Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative evaluation of methods for the prediction of protein-ligand binding sites

Javier S Utgés
University of Dundee
Geoffrey John Barton
g.jbarton@dundee.ac.uk

University of Dundee

Research Article

Keywords: Ligand binding site prediction, binding pocket, benchmark, reference dataset, machine learning, drug discovery

Posted Date: August 8th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4849153/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Journal of Cheminformatics on November 11th, 2024. See the published version at https://doi.org/10.1186/s13321-024-00923-z.

Comparative evaluation of methods for the prediction of protein-ligand binding sites

Javier S. Utgés and Geoffrey J. Barton*

Division of Computational Biology, School of Life Sciences,
University of Dundee, Dow Street, Dundee, DD1 5EH, Scotland, UK

*Correspondence to: gjbarton@dundee.ac.uk

Abstract

The accurate identification of protein-ligand binding sites is of critical importance in understanding and modulating protein function. Accordingly, ligand binding site prediction has remained a research focus for over three decades with over 50 methods developed since the early 1990s. Over this time, the paradigm has changed from geometry-based to machine learning. In this work, we collate 11 ligand binding site predictors, spanning 30 years, focusing on the latest machine learning-based methods such as VN-EGNN, IF-SitePred, GrASP, PUResNet, and DeepPocket and compare them to the established P2Rank or fpocket and earlier methods like PocketFinder, Ligsite and Surfnet. We benchmark the methods against the human subset of the new curated reference dataset, LIGYSIS. LIGYSIS is a comprehensive protein-ligand complex dataset comprising 30,000 proteins with bound ligands which aggregates biologically relevant unique protein-ligand interfaces across biological units of multiple structures from the same protein. LIGYSIS is an improvement for testing methods over earlier datasets like sc-PDB, PDBbind, binding MOAD, COACH420 and HOLO4K which either include 1:1 protein-ligand complexes or consider asymmetric units. Re-scoring of fpocket predictions by DeepPocket and PRANK display the highest recall (60%) whilst VN-EGNN (46%) and IF-SitePred (39%) present the lowest recall. We demonstrate the detrimental effect that redundant prediction of binding sites has on performance as well as the beneficial impact of stronger pocket scoring schemes, with improvements up to 14% in recall (IF-SitePred) and 30% in precision (Surfnet). Methods predicting few pockets per protein, e.g., GrASP and PUResNet are very precise (>90%) but are limited in recall. Finally, we propose recall as the universal benchmark metric for ligand binding site prediction and urge authors to share not only the source code of their methods, but also of their benchmark.

Keywords

Ligand binding site prediction, binding pocket, benchmark, reference dataset, machine learning, drug discovery.

Introduction

Identifying where ligands can bind to proteins is of critical importance in understanding and modulating protein function. While X-ray crystallography remains the gold-standard

to identify and characterise binding sites [1-5], over the last three decades, significant effort has been made to develop computational methods that predict binding sites from an apo three-dimensional protein structure [6].

Prediction methods exploit a variety of different techniques to suggest binding sites. Geometry-based techniques like fpocket [7], Ligsite [8] and Surfnet [9] identify cavities by analysing the geometry of the molecular surface of a protein and usually rely on the use of a grid, gaps, spheres, or tessellation [7-14]. Energy-based methods such as PocketFinder [15] rely on the calculation of interaction energies between the protein and chemical group or probe to identify cavities [15-20]. Conservation-based methods make use of sequence evolutionary conservation information to find patterns in multiple sequence alignments and identify conserved key residues for ligand site identification [21-23]. Template-based methods rely on structural information from homologues and the assumption that structurally conserved proteins might bind ligands at a similar location [24-29]. Combined approaches or meta-predictors combine multiple methods, or the use of multiple types of data to infer ligand binding sites, e.g., geometric features with sequence conservation [30-36]. Finally, machine learning methods utilise a wide range of machine learning techniques including random forest, as well as deep, graph, residual, or convolutional neural networks [37-55]. Machine learning methods comprise the bulk of the methods reviewed in this analysis and are exemplified by P2Rank [37, 39], DeepPocket [47], PUResNet [43, 55], GrASP [52], IF-SitePred [53] and VN-EGNN [54].

Table 1 summarises the methods evaluated in this work. VN-EGNN [54] combines virtual nodes with equivariant graph neural networks. Virtual nodes, represented by ESM-2 embeddings [56] are passed through a series of message-passing layers until they reach their final coordinates, which represent the centroid of predicted pockets. Pocket residues are not reported. IF-SitePred [53] represents protein residues with ESM-IF1 embeddings [57] and employs 40 different LightGBM models [58] to classify residues as ligand-binding if all forty models return a $p > 0.5$. It later utilises PyMOL [59] to place a series of cloud points which are clustered using DBSCAN [60] and a threshold of 1.7Å. Pocket centroids are obtained by averaging the clustered points' coordinates, scored and ranked based on the number of cloud points. Like VN-EGNN, no pocket residues are defined. GrASP [52] employs graph attention networks to perform semantic segmentation on all surface protein atoms, represented by 17 atom, residue and bond-

level features, scoring which are likely part of a binding site. Atoms with a score > 0.3 are clustered into binding sites using average linkage and a threshold of 15Å. Pocket scores are calculated as the sum of squares of binding site atom scores. PUResNet [43] combines deep residual and convolutional neural networks to predict ligand binding sites using a 18-element vector of atom-level features and one-hot encoding to represent grid voxels. Voxels with a score > 0.34 are clustered into binding sites using DBSCAN and a threshold of 5.5Å [55]. Pockets are represented by their residues, but neither pocket centroid, nor score or ranking are reported. Similarly to PUResNet, DeepPocket [47] exploits convolutional neural networks on grid voxels represented by 14 atom-level features to re-score and additionally extract new pocket shapes from fpocket candidates. P2Rank [39] relies on solvent accessible surface (SAS) points placed over the protein surface, represented by 35 atom and residue-level features, and a random forest classifier to score them based on their likelihood of binding a ligand. SAS points with a score > 0.35 are clustered into sites using single linkage and a threshold of 3Å. $P2Rank_{CONS}$ [61] works in the same manner but considers an extra feature: amino acid conservation as measured by Jensen-Shannon divergence [62]. Both report residue and pocket level scores, as well as pocket centroids and rank. PocketFinder [15] uses the Lennard-Jones [63] transformation on a 1Å grid surrounding the protein surface to predict protein cavities. PocketFinder does not report pocket centroid, score or rank. Finally, geometry-based methods: fpocket [7], Ligsite [8] and Surfnet [9] rely on the geometry of the molecular surface to find cavities. fpocket is the only one of these three methods that reports pocket centroid, score, rank and residues.

In this work, we have gathered eleven ligand binding site prediction tools, spanning three decades of research and compared them against our reference dataset LIGYSIS. LIGYSIS extends the work in Utgés *et al.*, 2024 [5] that identifies human protein-ligand binding sites for biologically relevant ligands, defined by BioLiP [64], from protein structures determined by X-ray crystallography. We compared the 11 methods to each other and to the LIGYSIS reference dataset according to a range of metrics including the number of ligand sites, their size, shape, proximity, overlap and redundancy. We have explored in detail and demonstrated the importance of a robust pocket scoring scheme and highlight how some of the methods can improve their performance significantly by re-scoring their predictions. The work here identifies the strengths and weaknesses of

prediction assessment metrics and leads to guidance for developing ligand binding site prediction tools or using these methods to understand protein function and in drug development.

Method	Approach	Features	# Features	P centroid	P residues	P score	P ranking	R score	R score threshold	Cluster	Algorithm	Threshold
VN-EGNN	EGNN + VN	ESM-2 embeddings	1,280	✓	×	✓	✓	×	-	-	-	-
IF-SitePred	LightGBM	ESM-IF1 embeddings	512	✓	×	✓	✓	×	0.5 (ALL 40)	cloud points	DBSCAN	1.7
GrASP	GAT - GNN	Atom, residue, bond…	17	✓	×	✓	✓	✓	0.3	atoms	Average	15
PUResNet	DRN + 3D - CNN	Atom + one-hot encoding	18	×	✓	×	×	×	0.34	atoms	DBSCAN	5.5
DeepPocket	3D-CNN	Atom	14	✓	✓	✓	✓	×	-	-	-	-
P2Rank$_{CONS}$	Random Forest	Atom and residue	36	✓	✓	✓	✓	✓	0.35	SAS points	Single	3
P2Rank	Random Forest	Atom and residue	35	✓	✓	✓	✓	✓	0.35	SAS points	Single	3
fpocket	α-spheres	-	-	×	✓	✓	✓	×	-	α-spheres	Multiple	?
PocketFinder	LJ potential	-	-	×	×	×	×	✓	-	-	-	-
Ligsite	Cubic grid	-	-	×	×	×	×	✓	-	-	-	-
Surfnet	Gap regions	-	-	×	×	×	×	✓	-	-	-	-

Table 1. Summary of ligand binding site prediction methods. Approach: the techniques applied by the method; Features/#Features: the Features and their number if the method is machine learning-based; P centroid/P residues/P score/P ranking/R score: whether the method reports the pocket centroid, pocket residues, pocket score, pocket ranking and residue *ligandability* score. Information about their clustering strategies is also relevant: whether the method uses a residue ligandability threshold (R score threshold), the instances they cluster (Cluster) to define the distinct pockets, the clustering algorithm used (Algorithm) and threshold employed (Threshold). For example, P2Rank uses a random forest classifier on SAS points represented by 35 atom and residue features. Points with a score > 0.35 are later clustered into binding sites using single linkage and a threshold of 3Å. EGNN + VN: equivariant graph neural network + virtual nodes; LightGBM: Light gradient boosting machine; GAT: graph attention network; GNN: graph neural network; DRN: deep residual network; 3D-CNN: three-dimensional convolutional neural network; LJ potential: Lennard-Jones potential.

Results

The LIGYSIS Dataset

The ligand binding site analysis, or LIGYSIS, dataset comprises protein-ligand complexes for 3,448 human proteins. For each protein, biologically relevant protein-ligand interactions, in accordance with BioLiP [64], are considered across the PISA-defined [65] biological assemblies of the multiple entries deposited in the PDBe [66]. Ligands are clustered using their protein interaction fingerprint to identify ligand binding sites as described in Utgés *et al.*, 2024 [5, 67]. The full LIGYSIS dataset includes ≈30,000 proteins

with known ligand-bound complexes. Here, we focus on the human subset of LIGYSIS as a manageable set to run all prediction methods on and refer to this as LIGYSIS for brevity.

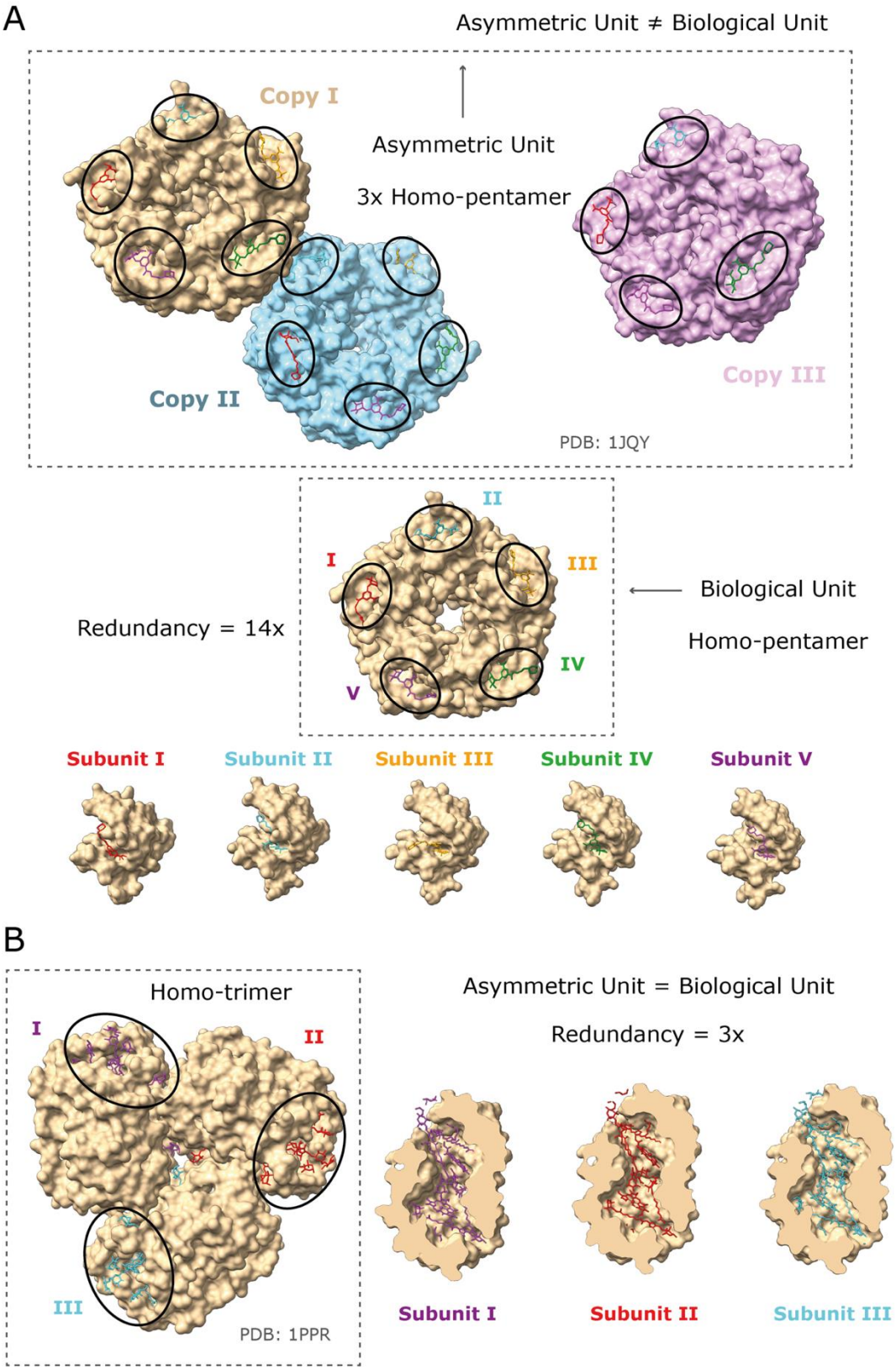


Figure 1. Redundancy in protein-ligand interfaces. Two examples of how the type of macromolecular assembly and difference between the asymmetric and biological unit of a protein-ligand complex results in redundant protein-ligand interfaces. (A) For PDB: 1QJY [68], the asymmetric unit comprises three copies

of a homo-pentamer, whereas the biologically functional assembly is a single pentamer. A BMSC-0010 ligand molecule binds to each copy, except for one, of each of the three pentamers. This results in the same protein-ligand interface repeated 14 times, i.e., 14x redundancy; (B) For PDB: 1PPR [69] both the asymmetric and biological units are a homo-trimer. Different molecules of the same ligands are binding to the same interfaces across the three copies of the trimer, i.e., 3x redundancy. Dashed rectangles indicate the asymmetric/biological units.

The LIGYSIS dataset differs from previous train and test sets for ligand binding sites by considering biological units, aggregating multiple structures of the same protein, and removing redundant protein-ligand interfaces. Biological units, or assemblies, as opposed to asymmetric units are considered in LIGYSIS in a consistent manner. This is important for any analysis that delves into molecular interactions at the residue or atomistic level. The crystal asymmetric unit often does not correspond to the biological unit and can result in artificial crystal contacts or redundant protein-ligand interfaces. An example of this illustrated in Figure 1A is PDB: 1JQY [68], present in the HOLO4K dataset, where the asymmetric unit is formed by three copies of a homo-pentamer, whereas the biological unit comprises a single pentamer. In this structure, 14 molecules of BMSC-0010 (A32) interact with 14 copies of *Escherichia coli* heat-labile enterotoxin B chain (P32890). This protein-ligand interface is the same repeated 14 times (Figure 1A). Protein-ligand interface redundancy can also be an issue when the asymmetric unit equals the biological assembly (Figure 1B). In PDB: 1PPR [69], also in HOLO4K, molecules of chlorophyll A (CLA), peridinin (PID) and digalactosyl diacyl glycerol (DGD) bind to the three copies of a peridinin-chlorophyll a-binding protein 1, chloroplastic, PCP, (P80484) trimer, resulting in a redundancy of 3x. To account for this, LIGYSIS considers unique non-redundant protein-ligand interfaces by retrieving the UniProt residue numbers of the residues the ligands interact with, so 1/14 interfaces would be retrieved for PDB: 1JQY and 12/36 for PDB: 1PPR. Finally, unique ligand interactions are aggregated across different structures for the same protein and ligand sites defined.

Figure 2 shows the comparison between PDB: 4GQQ [70], present in the PDBbind dataset, and the LIGYSIS entry for human pancreatic alpha-amylase (P04746), which representative structure is also 4GQQ. The entry in PDBbind represents a single protein-ligand complex, whereas LIGYSIS makes use of 51 structures, 195 ligands to define 13 different ligand binding sites. LIGYSIS aggregates all unique biologically relevant protein-ligand interactions for a protein in a non-redundant manner, thus representing the most complete and integrative protein-ligand binding dataset up to date. For this reason, we

propose LIGYSIS as a new benchmark dataset for the prediction of ligand binding sites and use it in this work to evaluate a set of 11 ligand binding site prediction and cavity identification tools.

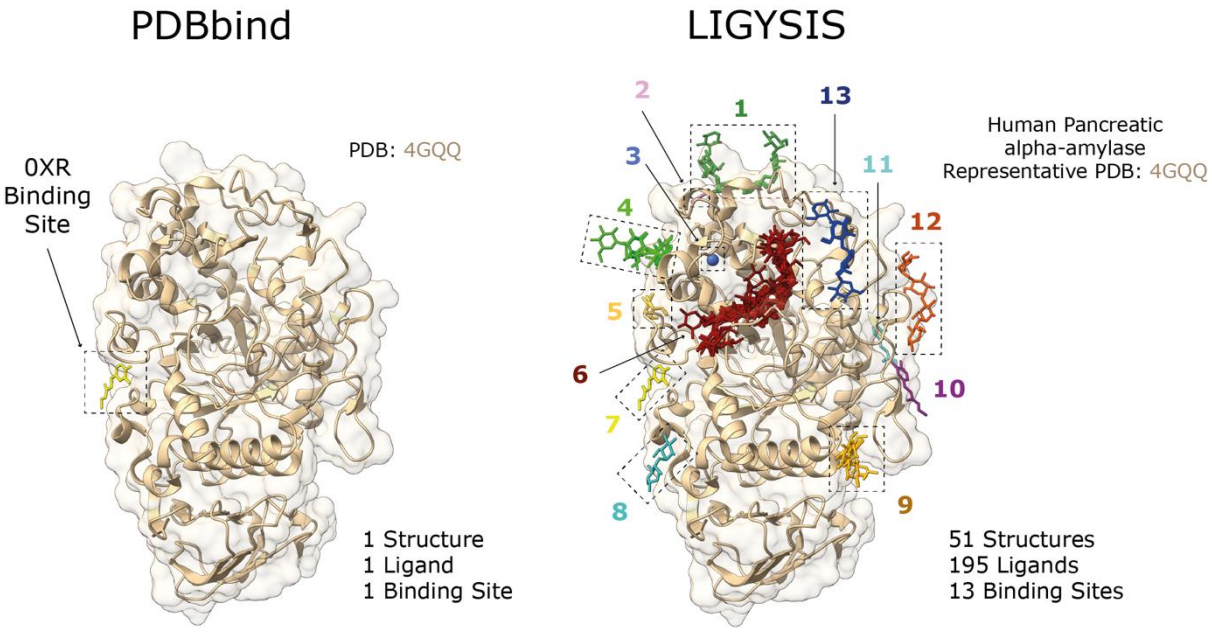


Figure 2. Comparison of PDBbind and LIGYSIS. PDBbind is comprised by complexes between a protein and the most biologically relevant ligand in a structure. For PDB: 4GQQ [70], this is ethyl caffeate (0XR). LIGYSIS considers all unique biologically relevant protein-ligand interactions across all the structures for a given protein. For human pancreatic alpha-amylase (P04746), which representative structure is 4GQQ, 13 ligand binding sites are defined from 195 ligands across 51 structures. LIGYSIS provides a better representation of the ligand-binding capabilities of a protein than a single protein-ligand complex and constitutes therefore a better benchmark for ligand binding site prediction tools.

Comparison of Datasets

To assess the scope and limitations of the predictive methods surveyed in this work, their training and test sets were compared with LIGYSIS by number of sites per protein, number of interacting protein chains per ligand site, ligand size, ligand site size, and ligand composition. sc-PDB$_{FULL}$ represents the full sc-PDB dataset used for training by DeepPocket, bMOAD$_{SUB}$ the subset of binding MOAD used for training by IF-SitePred and PDBbind$_{REF}$ the reference subset of PDBbind which VN-EGNN uses for testing. Table 2 summarises the size of the datasets, which methods employ them and their overlap with the LIGYSIS set. LIGYSIS differs from all other datasets since biologically relevant ions are considered, comprising ≈40% of the ligand sites. For additional reference, LIGYSIS$_{NI}$, a subset of LIGYSIS without ions, is also included in this analysis. Figure 3A shows the number of binding sites per entry across datasets. sc-PDB$_{FULL}$, PDBbind$_{REF}$ and SC6K only consider the most relevant ligand for each entry. COACH420 mostly presents single-

ligand entries (72%). $bMOAD_{SUB}$ (46%) and CHEN11 (58%) present more similar distributions to LIGYSIS, where 54% of the protein chains present more than one binding site. This percentage decreases for $LIGYSIS_{NI}$ (38%) as ion sites are removed. HOLO4K presents the highest proportion (62%) of multi-ligand entries. Both HOLO4K and COACH420 are based on asymmetric units, and not biological assemblies. For HOLO4K, 1,811 (40%) of structures present different numbers of chains between the asymmetric and biological units. This is even more frequent in COACH420: 234 (56%). Moreover, multimeric complexes might present the same protein-ligand interface repeated across the copies of the complex (Figure 1). Considering predictions of these interfaces as independent can lead to an overestimate of the performance of a predictor. Regarding the number of chains interacting with a given ligand (Figure 3B), CHEN11 and COACH420 present the smaller fraction of multimeric protein-ligand interactions: 3% and 6%, respectively, whereas SC6K presents the highest (44%). The rest of the methods range between 20-30%. There are no striking differences regarding the size of the interacting protein chains, represented by the number of residues (Figure 3C).

Dataset	Type	# Structures	# Sites	# Ligands	Overlap	Methods
LIGYSIS	NEW	3,448	8,244	**65,116**	-	-
$LIGYSIS_{NI}$	NEW	2,275	4,572	38,595	-	-
$sc\text{-}PDB_{FULL}$	TRAIN	17,594	17,594	17,594	9.7%	VN-EGNN, GrASP, PUResNet, DeepPocket
$bMOAD_{SUB}$	TRAIN	5,899	11,184	11,184	7.6%	IF-SitePred
CHEN11	TRAIN	244	479	479	0.5%	P2Rank
$PDBbind_{REF}$	TEST	5,316	5,316	5,316	3.8%	VN-EGNN
SC6K	TEST	6,147	6,147	6,147	3.1%	DeepPocket
HOLO4K	TEST	4,536	11,528	11,528	2.8%	ALL*
COACH420	TEST	413	624	624	0.5%	VN-EGNN, GrASP, DeepPocket, P2Rank, PUResNet

Table 2. Summary statistics of the different datasets analysed in this study. LIGYSIS is our reference dataset, $LIGYSIS_{NI}$ is a subset with no ion (NI) ligand binding sites, $sc\text{-}PDB_{FULL}$, $bMOAD_{SUB}$ and CHEN11 constitute the training datasets, whereas $PDBbind_{REF}$, SC6K, HOLO4K and COACH420 represent test sets. # Structures, # Sites and # Ligands represent the number of PDB structures, ligand sites and total number of ligands for each dataset. Note that for LIGYSIS and $LIGYSIS_{NI}$, 3,448 and 2,775, are the number of human structural segments considered, each represented by a single chain. For each segment, all biologically relevant ligand-binding structures were considered: N = 23,321 (LIGYSIS) and N = 19,012 ($LIGYSIS_{NI}$). The number of ligands is not equal to the number of sites for LIGYSIS, as data from multiple structures of the same protein are aggregated into unique sites, i.e., a LIGYSIS site often includes multiple ligands. Overlap is the percentage of protein-ligand complexes overlapping between these datasets and LIGYSIS. Methods

represents the ligand site predictors that use these datasets for training or test. Only the original version of each dataset is considered in the analysis, e.g., HOLO4K is analysed, but not HOLO4K$_{Mlig}$, nor HOLO4K$_{Mlig+}$ HAP, or HAP-small. The same goes for Mlig, Mlig+ versions of COACH420, sc-PDB$_{SUB}$ and sc-PDB$_{RICH}$. *ALL** represents all the methods compared in this work except for fpocket, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+}.

Figure 3D represents the ligand type composition of the datasets. Non-polymer ligands dominate all datasets (>66%), and the proportion of peptides and nucleic acids differ across datasets, with LIGYSIS and LIGYSIS$_{NI}$ presenting fewer ligands of these types (1.6% and 1%). sc-PDB$_{FULL}$ and SC6K are depleted in saccharides (<1%). Figure 3E depicts the difference in the number of atoms of the ligands in each dataset. LIGYSIS is of course different, due to its ion ligand content, however, there is no difference between LIGYSIS$_{NI}$ and the other datasets. Figure 3F conveys the difference in the number of ligand-interacting residues. LIGYSIS has the largest proportion of small sites, 1-10 residues, (56%). This is directly related to the prominent ion component, and the frequency decreases when ions are removed (LIGYSIS$_{NI}$: 36%). CHEN11, COACH420 and HOLO4K are more similar to LIGYSIS$_{NI}$, whereas sc-PDB$_{FULL}$, PDBbind$_{REF}$ and SC6K are clearly different and present almost exclusively large sites, larger than 20 amino acids, (>90%). Figure 3G explores the ligand diversity on each dataset by showing the top-5 most frequent ligands per dataset, the proportion of the total number of ligands they represent, as well as Shannon's entropy H'. Shannon's entropy is a measure of diversity. Larger numbers indicate a more evenly spread distribution of a larger number of different molecules, whereas small numbers indicate higher frequency of a few ligands. Whilst four out of the top-5 ligands of LIGYSIS are ions (Zn^{+2}, Ca^{+2}, Mg^{+2}, Mn^{+2}) and represent 19.2% of all ligands, its diverse composition is comparable to that of PDBbind$_{REF}$. Removing ions, LIGYSIS$_{NI}$ becomes the most diverse dataset with H' = 8.8 and its top-5 ligands only covering 5.3% of all ligands in the set. SC6K is the least diverse with its top-5 most frequent ligands covering 33% of all ligands. All datasets, except for LIGYSIS, LIGYSIS$_{NI}$, and PDBbind$_{REF}$, are dominated by co-factor ligands, such as flavin-adenine dinucleotide (FAD), nicotinamide-adenine dinucleotide (NAD), and haem (HEM) or energy carrier molecules such as adenine tri-, di- and monophosphate (ATP, ADP, AMP). Short peptides (<10 aas) are the most common ligands in PDBbind$_{REF}$ (5%), and energy carriers represent <2% of the top-5 ligands. Cholesterol, mannose and fucose are some of the most common ligands in LIGYSIS$_{NI}$. For all further analysis, including

characterisation of binding pockets and performance evaluation of the prediction methods, LIGYSIS, i.e., including ions, was utilised.

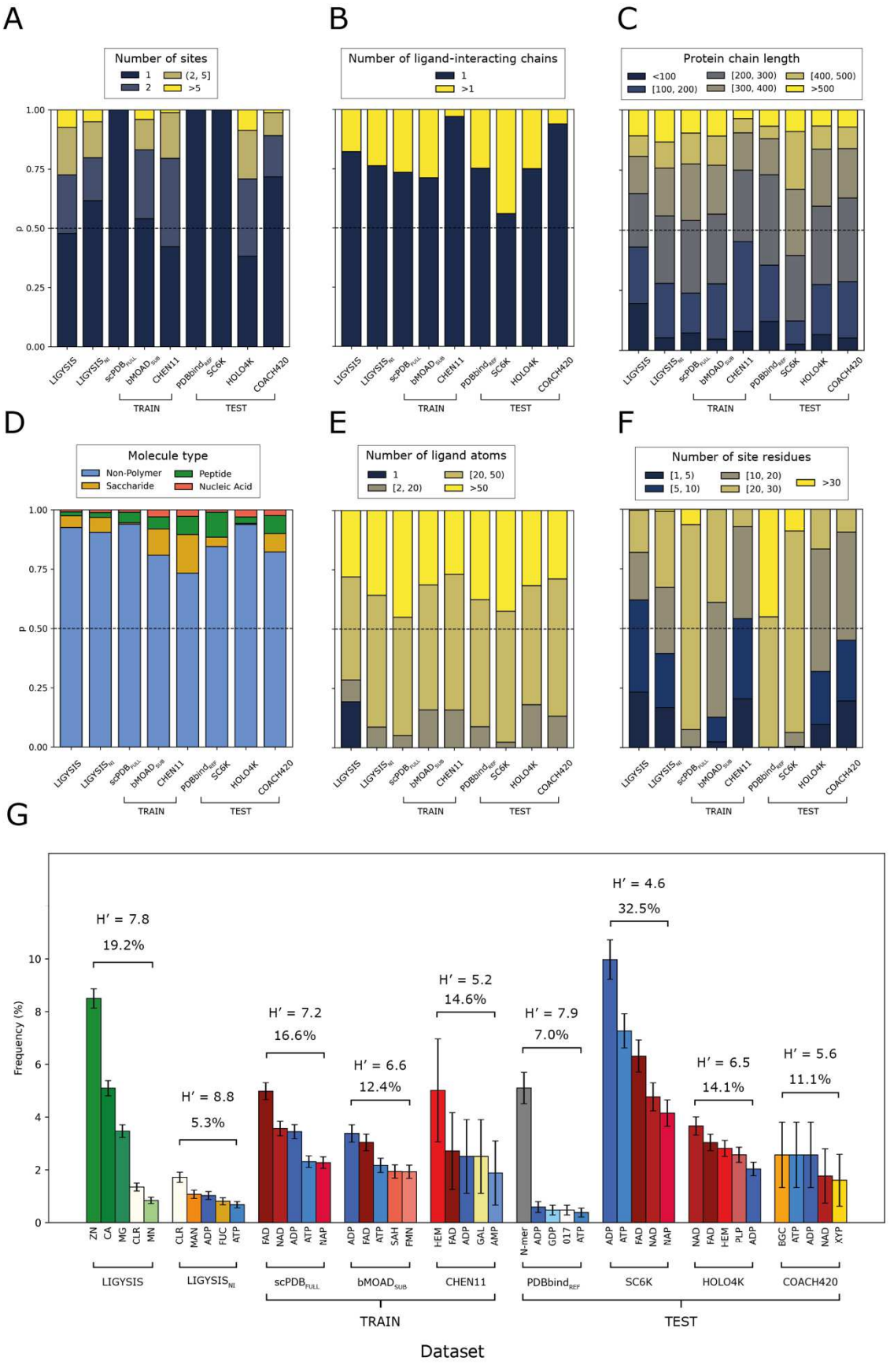


Figure 3. Comparison of datasets. Panels A, B, C, E and F plot the frequencies of binned intervals of a discrete variable, coloured using the *cividis* palette. Interval ranges were selected to facilitate data interpretation. (A) Number of ligand binding sites per dataset entry; (B) Number of ligand-interacting protein chains. This represents whether the ligand interacts with a single protein chain or more; (C) Length of ligand-interacting protein chains (number of amino acids); (D) Ligand molecule type frequency as described in the CCD; (E) Number of ligand atoms; (F) Binding site size, i.e., number of ligand-interacting protein residues. Dashed lines drawn at p = 0.5; (G) Five most frequent ligands per dataset. Error bars represent 95% confidence interval of the proportion [71]. Ligands of similar type are coloured in shades of the same colour: greens for ions, reds for co-factors, blues for energy-carrier molecules, yellows for sugars, grey for peptides and white for other non-polymeric ligands. Above the bars, Shannon's Entropy and the proportion of all ligands in each set covered by these top-5 can be found. Both are measures of ligand diversity within each dataset. A subset of LIGYSIS with no ions (NI), $\text{LIGYSIS}_{\text{NI}}$, is included in this analysis, as most training and test datasets do not consider ions. CLR: cholesterol; MAN: mannose; FUC: fucose; NAP: nicotinamide-adenine-dinucleotide phosphate; SAH: S-Adenosyl-L-homocysteine; FMN: Flavin mononucleotide; GAL: galactose; N-mer: protein peptides of *N* amino acids; 017: Darunavir; PLP: Vitamin B6 phosphate; BGC: glucose; XYP: xylose.

Characterisation of Binding Pockets

After removing backbone-only chains and those with missing residue mapping to UniProt, the final LIGYSIS set which was employed for the benchmark of the methods comprises 2,775 representative chains. Not all methods predict pockets on all the chains. VN-EGNN, GrASP, fpocket, PocketFinder^{+}, Ligsite^{+}, and Surfnet^{+} predict in >99% of the chains, $\text{P2Rank}_{\text{CONS}}$ on 93%, followed by P2Rank on 86%, PUResNet and $\text{DeepPocket}_{\text{SEG}}$ (85%), and finally IF-SitePred only predicts pockets on 75% of the chains. PUResNet, DeepPocket, $\text{P2Rank}_{\text{CONS}}$ and P2Rank often don't predict on smaller proteins (<100 amino acids) as well as non-globular or elongated proteins, representing 60-80% of proteins with no predicted pockets. However, for IF-SitePred larger globular proteins represent ≈50% of all proteins where this method fails to predict a pocket (See Supplementary Figure 1). Predicted residue ligandability scores for $\text{P2Rank}_{\text{CONS}}$, P2Rank and IF-SitePred (which we derived in this work), were examined for the proteins with no predicted pockets. Figure 4 illustrates 8 examples of proteins where residues with IF-SitePred (Eq. 10) high ligandability scores cluster in space into clear binding sites that are not reported as predictions by this method. This suggests that IF-SitePred is too strict in selecting only those residues predicted as ligand-binding by all 40 models. The cloud

point selection clustering approach or threshold in this method may also play a role in this.

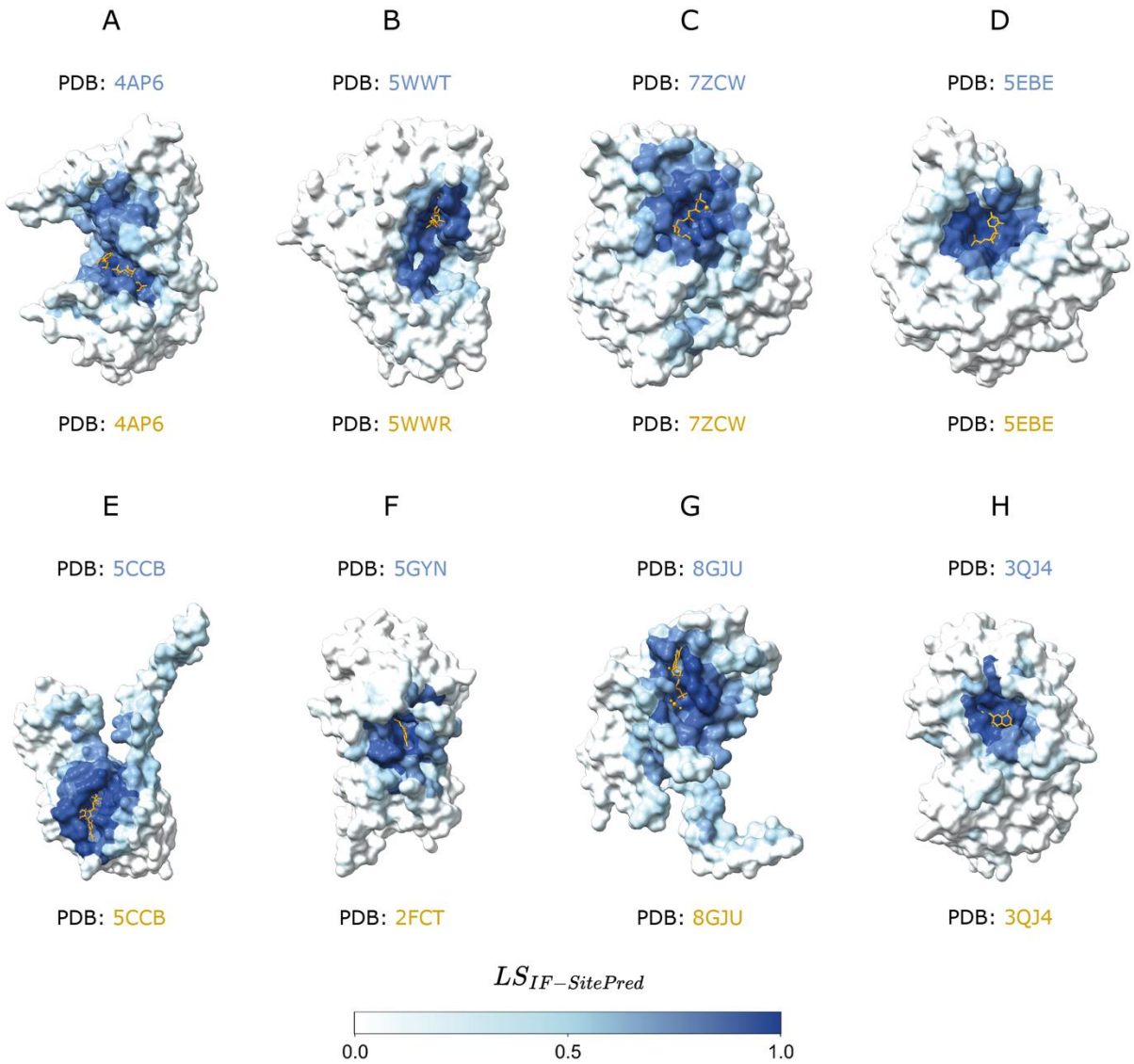


Figure 4. IF-SitePred "missed" predictions. Eight examples of human protein chains where IF-SitePred does not report any predicted ligand binding sites. Predictions are made on ligand-stripped chains. Ligand molecules, in orange, are superposed to illustrate how the ligandability scores recapitulate the observed binding site. These are protein representative chains and ligand molecules might not be observed in the same entry; (A) GDP-fucose protein O-fucosyltransferase 2, Q9Y2G5, with GFB superimposed (PDB: 4AP6) [72]; (B) tRNA (cytosine(72)-C(5))-methyltransferase NSUN6, Q8TEA1, (PDB: 5WWT) [73] with superposed SFG (PDB: 5WWR); (C) Tubulin beta-2B chain, Q9BVA1, with G2P (PDB: 7ZCW) [74]; (D) Cyclic GMP-AMP phosphodiesterase SMPDL3A, Q92484, with C5P (PDB: 5EBE) [75]; (E) tRNA (adenine(58)-N(1))-methyltransferase catalytic subunit, Q96FX7, with SAH (PDB: 5CCB) [76]; (F) Chronophin, Q96GD0, (PDB: 5GYN) (Lee HJ, Cho HJ, Kang BS, 2017) with PLP (PDB: 2FCT) [77] (G) Mitochondrial Methylmalonic aciduria type A protein, Q8IVH4, with GDP (PDB: 8GJU) [78]; (H) Renalase, Q5VYX0, (PDB: 3QJ4) with FAD [79]. Residues are coloured based on the ligandability score calculated by averaging the probabilities predicted by each of the 40 IF-SitePred prediction models. This is a score ranging 0-1 which is indicative of the likelihood of a given residue binding a ligand. Clear pockets can be observed formed by residues with high ligandability scores (darker blue colour), which agree with the sites where ligands bind. These "missed" pockets contribute to the lower recall of IF-SitePred and strongly suggest that a more permissive ligand-

binding residue selection threshold (currently all 40 models) or a different clustering scheme might be able to capture these predictions and increase the recall of this method.

Table 3 summarises the ligand site characterisation analysis. fpocket predicts the most sites out of all the methods, with 57,859, followed by IF-SitePred (44,948), DeepPocket$_{SEG}$ (21,718), VN-EGNN (13,582), P2Rank (12,412), P2Rank$_{CONS}$ (10,180), Surfnet^{+} (9,043), PocketFinder^{+} (8,913), Ligsite^{+} (6,903), GrASP (4,694) and PUResNet, which predicts fewest sites (2,621). LIGYSIS defines 6,882 binding sites from experimental data. Relative to LIGYSIS, the prediction methods have ratios of predicted/defined sites ranging from 8.4 (fpocket) to 0.4 (PUResNet) with P2Rank$_{CONS}$ in the middle, predicting 1.5 pockets per observed reference site. IF-SitePred, DeepPocket$_{SEG}$ as well as P2Rank and fpocket predict more pockets on larger protein chains, whereas the rest of methods do not (See Supplementary Note 1 and Figure 2). This effect is most clear with fpocket, which predicts 350 pockets for chain A of PDB: 7SUD [80], a structure of the DNA-dependent protein kinase catalytic subunit, DNPK1, (P78527) with 3,736 amino acid residues. In contrast, VN-EGNN, which initially places K = 8 virtual nodes, results in a maximum of 8 predicted pockets, regardless of protein chain size, and PUResNet predicts a single pocket in 90% of the protein chains.

Method	Coverage	Total Pockets	Pockets per protein	R_g	MCD	MRO	Redundancy
LIGYSIS	2,775	6,882	1, 1, 27	5.9	14.1	0	2.3%
VN-EGNN	2,764 (99.6%)	13,582 (×2.0)	1, 5, 7	5.9	**1.1**	**0.85**	**66.8%**
IF-SitePred	2,075 (**74.8%**)	44,948 (×6.5)	1, **20**, 129	5.9	3.4	0.55	49.5%
GrASP	2,771 (99.9%)	4,694 (×0.7)	1, 1, 12	7.9	21.4	0	0.0%
PUResNet	2,360 (85.1%)	2,621 (×0.4)	1, 1, 4	8.1	27.0	0	0.0%
DeepPocket$_{SEG}$	2,349 (84.7%)	21,718 (×3.2)	1, 6, 196	7.7	4.6	0.4	31.1%
P2Rank$_{CONS}$	2,759 (92.9%)	12,412(×1.8)	1, 3, 57	7.1	13.9	0.05	0.7%
P2Rank	2,402 (86.6%)	10,180 (×1.5)	1, 3, 85	7.1	13.8	0.05	0.6%
fpocket	2,759 (99.4%)	57,859 (×**8.3**)	1, 17, **349**	6.3	9.7	0.15	0.7%
PocketFinder^{+}	2775 (100%)	8,913 (×1.3)	1, 3, 23	8.6	18.7	0.05	0.0%
Ligsite^{+}	2775 (100%)	6,903 (×1.0)	1, 2, 12	**9.1**	16.7	0.09	0.0%
Surfnet^{+}	2775 (100%)	9,043 (×1.3)	1, 3, 40	8.4	17.2	0.07	0.0%

Table 3. Ligand site characterisation summary. LIGYSIS is not a ligand binding site predictor, but a reference dataset curated from experimentally determined structures of biologically relevant protein-ligand complexes. Coverage represents the number of protein chains the different methods return at least one prediction. Percentage is relative to LIGYSIS sites. Protein chains not covered by VN-EGNN, GrASP and fpocket represent errors in the programme execution, and not absence of predicted sites; Total number of pockets and ratio of predicted pockets by reference pockets in parenthesis, e.g., for each LIGYSIS site, fpocket predicts in average 8.3 pockets; Minimum, median and maximum number of pockets per protein; Median pocket radius of gyration R_g (Å); Minimum centroid distance (MCD) (Å). For proteins where multiple

pockets are predicted, MCD represents the distance to the closest pocket centroid for each of the different predicted pockets. This is a measure of how close predicted pockets are to each other; Maximum residue overlap (MRO). For a given pocket, MRO is the maximum overlap of the residues forming the pockets. MRO is a measure of how similar, in terms of shared residues, the predicted pockets are. For example, the median overlap between VN-EGNN predicted pockets is 85%. Redundancy represents the percentage of predicted pockets that are redundant, i.e., the closest pocket centroid is within 5Å or overlaps at least 3/4 (≥75%) residues. This is the case for 67% of VN-EGNN pockets on proteins, and 0% for GrASP pockets. Bold red font indicates the most extreme values within each column.

Figure 5 represents how the eleven ligand site predictors compare to each other as well to LIGYSIS, which defines ligand sites from experimentally determined biologically relevant protein-ligand complexes. Figure 5A shows how VN-EGNN, GrASP and PUResNet differ from the other methods with a maximum of 7, 12 and 4 predicted pockets, respectively. PocketFinder$^+$, Ligsite$^+$ and Surfnet$^+$ present narrow distributions like LIGYSIS and with medians of 1-3 pockets per protein. P2Rank$_{CONS}$ and P2Rank also present a median of 3 pockets per protein but display wider distributions as they can predict up to 60, and 80 pockets per protein, respectively. Overall, P2Rank$_{CONS}$ predicts fewer pockets than P2Rank. IF-SitePred, DeepPocket$_{SEG}$ and fpocket follow, with a median of 20, 6 and 17 pockets. The difference in number of pockets between DeepPocket$_{SEG}$ and DeepPocket$_{RESC}$ or fpocket is because 60% of fpocket candidates are not extracted by the CNN segmentation module implemented in DeepPocket.

Figure 5B shows the distribution of pocket radius of gyration, R_g. VN-EGNN and IF-SitePred differ from the rest of methods with narrow distributions and medians around 6Å. These two methods do not report pocket residues. Instead, they were obtained using a distance threshold of 6Å from the centroid, for VN-EGNN, and cloud points, for IF-SitePred. This is reflected by examining the percentage of pockets with $R_g > 10$Å which is 0% and 0.1% for VN-EGNN and IF-SitePred. This is a striking difference compared to the other methods: 1.8% (fpocket), 4.8% (P2Rank), 5.7% (DeepPocket$_{SEG}$), 6.4% (GrASP), 6.5% (P2Rank$_{CONS}$) PUResNet (11.6%), 12.6% (LIGYSIS), 16% (Surfnet$^+$), 21.4% (PocketFinder$^+$) and 33.5% (Ligsite$^+$). The latter three predict the sites with largest median $R_g \approx 9$Å. VN-EGNN, GrASP, PUResNet and DeepPocket$_{SEG}$ predict sites with $R_g = 0$Å. This is rather infrequent (7.8% GrASP) and <3% for the other three. These examples correspond to singletons, i.e., pockets formed by only one amino acid.

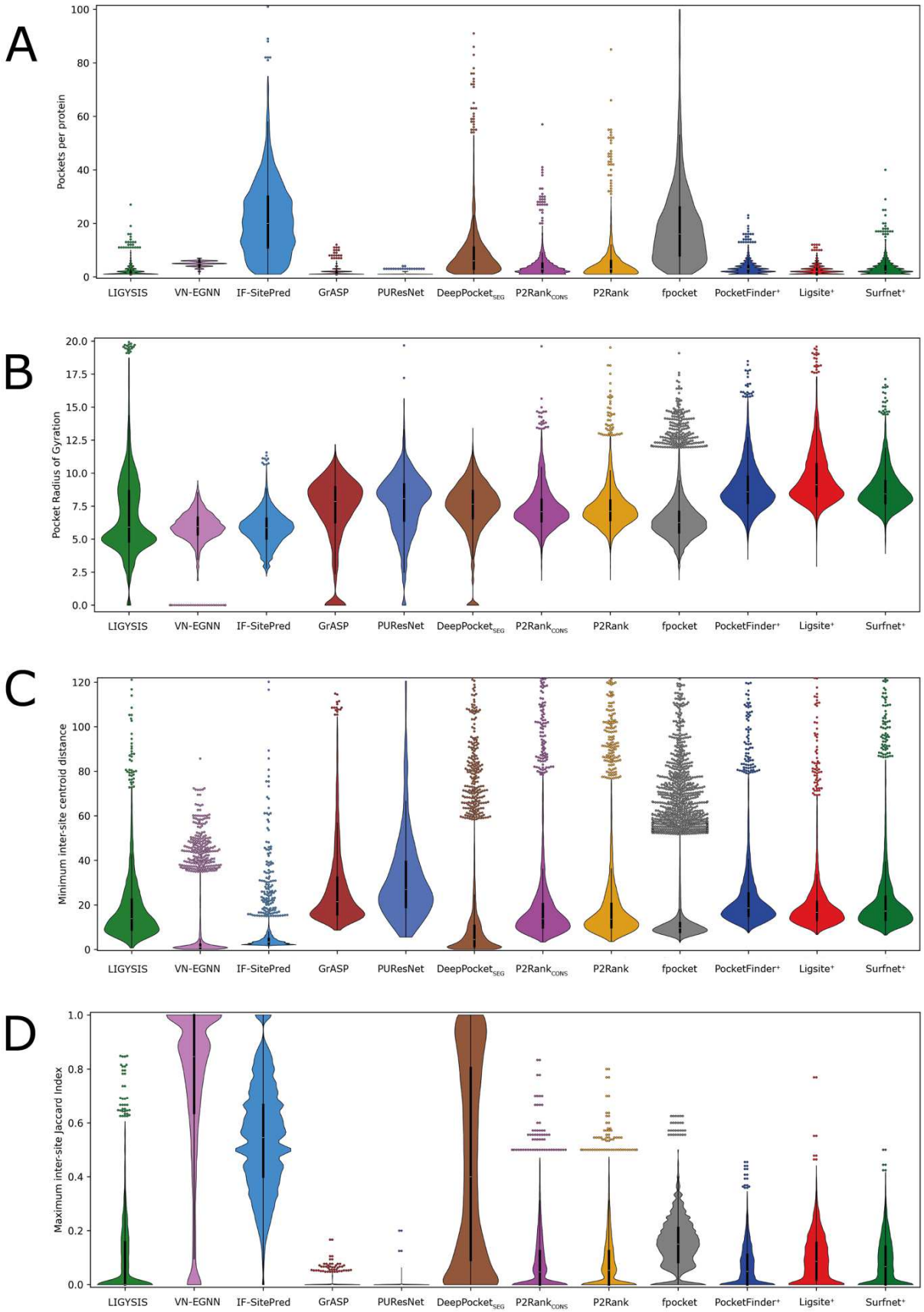


Figure 5. Binding pocket characterisation. Violin plots show the main distributions and swarm plots are used to show outliers. Data points farther than four standard deviations (SD) from the mean are considered outliers. The limit of the Y axis is the maximum *non-outlier* value plus a buffer value. This way, only the most extreme outliers are hidden, which maximises visual interpretation of the data whilst minimising the

number of data points not shown. Within the violin plots are box plots representing the underlying distribution. Line represents the median, box contains the interquartile range (IQR), and whiskers extend to 1.5 × IQR. (A) Number of pockets per protein; (B) Pocket radius of gyration R_g (Å); (C) Minimum inter-pocket centroid distance (MCD) (Å). This is a measure of how close predicted pockets are to each other within a protein; (D) Maximum inter-pocket residue overlap (MRO). Residue overlap was calculated as Jaccard Index. This is a measure of how much the pockets overlap in terms of binding residues.

Figure 5C illustrates how close predicted sites are to each other within a protein chain. Pairwise distances between the centroids of all ligand site pairs for a protein are calculated, and for each site, the minimum distance is taken. Sites predicted by VN-EGNN, IF-SitePred and DeepPocket$_{SEG}$ are very close to each other, with median distances ($\tilde{d}$) of 1.1, 3.4 and 4.6Å, respectively. fpocket follows with $\tilde{d}$ = 9.7Å. The rest of the methods present median distances ranging between 13-18Å (LIGYSIS, P2Rank$_{CONS}$, P2Rank, PocketFinder$^+$, Ligsite$^+$, Surfnet$^+$), and finally GrASP ($\tilde{d}$ = 21.7Å) and PUResNet ($\tilde{d}$ = 27Å). Both versions of P2Rank present the most similar distribution to what is observed on LIGYSIS.

Figure 5D depicts the overlap existing between residues that form the predicted pockets within a protein. All pairwise overlaps, i.e., Jaccard Index, are calculated between pockets in a chain, and for each pocket, the maximum is taken. This is a measure of how much the predicted pockets overlap. This is directly related to how close pockets are, and so VN-EGNN, IF-SitePred and DeepPocket$_{SEG}$ present very high overlaps $\tilde{o}$ = 0.85, $\tilde{o}$ = 0.55 and $\tilde{o}$ = 0.4, respectively. fpocket follows with $\tilde{o}$ = 0.15, Ligsite$^+$ ($\tilde{o}$ = 0.09), Surfnet$^+$ ($\tilde{o}$ = 0.07), P2Rank$_{CONS}$, P2Rank and PocketFinder$^+$ ($\tilde{o}$ = 0.05), and finally LIGYSIS, GrASP and PUResNet with $\tilde{o}$ = 0. GrASP is the only method of the 12 presented here that clusters atoms directly, and as a result, overlap between pockets is minimised. Other methods cluster ligands (LIGYSIS), cloud points (IF-SitePred), SAS points (P2Ranks) voxels (PUResNet, DeepPocket), alpha spheres (fpocket), or grid points (PocketFinder$^+$, Ligsite$^+$, Surfnet$^+$) but not residues, resulting consequently in higher overlapping.

Proximity in space between predicted sites as well as residue overlap are indicators of redundant ligand binding site predictions, i.e., duplicate predictions of a unique observed ligand site. This is the case for VN-EGNN, IF-SitePred and DeepPocket$_{SEG}$. This phenomenon can negatively impact the precision and recall of these methods. Accordingly, correcting for redundancy should have a significant impact on the performance of these methods. In contrast, GrASP and PUResNet which predict a small

number of pockets show low proximity and overlap of predicted sites and so redundancy is not an issue.

Evaluation of Predictive Performance

Redundancy in Ligand Site Prediction

We define pocket prediction redundancy as the prediction of pockets with centroids very close in space (d ≤ 5Å) or with overlapping residues (JI ≥ 0.75). This indicates multiple predictions of the same potential ligand binding site. Most ligand site prediction tools predict not only the location of the pocket by means of a centroid or pocket residues, but also a pocket confidence, and an associated rank among all the predicted pockets. Ligand site predictors tend to be evaluated by considering the top-N, or top-N+2 ranking pockets, where N is the number of observed sites for a given protein. Figure 6A-B illustrates how the redundant prediction of pockets can lead to an overestimation of precision and underestimation of recall. Figure 6C showcases human creatine kinase S-type, mitochondrial (PDB: 4Z9M) as an example of this phenomenon, where VN-EGNN and IF-SitePred redundantly predict the same pocket 9 and 33 times, whereas PUResNet returns a single prediction. All three methods correctly predict the site, just the difference is in the number of returned predictions.

Figure 5D shows that prediction redundancy is an issue particularly for VN-EGNN, IF-SitePred, and to a lesser extent DeepPocket$_{SEG}$. To assess the effect that redundancy has on the performance of these methods, non-redundant subsets of predictions were obtained and labelled with the subscript "$_{NR}$". Redundancy filtering was carried out for each method and adjacent pockets (d ≤ 5Å) *or* sharing more than ¾ of their residues (JI ≥ 0.75) were dropped, keeping always the higher scoring pocket. Redundancy (%) was calculated as the proportion of redundant pockets relative to the original total number of pockets. VN-EGNN presents the highest percentage of redundant pockets with 9,066/13,582 (67%) redundant pockets, followed by IF-SitePred with 22,232/44,948 (49%), and DeepPocket$_{SEG}$ with 6,744/21,718 (31%). For other methods, redundancy was minimal (<1%).

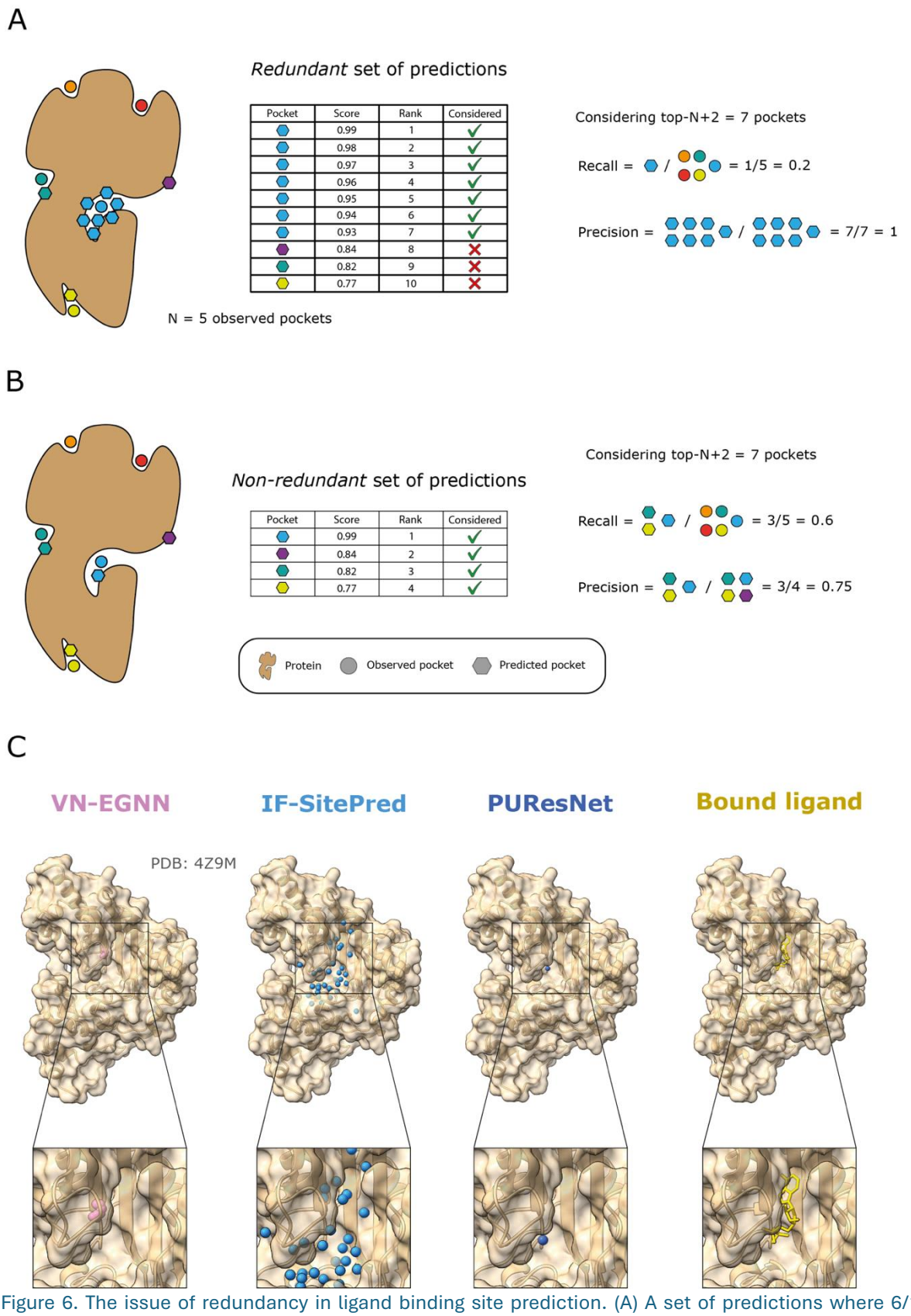


Figure 6. The issue of redundancy in ligand binding site prediction. (A) A set of predictions where 6/10 predictions are redundant, resulting in a low recall (1/5) and inflated precision (7/7); (B) When redundancy is removed, only four predictions remain and recall increases to 3/5 (60%) and precision decreases to ¾ (75%); (C) Predictions by VN-EGNN, IF-SitePred and PUResNet, on chain D of PDB: 4Z9M (Rabeh WM, Tempel W, Nedyalkova L, Landry R, Arrowsmith CH, Edwards AM, Bountra C, Bochkarev A, Park H,

Structural Genomics Consortium (SGC), 2015), where ADP binds. For this ADP binding site, VN-EGNN reports 7 predictions, IF-SitePred 33, and PUResNet a single prediction. These three methods correctly predict this site, however, VN-EGNN and IF-SitePred report redundant pocket predictions, which centroids are very close, ≤ 5Å, in space and residues overlap high (≥ 0.75).

Pocket Scoring Scheme

Newer ligand binding site prediction methods tend to score their pockets to rank them based on their likelihood of being a real binding site. A robust pocket scoring scheme is helpful to rank predicted pockets in a meaningful order, so pockets at the top are more likely to be real pockets. In other words, a pocket scoring scheme that captures the nature of ligand sites is key to maximising the benchmark performance and usability of these tools. IF-SitePred opts for a simplistic scheme, using the number of clustered cloud points as the pocket score. Other methods use more sophisticated approaches, such as taking the output of a machine learning model (VN-EGNN, DeepPocket), summing the model scores of atoms, residues or SAS points (GrASP, $P2Rank_{CONS}$, P2Rank) or using multiple pocket-level features (fpocket). Finally, methods like PUResNet, PocketFinder^{+}, Ligsite^{+} or Surfnet^{+} do not report a pocket score.

The effect that pocket scoring has on the performance of IF-SitePred, PUResNet, fpocket, PocketFinder^{+}, Ligsite^{+}, Surfnet^{+} was explored by (re-)scoring and (re-)ranking the pockets predicted by these methods. Three different approaches were investigated: 1. sorting by number of pocket amino acids "$_{AA}$" (larger pockets ranking higher); 2. running PRANK to (re-)score pockets [37] "$_{PRANK}$"; 3. re-calculating a score with the sum of square of cloud points (IF-$SitePred_{RESC}$) or grid points "$_{SS}$" for PocketFinder^{+}, Ligsite^{+} or Surfnet^{+}, in a similar manner as done in Krivák *et al.*, 2015 [37], and later Smith *et al.*, 2024 [52]. Refer to Supplementary Note 2 and Figure 3 for the effect that redundancy and different pocket scoring strategies has on pocket ranking.

Pocket Level Evaluation

The ideal ligand binding site predictor would have a high precision, i.e., most of the predictions it makes are correct, whilst maintaining a high recall, i.e., recapitulating most of the observed sites. Moreover, the ideal predictor returns predictions that are non-redundant, i.e., it does not predict the same pocket multiple times. Additionally, pockets are ranked in a systematic manner according to a strong and meaningful pocket scoring scheme which captures well the nature of existing ligand binding sites and therefore ranks the predicted pockets from more likely (high score, top) to least likely (low score,

bottom). A good predictor would also perform well at the residue level. This means it is able of capturing the likelihood for a residue to bind a ligand. This can be done by means of a residue ligandability score, which additionally might highlight key residues, the more "ligandable" within a binding site. Ligand site prediction methods were benchmarked with these criteria in mind. The following results concern the original version of the predictors. The non-redundant and re-scored variants of the methods will be discussed in a later section.

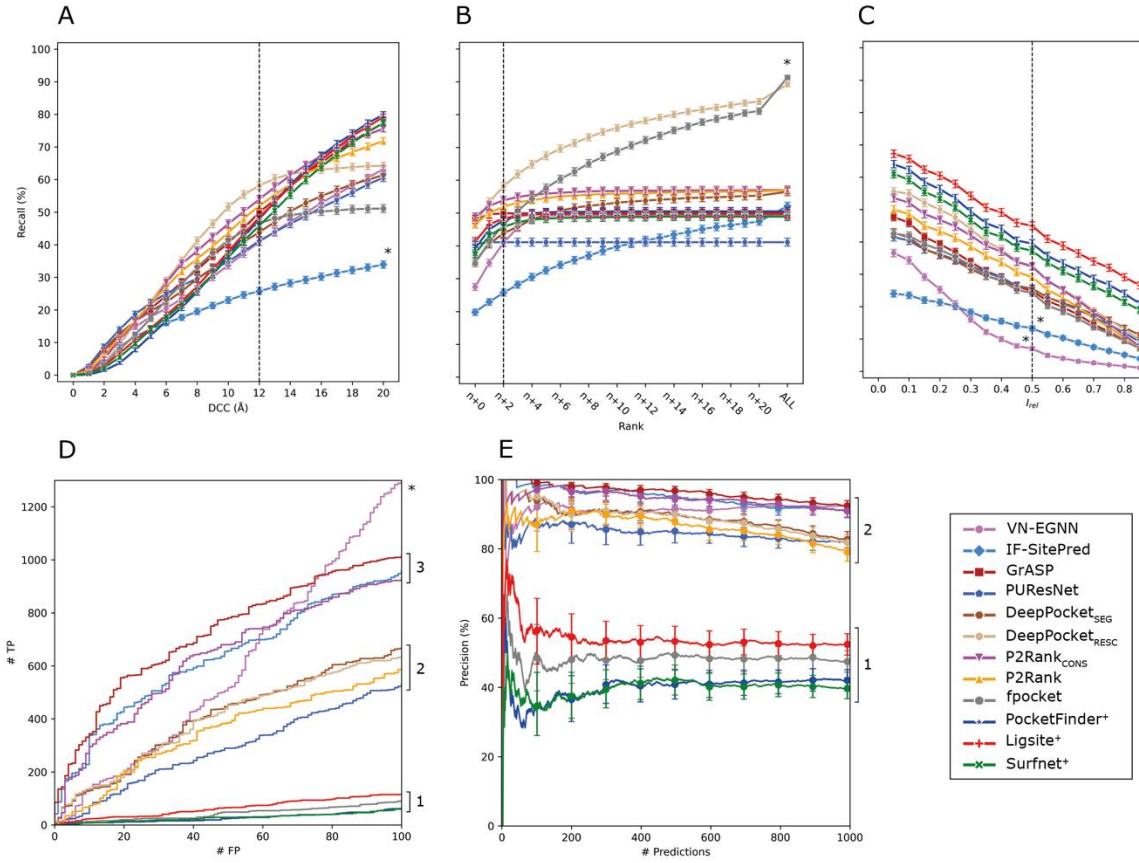


Figure 7. Ligand binding site prediction benchmark at the pocket level. These curves correspond to the default method predictions, not the variants generated in this work. (A) Recall, percentage of observed sites that are correctly predicted by a method according to a DCC = 12Å threshold; (B) Recall using DCC = 12Å but considering increasing rank thresholds, i.e., top-N, N+1, N+2, etc. *ALL* represents the maximum recall of a method, obtained by considering all predictions, regardless of their rank or score; (C) Recall curve using I_{rel} as a criterion; (D) ROC100 curve (cumulative TP against cumulative FP until 100 FP are reached); (E) Precision curve for the top-1,000 predictions of each method across the LIGYSIS dataset, $Precision_{1K}$. Error bars represent 95% CI of the recall (A-C) and precision (E), which are 100 × proportion. Numbers at the right of the panels indicate groups or blocks of methods that perform similarly for each metric. Asterisks (*) indicate outlier methods, or methods that perform very differently than the rest.

Figure 7A illustrates the recall curve for top-N+2 pockets for each method, where N is the number of observed sites for a target protein. Reported recall is obtained using DCC = 12Å. $DeepPocket_{RESC}$ yields the highest recall, 58.1%, closely followed by, $P2RANK_{CONS}$ (53.9%) and P2Rank (51.9%). The rest of the methods present recall <50%

with PUResNet, VN-EGNN and IF-SitePred with lowest recall of 41.1%, 40.9% and 25.7%, respectively (Table 4). Figure 7B, shows the recall curve considering different top-N+X predictions. Most methods reach a plateau by top-N+5, as they do not predict that many pockets. However, methods that predict more pockets per protein, such as IF-SitePred, fpocket, DeepPocket$_{SEG}$ and DeepPocket$_{RESC}$, which take fpocket predictions as a base, increase their recall as more predictions are considered. fpocket and DeepPocket$_{RESC}$ reach a maximum recall of ≈90% when *all* predictions are considered, regardless of rank. Other methods present maximum recall of ≈40-50%. Figure 7C illustrates the recall curve if residue overlap was used instead of DCC as a criterion. In this case, Ligsite^{+}, PocketFinder^{+}, and Surfnet^{+} come on top with recall ≈45% at $I_{rel} \geq 0.5$. This is explained by their prediction of massive cavities, that while often fully contain or overlap with the observed pocket, do not meet the DCC criterion, as their centroids are farther than 12Å from the observed site.

Figure 7D represents the cumulative number of TP against FP when predictions across the proteins in the reference dataset are sorted by score. This shows how effective the scoring scheme of each method is in ranking their predictions to reflect the nature of ligand binding sites. At 100 FP the #TP fall into three different blocks and one outlier: Ligsite^{+}, fpocket, Surfnet^{+} and PocketFinder^{+} are at the bottom with # TP ∈ (60, 120). Secondly, DeepPocket$_{SEG}$, DeepPocket$_{RESC}$, P2Rank, and PUResNet follow with # TP ∈ (530, 670). GrASP, IF-SitePred and P2Rank$_{CONS}$, present a high # TP ranging 900-1,000 at 100 FP. Finally, VN-EGNN sits at the top with 1,301 TP. However, this number is not representative, as the # TP are inflated due to the redundancy in the predictions of VN-EGNN. This is the same for IF-SitePred and DeepPocket$_{SEG}$. Redundant correct predictions of the same pocket will count as multiple TPs, whereas they should only count as 1 TP. Newer methods, e.g., GrASP, P2Rank$_{CONS}$, and VN-EGNN, and IF-SitePred, despite redundancy in their prediction for the latter two, are better at ranking their predicted pockets, presenting up to 1,000 more TP for 100 FP than earlier methods. This means their scoring schemes are significantly better at capturing the essence of a ligand binding site. Including evolutionary conservation in P2Rank (P2Rank$_{CONS}$) predictions results in an increase of +346 TP relative to default P2Rank, indicating that the fewer predicted pockets, and their scores are a more faithful representation of the observed LIGYSIS dataset.

Figure 7E provides insight into the precision of the methods by examining how this metric changes as more predictions are considered. In the same manner as for Figure 7D, predictions across proteins in the LIGYSIS dataset are sorted and cumulative precision is plotted for the top-1,000 scoring predictions. Methods group into two clear blocks. Newer (machine learning-based) methods VN-EGNN, GrASP, IF-SitePred, $P2Rank_{CONS}$, $DeepPocket_{SEG}$, $DeepPocket_{RESC}$ and PUResNet present a $Precision_{1K}$ of 80-95%. Earlier (geometry/energy-based) methods Ligsite$^+$, fpocket, PocketFinder$^+$ and Surfnet$^+$ present lower $Precision_{1K}$ of 40-50%. $DeepPocket_{SEG}$ takes fpocket (geometry-based) predictions as a starting point and achieves much higher # TP_{100FP} as well as $Precision_{1K}$. This is further evidence that performance can be boosted with a solid scoring scheme.

Method	% $Recall_{top-N}$	% $Recall_{top-N+2}$	% $Recall_{max}$	% $Precision_{1K}$	# $TP_{100\ FP}$	% RRO	% RVO
VN-EGNN	27.5	40.9	49.3	**92.5**	**1,301**	**28.0**	**20.0**
IF-SitePred	**19.8**	**25.7**	52.1	91.0	961	42.0	31.0
GrASP	48.0	49.9	50.0	**92.5**	1,017	56.0	69.0
PUResNet	40.6	41.1	**41.1**	81.6	534	65.0	73.0
$DeepPocket_{SEG}$	35.4	43.8	56.5	82.6	670	59.0	71.0
$DeepPocket_{RESC}$	46.6	**58.1**	89.3	81.7	637	50.0	41.0
$P2Rank_{CONS}$	**48.8**	53.9	57.0	90.7	932	57.0	50.0
P2Rank	46.7	51.9	57.0	79.2	586	56.0	65.5
fpocket	38.8	46.5	**91.3**	47.3	94	50.0	41.0
PocketFinder$^+$	39.2	47.8	50.5	42.0	64	79.0	97.0
Ligsite$^+$	41.3	48.4	49.7	52.3	115	**88.0**	**98.0**
Surfnet$^+$	37.7	45.8	48.9	**39.5**	**61**	76.0	92.0

Table 4. Pocket level evaluation summary table. Recall for each method considering top-N, N+2 and *all* predictions (max) without taking rank into consideration, i.e., maximum recall. Precision of the method for the top-1,000 scored predictions. Number of TP reached for the first 100 FP (# TP_{100FP}). Median relative residue overlap (RRO) for those sites correctly predicted and relative volume overlap (RVO) only for correctly predicted sites that have a volume, i.e., are pockets or cavities, and not exposed sites, which don't have a volume. These last two metrics represent the overlap in residues and volume relative to the observed site. Bold font indicates the best (green) and worst (red) performing methods for each metric.

Table 4 summarises these results and shows the median relative residue overlap (RRO) and relative volume overlap (RVO), which are qualitative measures of how well predicted sites align with observed ones in shape. VN-EGNN and IF-SitePred present the smallest RRO and RVO, but it is important to note that these methods do not report pocket residues and so residues were taken within 6Å of their centroid, or pocket spheres. PocketFinder$^+$, Ligsite$^+$, and Surfnet$^+$ present unusually high RRO and RVO. This is a consequence of the massive size of their predicted cavities, that rather than overlap

with the observed site, fully contain and are much larger than it. This might not be convenient in the context of pocket finding for drug discovery where more clearly defined drug-like sites might be of interest. GrASP, PUResNet and DeepPocket$_{SEG}$ present high values of RRO≈60% and RVO≈70% whilst presenting a size distribution more like LIGYSIS (Figure 5B) and provide the best representation of the observed sites in terms of shape similarity.

Residue Level Evaluation

Ligand binding site prediction tools can also be evaluated at the residue level. F1 Score as well as Matthews correlation coefficient (MCC) were utilised to assess prediction performance at the residue level. For each protein chain, F1 and MCC were obtained, distributions graphed, and medians reported (Table 5). IF-SitePred does not report a residue ligandability score beyond a binary label (0, 1). In this work, a score was computed by utilising the scores returned by the 40 prediction models of IF-SitePred. These scores range 0-1 and can be averaged as probabilities (Eq. 10). This will now be referenced as IF-SitePred ligandability score.

Method	F1	MCC	AUC	AP
VN-EGNN	0.29	0.26	-	-
IF-SitePred	0.29	0.24	**0.80**	**0.42**
GrASP	0.39	0.34	0.73	0.34
PUResNet	**0.41**	**0.39**	-	-
DeepPocket$_{SEG}$	0.27	0.21	-	-
DeepPocket$_{RESC}$	0.23	0.12	-	-
P2Rank$_{CONS}$	0.36	0.30	0.78	0.41
P2Rank	0.31	0.26	0.75	0.35
fpocket	**0.23**	**0.12**	-	-
PocketFinder^{+}	0.31	0.22	**0.71**	**0.25**
Ligsite^{+}	0.31	0.21	0.73	0.28
Surfnet^{+}	0.29	0.20	0.71	0.28

Table 5. Residue level evaluation summary table. Ligand binding site prediction benchmark at the residue level was calculated from 2,775 protein chains in the LIGYSIS dataset. Median F1 score, median Matthews correlation coefficient (MCC), ROC area under the curve (AUC) and precision recall (PR) curve average precision (AP). Bold font indicates the best (green) and worst (red) performing methods.

Figure 8A-B illustrate the distributions of the F1 score and MCC for each method on the 2,775 protein chains of the final LIGYSIS set. Both metrics agree that PUResNet (F1 = 0.41, MCC = 0.39), GrASP (F1 = 0.39, MCC = 0.33) and P2Rank$_{CONS}$ (F1 = 0.36, MCC

= 0.30) are the top-3 performing methods in this task of binary classification into pocket (1) and non-pocket residues (0). fpocket presents the lowest F1 = 0.23 and MCC = 0.12 since it predicts many unobserved pockets (residues) that will count as FP here.

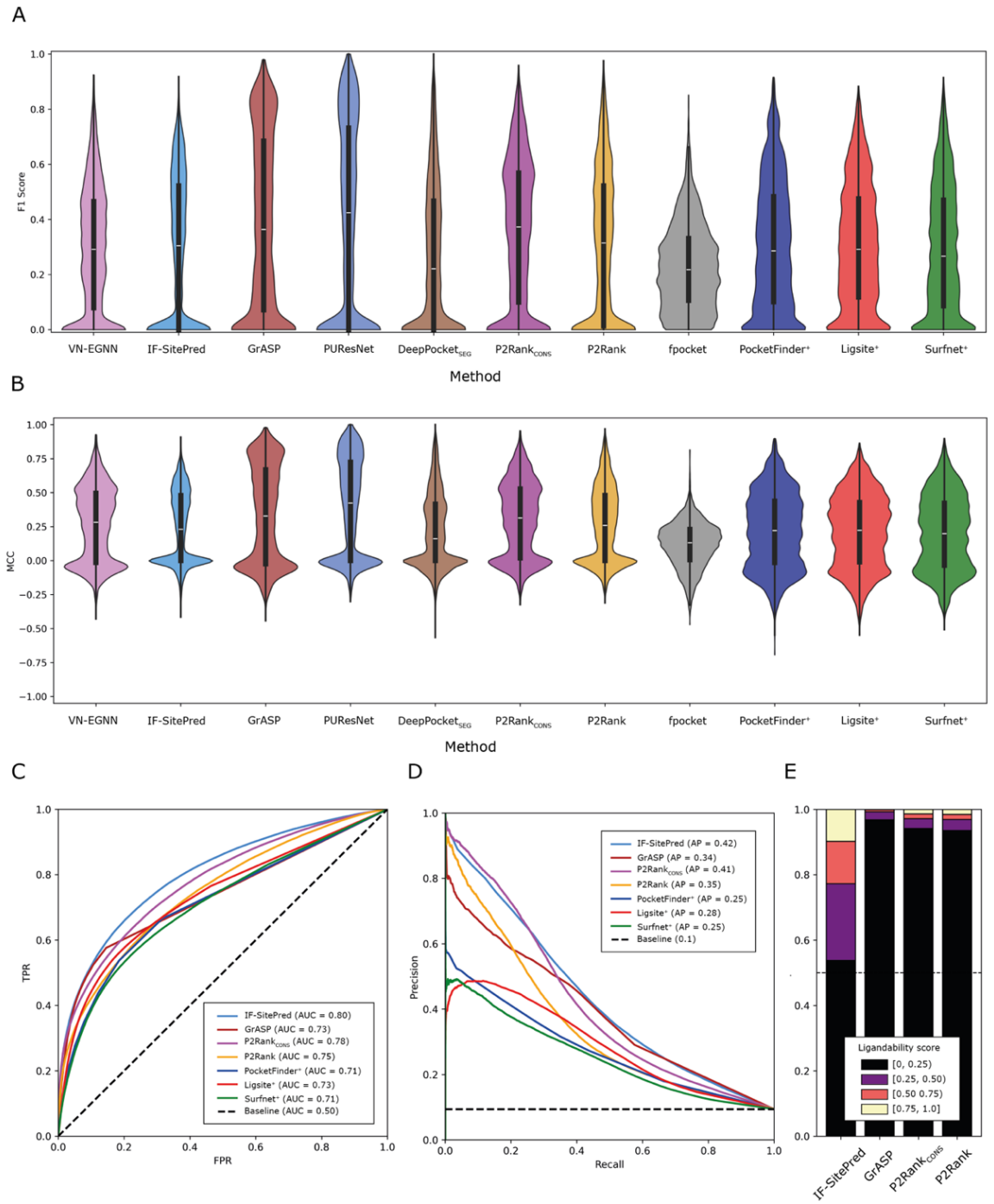


Figure 8. Ligand binding site prediction benchmark at the residue level. (A) F1 score distributions; (B) MCC distributions. In both panels, each data point in the distribution corresponds to the score obtained from all residues in a protein chain; (C) ROC curve for methods that report a residue score. Dashed line represents the baseline, 1 FP for each TP, i.e., diagonal and AUC = 0.5; (D) PR curve. Dashed line represents the baseline, i.e., proportion of observed binding residues (true positives) = 0.1; (E) Distribution of residue ligandability scores for IF-SitePred, GrASP and P2Ranks. PocketFinder+, Ligsite+ and Surfnet+ are not included as their scores do not range 0-1, and a small number of scores can reach values > 25.

For IF-SitePred, GrASP, P2Rank$_{CONS}$, P2Rank, PocketFinder^{+}, Ligsite^{+}, and Surfnet^{+}, which report a residue level score (beyond a binary label), ROC and PR curves were plotted (Figure 8C) and area under the curve (AUC) and average precision (AP) reported. IF-SitePred, with the ligandability score calculated in this work (Eq. 10), presents the highest AUC = 0.80, closely followed by P2Rank$_{CONS}$ (AUC = 0.78) and P2Rank (AUC= 0.75). Surfnet^{+} presents the lowest AUC = 0.71. Figure 8D shows the PR curves, which agree with ROC AUC and highlight IF-SitePred (with Eq. 10 scoring) as the method with the highest average precision = 0.42, followed by P2Rank$_{CONS}$ (AP = 0.41) and Surfnet^{+} the lowest with (AP = 0.25). Figure 8E shows IF-SitePred presenting a different residue ligandability score distribution to GrASP, P2Rank$_{CONS}$, and P2Rank. The IF-SitePred ligandability score, resulting from averaging the scores from the 40 IF-SitePred models, is the most "generous" with ≈20% of the residues presenting a score > 0.5, in contrast with GrASP which residue scoring is very strict (% (LS ≥ 0.5) = 0.75%) and P2Ranks (≈3%). This difference, combined with the ROC and PR curves further supports the use of the IF-SitePred ligandability score proposed in this work to define the predicted binding sites in this method. It also suggests that GrASP might benefit from a less strict residue level scoring scheme. PocketFinder^{+}, Ligsite^{+}, and Surfnet^{+} were not included in this analysis as their scores do not range 0-1 and very high scores (>25) can be obtained.

Improving Current Methods by Redundancy Removal and Pocket Re-scoring

Non-redundant variants of predictions were generated for VN-EGNN, IF-SitePred and DeepPocket$_{SEG}$, denoted with the "$_{NR}$" subscript. For other methods, pockets were re-scored using a variety of approaches, resulting in "$_{AA}$", "$_{PRANK}$", "$_{RESC}$", and "$_{SS}$" variants. In this section, we explore the effect that redundancy removal and pocket re-scoring have on Recall, Precision$_{1K}$, and # TP$_{100FP}$.

Figure 9 illustrates how removing redundant predictions and re-ranking those remaining resulted in a significant ($p \leq 0.05$) increase in recall of +5.2% for VN-EGNN (Figure 9A), +13.4% for IF-SitePred (Figure 9B) and +5.6% for DeepPocket$_{SEG}$ (Figure 9C). Re-scoring IF-SitePred predictions using Eq. 10 resulted in a further increase of +2.1%. After removing redundancy, the number of unique pockets by each method reduces drastically (−67%) for VN-EGNN, (−49%) IF-SitePred and (−31%) DeepPocket$_{SEG}$ (Figure 9D-F), thus limiting the recall of these methods. Re-scoring and re-ranking of PUResNet (0.0%), PocketFinder^{+} (1.1%), Ligsite^{+} (0.6%), and Surfnet^{+} (1.6%) predictions did not have

a significant effect ($p > 0.05$). Re-scoring of fpocket candidates using PRANK resulted in a striking increase in recall of +13.9% in agreement with previous studies [37, 39, 81]. For more details, see Supplementary Note 3, and Figures 4-5.

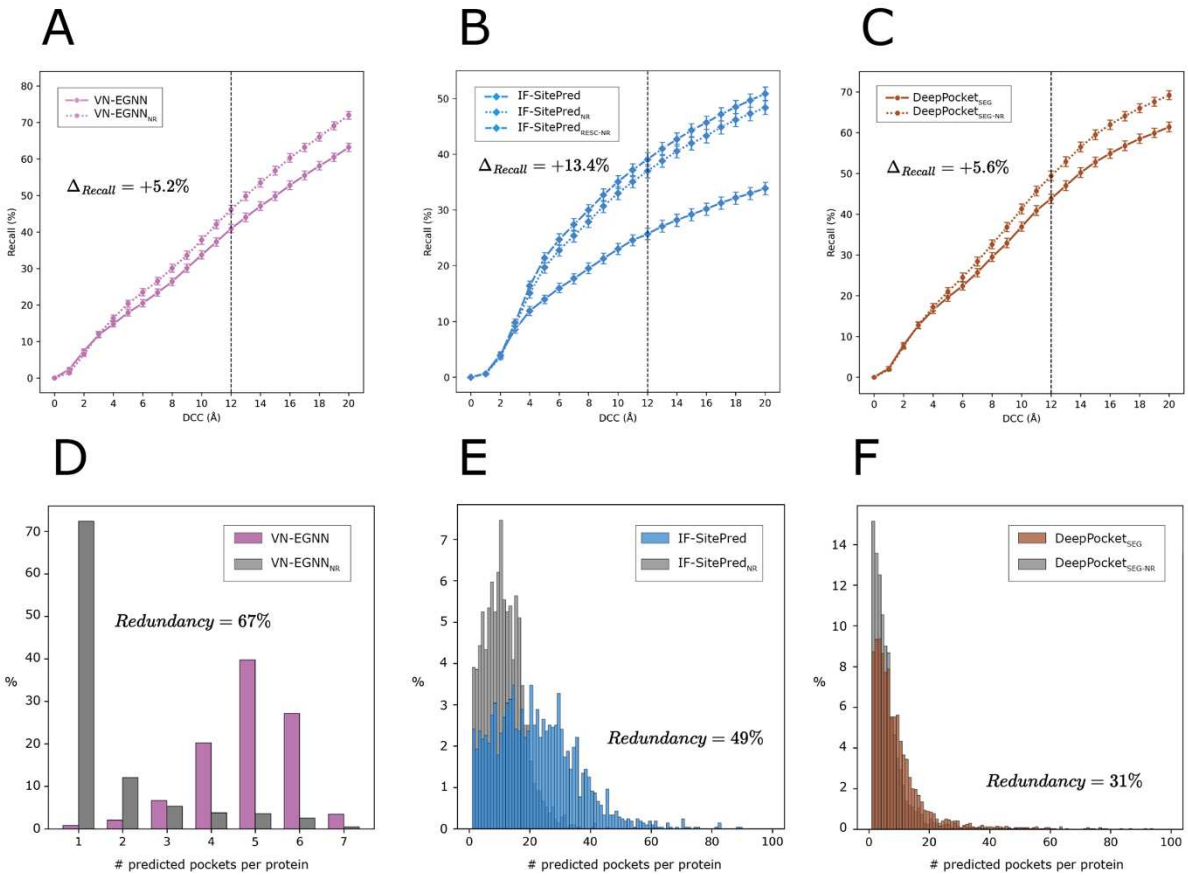


Figure 9. Improvement on the current methods. Solid lines represent recall, i.e., proportion of observed ligand binding sites that are correctly identified by each default version for VN-EGNN (A), IF-SitePred (B) and DeepPocket$_{SEG}$ (C). Dashed/dotted lines represent the non-redundant and re-scored (for IF-SitePred) variants of predictions by these methods generated in this work. Removing redundancy increases significantly the recall of VN-EGNN, IF-SitePred and DeepPocket$_{SEG}$ by > 5%. Vertical black dashed lines indicate the threshold used as reference in this work: DCC = 12Å. Comparison of number of predicted pockets per protein between standard and non-redundant predictions for VN-EGNN (D), IF-SitePred (E) and DeepPocket$_{SEG}$ (F).

Supplementary Note 4 and Figure 6 illustrate the powerful effect that removing redundancy, and a more robust pocket scoring scheme have when ranking predictions across multiple proteins, resulting in increases of up to +600 TP by 100 FP. Moreover, whilst re-scoring fpocket predictions also results in an increase in # TP, DeepPocket$_{RESC}$, DeepPocket$_{SEG-NR}$ and fpocket$_{PRANK}$, all starting from fpocket candidates, achieve no more than ≈600 TP, indicating that the maximum possible increase in recall achieved by pocket re-scoring is limited by the starting predictions, which are the same for these three methods. Finally, despite benefitting from a simple scoring of their pockets by taking the number of amino acids, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+} present the lowest # TP$_{100FP}$.

This is potentially because their purpose is the detection of cavities and not the prediction of ligand sites. Despite being similar tasks, they are not the same. This is reflected in the number of cavities they report and their size, which explain the low number of # TP_{100FP}. Moreover, these methods are not machine learning-based, and were not trained to learn from and recapitulate observed ligand binding sites.

Redundancy removal for VN-EGNN, IF-SitePred and $DeepPocket_{SEG}$ does not have a significant effect ($p \geq 0.05$) on the precision of these methods. However, (re-)scoring predictions of PUResNet, fpocket, $PocketFinder^{+}$, $Ligsite^{+}$, and $Surfnet^{+}$ increased precision on the top-1,000 predictions by +11.7%, +34.4%, 23.3%, 16.5% and 29.1%, respectively, further proving the importance of the employed pocket scoring scheme (Supplementary Note 5 and Figure 7). Supplementary Figure 8 and Table 1 summarise the results of this analysis.

Discussion

In this work we have carried out the most comprehensive comparative analysis of ligand binding site prediction methods to date, spanning three decades of methods development. First, predictions from the different methods as well as observed sites from our reference dataset, LIGYSIS, were compared in terms of the number of proteins methods predict on, the number of predicted sites per protein, their size, distance and overlap between the sites. This analysis provides insight into how the different methods work and hints at potential limitations or room for improvement, e.g., the prediction of a fixed number of sites per protein, or considerable proximity and overlap between the predictions.

We have shown how redundancy in prediction, i.e., predicting multiple times the same observed site, can underestimate the recall, and overestimate the precision of the methods, therefore providing a misleading assessment of the methods' performance. Redundancy removal and subsequent pocket re-ranking can yield a significant increase in recall. The importance of a robust pocket scoring scheme can have a strong impact in the performance, both in recall and precision of the methods and emphasis should be put into this area. Even if a single site is predicted per protein, a pocket score can be highly useful when ranking pockets in different proteins, e.g., when having a list of potential drug targets and deciding which protein might be best to target therapeutically.

Recall (% of sites that are correctly predicted) is more informative than precision (% of predictions that are correct). In a context of discovery where the true ligand binding sites of the target are unknown, it is more useful to have multiple predictions that might or might not correspond to real sites (lower precision), rather than a single or few predictions that are very precise but are missing other likely sites. Most methods do well in predicting the most obvious (orthosteric) site. This site, however, might not be available for therapeutic targeting and it is convenient to have other predicted sites that could modulate function acting as allosteric sites. Precision, though a metric that provides valuable insight, and covered in this work, must always be contextualised with recall. Our results show that the most precise methods do not correspond to higher recalling methods. A method predicting the most obvious site, that could be identified by eye, might be 90% precise, but present a lower recall, e.g., 30%. This matter is of even more importance if the "success rate" is defined as the precision of the top scoring 1-3 predictions, which is not a very representative performance assessment metric.

For this reason, we recommend that developers of methods not only share the code of the method, but also of the benchmarking analysis. Moreover, the definition of *success rate* must be standardised as recall, as some methods use recall, and others precision, both under the same term of "success rate". This can be confusing when comparing the results from different analyses. Alternatively, success rate should be clearly defined, so readers can fully understand the implications of the metric employed on a benchmark.

It is clear from our results that a DCC threshold of 4Å is too conservative, and a more flexible DCC threshold of 10-12Å should be used for comparable performance with DCA = 4Å. According to our results and our reference set, predictions with DCC 4-12Å overlap or are adjacent to observed sites and should be considered as correct predictions. The reason for this is the inherent noise in the ground truth, observed data or reference dataset, i.e., a ligand binding to a cavity might not be representative of all ligands that could bind to it. For most proteins, not all existing ligand sites are characterised and as different ligands can bind to the same region, it is unrealistic to use such a small DCC threshold. Our results show several examples of predictions of an observed cavity where DCC > 4Å.

fpocket$_{PRANK}$ (60%) and DeepPocket$_{RESC}$ (58%) present the highest recall of the methods reviewed in this work. P2Rank$_{CONS}$ and P2Rank follow closely with 54% and 52% recall, then GrASP (50%), DeepPocket$_{SEG-NR}$, Ligsite$^{+}_{AA}$ and PocketFinder$^{+}_{AA}$ with 49%, Surfnet$^{+}_{AA}$ (47%), VN-EGNN$_{NR}$ (46%), PUResNet$_{PRANK}$ (41%) and IF-SitePred$_{RESC-NR}$ (39%). fpocket is the method that predicts the most pockets per protein, reaching a maximum recall between 80-90% (considering all pockets regardless of the rank). P2Rank$_{CONS}$ comes second with a maximum recall of 50-60%. The rest of the methods range 40-55%. This indicates that whilst there are still some pockets un-predicted by fpocket (10-20%), the maximum recall of this method is 20-30% higher than any other method. However, considering top-N+2 pockets, fpocket only recalls 47% of the observed pockets. fpocket$_{PRANK}$ and DeepPocket$_{RESC}$ gain >10% in recall by simply re-scoring those predictions. This highlights the paramount importance of a robust scoring scheme, which captures well the nature of binding sites and places those with a higher probability of being real binding sites at the top of the ranking. Newer methods like VN-EGNN, IF-SitePred, GrASP and PUResNet are the most precise methods, however because of redundancy in predictions (VN-EGNN, IF-SitePred), or low number of predicted pockets per protein (VN-EGNN, GrASP and PUResNet) are limited in their recall. Their high precision indicates that their models learn and capture well the nature of ligand binding sites and so they represent a great venue to pursue in the field of ligand binding site prediction. Whilst removing redundancy post-prediction has a significant improvement in performance (VN-EGNN$_{NR}$ and IF-SitePred$_{NR}$), approaching this issue before prediction would be more beneficial. For VN-EGNN, which predicts a maximum of 8 sites, ensuring these 8 (or more) predictions are non-redundant is more desirable than removing redundant predictions ending up with 1/8 predictions. The same applies to IF-SitePred, where non-overlapping starting predictions are more convenient than dealing with redundancy post-prediction.

The usefulness of residue-level metrics as F1 score or MCC is limited, as methods that precisely and correctly predict the clearest sites, such as PUResNet (high precision, low recall) perform better on these metrics, whilst methods that predict more pockets (lower precision, higher recall) such as fpocket and re-scored versions, will obtain worse results. Pocket-level metrics, particularly recall, are more representative of the ability to predict ligand binding sites.

Whilst datasets like PDBbind, binding MOAD or the brand new PLINDER [82] (will) prove extremely useful to train, validate and test deep learning models tackling problems such as rigid body docking [83], flexible pocket docking [84], or pocket-conditioned ligand generation [85], they might not be ideal as a test set for ligand binding site prediction. LIGYSIS analyses all unique, biologically relevant protein-ligand interfaces, including ions, across the biological assembly from multiple experimentally determined structures of a given protein. It then clusters these ligands based on their interactions with the protein, resulting in the observed binding sites. Beyond considering biological assemblies and unique protein-ligand interfaces, the greatest innovation in LIGYSIS is leveraging the extensive structural data on the PDBe-KB to aggregate ligand-binding interactions across different structures of the same protein, thus better capturing the ligand-binding capabilities than just taking a single structure of a protein-ligand complex. In doing so, LIGYSIS represents the most complete and non-redundant protein-ligand complex dataset to date.

Aggregating protein-ligand interactions across structures of the same protein is likely to be beneficial not only for testing, but also when training these methods. Most current methods train on datasets where a protein is represented by a single structure interacting with a single ligand. For example, in 100% of sc-PDB and 50% of entries for binding MOAD training sets. Methods consider as ligand binding, and therefore TP, those residues within a certain distance of the ligand and TN all other residues. In doing so, residues of the same protein that bind ligands on other structures, but not the one present on these sets, will be incorrectly labelled as “non-ligand-binding” (FN). This mislabelling of residues could lead to a lower prediction performance. This issue is to a certain extent approached by P2Rank and GrASP, which enriched their training datasets by including ligands from other chains, or homologous structures. This noise in the training dataset might be more prevalent for DeepPocket, PUResNet and VN-EGNN, which seem to rely fully on 1:1 interactions. The usage of LIGYSIS, or any other data set that aggregates ligand interactions across structures, might alleviate this issue and hints at potential room for improvement in the field of ligand binding site prediction.

Conclusions

The conclusions resulting from our analysis are as follows:

- Ligand binding site prediction methods differ significantly in the number of predicted sites, their size, proximity and overlap, which offers insight into how the methods work.
- Redundancy in ligand binding site prediction leads to an underestimate of recall and an overestimate of precision. The removal of such redundancy and subsequent re-ranking of the remaining pockets results in a drastic increase in recall.
- A robust pocket scoring scheme is crucial for the correct ranking and prioritisation of predicted sites in downstream analysis, e.g., docking, simulation. Additionally, it has a significant positive effect on both precision and recall.
- Recall is a more informative measure of the performance of a ligand site prediction tool, rather than precision and so it must be reported. Precision, though a useful metric, should always be contextualised with recall.
- All authors of ligand site prediction tools should use recall as "success rate" for consistency. Benchmarking code should also be shared by the authors for the sake of reproducibility.
- Pocket-level metrics (recall, precision) are a more adequate representation of the capabilities of ligand site prediction methods than residue-level metrics (F1, MCC).
- A DCC threshold of 4Å is too conservative, and to obtain comparable results between DCA and DCC recall, a threshold of DCC of 10-12Å should be employed.
- Re-scoring of fpocket predictions, as $fpocket_{PRANK}$ or $DeepPocket_{RESC}$ present the highest recall (60%) among the methods reviewed in this analysis.
- Methods that systematically predict a low number of pockets, e.g., VN-EGNN, GrASP or PUResNet, are very precise (>90%), however their recall is low, and might not be as useful in a discovery context.
- IF-SitePred benefits significantly from pocket re-scoring, and suggests that protein embeddings, which aren't directly dependent of structure, represent great promise in the field of ligand site prediction.
- The use of duplicated protein-ligand interfaces in asymmetric units results in an overestimate of both precision and recall when benchmarking ligand site

predictors. Only unique protein-ligand interfaces in biological units should be considered for a more accurate benchmark of the performance of these methods.

- LIGYSIS aggregates non-redundant biologically relevant protein-ligand interactions across multiple structures for a protein and sets a new test set standard for the benchmark of ligand binding site prediction tools.

Methods

Derivation of the LIGYSIS dataset

There are 20,423 human reviewed proteins in UniProt [86]. Only 7,640 (37.4%) of these proteins present experimentally determined three-dimensional structures deposited in the Protein Data Bank [87-89], 5,455 (71.4%) of which present at least one ligand-binding structure [90]. Only biologically relevant ligands, as defined by BioLiP [64] were considered, yielding a set of 3,513 proteins including 4,037 structural segments, as defined by the PDBe-KB. These segments represent a UniProt sequence region with one or more structurally overlapping chains [91, 92]. A protein might be represented by multiple segments. Backbone only structures, as well as those with ligand-binding residues missing UniProt residue ID mappings were removed from our set, resulting in a final set of 3,048 human proteins, including 3,448 segments.

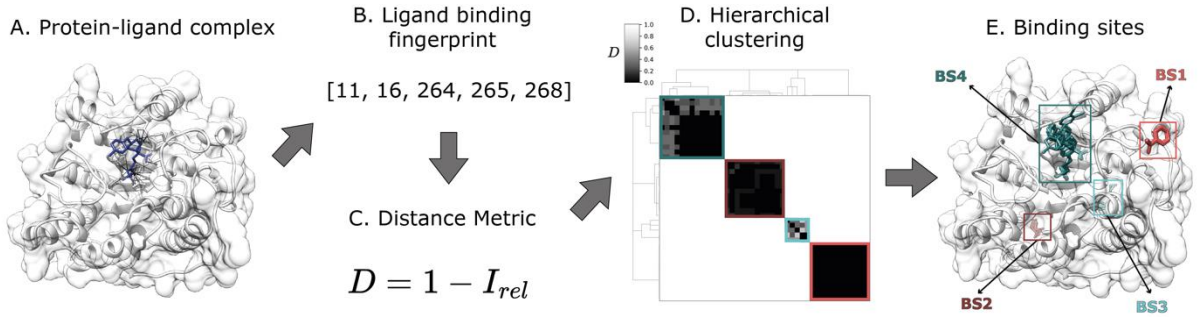


Figure 10. LIGYSIS ligand binding site definition algorithm. For a given protein-ligand interaction complex (A), a ligand binding fingerprint (B) is obtained as the set of unique UniProt sequence residue numbers interacting with the ligand as defined by pdbe-arpeggio; (C) Fingerprints from different ligands binding to the same protein can be compared and a distance calculated; (D) This distance can be used to perform hierarchical clustering, which groups the different ligands in distinct clusters binding to the same region of the protein or binding site (E).

Preferred biological assemblies, as defined by PISA [65], were downloaded from PDBe via ProIntVar [93]. Protein-ligand contacts were determined with pdbe-arpeggio [94]. Figure 10 illustrates the ligand site definition approach used to obtain our reference dataset: LIGYSIS. For a pair of ligands, L_A, L_B, fingerprints A, B are defined as sets containing the residue numbers of the amino acids interacting with each ligand. Relative

intersection, I_{rel}, is a similarity metric that quantifies how similar the fingerprints are [5]. Subtracting I_{rel} from 1 gives a distance, D (Eq. 1), which takes the value of 1 when A and B share all the binding residues and 0 when they share none. For a given protein segment, interacting with M biologically meaningful ligands across N chains, ligand fingerprints are clustered using average linkage with SciPy [95] and ligand sites obtained by cutting the tree at D = 0.5. This resulted in 8,244 ligand binding sites.

$$D = 1 - I_{rel} \quad (1)$$

Comparison of Datasets

Training and test datasets were downloaded for all machine learning based methods reviewed in this work. Datasets were compared to our reference set, LIGYSIS, in terms of number of sites per protein, ligand-interacting chains, chain lengths, size of the sites (number of amino acids), ligand composition, size and diversity. Ligand diversity was quantified by Shannon's Entropy [96] (Eq. 2) where p_i represents the proportion of each ligand i of the R ligands observed in the dataset. Ligand data was extracted from the Chemical Component Dictionary (CCD) [97]. An overlap (%) was calculated for each dataset as the proportion of LIGYSIS binding sites that were covered by at least one ligand in a test dataset. A simplistic approach was adopted by calculating the intersection of ligand IDs between LIGYSIS and each dataset. Ligand IDs were defined as a string of PDB ID + "_" + ligand ID, e.g., "6GXT_GTP" corresponds to the guanosine-5'-triphosphate of the PDB entry with ID: 6GX7 [98].

$$H' = -\sum_{i=1}^{R} p_i \ln(p_i) \quad (2)$$

Training Datasets

VN-EGNN trains on a subset [43] of the sc-PDB (v2017) [99-103] (sc-PDB_{SUB}). sc-PDB is a comprehensive database of pharmacological ligand-protein complexes. The database is comprised by proteins in complex with buried, biologically relevant synthetic or natural ligands deposited in the PDB. sc-PDB contains unique non-repeating protein-ligand pairs, which means that only one ligand is considered per PDB structure entry. Smith *et al.*, 2024 [52] enriched this dataset with 9,000 extra ligands resulting in a version of sc-PDB which we call sc-PDB_{RICH}, which GrASP trained on. Unfortunately, this dataset is not publicly accessible and therefore not considered in our analysis. DeepPocket used the

full sc-PDB set to train on, sc-PDB_{FULL}. IF-SitePred uses a sequence identity-filtered version of the non-redundant subset of the binding mother of all databases (MOAD) [104-107], which considers only protein family leaders. The binding MOAD, here referred to as $bMOAD_{SUB}$, is a large collection of crystal structures with clearly identified biologically relevant ligands with binding data extracted from the literature. Finally, P2Rank used the CHEN11 dataset to train, which aimed to cover all SCOP [108-110] families of ligand binding proteins in a non-redundant manner [111] and the JOINED dataset for validation. CHEN11 not only considers the ligands in each structure but is enriched with ligands binding to homologous structures. JOINED is a combined dataset formed by other smaller datasets: ASTEX [112], UB48 [30], DT198 [113] and MP210 [34], which represent diverse collections of protein-ligand complexes, including bound/unbound states, drug-target complexes and other ligand site predictor benchmark sets.

Test Datasets

The majority of ligand binding site predictors published since 2018 have been using two datasets that were first presented by Krivák *et al.*, 2018 [39]: COACH420 and HOLO4K, or subsets of them. COACH420 is comprised by a set of 420 single-chain structures binding a mix of drug-like molecules and naturally occurring ligands which is disjunct with the CHEN11 and JOINED datasets. COACH420 is a modified version of the original COACH test set [26, 27]. HOLO4K is a larger set, $N \approx 4{,}000$, based on the list by Schmidtke *et al.*, 2010 [114], which includes a mix of single- and multi-chain complexes, also disjunct with P2Rank training (CHEN11) and validation (JOINED) datasets. VN-EGNN, DeepPocket and GrASP use the Mlig and Mlig+ subsets of the COACH and HOLO4K datasets, which comprise strictly biologically relevant ligands as defined by the binding MOAD. IF-SitePred is tested on the HOLO4K-AlphaFold2 Paired (HAP) and HAP-small sets. HAP is a subset of the HOLO4K dataset which presents high quality models in the AlphaFold database [115]. HAP-small is a smaller subset of HAP that only contains proteins with sequence identity lower than 25% to proteins in the P2Rank training set. VN-EGNN uses the refined version of PDBbind (v2020), referred here as $PDBbind_{REF}$, [116-121] as a third test set. Like binding MOAD, the PDBbind database provides a comprehensive collection of experimentally measured binding affinity data for macromolecular complexes. Specifically, the refined set includes those protein-ligand complexes for which binding data was obtained with the literature and met certain experimental quality thresholds.

Lastly, SC6K is a dataset presented by Aggarwal *et al.*, 2021 [47] containing 6,000 protein-ligand pairs from PDB entries submitted from 01/01/2018 – 28/02/2020.

Protein Chain Alignment

For each protein chain, atomic coordinates were translated to be centred at the origin, O = (0, 0, 0), and rotated using a rotation matrix, R. The two principal components of the coordinate space pc_1 and pc_2 were obtained using principal component analysis (PCA) [122]. A third component, $pc_\perp$, was obtained with the cross-product of the other two, to ensure orthogonality. A rotation matrix P is constructed from these vectors (Eq. 3). By placing the main component pc_1 on the second row of P, we ensure the Y axis will be the major axis, representing the *height* of the protein chain. The second largest axis will be the X axis, representing the *width* of the protein, and lastly the *depth* will be represented by the smaller magnitude of the Z axis. The final rotation matrix R is obtained by multiplying P by the negative identity matrix NI (Eq. 4). This was done to maintain the left-handedness of the protein chains whilst ensuring a consistent alignment on the major axes.

$$pc_\perp = pc_1 \times pc_2, \qquad P = \begin{bmatrix} pc_2 \\ pc_1 \\ pc_\perp \end{bmatrix} \quad (3)$$

$$NI = -1 \cdot I_3 = -1 \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \qquad R = P \cdot NI \quad (4)$$

Protein Chain Characterisation

For a protein chain with N amino acid residues, the centre of mass, CM, was calculated by averaging the coordinates, r_i, of all atoms (Eq. 5), and from it, the radius of gyration, R_g, was derived (Eq. 6) [123]. As the protein chains are already aligned on the axis and centred on O = (0, 0, 0) the dimensions of the protein chain can be obtained as the magnitude of the PCA components or *eigenvectors*, i.e., the *eigenvalues*. The dimensions represent width, height, and depth for the X, Y and Z axes, respectively.

$$CM = \frac{1}{n}\sum_{i=1}^{n} r_i \rightarrow CM = O = (0,0,0) \quad (5)$$

$$R_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - CM)^2} \rightarrow R_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - O)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {r_i}^2} \quad (6)$$

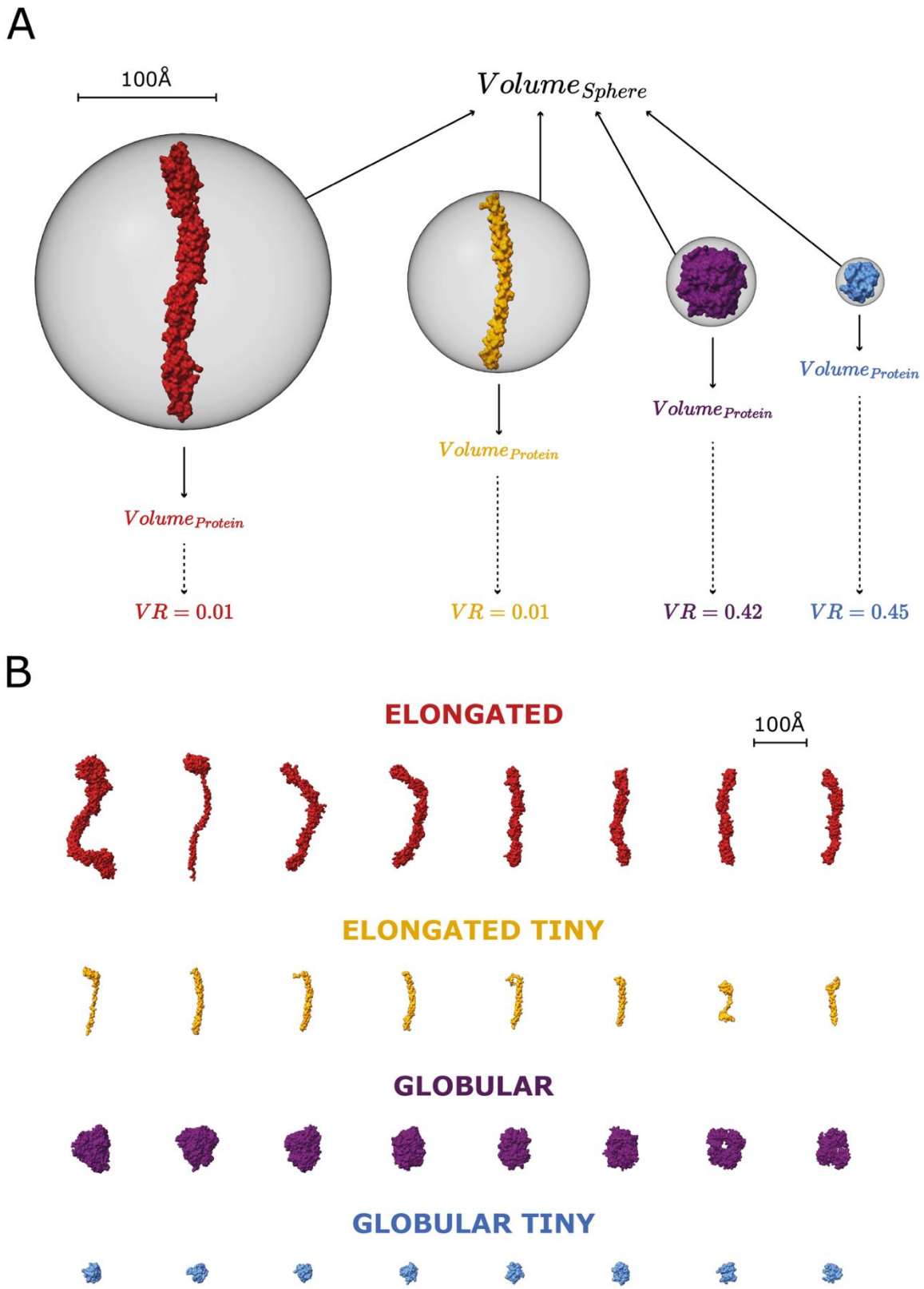


Figure 11. Protein chain shape and size classification approach. (A) The volume of the sphere enclosing the protein chain as well as the protein chain volumes are calculated, and their ratio obtained (VR). Globular proteins present more spherical shapes and therefore occupy a higher portion of the sphere volume, resulting in higher volume ratios. Non-globular, elongated or fibrous proteins on the other hand do not and present lower VRs. After extensive visual examination, a threshold was established at VR = 0.08, and so

proteins classified in these two groups. Proteins were classified as "tiny" if their chain was 100 amino acids; (B) Eight examples of each protein chain group to illustrate the outcome of the approach.

Protein chain volumes were calculated using ProteinVolume [124]. A sphere enclosing the protein and centred on the protein centre of mass was obtained. The radius of this sphere is the maximum Euclidean distance between the protein atoms and the CM (Eq. 7). The volume of the sphere is calculated using Eq. 8. Proteins were classified into four different groups based on their shape and size. Protein chains with ≤ 100 amino acids were classified as "tiny". Regarding the shape, protein chains were classified into "elongated" if their protein to sphere volume ratio ≤ 0.08 (VR) (Eq. 9), i.e., the protein volume contains no more than 8% of the sphere volume. This threshold was derived empirically by the visual inspection of all 3,448 protein chains on the LIGYSIS set. Otherwise, proteins were considered globular (Figure 11). In this manner, protein chains were classified into *globular* (N = 2,104; 61%), *elongated* (N = 670; 19%), *elongated tiny* (N = 341; 10%) and *globular tiny* (N = 333; 10%).

$$R = max_i \|r_i - CM\| \quad (7)$$

$$Volume_{Sphere} = \frac{4}{3}\pi R^3 \quad (8)$$

$$VR = Volume_{Protein}/Volume_{Sphere} \quad (9)$$

Ligand Binding Site Prediction

For each segment in the LIGYSIS dataset, the representative chain as defined in the PDBe-KB was selected. Structures were cleaned using the *clean_pdb.py* script [125]. Eleven different ligand binding site prediction tools were employed to predict on the 3,448 representative chains: VN-EGNN [54], IF-SitePred [53], GrASP [52], PUResNet [43], DeepPocket [47], P2Rank [37, 39], $P2Rank_{CONS}$ [61, 126], fpocket [7, 127], PocketFinder$^+$ [15], Ligsite$^+$ [8], and Surfnet$^+$ [9]. Conservation scores for P2Rank were obtained from PrankWeb https://prankweb.cz/. Re-implementations of Capra *et al.*, 2009 [33] were used for PocketFinder, Ligsite and Surfnet, indicated by the "$^+$" superscript. VN-EGNN, IF-SitePred, PocketFinder$^+$, Ligsite$^+$ and Surfnet$^+$ do not provide a list of residues for each pocket, but a list of centroids and their scores for the first two, and a list of grid points for each predicted pocket for the last three. For VN-EGNN, residues within 6Å of the virtual nodes were considered pocket residues. For 429 predicted pockets (≈3%) no residues were found within this threshold. For IF-SitePred, residues within 6Å of the clustered

cloud points that resulted on a predicted pocket centroid were considered as pocket residues. Pocket residues were obtained in a similar manner for PocketFinder⁺, Ligsite⁺ and Surfnet⁺, by taking those residues within 6Å of the pocket grid points. When running DeepPocket, the *-r* threshold was removed and so all fpocket candidates were passed to the CNN-based segmentation module for pocket shape estimation. fpocket predictions re-scored by DeepPocket will be referred as $\text{DeepPocket}_{\text{RESC}}$, whereas pockets extracted by the segmentation module of DeepPocket will be referred as $\text{DeepPocket}_{\text{SEG}}$.

Seven of the considered methods provide residue *ligandability* scores. P2Rank and $\text{P2Rank}_{\text{CONS}}$ report calibrated probabilities of residues being ligand-binding. Similarly, GrASP predicts the likelihood for any given heavy atom to be part of a binding site. A residue-level score was obtained for GrASP predictions by taking the maximum score of the residue atoms. For IF-SitePred, a residue ligandability score *LS* can be obtained by averaging the 40 independently predicted probabilities of a residue being ligand-binding (Eq. 10). Though calculated in a different way, these three scores range 0-1 and represent the likelihood of a residue binding a ligand and can therefore be compared. PocketFinder⁺, Ligsite⁺, and Surfnet⁺ also provide residue scores which maximum value can be > 1.

$$LS = \frac{1}{40}\sum_{i=1}^{40} p_i \quad (10)$$

The other three methods, i.e., VN-EGNN, PUResNet, DeepPocket, and fpocket do not report residue-level scores. However, binary labels represent whether a residue is part of a pocket (1) or not (0), in the same manner as all other methods.

Throughout this work the terms “site” and “pocket” are be used indistinctly. Across all figures, tables and legends, methods are sorted in chronological order.

Binding Site Characterisation

Radius of gyration, R_g, was calculated for pockets as it was done for whole protein chains (Eq. 6). Distance between pockets was calculated as the Euclidean distance between their centroids and overlap between pocket residues with the Jaccard Index (JI), or intersection over union (IOU) (Eq. 11) [128, 129]. POVME 2.0 was employed for pocket volume calculation [130-132]. A single inclusion region was used for each pocket. This region is defined by the smallest rectangular prism containing all pocket atoms. The

prism is centred on the pocket centroid and its dimensions are determined by the distance between the two farthest atomic coordinates on each axis. No exclusion regions were used. Points outside the convex hull were deleted. A contiguous-points region was defined as a 5Å-radius sphere on the pocket centroid (Figure 12).

$$JI(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (11)$$

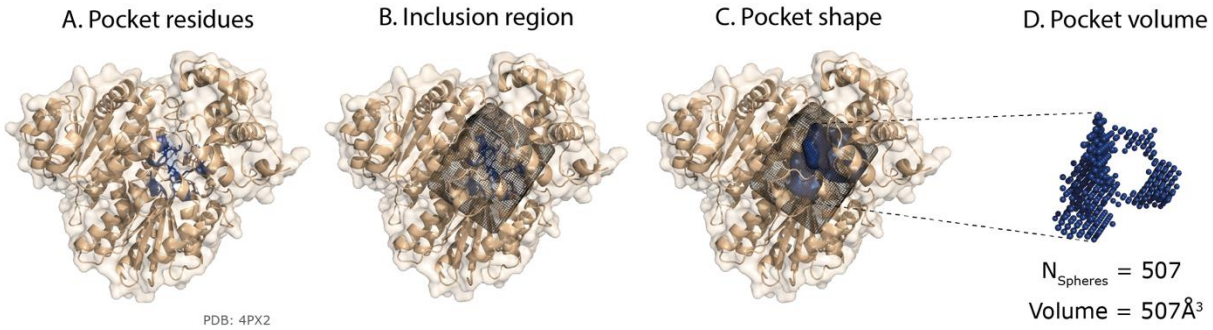


Figure 12. Pocket volume calculation algorithm. (A) PUResNet predicted pocket for PDB: 4PX2 (Jordan, S.R., Chmait, S., 2015). Pocket residues are coloured in blue and have their side chains displayed; (B) An inclusion region is determined using the coordinates of the pocket residue atoms; (C) POVME 2.0 calculates the shape of the pocket defined by the residues and contained within the inclusion region; (D) The pocket shape is defined by a series of unit volume (1Å^3) spheres. The volume of the pocket is calculated as the addition of the sphere volumes or the number of spheres within the pocket. Structure visualisation with PyMOL v2.5.2 [59].

Prediction Evaluation

After predicting on the 3,448 chains of the LIGYSIS dataset, only chains where all residues across all predicted sites presented UniProt residue mapping were kept. This resulted in a final set of 2,775 protein chains which was used for the performance assessment of the methods.

The performance of ligand binding site prediction methods can be evaluated at two different levels: *residue* level, and *pocket* level. Prediction at the *residue* level involves the discrimination of those residues that are likely to interact with a ligand, whereas the aim of *pocket*-level prediction is to define distinct regions on a protein, i.e., pockets where *a* ligand is likely to bind. This region can either be defined by a (pocket) centroid, a group of cloud/grid points, a set of pocket residues, or a combination of these. Some methods are *residue*-centric, and first predict at the residue-level, use a threshold to select high-probability ligand-binding residues, and then cluster them into pockets. Residue-centric methods include IF-SitePred, or GrASP. Other methods (*pocket*-centric) directly predict the location or shape of the pocket, without the need of predicting at the residue level first. Some of these methods can use their pocket-level prediction to report

residue *ligandability* scores, e.g., P2Rank$_{CONS}$, P2Rank, PocketFinder^{+}, Ligsite^{+} or Surfnet^{+}, and others, such as VN-EGNN, PUResNet, DeepPocket, or fpocket do not report residue ligandability scores.

Prior to evaluation, non-redundant sets of predictions were generated for VN-EGNN, IF-SitePred and DeepPocket$_{SEG}$, as these methods generate a considerable proportion of *redundant* predictions. A predicted pocket i is considered redundant if there exists a pocket $j \neq i$ so that the distance between their centroids $D_{i,j} \leq 5$Å or their residue overlap $JI_{i,j} > 0.75$, i.e., they share at least ¾ of their residues. Refer to Supplementary Figure 9 for the closest predicted sites for each method.

PUResNet, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+} do not score, nor explicitly rank their pockets, and so they were taken in the order given by their pocket ID. This means that when sorting across the dataset, the order of all pockets with the same rank is arbitrary. To obtain a score for these pockets, multiple strategies were employed. Firstly, a pocket score was obtained as the number of pocket amino acids, resulting in variants PUResNet$_{AA}$, PocketFinder$^{+}_{AA}$, Ligsite$^{+}_{AA}$ and Surfnet$^{+}_{AA}$. Secondly, PRANK pocket scoring was employed, resulting in variants PUResNet$_{PRANK}$, PocketFinder$^{+}_{PRANK}$, Ligsite$^{+}_{PRANK}$ and Surfnet$^{+}_{PRANK}$. fpocket predictions were also re-scored with PRANK, resulting in fpocket$_{PRANK}$. Additionally, for IF-SitePred pocket scores were calculated as the sum of squares (SS) of the ligandability scores (LS_i), calculated with (Eq. 10), of the K residues on a site (Eq. 12) resulting in IF-SitePred$_{RESC}$. For PocketFinder^{+}, Ligsite^{+}, and Surfnet^{+} the same was done but instead of residue scores, grid point scores (GS_i) were used (Eq. 13). This resulted in further variants PocketFinder$^{+}_{SS}$, Ligsite$^{+}_{SS}$, and Surfnet$^{+}_{SS}$. This is the same approach introduced by Krivák, *et al.*, 2015 [37] and later adopted by Smith *et al.*, 2024 [52].

$$SS_{IF-SitePred} = \sum_{i=1}^{K} LS_i^{\,2} \quad (12)$$

$$SS_{PocketFinder^{+},\, Ligsite^{+},\, Surfnet^{+}} = \sum_{i=1}^{K} GS_i^{\,2} \quad (13)$$

Residue Level Predictions

GrASP, P2RANK$_{CONS}$, P2Rank, PocketFinder^{+}, Ligsite^{+} and Surfnet^{+} all offer residue ligandability scores. Additionally, a ligandability score was derived for IF-SitePred using

Equation 10. Given a ligandability threshold t_{LS}, a residue with a ligandability score LS_i is classified as "positive" if $LS_i > t_{LS}$. Conversely, the residue is classed as "negative" if $LS_i \leq t_{LS}$. Further stratification results from comparing the predictions to the LIGYSIS reference dataset.

- True Positive (TP): residue classified as *positive* that binds a ligand according to the reference.
- False Positive (FP): residue classified as *positive* that does not bind a ligand in the reference.
- True Negative (TN): residue classified as *negative* that does not bind a ligand.
- False Negative (FN): residue classified as *negative* but is known to bind a ligand according to the data.

With these four classes, true positive rate (TPR), false positive rate (FPR) (Eq. 14), precision and recall (Eq. 15) can be calculated and receiver operating characteristic (ROC), precision-recall (PR) curves plotted. Area under the curve (AUC) for ROC and average precision (AP) for PR are reported. Baselines for these are 0.5 and the proportion of true positive labels, respectively.

$$TPR = \frac{TP}{TP + FN} \quad , \quad FPR = \frac{FP}{FP + TN} \tag{14}$$

$$precision = \frac{TP}{TP + FP} \quad , \quad recall = \frac{TP}{TP + FN} \tag{15}$$

Pocket binary labels (0: no pocket residue; 1: pocket residue) can determine TP, FP, TN and FN for each residue in a protein chain P_i. Across all residues in P_i, an F1 score $F1_i$ is computed, which combines precision and recall into a unified metric, capturing the accuracy and completeness of predictions at the residue level (Eq. 16). The Matthews Correlation Coefficient (MCC) [133] (Eq. 17) was also calculated. The median F1 score and MCC across the dataset proteins is reported for each method.

$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{16}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{17}$$

Pocket Level Predictions

There are no *negatives* predicted at the pocket level of ligand binding site prediction, only *positives*. A positive is a predicted pocket, which will be true (TP) or false (FP) depending

on whether it is observed in the reference data. False negatives are those pockets observed in the reference data that are not predicted. They are the pockets the method fails to predict, and therefore, are not scored. A true negative would be a "non-pocket" that is not predicted. This can't be quantified easily and even if it was, it would not be scored by the method, as it is not predicted. For this reason, in this context, neither TPR, nor FPR can be calculated. Consequently, ROC/AUC can't be utilised to assess ligand binding site prediction at the pocket level. False negatives are known, but not scored, and therefore PR/AUC is not an option either. What can be calculated is the recall given a certain criterion. In this case, because of the nature of the LIGYSIS dataset, where defined sites result from the clustering of multiple ligands, the distance between the predicted pocket centroid to the observed binding site (DCC) was chosen as a criterion.

For each observed binding site in our reference dataset, the "best" prediction for each method is chosen. This is defined as the prediction with the minimum Euclidean distance to the observed pocket centroid or DCC. Once the observed-predicted pairs were obtained, only those with DCC ≤ 12 Å were considered as correct predictions. A threshold of 12Å was chosen as 4Å is too strict a threshold when using DCC (See Supplementary Note 6 and Figures 10-14). A threshold of 4Å works well for the distance to closest ligand atom (DCA) but does not for DCC. The top-N and N+2 ranking predictions were considered to calculate success rate, or recall (Eq. 18), and maximum recall was calculated by considering *all* predictions, regardless of their score or rank. N represents the number of observed sites for a given protein.

$$Success\ rate\ (recall) = \frac{observed\ sites\ with\ predicted\ site\ DCC\ \leq 12\text{Å}}{observed\ sites} \quad (18)$$

Additionally, instead of conventional ROC, ROC100 [135, 136] can be used to measure the predictive performance of the methods. To do this, for each method, *all* predictions across dataset proteins were ranked based on the pocket score and cumulative true positives were plotted against cumulative false positives until 100 false positives are reached. In a similar way, a precision curve can be calculated by taking the top-N, in this case N = 1,000, predictions, which measures how precision changes as more predictions with lower scores are considered. This is indicative of how informative pocket scores are.

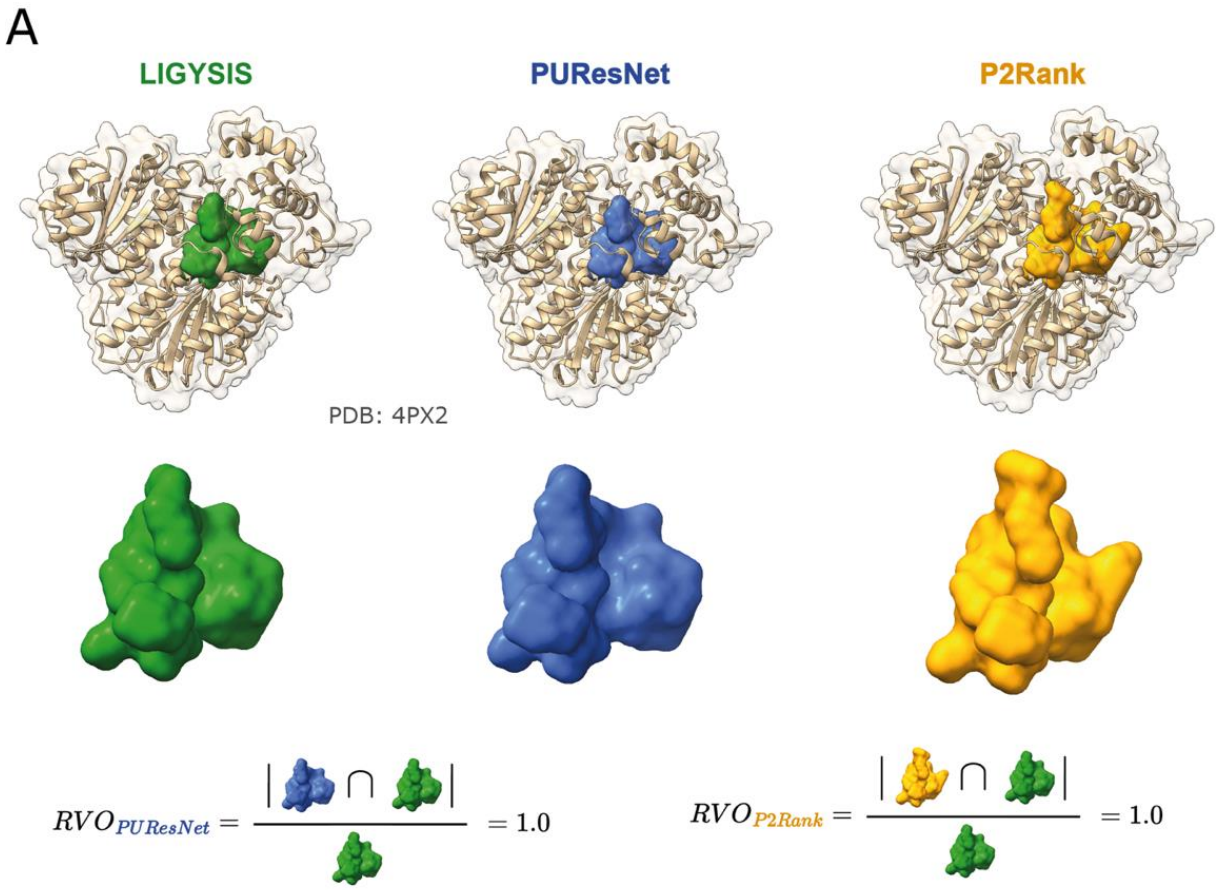


Figure 13. Relative Volume Overlap (RVO) calculation. (A) Example of two very accurate predictions by

PUResNet and P2Rank on PDB: 4PX2 (Jordan, S.R., Chmait, S., 2015). Pocket volumes are calculated with POVME 2.0 [131] and represented by coloured surfaces. These volumes result from the addition of unit volume spheres on a grid. To obtain the RVO, the intersection of these spheres between predicted and observed site is divided by the number of observed pocket spheres. Both predictions cover the entirety of the observed pocket volume; (B) GrASP and VN-EGNN predictions of a site on PDB: 2ZOX [134]. Th volumes of these predicted sites overlap less with the observed site: RVO = 0.67 for GrASP and RVO = 0.11 for VN-EGNN.

To measure the similarity in shape and residue membership between the predicted and observed pockets, relative residue overlap (RRO) and relative volume overlap (RVO) were employed. For an observed-predicted pocket pair, RRO represents the proportion of observed ligand-binding residues (R_o) that are covered by the predicted pocket residues (R_p) (Eq. 19). The POVME output was utilised for the calculation of RVO (Figure 13). POVME defines the volume of a pocket as a series of equidistantly spaced spheres of unit volume. As predictions by the different methods were on the same coordinate reference, these pocket volume spheres were already aligned, and the volume overlap was calculated simply as the proportion of spheres in the observed pocket (V_o) that overlap with the predicted pocket spheres (V_p) (Eq. 20).

$$RRO = \frac{|R_p \cap R_o|}{R_o} \quad (19)$$

$$RVO = \frac{|V_p \cap V_o|}{V_o} \quad (20)$$

Statistics and Reproducibility

VN-EGNN was installed and run locally from https://github.com/ml-jku/vnegnn. Likewise, for IF-SitePred: https://github.com/annacarbery/binding-sites. GrASP was obtained from https://github.com/tiwarylab/GrASP and predictions generated using their Google Colab Notebook. PUResNet predictions were obtained through the PUResNet v2.0 web server: https://nsclbio.jbnu.ac.kr/tools/jmol. DeepPocket was installed and executed locally: https://github.com/devalab/DeepPocket. P2Rank v2.4.2 was used to run all predictions as well as PRANK re-scoring: https://github.com/rdk/p2rank. fpocket v4.0 was installed via Conda: https://anaconda.org/conda-forge/fpocket. For PocketFinder, Ligsite and Surfnet, the ConCavity v0.1 "+" re-implementations were employed: https://compbio.cs.princeton.edu/concavity/.

Other recent methods including SiteRadar [49], NodeCoder [48], GLPocket [50], RefinePocket [137], EquiPocket [51], RecurPocket [46], PointSite [45], DeepSurf [44], Kalasanty [42], GRaSP [40], BiteNet [41] or DeepSite [38] were not included in this analysis due to technical reasons. Open-source methods with publicly accessible code, clear installation instructions, well defined dependencies, and accessible command line interfaces were prioritised in this work.

ChimeraX v1.7.1 [138] was used for structural visualisation in all figures unless otherwise stated, in which case PyMOL v2.5.2 was employed [59].

List of Abbreviations

017: Darunavir

3D: Three-Dimensional

AA: amino acid

ADP: Adenosine-5'-diphosphate

AP: Average Precision

ATP: Adenosine-5’-triphosphate

AUC: Area Under the Curve

BGC: Glucose

CI: Confidence Interval

CLR: Cholesterol

CM: Centre of Mass

CNN: Convolutional Neural Network

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

DCA: Distance to Closest ligand Atom

DCC: Distance Centroid to Centroid

DNN: Deep Neural Network

DRN: Deep Residual Network

EGNN: Equivariant Graph Neural Network

FAD: Flavin-adenine dinucleotide

FMN: Flavin mononucleotide

FN: False Negative

FP: False Positive

FUC: Fucose

GAL: Galactose

GAT: Graph Attention network

GNN: Graph Neural Network

HEM: Haem

IOU: Intersection Over Union

IQR: Interquartile Range

I_{rel}: Relative Intersection

JI: Jaccard Index

LightGBM: Light Gradient Boosting Machine

LJ: Lennard-Jones

MAN: Mannose

MCC: Matthews Correlation Coefficient

MCD: Minimum Centroid Distance

MRO: Maximum Residue Overlap

NAD: Nicotinamide-adenine-dinucleotide

NAP: Nicotinamide-adenine-dinucleotide phosphate

N-mer: protein peptide of N amino acids

NR: Non-redundant

PCA: Principal Component Analysis

PLP: Vitamin B6 phosphate

PR: Precision Recall curve

R_g: Radius of Gyration

ROC: Receiver Operating Characteristic

RRO: Relative Residue Overlap

RVO: Relative Volume Overlap

RSA: Relative Solvent Accessibility

SD: Standard Deviation

SAH: S-Adenosyl-L-homocysteine

SAS: Solvent Accessible Surface

SS: Sum of Squares

TN: True Negative

TP: True Positive

VN: Virtual Node

VR: Volume Ratio

XYP: Xylose

Data Availability

The main results tables and files necessary to replicate the analysis described in this paper can be found here: https://doi.org/10.5281/zenodo.13121414 [139].

Code Availability

Software developed to carry out this analysis is found in our GitHub repository: https://github.com/bartongroup/LBS-comparison.

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by grants to G.J.B. from UKRI-Biotechnology and Biological Sciences [BB/J019364/1; BB/R014752/1] and Wellcome Trust [101651/Z/13/Z; 218259/Z/19/Z]. J.S.U. was supported by a BBSRC EASTBIO Ph.D. Studentship [BB/J01446X/1].

Author Contributions

G.J.B. and J.S.U., conceived, designed, and developed the research. J.S.U. analysed the data. J.S.U. developed the software. J.S.U. and G.J.B. wrote, reviewed and edited the manuscript. G.J.B. secured funding and supervised.

Acknowledgements

We thank Florian Sestak, Rishal Aggarwal, Jeevan Kandel, Drs. Zachary Smith, Radoslav Krivák, and Profs. Günter Klambauer, Pratyush Tiwary and Deva Priyakumar, for their insightful discussions and help on how to run and understand their methods. We also appreciate the authors of the other methods reviewed in this paper for their contributions to the field. We extend our gratitude to Drs. James Abbot, Michele Tinti, Marek Gierliński, Stuart MacGowan, James Procter and James Webster for their suggestions. We also thank the IT service of the University of Dundee for their support of the HPC infrastructure this study was carried out on.

References

1. Congreve, M., et al., *A 'rule of three' for fragment-based lead discovery?* Drug Discov Today, 2003. **8**(19): p. 876-7.
2. Rees, D.C., et al., *Fragment-based lead discovery.* Nat Rev Drug Discov, 2004. **3**(8): p. 660-72.
3. Murray, C.W. and D.C. Rees, *The rise of fragment-based drug discovery.* Nat Chem, 2009. **1**(3): p. 187-92.
4. Schiebel, J., et al., *Six Biophysical Screening Methods Miss a Large Proportion of Crystallographically Discovered Fragment Hits: A Case Study.* ACS Chem Biol, 2016. **11**(6): p. 1693-701.
5. Utges, J.S., et al., *Classification of likely functional class for ligand binding sites identified from fragment screening.* Commun Biol, 2024. **7**(1): p. 320.
6. Volkamer, A., et al., *Analyzing the topology of active sites: on the prediction of pockets and subpockets.* J Chem Inf Model, 2010. **50**(11): p. 2041-52.
7. Le Guilloux, V., P. Schmidtke, and P. Tuffery, *Fpocket: an open source platform for ligand pocket detection.* BMC Bioinformatics, 2009. **10**: p. 168.
8. Hendlich, M., F. Rippmann, and G. Barnickel, *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.* J Mol Graph Model, 1997. **15**(6): p. 359-63, 389.
9. Laskowski, R.A., *SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.* J Mol Graph, 1995. **13**(5): p. 323-30, 307-8.
10. Levitt, D.G. and L.J. Banaszak, *POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids.* J Mol Graph, 1992. **10**(4): p. 229-34.
11. Kleywegt, G.J. and T.A. Jones, *Detection, delineation, measurement and display of cavities in macromolecular structures.* Acta Crystallogr D Biol Crystallogr, 1994. **50**(Pt 2): p. 178-85.
12. Liang, J., H. Edelsbrunner, and C. Woodward, *Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design.* Protein Sci, 1998. **7**(9): p. 1884-97.
13. Brady, G.P., Jr. and P.F. Stouten, *Fast prediction and visualization of protein binding pockets with PASS.* J Comput Aided Mol Des, 2000. **14**(4): p. 383-401.
14. Weisel, M., E. Proschak, and G. Schneider, *PocketPicker: analysis of ligand binding-sites with shape descriptors.* Chem Cent J, 2007. **1**: p. 7.
15. An, J., M. Totrov, and R. Abagyan, *Pocketome via comprehensive identification and classification of ligand binding envelopes.* Mol Cell Proteomics, 2005. **4**(6): p. 752-61.
16. Goodford, P.J., *A computational procedure for determining energetically favorable binding sites on biologically important macromolecules.* J Med Chem, 1985. **28**(7): p. 849-57.
17. An, J., M. Totrov, and R. Abagyan, *Comprehensive identification of "druggable" protein ligand binding sites.* Genome Inform, 2004. **15**(2): p. 31-41.
18. Laurie, A.T. and R.M. Jackson, *Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.* Bioinformatics, 2005. **21**(9): p. 1908-16.

19. Ghersi, D. and R. Sanchez, *EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures*. Bioinformatics, 2009. **25**(23): p. 3185-6.
20. Ngan, C.H., et al., *FTSite: high accuracy detection of ligand binding sites on unbound protein structures*. Bioinformatics, 2012. **28**(2): p. 286-7.
21. Armon, A., D. Graur, and N. Ben-Tal, *ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information*. J Mol Biol, 2001. **307**(1): p. 447-63.
22. Pupko, T., et al., *Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues*. Bioinformatics, 2002. **18 Suppl 1**: p. S71-7.
23. Xie, Z.R. and M.J. Hwang, *Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles*. Bioinformatics, 2012. **28**(12): p. 1579-85.
24. Zvelebil, M.J., et al., *Prediction of protein secondary structure and active sites using the alignment of homologous sequences*. J Mol Biol, 1987. **195**(4): p. 957-61.
25. Wass, M.N., L.A. Kelley, and M.J. Sternberg, *3DLigandSite: predicting ligand-binding sites using similar structures*. Nucleic Acids Res, 2010. **38**(Web Server issue): p. W469-73.
26. Roy, A., J. Yang, and Y. Zhang, *COFACTOR: an accurate comparative algorithm for structure-based protein function annotation*. Nucleic Acids Res, 2012. **40**(Web Server issue): p. W471-7.
27. Yang, J., A. Roy, and Y. Zhang, *Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment*. Bioinformatics, 2013. **29**(20): p. 2588-95.
28. Lee, H.S. and W. Im, *Ligand binding site detection by local structure alignment and its performance complementarity*. J Chem Inf Model, 2013. **53**(9): p. 2462-70.
29. Brylinski, M. and W.P. Feinstein, *eFindSite: improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands*. J Comput Aided Mol Des, 2013. **27**(6): p. 551-67.
30. Huang, B. and M. Schroeder, *LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation*. BMC Struct Biol, 2006. **6**: p. 19.
31. Glaser, F., et al., *A method for localizing ligand binding pockets in protein structures*. Proteins, 2006. **62**(2): p. 479-88.
32. Halgren, T.A., *Identifying and characterizing binding sites and assessing druggability*. J Chem Inf Model, 2009. **49**(2): p. 377-89.
33. Capra, J.A., et al., *Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure*. PLoS Comput Biol, 2009. **5**(12): p. e1000585.
34. Huang, B., *MetaPocket: a meta approach to improve protein ligand binding site prediction*. OMICS, 2009. **13**(4): p. 325-30.
35. Bray, T., et al., *SitesIdentify: a protein functional site prediction tool*. BMC Bioinformatics, 2009. **10**: p. 379.
36. Brylinski, M. and J. Skolnick, *FINDSITE: a threading-based approach to ligand homology modeling*. PLoS Comput Biol, 2009. **5**(6): p. e1000405.

37. Krivak, R. and D. Hoksza, *Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features*. J Cheminform, 2015. **7**: p. 12.
38. Jimenez, J., et al., *DeepSite: protein-binding site predictor using 3D-convolutional neural networks*. Bioinformatics, 2017. **33**(19): p. 3036-3042.
39. Krivak, R. and D. Hoksza, *P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure*. J Cheminform, 2018. **10**(1): p. 39.
40. Santana, C.A., et al., *GRaSP: a graph-based residue neighborhood strategy to predict binding sites*. Bioinformatics, 2020. **36**(Suppl_2): p. i726-i734.
41. Kozlovskii, I. and P. Popov, *Spatiotemporal identification of druggable binding sites using deep learning*. Commun Biol, 2020. **3**(1): p. 618.
42. Stepniewska-Dziubinska, M.M., P. Zielenkiewicz, and P. Siedlecki, *Improving detection of protein-ligand binding sites with 3D segmentation*. Sci Rep, 2020. **10**(1): p. 5035.
43. Kandel, J., H. Tayara, and K.T. Chong, *PUResNet: prediction of protein-ligand binding sites using deep residual neural network*. J Cheminform, 2021. **13**(1): p. 65.
44. Mylonas, S.K., A. Axenopoulos, and P. Daras, *DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins*. Bioinformatics, 2021. **37**(12): p. 1681-1690.
45. Yan, X., et al., *PointSite: A Point Cloud Segmentation Tool for Identification of Protein Ligand Binding Atoms*. J Chem Inf Model, 2022. **62**(11): p. 2835-2845.
46. Li, P., et al. *RecurPocket: Recurrent Lmser Network with Gating Mechanism for Protein Binding Site Detection*. in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2022.
47. Aggarwal, R., et al., *DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks*. J Chem Inf Model, 2022. **62**(21): p. 5069-5079.
48. Abdollahi, N., et al., *NodeCoder: a graph-based machine learning platform to predict active sites of modeled protein structures*. arXiv [q-bio.QM], 2023.
49. Evteev, S.A., A.V. Ereshchenko, and Y.A. Ivanenkov, *SiteRadar: Utilizing Graph Machine Learning for Precise Mapping of Protein-Ligand-Binding Sites*. J Chem Inf Model, 2023. **63**(4): p. 1124-1132.
50. Li, P., et al., *GLPocket: A Multi-Scale Representation Learning Approach for Protein Binding Site Prediction*, in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, E. Elkind, Editor. 2023, International Joint Conferences on Artificial Intelligence Organization. p. 4821-4828.
51. Zhang, Y., et al., *EquiPocket: an E(3)-Equivariant Geometric Graph Neural Network for Ligand Binding Site Prediction*. arXiv [q-bio.BM], 2023.
52. Smith, Z., et al., *Graph Attention Site Prediction (GrASP): Identifying Druggable Binding Sites Using Graph Neural Networks with Attention*. J Chem Inf Model, 2024. **64**(7): p. 2637-2644.
53. Carbery, A., et al., *Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures*. J Cheminform, 2024. **16**(1): p. 32.

54. Sestak, F., et al., *VN-EGNN: E(3)-Equivariant Graph Neural Networks with Virtual Nodes Enhance Protein Binding Site Identification*. arXiv [cs.LG], 2024.
55. Jeevan, K., et al., *PUResNetV2.0: a deep learning model leveraging sparse representation for improved ligand binding site prediction*. Journal of Cheminformatics, 2024. **16**(1): p. 66.
56. Rives, A., et al., *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences*. Proc Natl Acad Sci U S A, 2021. **118**(15).
57. Hsu, C., et al., *Learning inverse folding from millions of predicted structures*, in *Proceedings of the 39th International Conference on Machine Learning*, C. Kamalika, et al., Editors. 2022, PMLR: Proceedings of Machine Learning Research. p. 8946--8970.
58. Ke, G., et al., *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, in *Advances in Neural Information Processing Systems*, I. Guyon, et al., Editors. 2017, Curran Associates, Inc.
59. Schrödinger, L.L.C., *The PyMOL Molecular Graphics System, Version 1.8*. 2015.
60. Ester, M., et al., *A density-based algorithm for discovering clusters in large spatial databases with noise*, in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996, AAAI Press: Portland, Oregon. p. 226–231.
61. Jendele, L., et al., *PrankWeb: a web server for ligand binding site prediction and visualization*. Nucleic Acids Res, 2019. **47**(W1): p. W345-W349.
62. Capra, J.A. and M. Singh, *Predicting functionally important residues from sequence conservation*. Bioinformatics, 2007. **23**(15): p. 1875-82.
63. Jones, J.E. and S. Chapman, *On the determination of molecular fields.—I. From the variation of the viscosity of a gas with temperature*. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 1924. **106**(738): p. 441-462.
64. Yang, J., A. Roy, and Y. Zhang, *BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions*. Nucleic Acids Res, 2013. **41**(Database issue): p. D1096-103.
65. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state*. J Mol Biol, 2007. **372**(3): p. 774-97.
66. Armstrong, D.R., et al., *PDBe: improved findability of macromolecular structure data in the PDB*. Nucleic Acids Res, 2020. **48**(D1): p. D335-D343.
67. Utges, J.S., *bartongroup/FRAGSYS: Second release*. 2024, Zenodo.
68. Pickens, J.C., et al., *Anchor-based design of improved cholera toxin and E. coli heat-labile enterotoxin receptor binding antagonists that display multiple binding modes*. Chem Biol, 2002. **9**(2): p. 215-24.
69. Hofmann, E., et al., *Structural basis of light harvesting by carotenoids: peridinin-chlorophyll-protein from Amphidinium carterae*. Science, 1996. **272**(5269): p. 1788-91.
70. Williams, L.K., et al., *Order and disorder: differential structural impacts of myricetin and ethyl caffeate on human amylase, an antidiabetic target*. J Med Chem, 2012. **55**(22): p. 10177-86.
71. Wilson, E.B., *Probable Inference, the Law of Succession, and Statistical Inference*. Journal of the American Statistical Association, 1927. **22**(158): p. 209-212.

72. Chen, C.I., et al., *Structure of human POFUT2: insights into thrombospondin type 1 repeat fold and O-fucosylation*. EMBO J, 2012. **31**(14): p. 3183-97.
73. Liu, R.J., et al., *Structural basis for substrate binding and catalytic mechanism of a human RNA:m5C methyltransferase NSun6*. Nucleic Acids Res, 2017. **45**(11): p. 6684-6697.
74. Ramirez-Rios, S., et al., *VASH1-SVBP and VASH2-SVBP generate different detyrosination profiles on microtubules*. J Cell Biol, 2023. **222**(2).
75. Lim, S.M., et al., *The structure and catalytic mechanism of human sphingomyelin phosphodiesterase like 3a--an acid sphingomyelinase homologue with a novel nucleotide hydrolase activity*. FEBS J, 2016. **283**(6): p. 1107-23.
76. Finer-Moore, J., et al., *Crystal Structure of the Human tRNA m(1)A58 Methyltransferase-tRNA(3)(Lys) Complex: Refolding of Substrate tRNA Allows Access to the Methylation Target*. J Mol Biol, 2015. **427**(24): p. 3862-76.
77. Blasiak, L.C., et al., *Crystal structure of the non-haem iron halogenase SyrB2 in syringomycin biosynthesis*. Nature, 2006. **440**(7082): p. 368-71.
78. Mascarenhas, R., et al., *Architecture of the human G-protein-methylmalonyl-CoA mutase nanoassembly for B(12) delivery and repair*. Nat Commun, 2023. **14**(1): p. 4332.
79. Milani, M., et al., *FAD-binding site and NADP reactivity in human renalase: a new enzyme involved in blood pressure regulation*. J Mol Biol, 2011. **411**(2): p. 463-73.
80. Liu, L., et al., *Autophosphorylation transforms DNA-PK from protecting to processing DNA ends*. Mol Cell, 2022. **82**(1): p. 177-189 e4.
81. Comajuncosa-Creus, A., et al., *Comprehensive detection and characterization of human druggable pockets through novel binding site descriptors*. bioRxiv, 2024: p. 2024.03.14.584971.
82. Durairaj, J., et al., *PLINDER: The protein-ligand interactions dataset and evaluation resource*. bioRxiv, 2024: p. 2024.07.17.603955.
83. Stärk, H., et al., *EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction*. arXiv [q-bio.BM], 2022.
84. Qiao, Z., et al., *State-specific protein–ligand complex structure prediction with a multiscale deep generative model*. Nature Machine Intelligence, 2024. **6**(2): p. 195-208.
85. Schneuing, A., et al., *Structure-based Drug Design with Equivariant Diffusion Models*. arXiv [q-bio.BM], 2023.
86. UniProt, C., *UniProt: a worldwide hub of protein knowledge*. Nucleic Acids Res, 2019. **47**(D1): p. D506-D515.
87. Berman, H., K. Henrick, and H. Nakamura, *Announcing the worldwide Protein Data Bank*. Nat Struct Biol, 2003. **10**(12): p. 980.
88. ww, P.D.B.c., *Protein Data Bank: the single global archive for 3D macromolecular structure data*. Nucleic Acids Res, 2019. **47**(D1): p. D520-D528.
89. Varadi, M., et al., *PDBe and PDBe-KB: Providing high-quality, up-to-date and integrated resources of macromolecular structures to support basic and applied research and education*. Protein Sci, 2022. **31**(10): p. e4439.
90. consortium, P.D.-K., *PDBe-KB: a community-driven resource for structural and functional annotations*. Nucleic Acids Res, 2020. **48**(D1): p. D344-D353.
91. consortium, P.D.-K., *PDBe-KB: collaboratively defining the biological context of structural data*. Nucleic Acids Res, 2022. **50**(D1): p. D534-D542.

92. Ellaway, J.I.J., et al., *Identifying protein conformational states in the Protein Data Bank: Toward unlocking the potential of integrative dynamics studies*. Struct Dyn, 2024. **11**(3): p. 034701.
93. MacGowan, S.A., et al., *The Dundee Resource for Sequence Analysis and Structure Prediction*. Protein Sci, 2020. **29**(1): p. 277-297.
94. Jubb, H.C., et al., *Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures*. J Mol Biol, 2017. **429**(3): p. 365-371.
95. Virtanen, P., et al., *SciPy 1.0: fundamental algorithms for scientific computing in Python*. Nat Methods, 2020. **17**(3): p. 261-272.
96. Shannon, C.E., *A Mathematical Theory of Communication*. Bell System Technical Journal, 1948. **27**(3): p. 379-423.
97. Westbrook, J.D., et al., *The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank*. Bioinformatics, 2015. **31**(8): p. 1274-8.
98. Campanacci, V., et al., *Insight into microtubule nucleation from tubulin-capping proteins*. Proceedings of the National Academy of Sciences of the United States of America, 2019. **116**(20): p. 9859-9864.
99. Paul, N., et al., *Recovering the true targets of specific ligands by virtual screening of the protein data bank*. Proteins, 2004. **54**(4): p. 671-80.
100. Kellenberger, E., et al., *sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank*. J Chem Inf Model, 2006. **46**(2): p. 717-27.
101. Meslamani, J., D. Rognan, and E. Kellenberger, *sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins*. Bioinformatics, 2011. **27**(9): p. 1324-6.
102. Desaphy, J., et al., *sc-PDB: a 3D-database of ligandable binding sites--10 years on*. Nucleic Acids Res, 2015. **43**(Database issue): p. D399-404.
103. Mottin, M., et al., *Molecular dynamics simulations of Zika virus NS3 helicase: Insights into RNA binding site activity*. Biochem Biophys Res Commun, 2017. **492**(4): p. 643-651.
104. Hu, L., et al., *Binding MOAD (Mother Of All Databases)*. Proteins, 2005. **60**(3): p. 333-40.
105. Benson, M.L., et al., *Binding MOAD, a high-quality protein-ligand database*. Nucleic Acids Res, 2008. **36**(Database issue): p. D674-8.
106. Ahmed, A., et al., *Recent improvements to Binding MOAD: a resource for protein-ligand binding affinities and structures*. Nucleic Acids Res, 2015. **43**(Database issue): p. D465-9.
107. Smith, R.D., et al., *Updates to Binding MOAD (Mother of All Databases): Polypharmacology Tools and Their Utility in Drug Repurposing*. J Mol Biol, 2019. **431**(13): p. 2423-2433.
108. Hubbard, T.J., et al., *SCOP: a structural classification of proteins database*. Nucleic Acids Res, 1997. **25**(1): p. 236-9.
109. Hubbard, T.J., et al., *SCOP, Structural Classification of Proteins database: applications to evaluation of the effectiveness of sequence alignment methods and statistics of protein structural data*. Acta Crystallogr D Biol Crystallogr, 1998. **54**(Pt 6 Pt 1): p. 1147-54.
110. Lo Conte, L., et al., *SCOP: a structural classification of proteins database*. Nucleic Acids Res, 2000. **28**(1): p. 257-9.

111. Chen, K., et al., *A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds*. Structure, 2011. **19**(5): p. 613-21.
112. Hartshorn, M.J., et al., *Diverse, high-quality test set for the validation of protein-ligand docking performance*. J Med Chem, 2007. **50**(4): p. 726-41.
113. Zhang, Z., et al., *Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction*. Bioinformatics, 2011. **27**(15): p. 2083-8.
114. Schmidtke, P., et al., *Large-scale comparison of four binding site detection algorithms*. J Chem Inf Model, 2010. **50**(12): p. 2191-200.
115. Varadi, M., et al., *AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models*. Nucleic Acids Res, 2022. **50**(D1): p. D439-D444.
116. Wang, R., et al., *The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures*. J Med Chem, 2004. **47**(12): p. 2977-80.
117. Wang, R., et al., *The PDBbind database: methodologies and updates*. J Med Chem, 2005. **48**(12): p. 4111-9.
118. Cheng, T., et al., *Comparative assessment of scoring functions on a diverse test set*. J Chem Inf Model, 2009. **49**(4): p. 1079-93.
119. Li, Y., et al., *Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set*. J Chem Inf Model, 2014. **54**(6): p. 1700-16.
120. Liu, Z., et al., *PDB-wide collection of binding data: current status of the PDBbind database*. Bioinformatics, 2015. **31**(3): p. 405-12.
121. Liu, Z., et al., *Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions*. Acc Chem Res, 2017. **50**(2): p. 302-309.
122. Hotelling, H., *Analysis of a complex of statistical variables into principal components*. Journal of Educational Psychology, 1933. **24**(6): p. 417-441.
123. Fixman, M., *Radius of Gyration of Polymer Chains*. The Journal of Chemical Physics, 1962. **36**(2): p. 306-310.
124. Chen, C.R. and G.I. Makhatadze, *ProteinVolume: calculating molecular van der Waals and void volumes in proteins*. BMC Bioinformatics, 2015. **16**(1): p. 101.
125. Jubb, H., *pdbtools*. 2019.
126. Jakubec, D., et al., *PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures*. Nucleic Acids Res, 2022. **50**(W1): p. W593-W597.
127. Schmidtke, P., et al., *fpocket: online tools for protein ensemble pocket detection and tracking*. Nucleic Acids Res, 2010. **38**(Web Server issue): p. W582-9.
128. Jaccard, P., *Distribution de la florine alpine dans la Bassin de Dranses et dans quelques regiones voisines*. Bulletin de la Societe Vaudoise des Sciences Naturelles, 1901. **37**: p. 241-272.
129. Jaccard, P., *THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1*. New Phytologist, 1912. **11**(2): p. 37-50.
130. Durrant, J.D., C.A. de Oliveira, and J.A. McCammon, *POVME: an algorithm for measuring binding-pocket volumes*. J Mol Graph Model, 2011. **29**(5): p. 773-6.

131. Durrant, J.D., et al., *POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics*. J Chem Theory Comput, 2014. **10**(11): p. 5047-5056.
132. Wagner, J.R., et al., *POVME 3.0: Software for Mapping Binding Pocket Flexibility*. J Chem Theory Comput, 2017. **13**(9): p. 4584-4592.
133. Matthews, B.W., *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*. Biochim Biophys Acta, 1975. **405**(2): p. 442-51.
134. Noguchi, J., et al., *Crystal structure of the covalent intermediate of human cytosolic beta-glucosidase*. Biochem Biophys Res Commun, 2008. **374**(3): p. 549-52.
135. Webber, C. and G.J. Barton, *Increased coverage obtained by combination of methods for protein sequence database searching*. Bioinformatics, 2003. **19**(11): p. 1397-403.
136. Scott, M.S. and G.J. Barton, *Probabilistic prediction and ranking of human protein-protein interactions*. BMC Bioinformatics, 2007. **8**: p. 239.
137. Liu, Y., et al., *RefinePocket: An Attention-Enhanced and Mask-Guided Deep Learning Approach for Protein Binding Site Prediction*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023. **20**(5): p. 3314-3321.
138. Pettersen, E.F., et al., *UCSF ChimeraX: Structure visualization for researchers, educators, and developers*. Protein Sci, 2021. **30**(1): p. 70-82.
139. Sánchez Utgés, J., *LBS-Comparison results*. 2024, Zenodo.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Utges2024LBScomparisonSUPP.pdf
- Tables.docx