Protein Large Language Model-Powered 3D Ligand Binding Site Prediction from Protein Sequence

Shuo Zhang[1,2], **Lei Xie**[1,2,3*]

[1]Department of Computer Science, Hunter College, The City University of New York, New York, 10065, USA
[2]Helen & Robert Appel Alzheimer's Disease Research Institute, Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, Cornell University, New York, 10065, USA
[3]Ph.D. Program in Computer Science, The Graduate Center, The City University of New York, New York, 10016, USA
szhang4@gradcenter.cuny.edu, lei.xie@hunter.cuny.edu

Abstract

Prediction of ligand binding sites of proteins is a fundamental and important task for understanding the function of proteins and screening potential drugs. Most existing methods require experimentally determined protein holo-structures as input. However, such structures can be unavailable on novel or less-studied proteins. To tackle this limitation, we propose LaMPSite, which only takes protein sequences and ligand molecular graphs as input for ligand binding site predictions. The protein sequences are used to retrieve residue-level embeddings and contact maps from the pre-trained ESM-2 protein language model. The ligand molecular graphs are fed into a graph neural network to compute atom-level embeddings. Then we compute and update the protein-ligand interaction embedding based on the protein residue-level embeddings and ligand atom-level embeddings, and the geometric constraints in the inferred protein contact map and ligand distance map. A final pooling on protein-ligand interaction embedding would indicate which residues belong to the binding sites. Without any 3D coordinate information of proteins, our proposed model achieves competitive performance compared to baseline methods that require 3D protein structures when predicting binding sites. Given that less than 50% of proteins have reliable structure information in the current stage, LaMPSite will provide new opportunities for drug discovery.

Introduction

Identifying ligand binding sites of proteins is an essential step in the pipeline of drug discovery (Anderson 2003). The binding site serves as a key interface where proteins engage in fundamental cellular processes, making it an attractive target for drug molecules. To reduce the time and expense of binding site identification for protein-ligand complexes, computational methods have been proposed and achieved promising performance (Yuan, Pei, and Lai 2013; Macari, Toti, and Polticelli 2019; Dhakal et al. 2022). Non-machine learning computational methods include geometry-based ones (Laskowski 1995; Hendlich, Rippmann, and Barnickel 1997; Huang and Schroeder 2006; Xie and Bourne 2007; Weisel, Proschak, and Schneider 2007; Le Guilloux, Schmidtke, and Tuffery 2009; Xie, Xie, and Bourne 2009; Sael and Kihara 2012) that identify and rank hollow spaces on protein surfaces, probe-based ones (Laurie and Jackson 2005; Amari et al. 2006; Ghersi and Sanchez 2009; Hernandez, Ghersi, and Sanchez 2009) that strategically place probes on protein surfaces to compute energies for identifying binding locations, and template-based ones (Skolnick, Kihara, and Zhang 2004; Wass, Kelley, and Sternberg 2010; Yang, Roy, and Zhang 2013) that query protein templates in large databases with annotated binding sites. Machine learning methods have further advanced the field, leveraging various techniques, including classic machine learning algorithms (Krivák and Hoksza 2015, 2018) and deep learning models (Dhakal et al. 2022). Notably, 3D-convolutional neural networks have been widely used by treating the task of binding site identification as a segmentation problem within 3D space (Jiménez et al. 2017; Simonovsky and Meyers 2020; Stepniewska-Dziubinska, Zielenkiewicz, and Siedlecki 2020; Mylonas, Axenopoulos, and Daras 2021; Aggarwal et al. 2021; Kandel, Tayara, and Chong 2021; Yan et al. 2022).

Despite the progress in ligand binding site prediction, existing methods predominantly depend on the availability of experimentally determined ligand-bound (holo) protein structures, which pose certain limitations. Firstly, these methods encounter challenges when dealing with newly sequenced proteins lacking experimental structural data, hindering their applicability to a broader range of protein targets. Secondly, for proteins with only ligand-free (apo) structures available, detecting cryptic binding sites—those whose shapes may change upon ligand binding—can be a formidable task (Cimermancic et al. 2016). Recently, there have been breakthroughs in predicting highly accurate protein structures, which provide extra sources of structural information besides experimental techniques (Jumper et al. 2021; Baek et al. 2021; Wu et al. 2022; Lin et al. 2023). A growing number of works have applied binding site prediction methods on these computationally achieved protein structures (Akdel et al. 2022; Jakubec et al. 2022; Díaz-Rovira et al. 2023). However, due to the limited portion of predicted structures with high confidence and the ignorance of the dynamic nature of binding sites, experimentally determined structures still provide better and easier situations for binding site identification and docking (Karelina, Noh, and Dror 2023).

To address the aforementioned limitations, we have a two-fold objective. Firstly, we seek to leverage the computationally predicted protein information to expand our capability to

*Corresponding author.

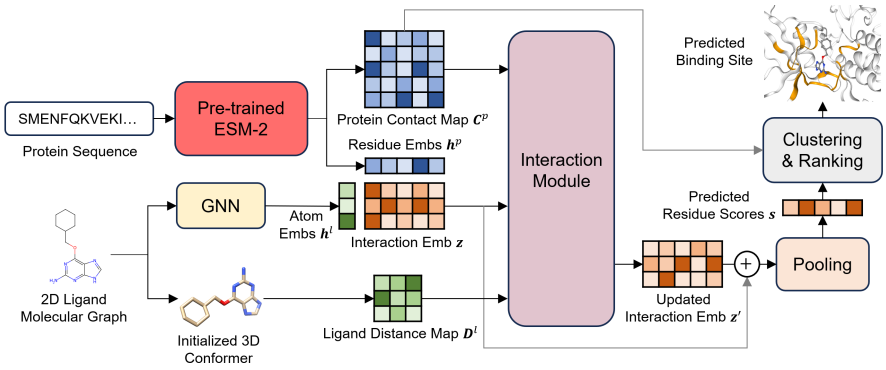

Figure 1: **Illustration of LaMPSite pipeline.** The overall pipeline involves taking a protein sequence and a 2D ligand molecular graph as input. Initially, the pre-trained ESM-2 model computes protein residue embeddings $\boldsymbol{h}^p$ and a contact map $\boldsymbol{C}^p$. Subsequently, ligand atom embeddings $\boldsymbol{h}^l$ are acquired through a GNN and are then combined with residue embeddings to calculate the interaction embedding $\boldsymbol{z}$ for the protein-ligand pair. $\boldsymbol{z}$ is refined in the interaction module using geometric constraints, including $\boldsymbol{C}^p$ and the ligand distance map $\boldsymbol{D}^l$ from the initialized 3D conformer. Then the interaction embeddings are aggregated, followed by pooling to generate scores $\boldsymbol{s}$ for each residue. Finally, residues are clustered based on the protein contact map, and these clusters are ranked to determine the binding site.

predict binding sites on a broader set of proteins, including those lacking experimental structural data. Secondly, we aim to reduce the heavy dependence on rigid 3D protein structures and instead incorporate the dynamic nature of binding sites. To achieve these goals, we propose Large Language Model-Powered Ligand Binding Site Predictor (LaMPSite), which capitalizes on the protein information computed from protein sequence based on the recent ESM-2 pre-trained protein large language model (LLM) (Lin et al. 2023). The computed protein information includes residue-level embeddings that implicitly capture the 3D structure of a protein. These embeddings are integrated with atom-level embeddings of ligands obtained from a graph neural network to compute the protein-ligand interaction embedding. We also utilize the predicted protein contact map from ESM-2 and the distance map of ligand conformer as geometric constraints to further update the protein-ligand interaction embedding to learn more realistic binding site structures. We then apply mean pooling to the interaction embedding, enabling us to quantify the score between each protein residue and the entire ligand, where a higher score indicates a greater likelihood that the residue is part of a binding site. Finally, we leverage the protein contact map information to guide the clustering of filtered residues, ultimately yielding our predictions for binding sites.

We evaluate our LaMPSite against several baselines for binding site prediction on a benchmark dataset. Notably, our method demonstrates competitive performance without 3D protein coordinate information, in contrast to baseline methods that heavily rely on 3D protein structures. The results highlight that LaMPSite provides a novel and promising direction for binding site prediction, driven by the capabilities of protein LLMs. Our main contributions are as follows:

- We introduce LaMPSite, a novel ligand binding site prediction method powered by a protein LLM. LaMPSite relies solely on protein sequences and ligand molecular graphs as initial input and eliminates the need for 3D protein coordinates throughout the prediction process.
- When compared to baseline methods that utilize experimental 3D protein holo-structures, our model demonstrates competitive performance on a benchmark dataset for binding site prediction.

Method

The overall architecture of LaMPSite is depicted in Figure 1. We will describe the relevant modules in this section. Additional details not covered here are included in the Appendix.

Involved Representations

Protein. Our model only takes protein sequences as the initial input to compute protein representations. Leveraging a pre-trained ESM-2 protein LLM, we generate protein residue embeddings $\boldsymbol{h}^p \in \mathbb{R}^{n_p \times d}$ from the provided protein sequence, where n_p is the number of residues in the sequence and d is the hidden dimension size. These embeddings are directly derived from ESM-2 and have demonstrated the emergence of 3D structural information inside (Lin et al. 2023). Additionally, we make use of the unsupervised contact prediction results obtained from ESM-2 to generate protein contact maps $\boldsymbol{C}^p \in \mathbb{R}^{n_p \times n_p}$, serving as a low-resolution estimate of the protein structural information.

Ligand. The 2D ligand molecular graphs excluding hydrogen atoms are initially provided to be fed into a Graph Neural Network (Zhang, Liu, and Xie 2023) to learn atom-level embeddings $\boldsymbol{h}^l \in \mathbb{R}^{n_l \times d}$, where n_l is the number of ligand atoms. Besides, we use RDKit (Landrum 2013) to initialize a 3D conformer for each ligand to compute the corresponding ligand distance map $\boldsymbol{D}^l \in \mathbb{R}^{n_l \times n_l}$.

Protein-ligand pair. Based on the protein residue embeddings $\boldsymbol{h}^p$ and ligand atom embeddings $\boldsymbol{h}^l$, we compute the interaction embedding $\boldsymbol{z} \in \mathbb{R}^{n_p \times n_l \times d}$ for the protein-ligand

pair, which models the interactions between protein residues and ligand atoms. For i-th protein residue and j-th ligand atom, we define $\boldsymbol{z}_{ij} = \boldsymbol{h}_i^p \odot \boldsymbol{h}_j^l$. This information inherently reflects the residues with stronger interactions with ligand atoms, thereby identifying residues within binding sites.

Interaction Module

While the aforementioned $\boldsymbol{z}$ can be directly employed to predict binding site residues, it's worth noting that the 3D structural information that emerged from ESM-2 may not be ideal holo-structures for detecting binding sites. Therefore, we want to incorporate the dynamic nature of binding sites into our model, which involves introducing further modifications to $\boldsymbol{z}$. To accomplish this, we use the trigonometry module in (Lu et al. 2022), which is a variant of the Evoformer block in (Jumper et al. 2021), as our interaction module to update $\boldsymbol{z}$ based on the geometric constraints in protein contact map $\boldsymbol{C}^p$ and ligand distance map $\boldsymbol{D}^l$.

Pooling Module

After having the updated interaction embedding $\boldsymbol{z}'$, we sum it together with the original $\boldsymbol{z}$ and perform a linear transformation to reduce the hidden dimension size to 1. The resulting final interaction embedding $\boldsymbol{z}^{final} \in \mathbb{R}^{n_p \times n_l \times 1}$ contains scores that indicate the likelihood of interaction between residues and ligand atoms. To compute a score $\boldsymbol{s}$ for each residue, representing the interaction likelihood between the entire ligand and the residue, we apply mean pooling specifically over the dimension related to ligand atoms on $\boldsymbol{z}^{final}$. For the i-th residue, $\boldsymbol{s}_i = \sum_j \boldsymbol{z}_{ij}^{final}$, where j represents the j-th ligand atom. A higher score indicates a stronger interaction between a residue and a ligand, concurrently signifying that the residue likely constitutes part of the binding site.

Clustering and Ranking Module

For the identification of binding sites on the test set, we first normalize the predicted scores $\boldsymbol{s}$ to the range of 0 to 1. Following this, we apply a threshold v to filter out residues with scores less than v. Subsequently, we cluster the remaining residues using the single linkage algorithm (Müllner 2011), utilizing the protein contact map generated by ESM-2. The clusters are finally ranked based on the squared sum of the associated residue scores. An illustration of these steps can be found in the Appendix.

Experiments

Datasets. For the training of our model, we use the scPDB v.2017 database (Desaphy et al. 2015), which is a widely used dataset for binding site identification. For the evaluation of our model, we use COACH420 (Krivák and Hoksza 2018), which is a data set with proteins that contain a mix of drug targets and natural ligands. More details of the datasets are provided in the Appendix.

Experimental Settings. For the training process, our model is optimized to minimize the binary cross-entropy loss between the predicted residue scores $\boldsymbol{s}$ and the ground-truth labels. An early-stopping strategy is adopted to decide the best epoch based on the validation loss.

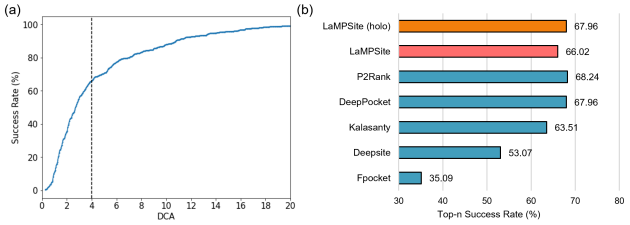

Figure 2: **Models' performance on COACH420.** (a) Success rate plot for different DCA thresholds for LaMPSite. (b) Comparison of identification performance (Top-n success rate) in terms of DCA.

For our evaluation criteria, we use the DCA criterion, which stands for the distance from the center of the predicted binding site to the closest ligand heavy atom. DCA criterion evaluates the model's ability to find the location of the binding site. We use the ground-truth protein structure to calculate the predicted binding site center by averaging the coordinates of all alpha-carbons in our predicted binding site residues. Predictions with DCAs < 4Å are considered successful. Specifically, we evaluate the model's ranking capability by measuring the success rates when considering the top n-ranked pockets, where n is the number of annotated binding pockets for a given protein. We calculate the success rates for each fold in the 10-fold cross-validation, and the final result is obtained by averaging these success rates. More details of the implementations can be found in the Appendix.

Baselines. We compare our model with the following baselines: Fpocket (Le Guilloux, Schmidtke, and Tuffery 2009), Deepsite (Jiménez et al. 2017), Kalasanty (Stepniewska-Dziubinska, Zielenkiewicz, and Siedlecki 2020), DeepPocket (Aggarwal et al. 2021), P2Rank (Krivák and Hoksza 2018). The results of the baselines are from (Aggarwal et al. 2021), which are evaluated on the same test set as ours. Fpocket is a geometry-based method, Deepsite, Kalasanty, and DeepPocket are 3D-CNN-based methods, and P2Rank is a random forest-based method. All baselines require the experimental protein holo-structures as input.

Results

To evaluate the binding site identification performance of LaMPSite, we first plot the relationship between DCA success rates and distance thresholds in Figure 2(a). LaMPSite returns an average DCA success rate of 66.02% at 4Å threshold. Comparing this result with baseline methods, we observe that LaMPSite significantly outperforms Fpocket, Deepsite, and Kalasanty, while trailing DeepPocket (67.96%) and P2Rank (68.24%), as illustrated in Figure 2(b). These results demonstrate that LaMPSite, which does not rely on 3D protein coordinate information for binding site prediction, effectively leverages the 3D structural information that emerged in the pre-trained ESM-2 model. Furthermore, it's noteworthy that

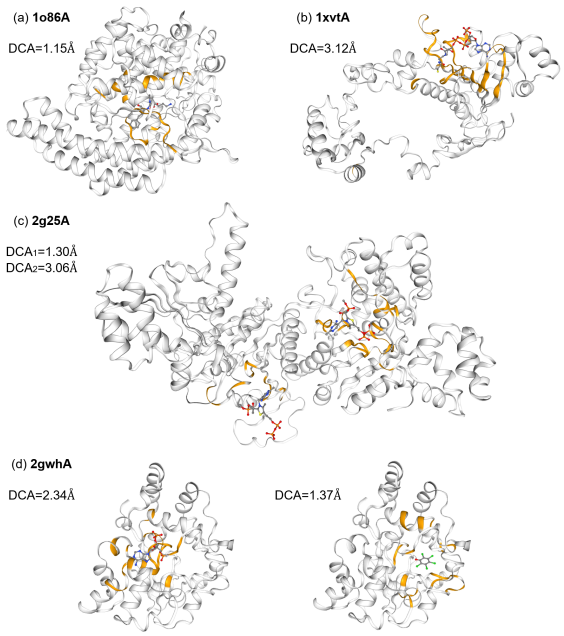

Figure 3: **Predictions and visualizations of binding sites in example complexes.** The binding site residues predicted by LaMPSite are highlighted in orange on the ground-truth 3D complex structures. The respective DCAs are also included.

LaMPSite consistently provides at least one binding site prediction for each protein in COACH420, a behavior similar to Fpocket, DeepPocket, and P2Rank. In contrast, Deepsite fails for one protein, and Kalasanty fails for 12 proteins. This indicates the accuracy and robustness of LaMPSite as a binding site prediction method. In terms of inference time, LaMPSite exhibits an impressive performance, requiring only approximately 0.2s per query protein. This efficiency highlights the potential for LaMPSite to be employed in virtual ligand screening applications.

We also explore how LaMPSite performs when replacing the protein contact maps with the corresponding experimental protein holo-structures. We denote the resulting model as LaMPSite (holo). As depicted in Figure 2(b), LaMPSite (holo) achieves a success rate of 67.96%, which is comparable to DeepPocket and P2Rank. This is expected, given that protein contact maps contain only low-resolution geometric information compared to experimentally determined protein structures, and the contact maps generated from ESM-2 are derived from unsupervised training on a limited set of 20 proteins, offering only a coarse estimation of protein structures.

Visualizations of Predicted Residues. In Figure 3, we show examples (PDB IDs: 1O86, 1XVT, 2G25, 2GWH) illustrating our predicted binding site residues. For 1O86, LaMPSite accurately identifies the correct binding site, which manifests as hollow spaces within the protein, as depicted in Figure 3(a). In the case of 1XVT, a protein with two domains connected by linkers, LaMPSite successfully identifies the binding site within the larger domain, situated in the upper right part of Figure 3(b). Regarding 2G25 that has two binding sites, LaMPSite can successfully detect all of them as shown in Figure 3(c). For 2GWH that binds to two different ligands on different binding sites, LaMPSite correctly predicts the binding site based on the corresponding ligand as input individually, which demonstrates the ability of LaMPSite to discriminate binding pockets on the same protein based on the ligand.

Table 1: **Results of ablation study.** The Top-n success rate (SR) results are compared.

Ablation	Top-n SR
LaMPSite	**66.02**
w/o clustering	65.18
w/o merging z & z'	63.50
w/o interaction module	62.40

Ablation Study. To evaluate the effectiveness of the modules in LaMPSite, we conduct an ablation study by creating three LaMPSite variants: one without the clustering step during inference, another without the merging of interaction embeddings z and z', and a third without the interaction module. The results are presented in Table 1, revealing that the original LaMPSite outperforms all ablated variants. This observation demonstrates the significance of incorporating all these modules into the model for optimal performance.

Limitations. While our model has demonstrated good performance, it is important to acknowledge several limitations. First, our model typically generates only one binding site candidate per prediction in most cases (a total of 383 candidates for 359 pairs in COACH420), constrained by the current filtering and clustering process during inference. Second, because our model takes a single chain as input, its performance in multi-chain scenarios may be limited, as it doesn't account for interactions between chains. Third, due to memory constraints, our model is not trained on relatively long protein sequences (length > 850), potentially impacting its performance. Lastly, the evaluation of our model could be expanded to encompass additional datasets and more challenging scenarios, such as the detection of cryptic binding sites in apo-structures.

Conclusion

In this work, we introduce LaMPSite, a ligand binding site prediction model empowered by a pre-trained protein LLM. By relying solely on protein sequences and ligand molecular graphs as initial input, LaMPSite eliminates the need for 3D protein coordinates during the prediction process, presenting a novel approach to ligand binding site detection model design. LaMPSite demonstrates competitive performance when compared to baselines that rely on experimental 3D protein holo-structures on benchmark dataset. Future directions for research include optimizing memory utilization, assessing scenarios involving multi-chain proteins, evaluating the model's ability to detect cryptic binding sites, and integrating computationally predicted protein structures into the model.

References

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

Akdel, M.; Pires, D. E.; Pardo, E. P.; Jänes, J.; Zalevsky, A. O.; Mészáros, B.; Bryant, P.; Good, L. L.; Laskowski, R. A.; Pozzati, G.; et al. 2022. A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*, 29(11): 1056–1067.

Amari, S.; Aizawa, M.; Zhang, J.; Fukuzawa, K.; Mochizuki, Y.; Iwasawa, Y.; Nakata, K.; Chuman, H.; and Nakano, T. 2006. VISCANA: visualized cluster analysis of protein- ligand interaction based on the ab initio fragment molecular orbital method for virtual ligand screening. *Journal of Chemical Information and modeling*, 46(1): 221–230.

Anderson, A. C. 2003. The process of structure-based drug design. *Chemistry & biology*, 10(9): 787–797.

Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; et al. 2021. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557): 871–876.

Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; et al. 2016. CryptoSite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. *Journal of molecular biology*, 428(4): 709–719.

Desaphy, J.; Bret, G.; Rognan, D.; and Kellenberger, E. 2015. sc-PDB: a 3D-database of ligandable binding sites—10 years on. *Nucleic acids research*, 43(D1): D399–D404.

Dhakal, A.; McKay, C.; Tanner, J. J.; and Cheng, J. 2022. Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. *Briefings in Bioinformatics*, 23(1): bbab476.

Díaz-Rovira, A. M.; Martin, H.; Beuming, T.; Díaz, L.; Guallar, V.; and Ray, S. S. 2023. Are deep learning structural models sufficiently accurate for virtual screening? Application of docking algorithms to AlphaFold2 predicted structures. *Journal of Chemical Information and Modeling*, 63(6): 1668–1674.

Ghersi, D.; and Sanchez, R. 2009. Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. *Proteins: Structure, Function, and Bioinformatics*, 74(2): 417–424.

Hendlich, M.; Rippmann, F.; and Barnickel, G. 1997. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, 15(6): 359–363.

Hernandez, M.; Ghersi, D.; and Sanchez, R. 2009. SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic acids research*, 37(suppl_2): W413–W416.

Huang, B.; and Schroeder, M. 2006. LIGSITE csc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC structural biology*, 6: 1–11.

Jakubec, D.; Skoda, P.; Krivak, R.; Novotny, M.; and Hoksza, D. 2022. PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures. *Nucleic Acids Research*, 50(W1): W593–W597.

Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; and De Fabritiis, G. 2017. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics*, 33(19): 3036–3042.

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

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

Karelina, M.; Noh, J. J.; and Dror, R. O. 2023. How accurately can one predict drug binding modes using AlphaFold models? *bioRxiv*, 2023–05.

Krivák, R.; and Hoksza, D. 2015. Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features. *Journal of cheminformatics*, 7(1): 1–13.

Krivák, R.; and Hoksza, D. 2018. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of cheminformatics*, 10: 1–12.

Landrum, G. 2013. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling.

Laskowski, R. A. 1995. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of molecular graphics*, 13(5): 323–330.

Laurie, A. T.; and Jackson, R. M. 2005. Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics*, 21(9): 1908–1916.

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

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

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

Macari, G.; Toti, D.; and Polticelli, F. 2019. Computational methods and tools for binding site recognition between proteins and small molecules: from classical geometrical approaches to modern machine learning strategies. *Journal of computer-aided molecular design*, 33: 887–903.

Müllner, D. 2011. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*.

Mylonas, S. K.; Axenopoulos, A.; and Daras, P. 2021. DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics*, 37(12): 1681–1690.

Nguyen, H.; Case, D. A.; and Rose, A. S. 2018. NGLview–interactive molecular graphics for Jupyter notebooks. *Bioinformatics*, 34(7): 1241–1242.

Sael, L.; and Kihara, D. 2012. Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins: Structure, Function, and Bioinformatics*, 80(4): 1177–1195.

Simonovsky, M.; and Meyers, J. 2020. DeeplyTough: learning structural comparison of protein binding sites. *Journal of chemical information and modeling*, 60(4): 2356–2366.

Skolnick, J.; Kihara, D.; and Zhang, Y. 2004. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins: Structure, Function, and Bioinformatics*, 56(3): 502–518.

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

Wass, M. N.; Kelley, L. A.; and Sternberg, M. J. 2010. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic acids research*, 38(suppl_2): W469–W473.

Weisel, M.; Proschak, E.; and Schneider, G. 2007. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central Journal*, 1(1): 1–17.

Wu, R.; Ding, F.; Wang, R.; Shen, R.; Zhang, X.; Luo, S.; Su, C.; Wu, Z.; Xie, Q.; Berger, B.; et al. 2022. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, 2022–07.

Xie, L.; and Bourne, P. E. 2007. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. In *BMC bioinformatics*, volume 8, 1–13. Springer.

Xie, L.; Xie, L.; and Bourne, P. E. 2009. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*, 25(12): i305–i312.

Yan, X.; Lu, Y.; Li, Z.; Wei, Q.; Gao, X.; Wang, S.; Wu, S.; and Cui, S. 2022. PointSite: a point cloud segmentation tool for identification of protein ligand binding atoms. *Journal of Chemical Information and Modeling*, 62(11): 2835–2845.

Yang, J.; Roy, A.; and Zhang, Y. 2013. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, 29(20): 2588–2595.

Yuan, Y.; Pei, J.; and Lai, L. 2013. Binding site detection and druggability prediction of protein targets for structure-based drug design. *Current pharmaceutical design*, 19(12): 2326–2333.

Zhang, S.; Liu, Y.; and Xie, L. 2023. A Universal Framework for Accurate and Efficient Geometric Deep Learning of Molecular Systems. *Scientific Reports*, 13(1): 19171.

Appendix

Details of Datasetss

The scPDB v.2017 database consists of 17594 binding sites, corresponding to 16612 PDB structures and 5540 UniProt IDs. To ensure a rigorous evaluation process, we adhere to the same data split strategy employed in (Stepniewska-Dziubinska, Zielenkiewicz, and Siedlecki 2020; Aggarwal et al. 2021), which prevents any data leakage, and utilizes a 10-fold cross-validation. Multi-chain 3D complexes are split into single chains with their own interacting ligands, and each split pair is treated as a protein-ligand pair. To reduce memory usage, we remove protein sequences longer than 850 amino acids. The ground-truth label of each residue is defined as whether the distance between the residue and the nearest ligand atom is < 8 Å. For COACH420, we use the version referenced in (Aggarwal et al. 2021), which excludes ligands that were improperly prepared or failed to be parsed, resulting in 291 protein structures and 359 ligands. To prevent any potential data leakage, we follow (Aggarwal et al. 2021) to exclude structures in scPDB that have either sequence identity $> 50\%$ or ligand similarity > 0.9 and sequence identity $> 30\%$ to any of the structures in COACH420. Consequently, our processed scPDB dataset contains 16270 protein-ligand pairs. We use the publicly available data for scPDB[1] and COACH420[2]. For the data split of scPDB and the screened data entries of COACH420, we use the source[3] provided by (Aggarwal et al. 2021).

Implementation Details

In LaMPSite, we use the ESM-2-650M model with 33 layers (Lin et al. 2023) for generating protein representations. The residue embeddings from the last layer are utilized. The Automatic Mixed Precision package (torch.amp) in Pytorch is used to enable mixed precision of our model for reducing memory consumption while maintaining accuracy. We use NGLview (Nguyen, Case, and Rose 2018) to generate the visualizations of our predicted residues in Figure 3. All of the experiments are done on an NVIDIA Tesla V100 GPU (32 GB). In Table 2, we list the typical hyperparameters used in our experiments.

Table 2: List of hyperparameters.

Hyperparameters	Value
Batch size	8
Hidden dim. in GNN	128
Hidden dim. in interaction module	32
Learning rate	5e-4
Number of GNN layers	4
Number of interaction modules	2
Max. Number of Epochs	30
Patience for early stopping	4
Dropout rate in interaction module	0.25
Threshold v	0.63

[1] http://bioinfo-pharma.u-strasbg.fr/scPDB/

[2] https://github.com/rdk/p2rank-datasets

[3] https://github.com/devalab/DeepPocket

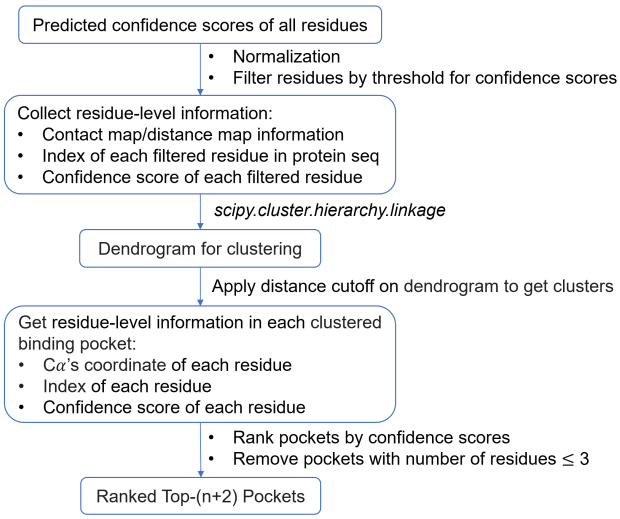

Figure 4: Illustration of the clustering and ranking steps in LaMPSite.

Details of the Clustering and Ranking Module

In Figure 4, we illustrate the detailed steps in the Clustering and Ranking Module of LaMPSite.