains very high-quality ligand-protein binding, a collection of high-quality holo-crystal structures resolved higher than 2.5Å or better. Currently MOAD release 2020 has 41,409 protein-ligand complexes and 15,223 entries of binding data. Coach420 [25] is an independent test dataset which consists of 420 protein structure with known ligands. We will use the dataset provided by PUResNet which utlizes the Coach420 dataset but removed 122 protein structures since they were present in PUResNet's training dataset. Therefore, 298 protein structures with ligands were selected. BU48 [25] is another test set consisting of 48 proteins with known ligands.

2.1.2 Evaluation Metrics

There are two standard metrics utilized for the evaluation: D_{CC} [26] and DVO [1],[26]. D_{CC} is the measured distance between the predicted and actual center of the binding pocket. It is typically used to describe a success rate for a method, i.e., the fraction of sites below the given D_{CC} threshold. DVO (discretized volume overlap) is a metric that compares the shapes of the predicted and actual pockets. It is the volume of the intersection of the predicted and the actual segmentations, divided by volume of their union. These two metrics complement one another, highlighting different aspects of prediction quality – correct location (D_{CC}) and shape (DVO). An O_{PL} matrix [26] indicates the intersection, at the atomic level, between the real and predicted binding sites divided by their union. It differs with the above distance-based metrics by also considering the shape of the binding sites, since it expresses a normalized spatial overlap between the predicted and the actual location of the binding pocket. F_1-score is a standard metric for detection tasks. It combines precision (positive predictive value) and recall (also called sensitivity) using the harmonic mean. Success rate (SR) is defined as the total number of successful predictions for all proteins divided by the corresponding total number of existing sites.

2.1.3 PLBP Methods for Evaluation

All three methods use ScPDB for model training. K-fold cross-validation is used to tune hyperparameters and the final model was trained using the full dataset for best performance [27]. Kalasanty presents a deep learning-based approach for finding binding pockets, which was inspired by semantic image segmentation instead of

classification. It used a 3D U-Net model structure [28]. The input to the model is the 3D coordinates of a protein, which is internally represented as a grid. Output is the corresponding ligandability scores of each vortex of the grid.

PUResNet improves Kalasanty by using a novel training data cleaning process based on structural similarity and a new model structure [2]. The data cleaning process groups protein structures according to the UniProtID and the Tanimoto index calculated for the high similarity of the protein structure in the cluster, followed by manual inspection using PyMol. Finally, 5,020 protein structures, out of 16,034, were selected from ScPDB dataset for training. Its model structure is derived from the concept of U-Net and ResNet by combining the convolutional segmentation of biomedical images with a skip connection between the blocks.

DeepSurf is a novel computational method for the prediction of potential binding sites [3]. It adopts ResNet architecture as its network backbone, which effectively combines the learning capabilities of advanced CNN architecture with a surface-based representation of the protein structure [29]. Two types of model architecture were investigated. One is the standard 3D-ResNet, and the other is a novel bottleneck 3D-LDS-ResNet, which shows better accuracy with a much smaller model size. We evaluated three methods, namely PUResNet, Kalasanty, and DeepSurf, on the Coach420 and BU48 sets as well as the PDBbind Core-Set (CASF-2016). The evaluation results are shown in Section 3.

2.2 Creation of Our EMD-Ligand Dataset

Cryo-EM technology has gained significant popularity since its introduction and an additional explosion in popularity following the advent of the "Resolution Revolution" that began in 2014, leading to an increase in experimentally-determined proteins in PDB banks using Cryo-EM. It produces 3D electron density maps of macromolecules and can explain the shape and interactions of protein complexes which forms the basis for modeling various pathologies, as well as assisting therapeutic drug design. Therefore, studying protein-ligand interactions based on Cryo-EM maps has become crucial. While ScPDB is the most popular database for protein ligand binding site prediction research, it mostly contains proteins with a few thousand atoms or fewer and higher resolution structures determined by x-ray crystallography (<2.5Å) and NMR. In contrast, protein structures found within the Cryo-EM database are typically large complexes with many chains, contain significantly more atoms, on the range of tens of thousands of atoms, and lower resolution (>3.0Å). Thus, available protein-ligand binding site prediction methods have not been utilized to test or report results on proteins with Cryo-EM maps.

To identify protein structures with both Cryo-EM maps and ligand-binding information, we used a database cross-query. Firstly, we queried the comprehensive data source for Cryo-EM maps in EMDataResource and found 24,505 Cryo-EM maps corresponding to 13,944 PDB coordinate entries in the PDB database. We downloaded the full PDB ID list (13,857) with detailed information, including 3D structures for

proteins and ligands. We searched each of these 13,857 PDB IDs of the Cryo-EM maps dataset in all publicly available protein-ligand binding datasets and found no matches in ScPDB, no matches in MOAD dataset, 19 matches in PDBbind, and 2,878 matches in the BioLip dataset.

Since many Cryo-EM proteins are large protein complexes with numerous chains, the BioLip dataset was used to provide information on each protein and its respective ligand(s) (which were split according to chain). Chains with found ligands were saved as individual pdb files, along with their corresponding ligands. In total, the 2,878 matched proteins corresponded to 41,542 protein chain files and 46,538 corresponding ligands. To focus on our study's objective, we removed all RNA, DNA, peptide, and ion (Zn, Mg, etc.) ligands, leaving 24,978 remaining ligands, corresponding to 1,621 proteins and 10,463 protein chains. Table 1 gives a breakdown of these statistics ranked by Cryo-EM map resolution, with 66% of Cryo-EM maps having resolutions lower than 3.0Å. This dataset forms the basis for our subsequent studies.

Table 1 Cryo-EM maps groups by resolution

Resolution(Å)	Number of Proteins	Number of Protein Chains	Number of Ligands
All	1621	10463	24978
0.5–3.0	400	2689	8318
3.0–3.5	562	3529	8496
3.5–4.0	358	2807	5793
4.0–4.5	149	613	1207

2.3 Benchmark on EMD-Ligand Dataset Using the Ground-Truth Protein Structure

DeepSurf is a surface-based method that can directly predict binding sites on relatively large protein complexes with multiple chains but it has a hard-coded limit on the size of the proteins in terms of the number of surface points. On the other hand, Kalasanty and PUResNet assume that the protein structure is within a 70Å x 70Å x 70Å box, with all residues outside the box being ignored. These constraints make these methods either fail or perform very poorly on larger proteins with many chains, so we split the proteins according to their respective chains and apply the PLBP method to each individual chain. We evaluated Kalasanty, PUResNet, and DeepSurf methods on our new EMD-Ligand Dataset. Unlike ScPDB and other datasets, which have only one ligand per protein, our dataset has multiple ligands per protein chain, with an average of nearly 2.5 ligands per protein chain. This means that the ligand-generation methods would generate multiple hypotheses as well, so we needed to modify our scoring algorithm to account for that.

We use the ground-truth protein chains (3D structures experimentally-determined from Cryo-EM maps) as input for all methods to predict ligand binding sites. Their

evaluation results are given in Section 3. As an illustration, Figure 2A shows the structure of EMD-5127 chain G (corresponding PDB ID: 3iyd) with a reference binding ligand (CMP) shows the DeepSurf-predicted ligand binding site (the pink dotted line on the right) with respect to potential interacting protein residues. Figure 2B shows the DeepTracer-predicted protein structure EMD-21391 chain M (corresponding PDB ID: 6vv5) with the reference ligand PAM showing the PUResNet predicted ligand binding site. As we can see, the binding site surrounds the ligand, and the center of the predicted pocket is close (<4.0Å) to the center of the ligand.

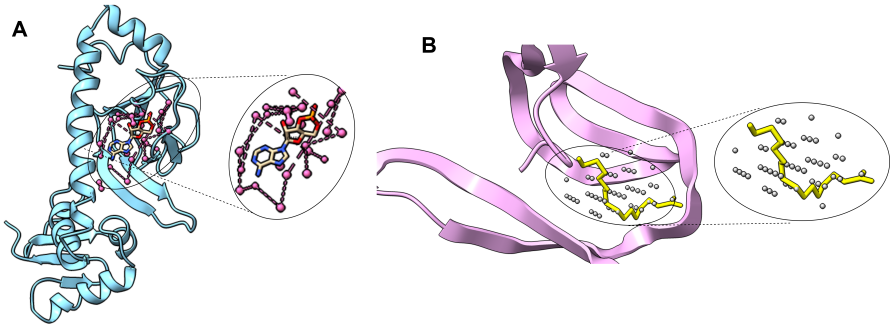

Fig. 2 A) The protein-ligand binding site prediction using ground truth protein structure. (PDB ID: 3iyd), chain G (EMD-5127) with its ligand CMP. The purple dotted line represents the predicted binding site by DeepSurf; CMP is depicted inside in the ground truth state. This is a correct prediction as the D_{CC} is less than 4.0Å and shows that the predicted site overlaps with the ligand (created in ChimeraX). B) The protein-ligand binding site prediction using DeepTracer-predicted protein structure. EMD-21391(PDB ID: 6vv5), chain-M with the PUResNet predicted pocket and the ligand PAM (yellow). Grey dots represent the PUResNet-predicted binding pocket (created in ChimeraX).

2.4 Integration with DeepTracer Protein Structure Prediction Pipeline

DeepTracer is a powerful deep learning-based method and service for protein structure modeling using Cryo-EM maps [18]. It uses a 3D convolutional neural network that allows for fast and accurate structure predictions. Compared to other state-of-the-art methods such as Phenix [30], Rosetta [31] [32] and MAINMAST [32], applied on a set of coronavirus-related Cryo-EM maps, there is on average 30% more residue coverage when compared to Phenix with an average RMSD improvement of 0.11Å; applied on a set of nine Cryo-EM maps, RMSD improvements range from 1.37Å to 0.85Å; yielding 57% more coverage than MAINMAST. To extend the functionality of this method, we integrate DeepTracer with two other PLBP methods, PUResNet and DeepSurf, which have demonstrated promising results on the EMD-Ligand Dataset. In doing so, we have built an end-to-end pipeline (DeepTracer-PLBP) that can directly predict protein-ligand binding sites from Cryo-EM maps without requiring experimentally-determined protein structures.

As shown in Figure 3, we present another example of this pipeline using the protein Holo-bacterioferritin-form-I (EMD-9913) in 3A the ground-truth state as well as

3B the DeepTracer-predicted structure of chain V. In 3C, we show an overlay of the ground-truth state and the DeepTracer-predicted structure with its corresponding ligands, Fe^{2+} (red) and HEM (cyan). Using PUResNet, the predicted ligand binding pocket (as shown with grey dots), which overlaps with the true ligand, is validated based on the distance between the center of the binding pocket and the center of the ligand itself with a distance <4.0Å.

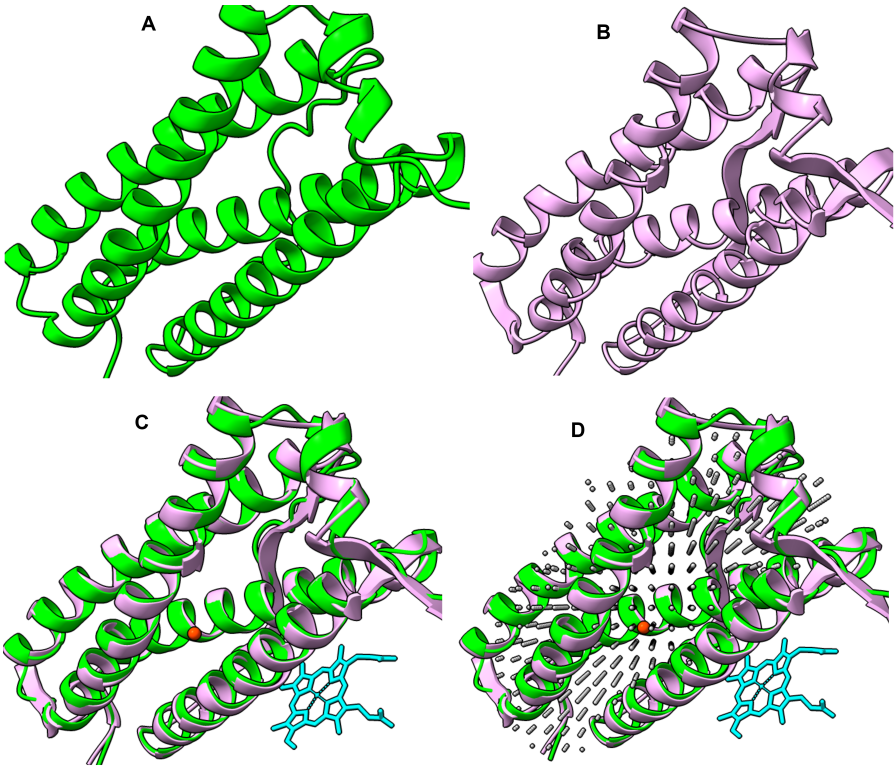

Fig. 3 The integration of ligand binding site prediction with DeepTracer. A) The ground truth state of the protein structure (PDB ID: 6K43, chain N). B) The DeepTracer-predicted protein structure of EMD-9913, chain V. C) The protein structure with two ligands, orange dot (Fe^{2+}), light blue (HEM). D: the gray dots represent the ligand binding site predicted by PUResNet, validated based on a D_{CC} value of less than 4.0Å to the ligands.

To demonstrate the effectiveness of DeepTracer-PLBP, we used the EMD-Ligand Dataset. However, we needed to make a small modification to our scoring algorithm since the predicted protein structures by DeepTracer may have a different number of chains compared to the ground-truth structures, as the chain IDs may not correspond to the same section of the protein of interest. Therefore, we compare each ligand prediction to all reference ligands of the full protein complex, not just one chain.

As depicted in Figure 4, the pipeline begins by taking a de novo multichain protein complex with a Cryo-EM map as input to DeepTracer, which predicts the backbone structure. Next PyMol is used to split all predicted protein chains into individual units. Lastly, each chain is fed into either PUResNet or DeepSurf to predict possible ligand-binding sites.

2.5 DeepTracer Drug-like Ligand Generator

The ligand binding site prediction capability of DeepTracer can be utilized for in silico drug design through integration with the Pocket2Mol pipeline [19]. Pocket2Mol is an E(3)-equivariant generative network composed of a new GNN that captures both spatial and bonding relationships between atoms of the binding pocket and a new efficient algorithm that samples new drug candidates conditioned on the pocket representations from a tractable distribution without relying on conventional MCMC algorithms [33]. According to Peng et al., evaluation of ligand-generation performance of Pocket2Mol is completed through comparison to other existing, state-of-the-art methods such as a CVAE-base model [19] and another auto-regressive generative model [33]. The result shows that Pocket2Mol can sample drug candidates with more accurate binding properties including binding affinity, drug-likeness, and synthetic accessibility.

To demonstrate that DeepTracer-PLBP pipeline can be extended to generate drug-like molecules, we appended the Pocket2Mol pipeline to it. As shown in Figure 4 below, the DeepTracer-predicted protein structure is provided to PUResNet to predict possible ligand-binding sites. The coordinates of the center of the predicted binding site and the predicted protein structures are then provided to Pocket2Mol which uses the chemical and geometric attributes of the binding pockets to generate candidate ligands for the predicted binding site.

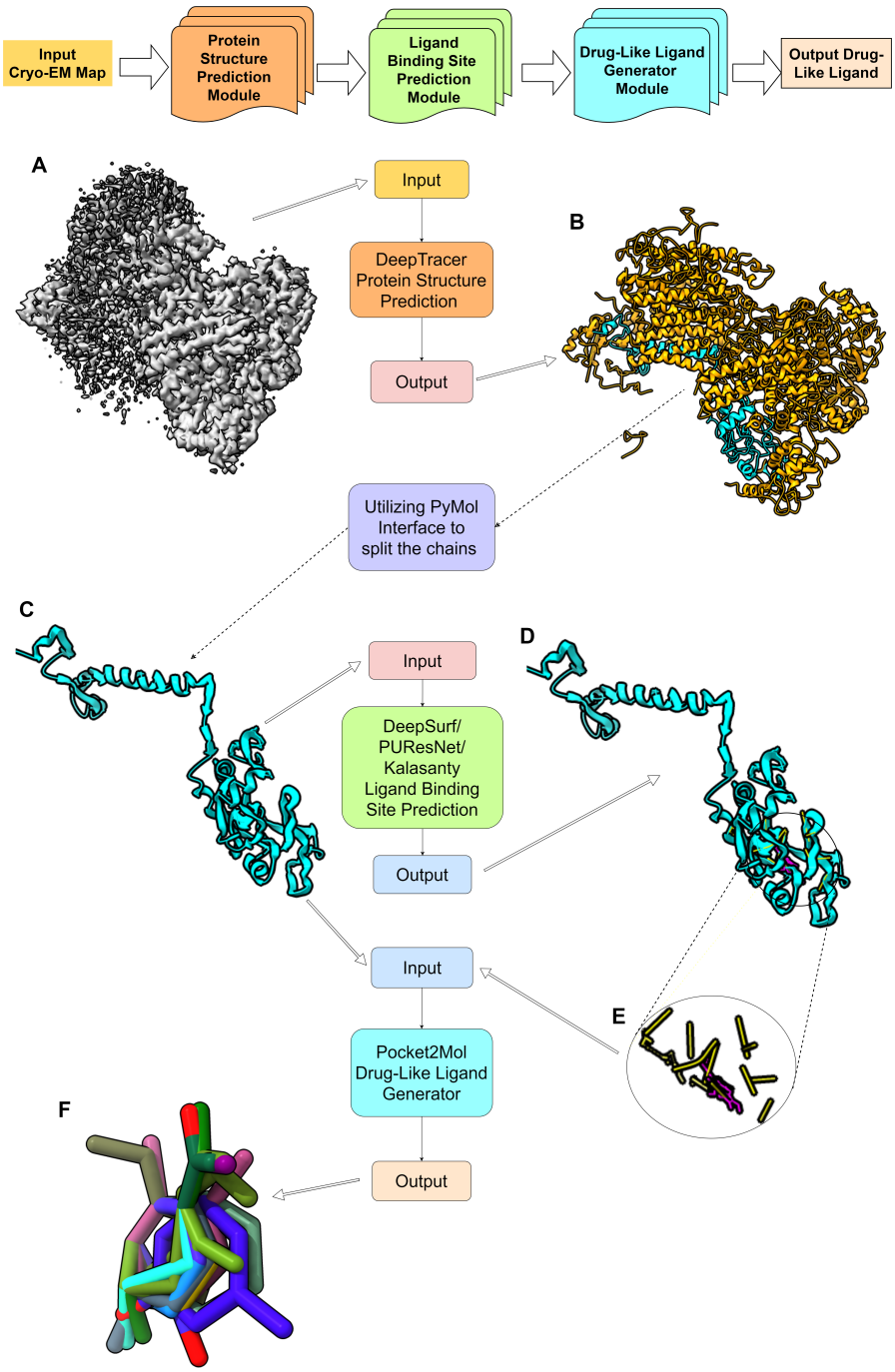

Fig. 4 The high-level workflow pipeline of DeepTracer-PLBP and DeepTracer-DLMGen from an input Cryo-EM map. A: Input Cryo-EM map EMD4165. B: DeepTracer-predicted protein structure. C: Chain A of the DeepTracer-predicted protein structure. D and E: DeepSurf predicted pocket (yellow) with the ligand (purple) on chain A. F: seven generated drug-like ligands (overlapped in different colors).

Figure 5 demonstrates the capability of the DeepTracer-DLMGen pipeline. Panel A depicts the predicted structure of Alternative complex III (PDB ID: 6F0K - EMD-4165) chain A. Panel B demonstrates several Pocket2Mol-generated drug-like molecules (white) located in the ligand binding pocket predicted by DeepSurf. This pipeline produces various DLM candidates, though for space considerations only a few are plotted in the figure. Panel C and D show the DeepTracer-predicted structure, PUResNet-predicted binding site, and the Pocket2Mol-generated DLMs of Human IMPDH2 (PDB ID: 6UAJ - EMD-20707) chain R.

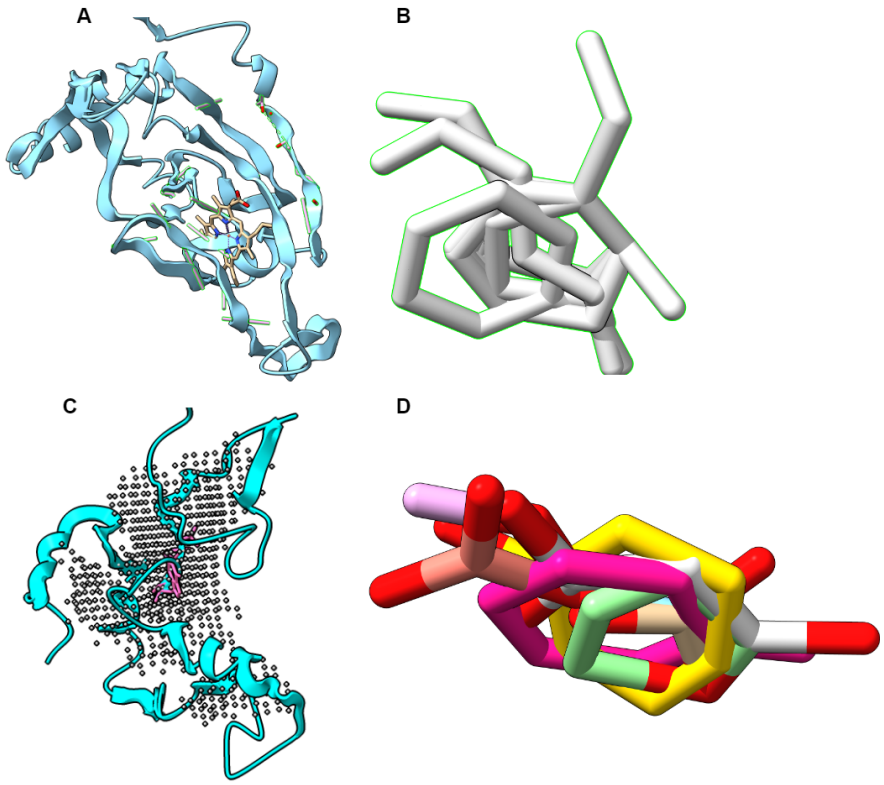

Fig. 5 DeepTracer-BindPRED and DLMGen of Alternative complex III (PDB ID: 6F0K - EMD-4165) chain A and Human IMPDH2 (PDB ID: 6UAJ – EMD-20707). A) The DeepTracer-predicted structure of Alternative complex III with the DeepSurf-predicted ligand binding site bound with HEC (represented by sticks). B) The DLMs generated by Pocket2Mol (white). C) DeepTracer-predicted structure of Human IMPDH2, chain R with the binding site predicted by PUResNet (shown in gray dots) and bound GTP (purple). D) The DLMs generated by Pocket2Mol.

3 Experiments and Results

3.1 Evaluation of PLBP Methods' Performance on the Traditional Protein-Ligand Binding Databases

We evaluate the effectiveness of PUResNet, Kalasanty, and DeepSurf on the Coach420, BU48, and PDBbind Core-Set (CASF-2016) datasets to provide a comprehensive comparison of their binding-site prediction capabilities. It is important to note that PUResNet expresses its predicted binding site in terms of the grid in the pocket or cavity where the ligand is located while Kalasanty expresses its predicted binding sites in terms of the atoms and/or residues that are near the predicted site (using a distance cut-off threshold of 4.0Å). Therefore, the predicted sites of these two methods cannot be directly compared. To ensure a fair comparison, we modifed the Kalasanty algorithm to express its prediction in terms of the pocket or cavity as well. Figure 6 (left) demonstrates that the PUResNet prediction produces a slightly better D_{CC} curve on the Coach420 dataset. However, the DeepSurf method utilizes surface-point based prediction and thus can only express binding sites in terms of the atoms/residues on the protein surface. To compare DeepSurf with Kalasanty and PUResNet, we modifed the PUResNet method to output atoms and/or residues close to its predicted binding sites in the exact same way as Kalasanty does. Figure 6 (right) shows the D_{CC} curves of predicted binding sites (not pockets/cavities) of these three methods on the Coach420 dataset.

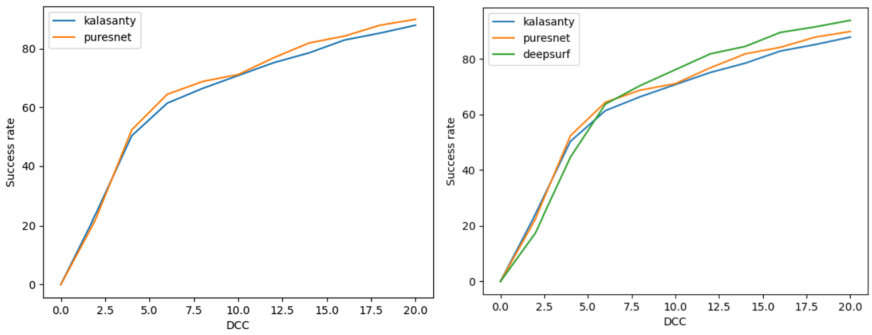

Fig. 6 (left) D_{CC} performance of Kalasanty (cavity) and PUResNet on testing dataset Coach420. (right) D_{CC} performance of Kalasanty, PUResNet and DeepSurf on testing dataset Coach420, comparing predicted sites (residuals).

When D_{CC} falls below 4.0Å, it is considered a valid prediction. Using this criterion, we can calculate True Positive (TP), False Positive (FP), False Negative (FN), F_1-score, and Success Rate (SR). Table 2 shows the results comparison of these three methods applied to three chosen test datasets: Coach420, BU48, and CASF-2016. For the Coach420 dataset, DeepSurf achieves the best F_1-score and SR, while PUResNet is only slightly better than Kalasanty. For the BU48 dataset, DeepSurf and PUResNet both achieve a ~60% SR, and DeepSurf has the best F_1-score due to significantly fewer FPs. However, Kalasanty performs the worst on this dataset, demonstrating a clear gap

from PUResNet even though they perform similarly on the larger Coach420 dataset. On the PDBbind Core Set, Kalasanty performs significantly better than PUResNet. In summary, no one method clearly outperforms the other methods on all test sets.

Table 2 Comparison of Kalasanty and PUResNet on predicted binding sites on the Coach420 dataset. Performance comparison of predicted binding sites (represented by atoms/residues on the protein, not by pockets/cavities) on Coach420 set, containing 298 protein-ligand complexes. Evaluation results of each method on the BU48 dataset, containing 62 protein-ligand complexes. Evaluation results of each method on the PDBbind Core Set (CASF-2016), containing 285 protein-ligand complexes. (DeepSurf results are not included in last row due to unexpected results – see explanation above).

Test Dataset	Model	TP	FP	FN	F_1-score	SR	Declared SR[1]
Coach420	Kalasanty	119	174	25	0.54	40.0%	51%
	PUResNet	122	175	19	0.56	40.9%	53%
	DeepSurf	133	184	11	0.58	44.6%	N/A
BU48	Kalasanty	27	33	7	0.574	43.5%	57%
	PUResNet	37	34	3	0.667	59.7%	62%
	DeepSurf	36	27	2	0.713	58.1%	N/A
CASF-2016	Kalasanty	165	110	30	0.702	57.9%	N/A
	PUResNet	130	167	17	0.586	45.6%	N/A

[1]from Peng et al, 2022

3.2 Evaluation of PLBP Methods' Performance on the Novel EMD-Ligand Dataset

We utilize the ground-truth all-chain structures as input for all methods to predict ligand binding sites. The evaluation results are given in Table 3. The reference ligand information is given and the default D_{CC} threshold of 4.0Å is used. To understand the impact of Cryo-EM map resolution, we broke down the results into groups based on map resolution.

As seen in Table 3, PUResNet performs the best overall (SR=16.1%) compared to DeepSurf (SR=15.6%) and Kalasanty (SR=6.4%). It is worth noting that all methods show large variations across different resolution groups, as expected. The highest resolution group (0.5Å-3.0Å) had the worst accuracy among all resolutions, which is surprising, especially considering that all these methods were trained using the ScPDB database in which all protein images are in that same resolution range. PUResNet had the best accuracy for all resolutions excluding the poor performance noted in the 4.0Å-4.5Å group in which DeepSurf performed better (SR=21.9%). PUResNet showed significantly better performance than the other two methods on the 3.5Å-4.0Å resolution group (SR=23.5%). Strikingly, the performances of all tested methods on this dataset, when compared to the non-Cryo-EM maps protein/ligand dataset reported in Section 3.1 (Table 2), are significantly worse. This clearly demonstrates the fragility of these PLBP methods across different datasets, though

the reasons are unclear. To improve this methodology, more research is required.

The scoring algorithm used in our evaluation considers multiple hypotheses since the EMD-Ligand Dataset has multiple ligands per protein chain. However, all methods still produced many false negatives. We believe that this is the primary cause of their poor performance compared to the non-Cryo-EM maps protein/ligand dataset reported in Table 2. This may be due to the fact that all three examined methods were trained on the ScPDB database, which is very different from the EMD-Ligand Dataset in terms of map resolution and protein size. We hypothesize that adding more diverse training data into these methods will greatly increase their performance. To test this, we conducted a preliminary using Kalasanty. We aimed to refine or adapt the original Kalasanty model using protein-ligand data from the resolution 3.0Å-3.5Å subset and evaluate the adapted model on the 3.5Å-4.0Å subset. The results are given in Table 3. It is promising to observe a 6.1% increase in SR for the adapted Kalasanty model compared to the original model. This suggests that adding more diverse training data could improve the performance of these PLBP methods.

Table 3 Evaluation results of Kalasanty, DeepSurf, and PUResNet on the novel EMD-Ligand dataset. DeepSurf performance on resolution 4.0Å-4.5Å subset with different D_{CC} thresholds in scoring. Results of the original Kalasanty model vs. the adapted Kalasanty model (Trained on the 3.0Å-3.5Å subset and evaluated on the 3.5Å-4.0Å subset)

Resolution (Å)	Model	D_{CC} threshold	Pred-Ligand Number	FN	TP	FP	F_1-score	SR (%)
0.5-3.0	Kalasanty	4.0Å	1349	7939	378	971	0.078	4.60
	DeepSurf	4.0Å	3281	7798	517	2764	0.089	6.30
	PUResNet	4.0Å	1689	7718	597	1092	0.119	7.20
3.0-3.5	Kalasanty	4.0Å	2729	7758	738	1991	0.132	8.70
	DeepSurf	4.0Å	3901	7663	830	3071	0.134	9.80
	PUResNet	4.0Å	3274	6944	1546	1728	0.263	18.30
3.5-4.0	Original-Kalasanty	4.0Å	2620	5053	740	1880	0.176	12.80
	Adapted-Kalasanty	4.0Å	2160	4696	1097	1063	0.276	18.90
	DeepSurf	4.0Å	2588	5297	496	2092	0.118	8.60
	PUResNet	4.0Å	2798	4430	1362	1436	0.317	23.50
4.0-4.5	Kalasanty	4.0Å	352	1117	90	262	0.116	7.50
	DeepSurf	4.0Å	1250	942	264	986	0.215	21.90
	DeepSurf	5.0Å	–	891	314	936	0.256	26.10
	DeepSurf	6.0Å	–	852	351	899	0.286	29.40
	PUResNet	4.0Å	485	1020	187	298	0.221	15.50

Note for Table 3: We did the above experiments for each resolution subset. The results prove promising for the continued research on our EMD-Ligand Dataset. There is evidence that the improving success rate from 12.8% to 18.9% following 3.0Å-3.5Å subset training and evaluation on the 3.5Å-4.0Å group.

3.3 Evaluation of De Novo Drug-Like Molecule Generation

Several metrics are widely used to evaluate the quality of generated drug candidates [25]. The Vina score estimates the binding affinity between the generated molecules and binding site, with higher affinity being calculated as the percentage of sites in which the generated molecules have higher or equal affinity to those in the test dataset [34]. The Quantitative Estimation of Drug-likeness (QED) metric measures how likely a molecule is to be a potential drug candidate based on several criterion [17]. The Synthetic Accessibility Score (SAscore) represents the difficulty of drug synthesis [35]. The LogP value represents the octanol-water partition coefficient and indicates ideal drug candidates based on values ranging from -0.4 to 5.6 [36]. Finally, Lipinski's rule of five measures how many rules a ligand follows to be considered drug-like [37].

According to Roy et al., 2012, Pocket2Mol has shown better scores in all these metrics when tested on the CrossDocked dataset compared to other methods [25]. The generated molecules not only demonstrate higher binding affinities and more stable chemical properties, but also produce more realistic and accurate structures. In our pipeline, we use the same metrics (excluding Vina Score and High affinity, which are either unavailable or inapplicable) to evaluate the quality of generated molecules. Results for four binding sites are provided in Table 4, where higher scores indicate better results. The scores are averaged over all generated molecules for each site. Compared to Roy et al., 2012, our results are very competitive, though analysis was run on a different dataset and average over many proteins structures [25].

To further compare these results, we also used the ground-truth protein structures and ligand-binding information as inputs for Pocket2Mol-based DLMGen. The results of this examination can be viewed in Table 4 below. For each sample, row 1 represents the DeepTracer-generated structure and the predicted ligand-binding site while the second row represents the corresponding ground-truth structure and ground-truth ligand-binding site. Using DeepTracer-predicted structures and ligand binding sites was as effective as using ground-truth structures, as shown by comparable validation criterion values. This demonstrates that DeepTracer-DLMGen is a valuable tool for drug design when applied to Cryo-EM maps.

Table 4 Evaluation of Pocket2Mol-generated molecules using DeepTracer-predicted protein structures and PUResNet/DeepSurf-predicted ligand-binding sites. A comparison of generated molecules to ground-truth protein structures.

Drug-like Molecules Sample #	EMD ID-chain pdb chain	QED	SA	Lipinski	LogP
Sample 1	EMD20707-R	0.553	0.877	5.000	1.591
	6uaj-A	0.562	0.763	4.809	-0.099
Sample 2	EMD21391-M	0.553	0.901	5.000	1.734
	6vv5-C	0.599	0.884	5.000	2.274
Sample 3	EMD6772-E	0.556	0.857	5.000	2.206
	5xtc-U	0.573	0.885	5.000	1.844
Sample 4	EMD7942-E	0.540	0.748	5.000	2.449
	6dju-B	0.444	0.682	4.713	0.389
Test Set	**Declared score (Pocket2Mol paper)**	**0.563**	**0.765**	**4.902**	**1.586**

Note for the standard metrics used in Table 4: QED is a measure of how likely a molecule is to be a potential drug-candidate; SA represents the difficulty of drug synthesis; QED and SA score were normalized between 0 and 1 where higher values indicate easier synthesis; LogP represents the octanol-water partition coefficient, and in general logP values should be between -0.4 and 5.6 to be good drug candidates; Lipinski measures how many rules the drug follow the Lipinski's rule of five, so five is the highest score.

4 Discussion and Conclusions

We conducted a comprehensive comparative study of three advanced deep learning-based methods for predicting protein-ligand binding sites (Kalasanty, PUResNet, and DeepSurf) on publicly available databases (Coach420, BU48 and CASF-2016). We then initiated the development of ligand-binding site prediction for proteins using Cryo-EM maps by introducing our DeepTracer-PLBP pipeline and created a novel EMD-Ligand Dataset specifically for this purpose. To our knowledge, this is the first database of its kind for predicting ligand binding in protein complexes with Cryo-EM maps. We evaluated the performance of these methods on this database and found that their accuracy is very limited compared to single-chain proteins with high-resolution structures determined by X-ray crystallography and NMR. This finding calls for more research efforts from the community to address this issue. Our preliminary experiments demonstrated that the addition of more training data, particularly proteins with Cryo-EM maps, can significantly improve their performance, even though there is still a large difference between them with their performance on the conventional datasets.

We also applied these methods to structures predicted by DeepTracer using Cryo-EM maps and successfully demonstrated that the integration of these methods with DeepTracer could create an efficient pipeline to predict ligand binding sites directly from Cryo-EM maps, eliminating the need for experimental methods and processes

to determine protein structures, which can be costly and time-consuming.

We further extended the DeepTracer pipeline with Pocket2Mol to generate de novo drug-like molecules (DeepTracer-DLMGen) automatically and efficiently from Cryo-EM maps, which could be valuable for drug discovery and design. Our future work involves gathering more relevant datasets, particularly those with Cryo-EM maps, to train better models for this task. We plan to improve the existing deep learning models by introducing better model structures, such as attention mechanisms, better features, and model ensembles. Additionally, we believe that training models with DeepTracer-predicted protein structures, rather than only experimentally-determined structures, will significantly improve their performance. The success of the DeepTracer-PLBP could improve the accuracy and ease of drug discovery and generation and make the entire pipeline more impactful overall.

References

[1] Stepniewska-Dziubinska, M.M., Zielenkiewicz, P., Siedlecki, P.: Improving detection of protein-ligand binding sites with 3d segmentation. Sci Rep **10**, 5035 (2020) https://doi.org/10.1038/s41598-020-61860-z

[2] Kandel, J., Tayara, H., Chong, K.T.: Puresnet: prediction of protein-ligand binding sites using deep residual neural network. J Cheminform **13**, 65 (2021) https://doi.org/10.1186/s13321-021-00547-7

[3] Mylonas, S.K., Axenopoulos, A., Daras, P.: Deepsurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. Bioinformatics **37**, 1681–1690 (2021) https://doi.org/10.1093/bioinformatics/btab009

[4] Desaphy, J., Bret, G., Rognan, D., Kellenberger, E.: sc-pdb: a 3d-database of ligandable binding sites–10 years on. Nucleic Acids Res **43**, 399–404 (2015) https://doi.org/10.1093/nar/gku928

[5] Wang, R., Fang, X., Lu, Y., Yang, C.Y., Wang, S.: The pdbbind database: methodologies and updates. J Med Chem **48**, 4111–4119 (2005) https://doi.org/10.1021/jm048957q

[6] Mehio, W., Kemp, G.J., Taylor, P., Walkinshaw, M.D.: Identification of protein binding surfaces using surface triplet propensities. Bioinformatics **26**, 2549–2555 (2010) https://doi.org/10.1093/bioinformatics/btq490

[7] Gainza, P.e.a.: Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods **17**, 184–192 (2020) https://doi.org/10.1038/s41592-019-0666-6

[8] Cui, Y., Dong, Q., Hong, D., Wang, X.: Predicting protein-ligand binding residues with deep convolutional neural networks. BMC Bioinformatics **20**, 93 (2019)

https://doi.org/10.1186/s12859-019-2672-1

[9] Xia, C.Q., Pan, X., Shen, H.B.: Protein-ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data. Bioinformatics **36**, 3018–3027 (2020) https://doi.org/10.1093/bioinformatics/btaa110

[10] Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. Int Conf Med image Comput Interv **13**, 234–241 (2015) https://doi.org/10.48550/arXiv.1505.04597

[11] Kühlbrandt, W.: The resolution revolution. Science **343**, 1443–1444 (2014) https://doi.org/10.1126/science.1251652

[12] Verma, N.e.a.: Ssnet: A deep learning approach for protein-ligand interaction prediction. Int J Mol Sci **22**, 1392 (2021) https://doi.org/10.3390/ijms22031392

[13] Jacob, L., Vert, J.P.: Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics **24**, 2149–2156 (2008) https://doi.org/10.1093/bioinformatics/btn409

[14] 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. KDD '21, pp. 975–985. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3447548.3467311 . https://doi.org/10.1145/3447548.3467311

[15] Lim, J.e.a.: Predicting drug-target interaction using a novel graph neural network with 3d structure-embedded graph representation. J Chem Inf Model **59**, 3981–3988 (2019) https://doi.org/10.1021/acs.jcim.9b00387

[16] Robertson, M.J., Zundert, G.C.P., Borrelli, K., Skiniotis, G.: Gemspot: A pipeline for robust modeling of ligands into cryo-em maps. Structure **28**, 707–716 (2020) https://doi.org/10.1016/j.str.2020.04.018

[17] Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., Hopkins, A.L.: Quantifying the chemical beauty of drugs. Nat Chem **4**, 90–98 (2012) https://doi.org/10.1038/nchem.1243

[18] Pfab, J., Phan, N.M., Si, D.: Deeptracer for fast de novo cryo-em protein structure modeling and special studies on cov-related complexes. Proc Natl Acad Sci U S A **118** (2021) https://doi.org/10.1073/pnas.2017525118

[19] Peng, X., Luo, S., Guan, J., Xie, Q., Peng, J., Ma, J.: Pocket2Mol: Efficient molecular sampling based on 3D protein pockets. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th

International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 17644–17655. PMLR, United States (2022). https://proceedings.mlr.press/v162/peng22b.html

[20] Yang, J., Roy, A., Zhang, Y.: Biolip: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res **41**, 1096–1103 (2013) https://doi.org/10.1093/nar/gks966

[21] Hu, L., Benson, M.L., Smith, R.D., Lerner, M.G., Carlson, H.A.: Binding moad (mother of all databases). Proteins **60**, 333–340 (2005) https://doi.org/10.1002/prot.20512

[22] Yang, J., Roy, A., Zhang, Y.: Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics **29**, 2588–2595 (2013) https://doi.org/10.1093/bioinformatics/btt447

[23] Huang, B., Schroeder, M.: Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. BMC Structural Biology **6**, 19 (2006) https://doi.org/10.1186/1472-6807-6-19

[24] Su, M.e.a.: Comparative assessment of scoring functions: The casf-2016 update. J Chem Inf Model **59**, 895–913 (2019) https://doi.org/10.1021/acs.jcim.8b00545

[25] Roy, A., Yang, J., Zhang, Y.: Cofactor: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res **40**, 471–477 (2012) https://doi.org/10.1093/nar/gks372

[26] Chen, K., Mizianty, M.J., Gao, J., Kurgan, L.: A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. Structure **19**, 613–621 (2011) https://doi.org/10.1016/j.str.2011.02.015

[27] Vanwinckelen, G., Blockeel, H.: Look before you leap: Some insights into learner evaluation with cross-validation. In: Hämäläinen, W., Petitjean, F., Webb, I. (eds.) Proceedings of the Workshop on Statistically Sound Data Mining at ECML/PKDD. Proceedings of Machine Learning Research, vol. 47, pp. 3–20. PMLR, Nancy, France (2015). https://proceedings.mlr.press/v47/vanwinckelen14a.html

[28] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation. Medical Image Computing and Computer-Assisted Intervention **9901**, 424–432 (2016) https://doi.org/10.1007/978-3-319-46723-8_49

[29] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[30] Liebschner, D.e.a.: Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. Acta Crystallographica **75**, 861–877 (2019) https://doi.org/10.1107/S2059798319011471

[31] Frenz, B., Walls, A.C., Egelman, E.H., Veesler, D., DiMaio, F.: Rosettaes: a sampling strategy enabling automated interpretation of difficult cryo-em maps. Nat Methods **14**, 797–800 (2017) https://doi.org/10.1038/nmeth.4340

[32] Terashi, G., Kihara, D.: De novo main-chain modeling for em maps using mainmast. Nat Commun **9**, 1618 (2018) https://doi.org/10.1038/s41467-018-04053-7

[33] Ragoza, M., Masuda, T., Koes, D.R.: Generating 3d molecules conditional on receptor binding sites with deep generative models. Chem Sci **13**, 2701–2713 (2022) https://doi.org/10.1039/d1sc05976a

[34] Quiroga, R., Villarreal, M.A.: Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. PLOS ONE **11**(5), 1–18 (2016) https://doi.org/10.1371/journal.pone.0155183

[35] Ertl, P., Schuffenhauer, A.: Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Cheminform **1**, 8 (2009) https://doi.org/10.1186/1758-2946-1-8

[36] Daina, A., Michielin, O., Zoete, V.: ilogp: A simple, robust, and efficient description of n-octanol/water partition coefficient for drug design using the gb/sa approach. Journal of Chemical Information and Modeling **54**, 3284–3301 (2014) https://doi.org/10.1021/ci500467k

[37] Benet, L.Z., Hosey, C.M., Ursu, O., Oprea, T.I.: Bddcs, the rule of 5 and drugability. Adv Drug Deliv Rev **101**, 89–98 (2016) https://doi.org/10.1016/j.addr.2016.05.007

Acknowledgments

We express our gratitude to the authors of DeepSurf and PUResNet for their generous help in answering our questions throughout this work. We would also like to extend our appreciation to the members of the DAIS team, including Jason Chen, Haowen Guan, and Natha Chiu, for their invaluable support. Additionally, we are grateful to Dr. Wooyoung Kim and Dr. Minglei Zhao for their insightful discussions and advice.

Additional Information

Code availability: The DeepTracer-PLBP pipeline will be available to use in the near future at https://deeptracer.uw.edu/
Competing interests: The authors declare no competing interests.

How to cite this article: Lu, C. et al. Protein-Ligand Binding Site Prediction and de Novo Ligand Generation from Cryo-EM Maps (2023).

Appendix

The scoring algorithm's pseudo-code:
1. For each predicted pocket
2. Find the closest reference ligand (among all reference ligands for this protein chain)
3. If the distance D_{CC} <= 4.0Å, TP += 1
4. Else FP += 1
5. For each reference ligand
6. Find the closest pocket (among all predicted pockets for this protein chain)
7. If the distance D_{CC} >4.0Å, FN += 1
8. F_1-score = 2 * TP / (2 * TP + FP + FN)
9. Success Rate = TP / Total_ligands.

The following tables are sample lists of EMD maps (with chain) whose ligand binding sites are correctly predicted by our DeepTracer Ligand-Binding Site Prediction Pipeline (shown in Figure 6). Notation is as follows: "EMD-20707-R" means chain R of EMD map 20707. Reference "6uaj_GTP_A_1" is the ground truth ligand, 6uaj is the PDB id for EMD-20707. GTP is the ligand name, A is the chain id, "1" is the ligand index (6uaj chain A could have more than 1 ligand).

Table 1 For resolution group 3.5Å–4.0Å, the correct prediction by PUResNet.

Hypothesis	Reference
EMD-20707-R	6uaj_GTP_A_1.pdb
EMD-21391-M	6vv5_PAM_C_1.pdb
EMD-6772-E	5xtc_PLX_U_1.pdb
EMD-7942-E	6dju_AGS_B_2.pdb
EMD-7942-M	6dju_AGS_C_1.pdb
EMD-9046-E	6ef8_HEC_C_4.pdb
EMD-9913-A	6k43_FE2_R_1.pdb
EMD-9913-C	6k43_FE2_K_1.pdb
EMD-9913-D	6k43_FE2_W_1.pdb
EMD-9913-E	6k43_FE2_P_1.pdb
EMD-9913-G	6k43_FE2_C_1.pdb
EMD-9913-H	6k43_FE2_O_1.pdb
EMD-9913-J	6k43_FE2_T_1.pdb
EMD-9913-N	6k43_FE2_F_1.pdb
EMD-9913-O	6k43_FE2_B_1.pdb
EMD-9913-Q	6k43_FE2_U_1.pdb
EMD-9913-T	6k43_FE2_S_1.pdb
EMD-9913-V	6k43_FE2_N_1.pdb
EMD-9913-W	6k43_FE2_A_1.pdb

Table 2 For resolution group 3.5Å–4.0Å, the correct predictions by DeepSurf.

Hypothesis	Reference
EMD-20707-P	6uaj_GTP_B_1.pdb
EMD-30074-A	6m49_HC3_A_1.pdb
EMD-31350-H	7ewk_CLA_6_3.pdb
EMD-31350-O	7ewk_CLA_B_14.pdb
EMD-31350-X	7ewk_CLA_1_6.pdb
EMD-4165-A	6f0k_HEC_A_4.pdb
EMD-6929-V	5zgb_CLA_4_1.pdb
EMD-9610-D	6adq_HEM_N_2.pdb
EMD-9791-E	6jbj_ATP_A_1.pdb
EMD-9913-G	6k43_FE2_C_1.pdb
EMD-9913-O	6k43_FE2_B_1.pdb

Table 3 For resolution group 4.0Å–4.5Å, the correct predictions by PUResNet.

Hypothesis	Reference
EMD-4942-U	6rn4_AGS_C_2.pdb

Table 4 For resolution group 4.0Å–4.5Å, the correct predictions by DeepSurf.

Hypothesis	Reference
EMD-11342-M	6zpo_CDL_b_1.pdb
EMD-23773-A	7mdi_CDP_B_1.pdb

Table 5 For resolution group 4.0Å–4.5Å, the correct predictions by PUResNet.

Hypothesis	Reference
EMD-0726-h	6kmw_SF4_cC_2.pdb
EMD-0726-J	6kmw_CLA_aB_34.pdb
EMD-0903-C	6li9_ARG_B_1.pdb
EMD-10468-U	6tdv_TRT_p_1.pdb
EMD-10746-S	6yac_CLA_4_1.pdb
EMD-12180-H	7bgi_CLA_A_26.pdb
EMD-13199-D	7p5h_SF4_A_2.pdb
EMD-13429-H	7pi0_CLA_C_3.pdb
EMD-13976-A	7qhm_HEM_O_1.pdb
EMD-14224-A	7r0w_ECH_A_1.pdb
EMD-14867-D	7zq9_CLA_A_44.pdb
EMD-14971-C	7zub_JY6_D_1.pdb
EMD-21320-G	6vpv_CLA_B_3.pdb

Table 6 For resolution group 0.5Å–3.0Å, the correct prediction by DeepSurf.

Hypothesis	Reference
EMD-0726-d	6kmw_CLA_cB_35.pdb
EMD-0726-N	6kmw_CLA_aA_12.pdb
EMD-0726-P	6kmw_CLA_cA_14.pdb
EMD-0726-Q	6kmw_CLA_cA_20.pdb
EMD-0727-m	6kmx_CLA_bA_14.pdb
EMD-0727-P	6kmx_CLA_cA_24.pdb
EMD-0834-O	6l4t_CLA_14_8.pdb
EMD-0834-P	6l4t_CLA_13_5.pdb
EMD-0834-U	6l4t_KC1_13_1.pdb
EMD-10236-C	6sl5_CLA_B_13.pdb
EMD-10236-T	6sl5_CLA_O_3.pdb
EMD-10236-Z	6sl5_CLA_B_42.pdb
EMD-11588-P	6zzx_CLA_9_2.pdb
EMD-11588-Y	6zzx_CLA_6_8.pdb
EMD-12180-P	7bgi_CHL_6_4.pdb
EMD-13199-a	7p5h_FES_E_1.pdb
EMD-14536-E	7z7s_FMN_F_1.pdb
EMD-23023-R	7ksq_CLA_4_6.pdb
EMD-24821-R	7s3d_CLA_b_19.pdb

Table 7 For resolution group 3.0Å – 3.5Å, the correct prediction by PUResNet.

Hypothesis	Reference
EMD-0821-I	6l35_CLA_A_1.pdb
EMD-0821-R	6l35_CLA_5_6.pdb
EMD-12209-J	7bkc_FAD_e_1.pdb
EMD-12209-U	7bkc_SF4_m_1.pdb
EMD-12210-B	7bkd_FAD_E_1.pdb
EMD-12606-B	7nvm_AF3_E_1.pdb
EMD-13863-H	7q8b_ADP_C_1.pdb
EMD-20846-L	6uqf_CMP_C_1.pdb
EMD-24486-Q	7rje_HEC_N_1.pdb
EMD-30450-F	7crc_ATP_D_1.pdb
EMD-30926-O	7dz8_CLA_2_2.pdb
EMD-7286-H	6btm_HEC_A_2.pdb
EMD-2513-u	4ci0_SF4_B_2.pdb

Table 8 For resolution group 3.0Å – 3.5Å, the correct prediction by DeepSurf.

Hypothesis	Reference
EMD-0094-H	6gyo_CMP_B_1.pdb
EMD-0821-N	6l35_CLA_K_1.pdb
EMD-0821-T	6l35_CLA_6_7.pdb
EMD-0821-V	6l35_CLA_A_41.pdb
EMD-0821-W	6l35_CLA_B_33.pdb
EMD-0849-E	6l7o_SF4_I_2.pdb
EMD-12209-S	7bkc_SF4_l_1.pdb
EMD-12209-X	7bkc_SF4_K_6.pdb
EMD-14169-e	7qv7_SF4_D_3.pdb
EMD-14169-E	7qv7_SF4_B_2.pdb
EMD-14169-N	7qv7_SF4_C_3.pdb
EMD-15017-B	7zxy_HEM_A_2.pdb
EMD-15313-A	8ab7_HEM_C_2.pdb
EMD-20846-H	6uqf_CMP_D_1.pdb
EMD-22445-I	7jrg_HEC_P_1.pdb
EMD-24455-T	7rh5_HEC_O_2.pdb
EMD-24482-B	7rja_HEM_T_2.pdb
EMD-24482-D	7rja_HEM_K_2.pdb
EMD-24486-A	7rje_HEM_K_1.pdb
EMD-25665-B	7t3n_I7Y_A_1.pdb
EMD-26469-Q	7uea_SF4_B_1.pdb
EMD-30926-M	7dz8_CLA_9_4.pdb
EMD-30926-P	7dz8_CLA_2_4.pdb
EMD-30926-y	7dz8_CLA_1_10.pdb
EMD-32462-r	7wfd_CLA_A6_7.pdb
EMD-32463-b	7wfe_CLA_B5_7.pdb
EMD-32463-T	7wfe_CLA_B3_2.pdb
EMD-33112-M	7xc2_ATP_G_1.pdb
EMD-4873-J	6rfr_ZMP_Q_1.pdb
EMD-6932-T	5zji_CHL_Y_5.pdb
EMD-7286-H	6btm_HEC_A_1.pdb