PICKPOCKET ENABLES BINDING SITE PREDICTION AT THE PROTEOME SCALE

Stelina Tarasi[1], **Laura Malo**[2], **Alexis Molina**[1,3*]

[1]Department of Artificial Intelligence, Nostrum Biodiscovery, Barcelona, Spain
[2]Department of Drug Discovery, Nostrum Biodiscovery, Barcelona, Spain
[3]Atlas Labs, Barcelona, Spain
*alexis.molina@nostrumbiodiscovery.com, @theatlaslabs.com

ABSTRACT

Accurately identifying protein binding sites is essential for drug discovery, yet existing computational methods often struggle to balance precision, recall, and scalability. We introduce PickPocket, a deep learning model that integrates sequence-derived evolutionary embeddings from ESM-2 with geometric structural representations from GearNet to predict ligand-binding sites at the proteome scale. PickPocket leverages both residue-level sequence context and graph-based spatial relationships, enabling it to generalize across diverse protein families while maintaining high precision. Evaluated on the LIGYSIS benchmark, PickPocket outperforms state-of-the-art methods, achieving the highest F1 score (0.42) and maintaining a competitive Matthews Correlation Coefficient (MCC) (0.37). PickPocket effectively predicts cryptic pockets, surpassing specialized models like PocketMiner even without explicit training on ligand-induced conformational changes. Our large-scale analysis of 356,711 proteins further demonstrates PickPocket's ability to identify novel binding sites across human drug targets. By combining evolutionary and geometric learning, PickPocket represents a scalable, data-driven approach for structure-based drug discovery.

1 INTRODUCTION

Protein function is often modulated by small molecules or peptides that bind to specific surface regions, altering structural dynamics and biochemical activity (Bedard et al., 2020). Identifying these binding sites is a key step in rational drug discovery, as it guides molecular design and ensures target selectivity (Ehrt et al., 2016). Traditional computational approaches for binding site detection include molecular docking, geometric analyses, and physicochemical mapping, but these methods often require prior knowledge of the binding region and may struggle with cryptic or allosteric sites (Zhang et al., 2022; Davis et al., 2009). Advances in artificial intelligence and geometric deep learning have enabled data-driven models to predict binding sites by leveraging protein structural features, sequence conservation, and physicochemical properties, enhancing accuracy beyond rule-based algorithms.

We introduce PickPocket, a model that integrates structural and evolutionary insights for binding site prediction. By combining evolutionary embeddings from ESM-2 (Lin et al., 2023) with geometric representations from GearNet (Zhang et al., 2023b), our approach captures both sequence-derived functional patterns and spatial relationships between residues. This fusion of protein language models and graph neural networks (GNNs) enables robust binding site identification, particularly for conserved and functionally relevant cavities. We demonstrate that PickPocket outperforms existing methods on relevant benchmarks, offering a scalable solution for proteome-wide binding site prediction.

*alexis.molina@nostrumbiodiscovery.com, @theatlaslabs.com

2 Related Work

Binding site prediction has traditionally relied on geometric and biophysical methods. Fpocket (Guilloux et al., 2009) employs Voronoi tessellation and alpha spheres to detect ligand-accessible cavities, while Ligsite (Hendlich et al., 1997) uses a cubic Cartesian grid for solvent-accessible region detection. PocketFinder (An et al., 2005) applies a Lennard-Jones potential map to identify favorable interaction sites. Though efficient and interpretable, these methods depend on predefined heuristics and struggle with conformational flexibility. Machine learning techniques improve upon these approaches by leveraging statistical patterns. Classical methods such as PRANK (Krivák & Hoksza, 2015) and P2Rank (Krivák & Hoksza, 2015; 2018) use decision trees and feature-based classification, achieving higher accuracy than heuristic-based models but remain limited by hand-engineered features. Deep learning advances binding site prediction further, with PUResNet (Kandel et al., 2021; 2024) using 3D CNNs for voxelized protein structures, and GNN-based models like VN-EGNN (Sestak et al., 2024) and GrASP (Smith et al., 2023) leveraging graph attention mechanisms to capture molecular interactions. With the rise of representation learning in biology, methods such as IF-SitePred (Carbery et al., 2024) leverage evolutionary information for residue classification, enhancing binding site prediction by combining ESM-IF1 embeddings (Hsu et al., 2022) with LightGBM. This sequence-informed approach mitigates structural inaccuracies, offering a scalable and generalizable alternative. A more detailed outline of relevant works can be found in Appendix B.

3 Methods

PickPocket's architecture. PickPocket combines sequence and structural information using ESM-2 (Lin et al., 2023) and GearNet (Zhang et al., 2023b). ESM-2, a transformer-based protein language model, processes the sequence through self-attention layers to generate residue embeddings. Following serial fusion (Zhang et al., 2023a), these embeddings serve as node features for GearNet, which processes the protein structure as a graph where nodes represent residues and edges capture their structural relationships. Both components leverage pretrained weights: ESM-2 from large-scale sequence data and GearNet from residue type prediction (Zhang et al., 2023a). For binding site prediction, we concatenate the outputs of both models and pass them through a two-layer MLP classifier to generate per-residue binding probabilities. To assemble neighboring predicted residues into cohesive pockets, we apply DBSCAN clustering. See Appendix C for a more in-depth explanation of PickPocket's architecture.

Training. We use the 2017 release of sc-PDB (Desaphy et al., 2015) as our training dataset, which contains 4,782 proteins, and 6,326 ligands. Following Kandel et al. (2021) and Zhang et al. (2024), we used their pre-clustered dataset based on UniProt IDs. We additionally filtered for single-chain proteins under 1,022 residues, yielding 3,520 protein chains for training and validation.

4 Results and Discussion

4.1 PickPocket Effectively Learns to Predict Binding Pockets

To benchmark our model, we used LIGYSIS (Utgés & Barton, 2024), a curated resource for protein-ligand binding site prediction. Unlike previous benchmarks, LIGYSIS aggregates binding interfaces across multiple protein structures, includes diverse ligands, and prioritizes biological units for functional relevance. After filtering chains with missing UniProt (Consortium, 2024) mappings, the final set comprises 2,775 protein chains. To meet ESM-2 embedding constraints, proteins over 1,022 residues were cropped, excluding 7 pockets from evaluation.

The results show that PickPocket achieves the highest F1 score (0.42) and maintains a competitive MCC of 0.37, outperforming existing deep learning-based methods such as PUResNet, GrASP, and P2Rank_CONS (Jakubec et al., 2022), as well as classical geometric approaches like fpocket, PocketFinder, and Ligsite (Table 2). Additionally, we compute global rankings from metrics in Tables 2 and 3 for all methods. PickPocket demonstrates a well-balanced performance across the evaluated metrics, securing the top rank in Top-N and F1 while maintaining strong rankings in MCC (2nd place) and Top-N+2 (4th place). With an overall total ranking score of 8, PickPocket outperforms other methods, highlighting its ability to consistently identify binding pockets with high precision.

Table 1: Overall ranking of methods grouped by core method names (best rank per group), ordered by total rank.

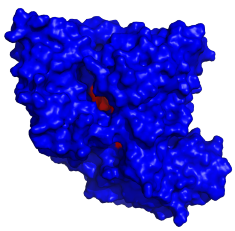
Method Group	Top-N	Top-N+2	F1	MCC	Total
PickPocket	**1**	4	**1**	2	**8**
P2Rank (incl. CONS)	2	3	4	4	13
GrASP	4	5	3	3	15
PUResNet (incl. PRANK)	10	10	2	**1**	23
fpocket (incl. PRANK)	2	**1**	11	11	25
DeepPocket (incl. SEG, RESC)	5	2	10	8	25
Ligsite (incl. AA)	6	6	5	8	25
PocketFinder (incl. AA)	7	7	5	7	26
VN-EGNN (incl. NR)	7	9	7	5	28
Surfnet (incl. AA)	9	8	7	10	34
IF-SitePred (incl. RESC-NR)	11	11	7	6	35

Unlike methods such as fpocket, which achieves the best Top-N+2 ranking but struggles in F1 and MCC, or PUResNet, which excels in MCC but ranks lower in other categories, PickPocket maintains a competitive standing across all metrics. This balance suggests that PickPocket provides robust and reliable pocket predictions without significant trade-offs in different performance aspects (Table 1).

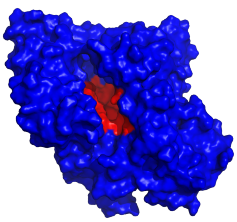
4.2 PickPocket Predicts Cryptic Sites Without Specific Training

Identifying hidden binding pockets, such as cryptic or allosteric sites, remains a challenge in drug discovery, as they often go undetected in static protein structures. These sites may be transient, hidden in unbound states, or form only under specific conformations. Detecting them requires integrating evolutionary, structural, and physicochemical information. We assessed PickPocket on 24 apo structures from the PocketMiner (Meller et al., 2023) test set, comparing predictions to holo structures. Unlike PocketMiner, which relies on molecular dynamics-derived labels, PickPocket predicts cryptic sites without direct supervision.

Analysis using precision-recall curves showed PickPocket outperformed PocketMiner for both apo and holo structures, achieving higher AUC scores (0.617 vs 0.438 for apo; 0.656 vs 0.539 for holo). PickPocket maintained superior precision across a broader recall range, particularly in apo structures where cryptic pockets are harder to detect. This suggests effective capture of structural and evolutionary signals correlating with cryptic site formation, even in unbound states. Interestingly, PickPocket's predictions were particularly enriched in regions undergoing substantial conformational rearrangements upon ligand binding (Figure 1), aligning well with experimentally validated cryptic pockets. This observation suggests that PickPocket effectively identifies pockets that are structurally predisposed to ligand binding, reinforcing its utility in cryptic pocket discovery.

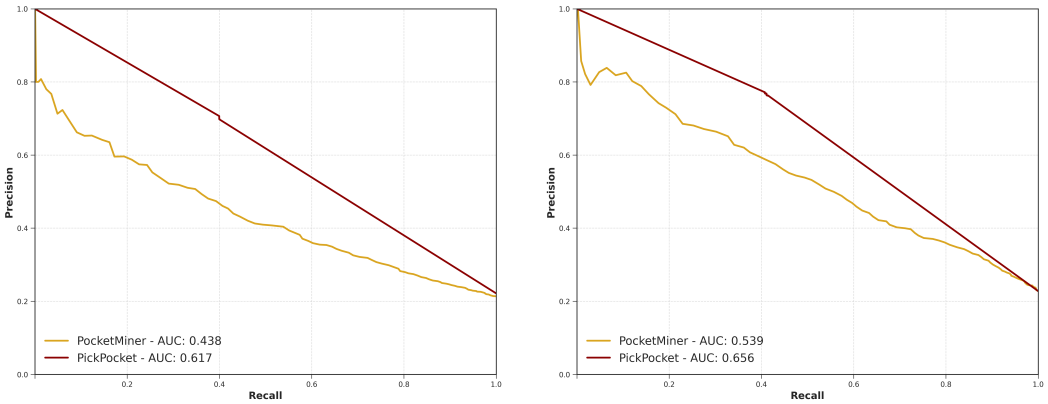


(a) Structural arrangement of LTA4H with cryptic pocket in closed form. Region predicted by PickPocket in red.



(b) Structural arrangement of LTA4H with cryptic pocket in open form. Region predicted by PickPocket in red.

Figure 1: PickPocket accurately identifies binding sites in both conformations: (a) closed pocket (PDB: 5NI6) and (b) open pocket (PDB: 5NIA).



(a) Precision-Recall curve for apo structures.

(b) Precision-Recall curve for holo structures.

Figure 2: Comparison of PickPocket and PocketMiner on apo (unbound) and holo (bound) structures for cryptic pocket prediction. PickPocket consistently achieves superior performance, generalizing across different protein conformations without requiring MD-derived labels.

This underscores the scalability and efficiency of PickPocket for proteome-wide cryptic site discovery, offering a valuable alternative to simulation approaches while reducing computational costs.

4.3 Scaling Binding Site Prediction to All Human Protein Structures

To evaluate the scalability and applicability of PickPocket, we predicted ligand-binding pockets across all single-chain protein structures available in the Protein Data Bank (PDB) as of November 2024 (Berman et al., 2000; 2003). In total, we analyzed 356,711 individual chains, treating each as an independent system. PickPocket identified at least one pocket in 76.2% of chains, with an average of 3.08 pockets per target when at least one was detected. These values are influenced by the choice of aggregation and clustering parameters, which can be adjusted to modulate sensitivity to smaller pocket-like regions. Here, we adopted the sc-PDB distance threshold to ensure consistency with established datasets. To assess PickPocket's performance on predicted structures, we repeated the analysis using AlphaFold2-generated (Jumper et al., 2021; Varadi et al., 2021) models of the same protein chains. Predictions were obtained for 74.8% of structures, with 98.9% overlap in pocket detection between experimental and predicted conformations. This consistency suggests that PickPocket is robust to structural variations, making it applicable to both crystallographic and computationally derived protein models.

To assess the conservation and structural variability of binding pockets across evolutionarily related proteins, we analyzed three homologous superfamilies, selecting a representative structure for each family. We retrieved all available homologs for these reference proteins, predicted binding pockets using PickPocket, and superimposed the structures to evaluate pocket conservation based on centroid distances. For metalloproteases (TldD/PmbA, N-terminal domain), we selected 1VPB, a putative modulator of DNA gyrase (BT3649) from Bacteroides, and analyzed 11 homologous structures within the superfamily. For GPCRs (family 2, extracellular hormone receptor domain), we used 7UZO, the parathyroid hormone 1 receptor extracellular domain complexed with a peptide ligand, comparing it to 191 homologous structures. For kinases (protein kinase-like domain superfamily), we selected 6OQO, a CDK6 complex with an experimental inhibitor, and examined 8,000 homologs from the PDB. Binding pocket predictions were performed on the reference structures, while homologous proteins were superimposed to compute the distance between pocket centroids. Among aligned structures (defined as those with no missing atoms and at least 10 matching Cα residues), the average RMSD between pocket centroids was 4.4 Å for metalloproteases, 5.92 Å for GPCRs, and 3.8 Å for kinases.

The large-scale assessment of PickPocket across experimental and predicted structures demonstrates its capacity to detect ligand-binding sites in diverse protein architectures. The high overlap between pocket predictions in crystallographic and AlphaFold2-derived structures suggests that PickPocket generalizes well across structural conformations, even in the absence of explicit holo-state training.

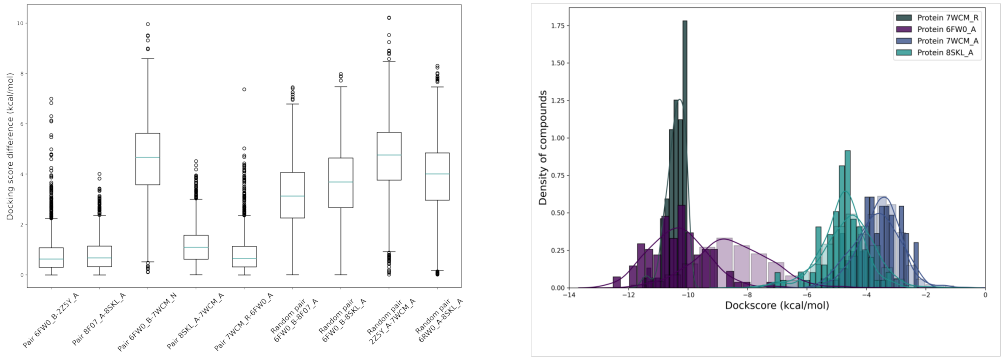
This robustness is particularly relevant given the increasing reliance on computationally predicted structures in drug discovery, where experimental data may be unavailable. However, the small fraction of discrepancies highlights the potential influence of conformational flexibility, particularly in cases where ligand-induced pocket formation is essential. The analysis of homologous superfamilies further reveals how PickPocket captures both conserved and variable aspects of binding site topology. The lower RMSD observed for kinases compared to GPCRs aligns with known functional constraints, as kinases exhibit strong evolutionary pressure to maintain ATP-binding sites, whereas GPCR extracellular domains undergo more structural variation to accommodate different ligands. These findings suggest that PickPocket's predictions are influenced by both structural rigidity and evolutionary conservation.

4.4 Predicted Pocket Embeddings Enable Binding Site Pairing

To assess whether pocket embedding similarity correlates with ligand compatibility, we docked 4,300 diverse compounds across selected receptor pairs and analyzed their embedding distances alongside ligand-binding distributions (See Appendixes G.2 and G.3). Our analysis revealed that pockets with lower embedding distances generally accommodate overlapping ligands, while those with higher distances show minimal shared ligand preferences. More details in Appendix G.4

Low-distance pocket pairs demonstrated high ligand-binding overlap. For example, 6FW0_B – 2Z5Y_A (Monooxygenase-Monooxygenase, Distance = 6.04) and 8F07_A – 8SKL_A (Hydrolase-Hydrolase, Distance = 14.82) showed strong ligand correlation. Interestingly, 7WCM_R – 6FW0_A (GPCR-Monooxygenase, Distance = 13.73) exhibited unexpected ligand compatibility, suggesting conserved physicochemical properties. Most high-distance pairs showed minimal ligand overlap, though 8SKL_A – 7WCM_A (Hydrolase-G-Protein, Distance = 37.96) was a notable exception, showing strong ligand compatibility despite its high embedding distance. Further investigation of this case can be found in Appendix G.4.

These findings demonstrate that embedding similarity can guide target expansion, particularly for structurally related proteins. However, the results also indicate that additional descriptors may be needed to refine ligand-based predictions, especially when considering cross-family compatibility.



(a) (b)

Figure 3: Docking scores for the selected proteins. (a) The docking difference for selected and random pairs, demonstrating that smaller Euclidean distances correlate with more similar docking scores. (b) The distribution of docking scores for the selected pairs, showing a concentration of similar docking scores for these cases. Further details in Appendix G.3

5 Conclusions

PickPocket integrates evolutionary embeddings with graph-based structural representations to achieve precise and scalable binding site prediction, addressing the limitations of traditional methods that overpredict or rely on predefined heuristics. It consistently outperforms existing approaches in identi-

fying ligand-binding sites while maintaining competitive performance in cryptic pocket detection, surpassing PocketMiner despite the absence of explicit training on conformational rearrangements.

Beyond structural resemblance, PickPocket quantifies physicochemical compatibility using embedding-based similarity metrics, providing a robust framework for assessing binding site conservation and functional overlap. With high recall and precision, it achieves a 98.9% structural overlap between PDB and AlphaFold2-derived models, demonstrating reliability across diverse protein structures. Its scalability and ability to infer ligand-binding potential from sequence and structural data make it a powerful tool for data-driven drug discovery and functional proteome annotation. By leveraging pocket embeddings, PickPocket enables large-scale binding site comparisons, advancing target deorphanization, polypharmacology, and ligand repurposing.

REFERENCES

Jianghong An, Maxim Totrov, and Ruben Abagyan. Pocketome via comprehensive identification and classification of ligand binding envelopes*. *Molecular & Cellular Proteomics*, 4(6):752–761, 2005. ISSN 1535-9476. doi: https://doi.org/10.1074/mcp.M400159-MCP200. URL https://www.sciencedirect.com/science/article/pii/S1535947620314742.

Philippe L Bedard, David M Hyman, Matthew S Davids, and Lillian L Siu. Small molecules, big impact: 20 years of targeted therapy in oncology. *The Lancet*, 395(10229):1078–1088, 2020.

Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. *Nature structural & molecular biology*, 10(12):980–980, 2003.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Anna Carbery, Martin Buttenschoen, Rachael Skyner, Frank von Delft, and Charlotte M Deane. Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures. *Journal of Cheminformatics*, 16(1):32, 2024.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 11 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1010. URL https://doi.org/10.1093/nar/gkae1010.

Ian W Davis, Kaushik Raha, Martha S Head, and David Baker. Blind docking of pharmaceutically relevant compounds using rosettaligand. *Protein science*, 18(9):1998–2002, 2009.

Jérémy Desaphy, Guillaume Bret, Didier Rognan, and Esther Kellenberger. sc-pdb: a 3d-database of ligandable binding sites—10 years on. *Nucleic acids research*, 43(D1):D399–D404, 2015.

Christiane Ehrt, Tobias Brinkjost, and Oliver Koch. Impact of binding site comparisons on medicinal chemistry and rational molecular design. *Journal of medicinal chemistry*, 59(9):4121–4151, 2016.

Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996.

Vincent Le Guilloux, Peter Schmidtke, and Pierre Tufféry. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:168 – 168, 2009. URL https://api.semanticscholar.org/CorpusID:6633082.

Thomas A Halgren, Robert B Murphy, Richard A Friesner, Hege S Beard, Leah L Frye, W Thomas Pollard, and Jay L Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of medicinal chemistry*, 47(7):1750–1759, 2004.

Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

Manfred Hendlich, Friedrich Rippmann, and Gerhard Barnickel. Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, 15(6):359–363, 1997. ISSN 1093-3263. doi: https://doi.org/10.1016/S1093-3263(98)00002-3. URL https://www.sciencedirect.com/science/article/pii/S1093326398000023.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

Liegi Hu, Mark L Benson, Richard D Smith, Michael G Lerner, and Heather A Carlson. Binding moad (mother of all databases). *Proteins: Structure, Function, and Bioinformatics*, 60(3):333–340, 2005.

David Jakubec, Petr Škoda, Radoslav Krivák, Marian Novotný, and David Hoksza. Prankweb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures. *Nucleic Acids Research*, 50, 05 2022. doi: 10.1093/nar/gkac389.

John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583 – 589, 2021. URL https://api.semanticscholar.org/CorpusID:235959867.

Jeevan Kandel, Hilal Tayara, and Kil Chong. Puresnet: prediction of protein-ligand binding sites using deep residual neural network. *Journal of Cheminformatics*, 13, 09 2021. doi: 10.1186/s13321-021-00547-7.

Jeevan Kandel, Shrestha Palistha, Hilal Tayara, and Kil Chong. Puresnetv2.0: a deep learning model leveraging sparse representation for improved ligand binding site prediction. *Journal of Cheminformatics*, 16:66, 06 2024. doi: 10.1186/s13321-024-00865-6.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Esther Kellenberger, Pascal Muller, Claire Schalon, Guillaume Bret, Nicolas Foata, and Didier Rognan. sc-pdb: an annotated database of druggable binding sites from the protein data bank. *Journal of chemical information and modeling*, 46(2):717–727, 2006.

Radoslav Krivák and David Hoksza. P2rank: Knowledge-based ligand binding site prediction using aggregated local features. In *International Conference on Algorithms for Computational Biology*, 2015. URL https://api.semanticscholar.org/CorpusID:42896652.

Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10, 2018. URL https://api.semanticscholar.org/CorpusID:52004658.

Radoslav Krivák and David Hoksza. Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features. *Journal of cheminformatics*, 7:12, 04 2015. doi: 10.1186/s13321-015-0059-5.

Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10, 08 2018. doi: 10.1186/s13321-018-0285-8.

David G. Levitt and Leonard J. Banaszak. Pocket: A computer graphies method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of Molecular Graphics*, 10(4):229–234, 1992. ISSN 0263-7855. doi: https://doi.org/10.1016/0263-7855(92)80074-N. URL https://www.sciencedirect.com/science/article/pii/026378559280074N.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574. Earlier versions as preprint: bioRxiv 2022.07.20.500902.

Artur Meller, Michael Ward, Jonathan Borowsky, Jeffrey Lotthammer, Felipe Oviedo, Juan Lavista Ferres, and Gregory Bowman. Predicting locations of cryptic pockets from single protein structures using the pocketminer graph neural network. *Nature Communications*, 14, 03 2023. doi: 10.1038/s41467-023-36699-3.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Florian Sestak, Lisa Schneckenreiter, Johannes Brandstetter, Sepp Hochreiter, Andreas Mayr, and Günter Klambauer. Vn-egnn: E(3)-equivariant graph neural networks with virtual nodes enhance protein binding site identification, 2024. URL https://arxiv.org/abs/2404.07194.

Richard D. Smith, Jordan J. Clark, Aqeel Ahmed, Zachary J. Orban, James B. Dunbar, and Heather A. Carlson. Updates to binding moad (mother of all databases): Polypharmacology tools and their utility in drug repurposing. *Journal of Molecular Biology*, 431(13):2423–2433, 2019. ISSN 0022-2836. doi: https://doi.org/10.1016/j.jmb.2019.05.024. URL https://www.sciencedirect.com/science/article/pii/S0022283619302967. Computation Resources for Molecular Biology.

Zachary Smith, Michael Strobel, Bodhi Vani, and Pratyush Tiwary. Graph attention site prediction (grasp): Identifying druggable binding sites using graph neural networks with attention, 07 2023.

Benjamin Tingle, Khanh Tang, Jose Castanon, John Gutierrez, Munkhzul Khurelbaatar, Chinzorig Dandarchuluun, Yurii Moroz, and John Irwin. Zinc-22 - a free multi-billion-scale database of tangible compounds for ligand discovery. 10 2022. doi: 10.26434/chemrxiv-2022-82czl.

Javier S Utgés and Geoffrey J Barton. Comparative evaluation of methods for the prediction of protein–ligand binding sites. *Journal of Cheminformatics*, 16(1):126, 2024.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022.

Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The pdbbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *Journal of medicinal chemistry*, 47(12):2977–2980, 2004.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Baohua Zhang, Hui Li, Kunqian Yu, and Zhong Jin. Molecular docking-based computational platform for high-throughput virtual screening. *CCF Transactions on High Performance Computing*, pp. 1–12, 2022.

Yang Zhang, Zhewei Wei, Ye Yuan, Chongxuan Li, and Wenbing Huang. Equipocket: an e(3)-equivariant geometric graph neural network for ligand binding site prediction, 2024. URL https://arxiv.org/abs/2302.12177.

Zuobai Zhang, Chuanrui Wang, Minghao Xu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. A systematic study of joint representation learning on protein sequences and structures, 2023a. URL https://arxiv.org/abs/2303.06275.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining, 2023b. URL https://arxiv.org/abs/2203.06125.

A DATASETS

Training. The sc-PDB (Structural Chemogenomics Protein Data Bank) (Kellenberger et al., 2006; Desaphy et al., 2015) is a curated dataset of druggable protein-ligand complexes extracted from the PDB. Compared to datasets like PDBBind (Wang et al., 2004) and Binding MOAD (Hu et al., 2005; Smith et al., 2019), sc-PDB emphasizes druggable sites and ligand diversity, making it a valuable resource for computational drug design. The dataset used for training and validation is the 2017 release of sc-PDB database (Desaphy et al., 2015), which comprises 17,594 structures, 16,034 entries, 4,782 proteins, and 6,326 ligands. We used a subset of this dataset following PUResNet (Kandel et al., 2021) and EquiPocket (Zhang et al., 2024), where structures were clustered based on their Uniprot (Consortium, 2024) IDs, and protein structures with the longest sequences were selected from each cluster. The training split follows VN-EGNN but with two additional constraints: proteins longer than 1,022 residues were excluded due to ESM embedding limitations, and the dataset was restricted to single-chain proteins only. These preprocessing steps resulted in a final dataset of 3,520 protein chains for training and validation.

Testing. We utilized the LIGYSIS dataset (Utgés & Barton, 2024), a curated resource for evaluating protein–ligand binding site prediction methods. LIGYSIS comprises approximately 30,000 biologically relevant protein–ligand complexes, aggregating unique ligand-binding interfaces across multiple structures of the same protein. The dataset includes diverse ligand types—ions, peptides, nucleic acids, and small molecules—accounting for approximately 40% ion-binding sites, which are largely absent in other benchmarks. LIGYSIS aggregates data from multiple structures of the same protein, offering a comprehensive view of ligand-binding diversity.

For instance, human pancreatic alpha-amylase, represented by PDB entry 4GQQ in PDBbind, is expanded in LIGYSIS to include 13 unique binding sites derived from 51 structures. This approach significantly enhances dataset diversity, as indicated by a Shannon entropy of 8.8 in the ion-excluded subset ($\text{LIGYSIS}_{\text{NI}}$), surpassing all other benchmark datasets. Compared to $\text{sc-PDB}_{\text{FULL}}$, which limits each protein to the most relevant ligand, LIGYSIS provides a richer representation of binding site diversity.

Unlike prior datasets such as $\text{sc-PDB}_{\text{FULL}}$, $\text{bMOAD}_{\text{SUB}}$, CHEN11, $\text{PDBbind}_{\text{REF}}$, SC6K, HOLO4K, and COACH420, LIGYSIS considers biological units instead of crystallographic asymmetric units, ensuring that only functional macromolecular assemblies are represented. Redundant protein–ligand interfaces, often inflated because of symmetry in crystallographic data, were systematically removed by clustering ligand interaction sites based on their protein interaction fingerprints.

LIGYSIS diverges from HOLO4K and COACH420, which rely on asymmetric units, by avoiding redundancy through a focus on biologically meaningful assemblies, improving the reliability of benchmarking results. Additionally, it corrects overrepresentation in datasets such as SC6K by ensuring non-redundant multimeric interactions and ligand diversity.

The final LIGYSIS benchmark set, after removing chains with missing residue mappings to UniProt, comprises 2,775 protein chains. To accommodate the maximum input size for ESM-2 embeddings, we crop proteins longer than 1022 residues to the first 1022 residues. Residues that were cropped are excluded from the ground truth. This preprocessing step led to the exclusion of 7 pockets in total from the LIGYSIS dataset.

B RELATED WORK

B.1 GEOMETRICAL AND BIOPHYSICAL METHODS

Geometrical methods have been extensively developed to address this challenge. One prominent example is the fpocket algorithm (Guilloux et al., 2009), an open-source tool that employs Voronoi tessellation and alpha spheres to detect binding cavities. The key concept behind fpocket is the use of alpha spheres, which are defined as spheres that contact four atoms on their boundary without any internal atoms. These alpha spheres serve as proxies for identifying regions likely to accommodate ligands. The fpocket workflow begins with Voronoi tessellation to determine the alpha spheres from the protein's atomic coordinates. The identified alpha spheres are then filtered based on size and clustered using proximity-based methods to form potential binding pockets. Finally, the

detected pockets are scored and ranked using geometric and physicochemical descriptors, such as hydrophobicity and alpha sphere density, to prioritize the most promising binding sites.

Another widely used geometrical method is Ligsite (Hendlich et al., 1997), which builds upon the earlier POCKET (Levitt & Banaszak, 1992) algorithm to enhance accuracy and efficiency. Ligsite utilizes a cubic Cartesian grid to identify pockets by detecting solvent-accessible regions enclosed by protein atoms along the x, y, and z axes, as well as along the cubic diagonals. This approach reduces the dependency on protein orientation, making it more robust. The grid points are scored based on Protein-Solvent-Protein (PSP) events, where higher scores indicate deeper pockets. Ligsite allows users to adjust parameters like grid resolution and pocket size thresholds to refine the detection process. Its rigorous scanning method ensures precise identification of pocket shapes while filtering out irrelevant surface irregularities. Notably, Ligsite is computationally efficient, capable of processing medium-sized proteins in 5-10 seconds with a 0.5 Å grid resolution, making it suitable for large-scale analyses.

PocketFinder (An et al., 2005) is a computational method that identifies and classifies ligand-binding envelopes by leveraging a transformation of the Lennard-Jones potential derived from protein structures. Unlike traditional approaches, PocketFinder predicts binding envelopes rather than surface binding sites, without requiring prior knowledge of ligand identity. The method was tested on large datasets of liganded (5,616) and unliganded (11,510) structures, achieving 96.8% accuracy in identifying experimental binding sites with over 50% overlap in liganded structures and 95% in unliganded ones, demonstrating robustness against conformational changes. PocketFinder combines geometric and physicochemical principles, calculating a van der Waals potential map using a probe atom, smoothing it iteratively, and contouring the resulting map to define binding envelopes. These envelopes are filtered by size and sorted by volume, prioritizing significant binding sites. The method also introduces a hierarchical clustering of predicted envelopes into a ”pocketome”, enabling the analysis of binding site diversity across structural proteomes.

B.2 Learning-based methods

Computational methods have evolved from traditional geometric and biophysical approaches to more advanced machine learning-based techniques.

Classical methods. PRANK (Krivák & Hoksza, 2015) is a post-processing algorithm designed to enhance the ranking of predicted protein-ligand binding pockets by addressing the limitations of traditional heuristic methods, which often fail to distinguish true binding pockets from false positives. PRANK represents binding pockets as ”inner pocket points” sampled from the protein’s Connolly surface within 4 Å of pocket-defining atoms. Each point is assigned a feature vector combining residue-level properties, such as hydropathy, with atomic-level properties like partial charges and ligand-binding propensities, along with additional metrics such as protrusion index and H-bond information. A distance-weighted function aggregates these features into final point descriptors, which are then evaluated by a Random Forest classifier trained to predict ligandability. Points near known ligands serve as positives, while all others are negatives. Pockets are scored based on the cumulative squared probabilities of their points being ligandable, enabling PRANK to effectively differentiate true binding pockets from false positives. P2Rank (Krivák & Hoksza, 2018) builds upon a machine learning framework to predict ligand-binding sites by leveraging chemical and geometric features of protein solvent-accessible surfaces. Unlike template-based methods that rely on known protein-ligand complexes, P2Rank infers ligand-binding potential directly from local surface properties. This approach enables P2Rank to predict novel binding sites with high accuracy and speed, achieving state-of-the-art performance on benchmark datasets like COACH420 and HOLO4K. Its minimal preprocessing requirements, standalone nature, and ability to process proteins in seconds make it highly suitable for automated pipelines and large-scale applications.

Convolutional Neural Networks. PUResNet (Kandel et al., 2021; 2024) advances protein-ligand binding site prediction through its integration of 3D convolutional neural networks and residual learning. Tackling challenges such as data imbalance and structural redundancy, PUResNet employs a rigorous preprocessing pipeline that refines the sc-PDB dataset (Desaphy et al., 2015) by clustering proteins based on UniProt IDs and selecting representative sequences to ensure diversity and eliminate redundancy. Proteins are voxelized into 3D grids with each voxel encoding 18 atomic features, and a U-Net-inspired architecture processes these grids using 3D convolutions. Residual connections

throughout the encoder-decoder structure address vanishing gradient issues while preserving fine-grained spatial details. The model is trained using Dice loss, optimized for imbalanced datasets, and evaluated through k-fold cross-validation. As stated in Kandel et al. (2021; 2024), PUResNet demonstrates state-of-the-art performance in metrics such as Distance Center-Center (DCC) and Discretized Volume Overlap (DVO), excelling in predicting binding sites for unbound proteins.

Graph Neural Networks. VN-EGNN (Sestak et al., 2024) addresses key limitations of graph neural networks (GNNs) (Scarselli et al., 2008; Satorras et al., 2021; Xu et al., 2018) in ligand-binding site prediction, such as poor learning dynamics due to oversquashing and the absence of nodes dedicated to geometric representations of binding pockets. By introducing virtual nodes, VN-EGNN enhances information flow and captures complex spatial relationships through an extended heterogeneous message-passing scheme. Its E(3)-equivariance ensures predictions remain consistent under geometric transformations, critical for the irregular nature of protein structures. Unlike traditional methods that infer binding site centers from segmented regions, VN-EGNN directly predicts these centers, aligning virtual node coordinates to physical binding positions. This approach improves predictive accuracy, as demonstrated by its superior performance in benchmarks like COACH420 and HOLO4K. GrASP (Smith et al., 2023) transforms ligand-binding site prediction by leveraging graph attention networks (GATs) to dynamically learn atomic and residue-level features. Proteins are encoded as graphs where nodes represent atoms and edges capture spatial relationships. GrASP integrates multi-head attention, advanced regularization techniques, and rotationally invariant GNNs to address challenges like oversmoothing and enhance predictive precision. Trained on an expanded sc-PDB dataset, GrASP employs an encoder-processor-decoder architecture to predict ligandable atoms, clustering high-scoring atoms into discrete binding sites and ranking them with a refined metric adapted from P2Rank. GrASP consistently outperforms competing methods in precision and computational efficiency, excelling in metrics tailored for drug discovery.

Protein embeddings. IF-SitePred (Carbery et al., 2024) introduces a novel approach to protein-ligand binding site prediction by combining embeddings from the ESM-IF1 protein language model (Hsu et al., 2022) with LightGBM classifiers (Ke et al., 2017). Unlike traditional methods that rely on all-atom features, IF-SitePred focuses on backbone-derived geometric properties, making it robust to inaccuracies in predicted structures. The model classifies residues as binding or non-binding using 512-dimensional embeddings that capture local residue environments. Binding residues are mapped into three-dimensional space, where DBSCAN clustering identifies potential binding site centers ranked by binding-labeled point density. This backbone-focused approach ensures resilience to side-chain inaccuracies, which often hinder existing methods. Evaluation on paired PDB and AlphaFold2 (Varadi et al., 2022) structures demonstrates IF-SitePred's robust performance, achieving a 93% top-3 success rate (DCA $\leq$ 4A) and outperforming traditional tools even on low-accuracy structures (Carbery et al., 2024). Additionally, ensemble strategies leveraging molecular dynamics conformations further enhance accuracy, highlighting the model's capacity to exploit structural diversity and generalize across novel ligand-binding sites.

C PICKPOCKET

C.1 ESM-GEARNET: JOINT SEQUENCE-STRUCTURE LEARNING

We propose PickPocket, a joint sequence-structure protein binding site prediction model that leverages serial fusion of ESM-2 and GearNet (Zhang et al., 2023a). Serial fusion enables effective integration of evolutionarily conserved patterns from sequences with geometric structural relationships by using ESM-2's contextual representations to initialize GearNet's node features. This fusion strategy has been shown to outperform parallel and cross fusion approaches while maintaining architectural simplicity. By initializing GearNet with ESM-2's pretrained sequence knowledge and using a reduced learning rate for ESM-2 to preserve its pretrained representations during training, our model effectively combines the complementary strengths of both sequence-based language modeling and structure-based geometric learning for accurate pocket identification.

C.1.1 SEQUENCE REPRESENTATION WITH ESM-2

ESM-2 (Lin et al., 2023), a transformer-based protein language model, generates sequence embeddings for each residue in a protein sequence $\mathcal{R} = [r_1, r_2, \ldots, r_n]$. The sequence embedding process

begins by encoding each residue into an initial feature vector:

$$h_i^{(0)} = \text{Embedding}(r_i) \in \mathbb{R}^d,$$

where d is the embedding dimension. Through a stack of transformer layers, these embeddings are refined using multi-head self-attention:

$$\alpha_{ij}^{(l)} = \text{Softmax}_j \left(\frac{W_q h_i^{(l)\top} W_k h_j^{(l)}}{\sqrt{d}} \right)$$

$$h_i^{(l+0.5)} = h_i^{(l)} + \sum_{j=1}^{n} \alpha_{ij}^{(l)} W_v h_j^{(l)}$$

The final layer produces sequence embeddings $h^{(L)}$ that capture contextual and evolutionary information.

C.1.2 Structure Representation with GearNet

GearNet (Zhang et al., 2023b) processes the protein structure as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$, where nodes represent residues and edges define their relationships. We establish three types of edges: sequential edges connecting residues within distance of 3 in the primary sequence, spatial edges connecting residues within 10Å, and k-nearest neighbor edges (k=10). Following serial fusion, each node is initialized with its corresponding ESM-2 sequence embedding:

$$u_i^{(0)} = h_i^{(L)}$$

The structural information is integrated through 6 layers of message passing networks (hidden dimension 512):

$$u_i^{(l+1)} = u_i^{(l)} + \sigma \left(\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_r(i)} W_r u_j^{(l)} \right)$$

where $\mathcal{N}_r(i)$ represents neighbors of node i connected by edge type r. The network incorporates batch normalization and residual connections throughout these layers.

C.2 Binding Site Prediction Architecture

C.2.1 Model Initialization

ESM-2 is initialized with its pretrained weights from large-scale protein sequence data, while GearNet is initialized with the weights from residue type prediction (Zhang et al., 2023a). This pretraining approach allows both models to leverage their respective pretrained knowledge of sequence and structure.

C.2.2 Residue-Level Classification

The final residue representations are created by concatenating outputs from both ESM-2 and GearNet. These are fed into a two-layer multilayer-perceptron (MLP) (Haykin, 1994) classifier:

$$h_i = \text{GELU}(W_1 x_i + b_1)$$

$$y_i = W_2 h_i + b_2$$

where y_i represents the predicted binding probability for residue i.

C.2.3 Pocket Extraction

After obtaining residue-level predictions, we employ DBSCAN (Ester et al., 1996) clustering to identify cohesive binding pockets. Residues with binding probabilities exceeding 0.5 are considered for clustering, using parameters eps = 5Å (maximum distance between two residues to be considered neighbors) and min_samples = 3 (minimum number of residues required to form a cluster). This post-processing step helps identify contiguous regions that are likely to form functional binding sites.

C.3 TRAINING-TESTING STRATEGY

C.3.1 TRAINING PROCESS

The training process jointly optimizes all components of the model using different learning rates: a lower learning rate of 10^{-5} for the ESM-2 parameters, and 10^{-4} for both GearNet and MLP classifier parameters. The model is trained using a smooth F1 loss function:

$$\mathcal{L}_{\text{smooth-F1}} = 1 - \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN} + \epsilon}, \tag{1}$$

where TP, FN, and FP represent true positives, false negatives, and false positives respectively. We use the Adam optimizer with a batch size of 4. The smooth F1 loss helps balance the treatment of positive and negative cases, making it particularly suitable for imbalanced datasets.

D LIGYSIS RESULTS

PickPocket was evaluated against state-of-the-art binding site prediction methods using the LIGYSIS benchmark, which provides a comprehensive dataset of ligand-binding sites across diverse protein structures. For evaluating residue classification, specifically determining whether residues belong to a binding site, we used standard metrics such as F1 score and Matthews Correlation Coefficient (MCC). The results indicate that PickPocket achieves the highest F1 score (0.42) and maintains a competitive Matthews Correlation Coefficient (MCC) of 0.37, outperforming existing deep learning-based methods such as PUResNet, GrASP, and P2Rank_CONS, as well as classical geometric approaches like fpocket, PocketFinder, and Ligsite (Table 2). These findings highlight PickPocket's ability to balance precision and recall, making it a robust approach for identifying functionally relevant binding pockets.

To assess its effectiveness in binding site detection, PickPocket was evaluated using Top-N recall as a metric. Top-N recall measures the model's ability to identify true binding sites among its top-ranked predictions. A prediction is considered correct if the distance between the predicted pocket centroid and the observed binding site centroid (DCC) is ≤ 12 Å. When evaluating recall, we consider both the Top-N and Top-(N+2) ranked predictions, where N is the number of true binding sites in the protein. Following the approach used in LIGYSIS, we compute the recall as:

$$\text{Recall} = \frac{\text{\# of observed sites with a predicted site at DCC} \leq 12 \text{ Å}}{\text{\# of total observed sites}}$$

PickPocket achieves a Top-N recall of 48.9%, the highest among the benchmarked methods, outperforming state-of-the-art approaches such as P2Rank, GrASP, and DeepPocket. Additionally, it maintains a Top-N+2 recall of 53.3%, demonstrating consistent performance across different ranking cutoffs (Table 3). These results highlight PickPocket's superior ability to retrieve relevant binding sites compared to previous methods.

Compared to other methods, PickPocket benefits from a combination of protein language models and graph-based structural learning. The integration of ESM-2 embeddings allows it to incorporate evolutionary information, which is known to be relevant for binding site detection. At the same time, the use of GearNet enables it to capture spatial relationships between residues, complementing the sequence-based information. This dual approach helps improve both recall and precision by identifying functionally relevant binding sites while reducing the likelihood of detecting non-functional surface cavities.

Table 2: Performance comparison ordered by F1 score.

Method	F1	MCC
PickPocket	**0.42**	0.37
PUResNet	0.41	**0.39**
GrASP	0.39	0.34
P2Rank_CONS	0.36	0.30
P2Rank	0.31	0.26
PocketFinder	0.31	0.22
Ligsite	0.31	0.21
VN-EGNN	0.29	0.26
IF-SitePred	0.29	0.24
Surfnet	0.29	0.20
DeepPocket_SEG	0.27	0.21
fpocket	0.23	0.12

E CRYPTIC SITES PREDICTION

Identifying binding pockets that are not immediately apparent from static protein structures remains a significant challenge in drug discovery. Many functionally relevant sites, such as cryptic pockets that

Table 3: Recall performance ordered by Top-N recall.

Method	Top-N	Top-N+2
PickPocket	**48.9**	53.3
P2Rank_CONS	48.8	53.9
fpocket_PRANK	48.8	**60.4**
GrASP	48.0	49.9
P2Rank	46.7	51.9
DeepPocket_RESC	46.6	58.1
Ligsite_AA	44.9	49.0
VN-EGNN_NR	44.5	46.1
PocketFinder_AA	44.5	48.9
DeepPocket_SEG-NR	43.4	49.4
Surfnet_AA	43.3	47.4
PUResNet_PRANK	40.8	41.1
fpocket	38.8	46.5
IF-SitePred_RESC-NR	29.7	39.1

emerge upon ligand binding or allosteric sites that regulate protein activity from a distance, are often missed by traditional structure-based methods. These pockets may be transient, hidden in unbound structures, or only form under specific conformational states. Accurately detecting such sites requires a model capable of integrating sequence-derived evolutionary information, structural flexibility, and binding-relevant physicochemical features.

To evaluate PickPocket's ability to predict cryptic binding sites, we analyzed 24 apo structures from the PocketMiner test set and compared its predictions to the corresponding holo structures. Unlike PocketMiner, which is explicitly trained to detect cryptic pockets using molecular dynamics derived labels, PickPocket operates without direct supervision for cryptic site prediction.

Quantitative assessment of PickPocket's capability in identifying cryptic pockets was performed using Precision-Recall (PR) curves, comparing its performance against PocketMiner (Figures 2a and 2b). PickPocket consistently outperformed PocketMiner in both apo and holo structures, achieving significantly higher AUC scores (0.617 and 0.656) compared to PocketMiner's (0.438 and 0.539), respectively.

The PR curves highlight that PickPocket maintains superior precision across a broader recall range, particularly in the apo structures (Figure 2a), where cryptic pockets are inherently more challenging to detect due to the absence of ligand-induced conformational rearrangements. The improved AUC suggests that PickPocket effectively captures latent structural and evolutionary signals that correlate with cryptic site formation, even in unbound states.

Table 4: Total Inference Time Comparison in CPU

Length	PickPocket (s)	PocketMiner (s)
150 residues	0.34	1.41
350 residues	0.51	1.34
1000 residues	1.44	1.52

In the holo structures (Figure 2b), where ligand-induced conformational changes reveal binding pockets more explicitly, PickPocket continues to outperform PocketMiner. Its higher AUC (0.656) indicates that it is capable of accurately localizing binding pockets even in cases where the pocket has undergone a cryptic-to-open transition. The consistency in its performance across both apo and holo structures suggests that PickPocket generalizes well to cryptic pocket identification without requiring explicit cryptic pocket training.

Interestingly, PickPocket's predictions were particularly enriched in regions undergoing substantial conformational rearrangements upon ligand binding, aligning well with experimentally validated

cryptic pockets. This observation suggests that PickPocket effectively identifies pockets that are structurally predisposed to ligand binding, reinforcing its utility in cryptic pocket discovery.

This underscores the scalability and efficiency of PickPocket for proteome-wide cryptic site discovery, offering a valuable alternative to MD-based approaches while significantly reducing computational costs.

F Ablation Studies

Table 5: Performance Metrics for Different Ablation

Ablation	F1 Score	MCC	Top-N (%)	Top-N+2 (%)
Freeze ESM	0.35	0.32	37.5	40.3
Freeze GearNet	0.40	0.35	47.3	51.5
No Freezing	0.42	0.37	48.9	53.3

To assess the relative contributions of sequence-based and structure-based features in PickPocket, we conducted an ablation study by systematically freezing different components of our model during training. Specifically, we evaluated three conditions: (i) freezing the ESM embeddings while allowing GearNet and the classification head to update, (ii) freezing the GearNet encoder while updating ESM embeddings and the classifier, and (iii) training all components without freezing any parameters. The results of this experiment are summarized in Table 5.

When freezing ESM, the model achieved an F1 score of 0.35, an MCC of 0.32, and a Top-N recall of 37.5%, the lowest among all tested configurations. This decline in performance suggests that evolutionary embeddings from ESM provide essential residue-level contextual information that significantly enhances binding site prediction. Without updating these embeddings, the model primarily relies on structural features extracted by GearNet, which alone are insufficient for maximizing prediction accuracy. The particularly low Top-N and Top-N+2 recall further indicate that the model struggles to consistently rank correct binding sites among the top candidates when sequence-derived information is not actively incorporated.

Freezing GearNet while allowing ESM embeddings to update resulted in improved performance compared to freezing ESM. The F1 score increased to 0.40, and the MCC reached 0.35, while Top-N recall improved to 47.3%. This result suggests that while structural representation learning is beneficial, evolutionary embeddings alone capture a substantial portion of the predictive signal. The relatively minor performance drop compared to the fully trainable model highlights the robustness of ESM's pretrained features, which retain essential sequence-derived patterns even when the structural encoder remains static. However, the lower Top-N recall compared to the no-freezing condition implies that structural refinement via GearNet enhances the model's ability to prioritize true binding pockets.

Allowing both ESM and GearNet to update during training yielded the highest overall performance, with an F1 score of 0.42, an MCC of 0.37, and a Top-N recall of 48.9%. This configuration demonstrated the strongest ability to correctly rank binding sites and identify residue-level features indicative of ligand interaction. The improvements in Top-N and Top-N+2 recall indicate that integrating both evolutionary and structural information leads to more reliable predictions. These results confirm that PickPocket benefits from joint optimization of sequence and structure representations, where evolutionary embeddings guide feature extraction while geometric learning refines residue interactions.

The ablation study highlights the complementary nature of protein language models and geometric graph learning for binding site prediction. ESM embeddings provide deep contextual insights derived from large-scale evolutionary training, enabling the model to recognize conserved functional motifs that may not be immediately apparent from structural data alone. On the other hand, GearNet captures local residue-residue interactions, geometric constraints, and physicochemical properties, refining predictions based on structural context.

Our findings suggest that sequence features alone provide a strong baseline for binding site identification, as demonstrated by the relatively high performance of the freezing GearNet embeddings ablation.

However, the inclusion of structural learning improves recall and pocket ranking, demonstrating the added value of geometric deep learning in fine-tuning binding predictions. Conversely, removing trainable sequence embeddings weakens the model's ability to generalize, reinforcing the necessity of evolutionary signals in identifying functionally relevant sites.

G PAIRING BINDING SITES WITH PICKPOCKET EMBEDDINGS

G.1 DOCKING PROTOCOL

We used the Glide (Halgren et al., 2004) software to perform SP docking on the selected structures. In most cases, we positioned the grid by selecting the nearest point to the center of mass based on predictions from PickPocket. In certain instances, we included additional residues to define a more meaningful docking region for the compounds.

G.2 DIVERSITY DATASET

We selected compounds from the ZINC22 database Tingle et al. (2022) based on specific physicochemical properties and availability. The filtering criteria were:

- Molecular weight (MW) < 425 Da
- LogP < 5
- In stock as of 2022 (when the dataset was downloaded)

This initial filtering yielded approximately 9 million compounds. To obtain a representative subset, we performed stratified sampling based on molecular weight. From this sampled set, additional manual selection was conducted to ensure structural diversity.

G.3 SELECTED PAIRS AND EUCLIDEAN DISTANCES

To investigate the relationship between pocket embedding similarity and functional binding site overlap, we selected a diverse set of protein chain pairs and computed their Euclidean distances in the PickPocket embedding space. This analysis aimed to assess whether lower embedding distances correlate with greater ligand-binding site similarity, potentially enabling target expansion and ligand repurposing strategies.

We selected protein chains from distinct structural and functional categories to include a variety of ligand-binding domains. Pairs were chosen based on:

- Protein functional annotation from UniProt and PDB metadata.
- Presence of at least one PickPocket-predicted binding site with a confidence score above 0.5.
- Structural classification based on SCOP/CATH domains to ensure representation across different protein families.
- Diversity in molecular function, spanning hydrolases, monooxygenases, G-protein coupled receptors (GPCRs), and non-binding stabilizers (NB-stabilizers).

For each selected protein pair, we extracted the PickPocket-generated pocket embeddings and computed their Euclidean distances to quantify binding site similarity. The process involved the following steps:

1. **Pocket Representation**: For each protein chain, we identified the top-ranked predicted pocket based on PickPocket's binding probability scores.
2. **Feature Extraction**: The pockets were represented as 4352-dimensional embeddings derived from the final layer of the PickPocket model, averaging the representation.

3. **Distance Calculation**: The Euclidean distance between the pocket embeddings of two protein chains was computed as:

$$d(P_1, P_2) = \sqrt{\sum_{i=1}^{4352} (e_{P_1,i} - e_{P_2,i})^2} \tag{2}$$

where $e_{P_1,i}$ and $e_{P_2,i}$ represent the i-th feature in the respective pocket embeddings.

4. **Functional Comparison**: The functional annotations of the protein chains were retrieved from UniProt, and their biological roles were compared to examine potential ligand-binding overlap.

Table 6 presents the computed Euclidean distances for selected protein pairs alongside their functional annotations. The distance values range from 6.04 (highly similar monooxygenases) to 37.96 (divergent hydrolase-G-Protein pair). Low embedding distances generally corresponded to functionally related proteins, supporting the hypothesis that PickPocket embeddings capture meaningful binding site relationships. However, certain high-distance pairs (e.g., 7WCM_R – 6FW0_A) exhibited unexpected ligand compatibility.

Chain 1	Chain 2	Distance	Function 1	Function 2
6fw0_B	2z5y_A	6.04	monooxigenase	monooxigenase
8f07_A	8skl_A	14.82	hydrolase	hydrolase
6fw0_B	7wcm_N	37.07	monooxigenase	nb-stabilizer
8skl_A	7wcm_A	37.96	hydrolase	G-protein
7wcm_R	6fw0_A	13.73	gpcr	monooxigenase
6fw0_B	8f07_A	20.28	monooxigenase	hydrolase
6fw0_B	8skl_A	28.06	monooxigenase	hydrolase
2z5y_A	7wcm_A	31.78	monooxigenase	G-protein
6fw0_A	8skl_A	19.21	monooxigenase	hydrolase

Table 6: Chain Pairwise Scores and Functional Annotations

G.4 Results for 8SKL_A – 7WCM_A

After superimposing the structures, we found that the high distance was influenced by a poor selection of residues from PickPocket. While these residues were near the binding site, they were insufficient to provide the resolution needed for accurate representations. Additionally, in the docking distance plot, we observe that the compounds contributing the most to the low distance in scores are those with high (closer to zero) docking scores. This suggests that even though the pocket exists in these pairs, the ligands tend not to dock with enough meaningful contacts to obtain a better score.

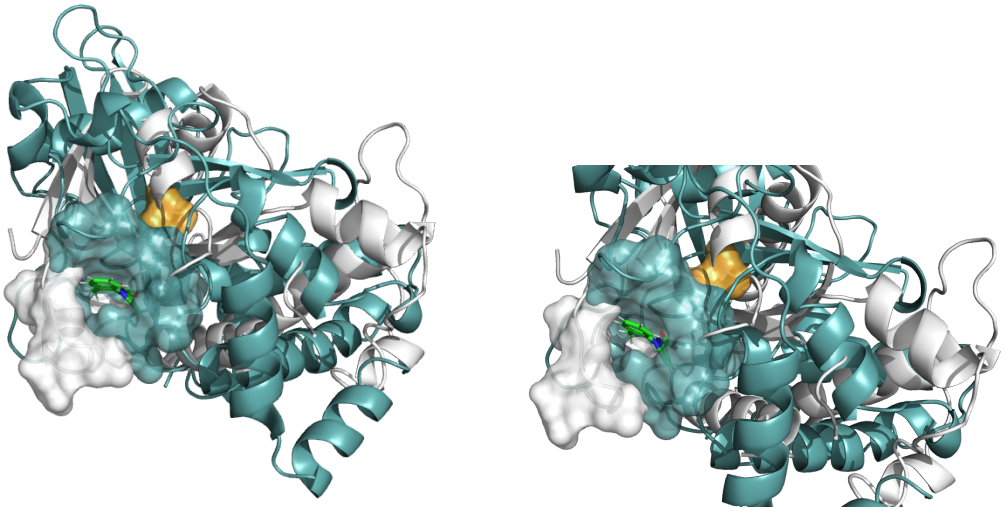
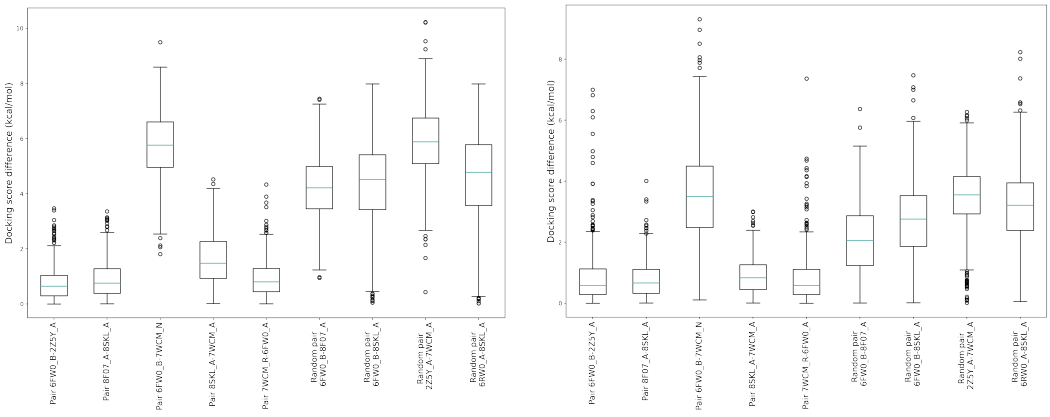


Figure 4: Pocket superposition of structures 8SKL_A (blue) and 7WCM_A (white). White surface indicates the pocket obtained from PickPocket prediction and in orange the one added to successfully perform docking.



(a) Distribution for the best 500 compounds in the pair.

(b) Distribution for the worst 500 compounds in the pair.

Figure 5: Distribution of docking score differences for the same compound across selected pairs. We evaluate this using the top 500 and bottom 500 compounds based on docking scores.