PDBspheres - a method for finding 3D similarities in local regions in proteins

Adam Zemla[1*], Jonathan E. Allen[1], Dan Kirshner[2], Felice C. Lightstone[2]

[1]Global Security Computing Applications, Lawrence Livermore National Laboratory, Livermore, CA
[2]Biosciences and Biotechnology Division, Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA

*Corresponding author: Adam Zemla, zemla1@llnl.gov

Abstract

We present a structure-based method for finding and evaluating structural similarities in protein regions relevant to ligand binding. PDBspheres comprises an exhaustive library of protein structure regions ("spheres") adjacent to complexed ligands derived from the Protein Data Bank (PDB), along with methods to find and evaluate structural matches between a protein of interest and spheres in the library. Currently, PDBspheres' library contains more than 2 million spheres, organized to facilitate searches by sequence and/or structure similarity of protein-ligand binding sites or interfaces between interacting molecules. PDBspheres uses the LGA structure alignment algorithm as the main engine for detecting structure similarities between the protein of interest and library spheres. An all-atom structure similarity metric ensures that sidechain placement is taken into account in the PDBspheres' primary assessment of confidence in structural matches. In this paper, we (1) describe the PDBspheres method, (2) demonstrate how PDBspheres can be used to detect and characterize binding sites in protein structures, (3) compare PDBspheres use for binding site prediction with seven other binding site prediction methods using a curated dataset of 2,528 ligand-bound and ligand-free crystal structures, and (4) use PDBspheres to cluster pockets and assess structural similarities among protein binding sites of the 4,876 structures in the "refined set" of PDBbind 2019 dataset. The PDBspheres library is made publicly available for download at https://proteinmodel.org/AS2TS/PDBspheres

1. Introduction

Interactions between proteins and small molecule ligands are a cornerstone of biochemical function. Modern drug discovery often relies on structure-based drug discovery, which requires structural information about the target of interest (typically a protein). When a new structure is obtained, it may be the case that little is known with regard to potential binding sites on that structure. A number of binding site prediction methods attempt to address this issue [25-34]. These can be categorized in three broad sets: (1) template-based methods that use known protein information, (2) physics-based methods that rely on geometry (for example, cavity detection) and/or physicochemical properties (for example, surface energy interactions with probe molecules), and (3) machine learning (ML) - rapidly developing in recent years methods capable of efficiently processing information collected in their training data sets. For ML methods data can come from both experiments and in silico data processing, as described for example in [14], [24], and [35]. PDBspheres can be classified as a template-based method, however it relies solely on local structure conformation similarity – i.e., doesn't utilize any prior information from libraries of sequences, motifs or residues forming binding sites. In this structure template-based prediction method, binding sites are identified exclusively based on the structure similarity between regions from the query protein and pocket spheres from the PDBspheres library. Ligand placement within a predicted pocket is calculated based on a protein-sphere superposition, i.e., an agreement in structure conformation between atoms from the query protein and protein atom coordinates from the template sphere. The main premise of structure template-based binding site prediction methods is that the number of pockets is limited [18], therefore each one of them may serve as a binding site for a large diversity of ligands [19]. However, possible conformations of ligands bound in the pocket are also limited. Indeed, when we predict a ligand-protein conformation we focus on what parts of the ligand (e.g., its core region) need to be in a correct conformation, i.e., a conformation that can be confirmed by experiment and from which the binding affinities can be estimated. Usually such “correct” conformations are shared between ligands that bind a given pocket (at least shared by their “core” regions.) The evaluation of the correctness of the ligand placement within a pocket is not an easy task. For example, when for a given protein two pocket predictions with different ligands inserted need to be evaluated a simple calculation and comparison of the ligand centroids (the center of mass of the bound ligands) can be misleading. This is because the two predictions may differ in assigned placements of the ligands within a pocket (especially when the pocket size is large allowing the ligands to fit in different areas within the cavity), or the ligand sizes and their shapes are different (e.g., some parts of the ligands can be exposed outside the pocket with different orientations [23]; see Fig.1). These difficulties in the assessment of the accuracy of binding site predictions and ligand placements within predicted pockets could be overcome when we focus on the evaluation of residues interacting with the bound ligands.

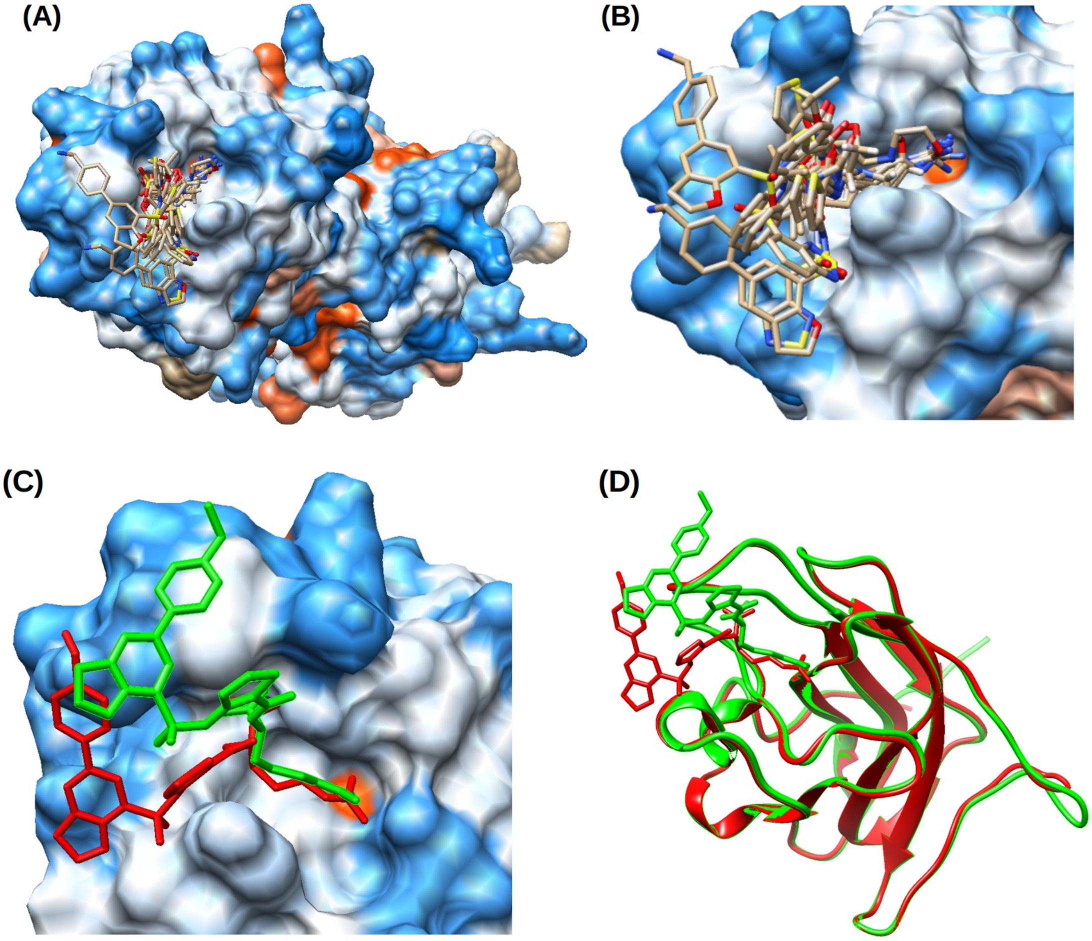


Fig.1. (A) Example of the b1 domain from the Human Neuropilin-1 (PDB 2qqi) where a single pocket can serve as a binding site for many different ligands. (B) Zoom-in to the pocket that can accommodate

ligands of different sizes. A list of ligands includes: 6JY.20 (Arg-7), 6K8.24 (Arg-6), 8DR.32 (EG00229), AAG.15 (M45), AR5.19 (Arg-5), BCN.23 (Bicine), DUE.40 (EG01377), HRG.13 (Arg-1), R40.22 (Arg-4), where ligand information (PDBid.size) is provided with a size representing a number of heavy atoms. (C) Two different poses of the ligand DUE.40 (EG01377) are reported in Powell et al. 2018 [23] and identified by PDBspheres based on two templates: 6fmc and 6fmf. (D) Superposition of two experimentally solved structures of Neuropilin1-b1 domain in complex with EG01377 (in green: PDB 6fmc at resolution 0.9 Å, and in red: PDB 6fmf at resolution 2.8 Å) shows almost identical protein structure conformations while poses of bound ligands differ significantly outside the core. The distance between centroids of the ligand's two orientations when placed in the pocket is more than 4.6 Å.

Knowing the correct "core" conformation of the ligand within the pocket (i.e., the pose of the conserved part) can significantly improve "*in silico*" drug discovery; it can be used in compound screening/docking efforts as a pre-filter for selection of most promising compounds for a more detailed and expensive computational evaluation. Currently, there are millions of compounds screened using docking/MMGBSA approaches to find good candidates for a further (experimental) analysis of potential inhibitors for given targeted proteins. Better predictions of the initial "core" conformation and protein-ligand residue contacts for most suitable compounds can significantly reduce expensive calculations and limit them to the carefully preselected most "promising" compounds only.

Recently, to address a need in proper evaluation of the binding site prediction systems a benchmark dataset (LBPs dataset) of ligand-bound and ligand-free crystal structures for 304 unique protein families (2,528 structures in 1456 "holo" and 1082 "apo" conformations) has been developed [1], [9]. The main criterion for the assessment of the accuracy of predictions of the binding site is the agreement in the set of residues predicted to be in contact with experimentally confirmed ligands (residue contacts derived from protein-ligand co-crystals from PDB [21].) It means that any residue predicted to be part of the binding site which is confirmed in the experimental data is denoted "true positive" (TP), and any residue predicted to be part of the binding site which is not confirmed in experimental data is denoted "false positive" (FP) or over-predictions. Any remaining residues in the experimental data not accounted for in an algorithm's predicted binding site are denoted "false negative" (FN) or under-predictions, and all remaining residues (which are not predicted as part of the binding site and not confirmed in experimental data) are denoted as "true negative" (TN). Both Matthew's Correlation Coefficients (MCCs), and F scores as calculated by formulas (Eqs. 1-2, Section 2.3) have proven to be useful as metrics to represent the predictive power of the various methods.

2. Methods

The PDBspheres system is designed to help assess the similarity between proteins based on their structure similarity in selected local regions (e.g., binding sites, protein interfaces, or any other local regions that might be structurally characteristic of a particular group of proteins). The PDBspheres system has three main components: (1) the PDBspheres library of binding site templates, (2) a structure similarity search algorithm to detect similarity between evaluated local structural regions (e.g., binding sites), and (3) numerical metrics to assess confidence in detected similar regions.

Currently, the PDBspheres library (ver. 2021/10/13) contains 2,002,354 compound binding site models and 67,445 short-peptide binding site models.

The LGA program [10] is used to perform all structure similarity searches. In the process of detecting pocket candidates within proteins structure similarity searches can be performed using all templates in the PDBspheres library (i.e, exhaustive search, testing all 2.0 M pockets from the library), or searches can be performed on a preselected subset of the sphere templates. There are two standard preselection approaches implemented in the PDBspheres system: (1) a set of template spheres can be preselected based on specific targeted ligands (e.g., their names or sizes), or (2) template spheres can be preselected based on sequence similarity between a query protein and protein pockets (sphere templates) in the PDBspheres library. The computational time of the structure processing can vary depending on the size of the protein or the number of template spheres from the PDBspheres library preselected by the system for the pocket detection and similarity evaluation. An exhaustive search (against entire PDBspheres library) can take more than one day on a single processor machine, however with standard preselection procedures the calculations can be completed in less than one hour for medium size proteins. For example, the processing of three protein structures discussed in the manuscript and illustrated on Fig.1 (Neuropilin-1, 158aa), Fig.2 (PL2pro, 318aa), and Fig.6 (Tryptase, 248aa)) took 7 min, 52 min, and 50 min respectively on a single processor with 16 cores machine.

The main structure similarity metrics used to assess confidence in detected pockets are LGA_S (combination of GDT and LCS measures [10]), which evaluates structure similarities on Calpha and/or Cbeta levels, and GDC (Global Distance Calculations [11]), which allows evaluation on an all-atoms level, i.e., extending structure similarity evaluation to the conformation of side-chain atoms.
When applied to predict protein binding sites in a given protein structure, PDBspheres detects ligand-protein binding regions using the pocket/sphere templates in the PDBspheres library constructed from all available structures deposited in the PDB database. After a binding pocket is detected the ligand(s) from matching template(s) is/are inserted into the identified pocket in a query protein to illustrate an approximate location of the ligand. The location is approximate because it is based on the alignment of the query protein with the template protein spheres, and no docking or any energy minimization or structure relaxation are undertaken. Thus, it should be noted that PDBspheres is not a docking system to predict *de novo* ligand poses within a binding site. However, further relaxation of the predicted protein-ligand complex can be considered as the next step to improve a ligand placement within a pocket (which is a subject of continuing PDBspheres system development.)

The clustering of identified pocket-spheres matches and their corresponding ligands characterizes distinct pocket regions within the protein by the sets of residues interacting with inserted ligands. The clustering of interacting residues allows identify overlapping parts of ligands that approximate the ligand's "core" conformation within detected pockets and for each cluster defines a representative set of residues forming a consensus pocket. An example of an identified cluster of pocket-spheres is shown in Fig 2. Each protein may have more than one consensus pocket.

2.1 PDBspheres library

Each entry in the PDBspheres library is a subset of the records in a PDB entry, consisting of the coordinates of all atoms belonging to a "query ligand," and coordinates of all protein atoms belonging to residues near that query ligand (water atoms are excluded.) The protein residue is included if at least one of its atoms is within 12.0 Å of any atom of the query ligand. Previous research indicated that distances of 7.5 Å are sufficient to capture informative functional properties for clustering purposes [12]; however, based on our experimentation using the LGA program for detecting local structure similarities we expanded the distance to 12.0 Å. This size of "template sphere" is sufficient to capture the structural environment of the query ligand for structure comparative needs. Our tests indicated that larger than ~12 Å distance spheres would affect accuracy in calculated structure conformation-based local residue-residue correspondences between template spheres and the query protein within evaluated pockets. On the other hand, the smaller than ~12 Å distance criteria may not provide sufficient fold constraints to capture the uniqueness of searched pockets. To assist in the identification of functional residues, PDBspheres also collects and reports information on protein-ligand interface residues. These residues are identified as those of which at least one atom is within 4.5 Å of any ligand atom.

In the PDBspheres library the 12.0 Å “sphere” entries have been constructed for each ligand in PDB, including peptides, metals, and ions, although the library includes only peptides containing 25 or fewer residues. The library is updated weekly in coordination with new PDB releases. As of 2021/10/13 the library consists of 2,069,796 spheres (binding site templates).

The primary use of the PDBspheres library is to identify “sphere” protein structures that are structurally similar to regions of a query protein structure. The query protein structure may be a complete, multimeric assembly, or it may be a single sub-unit of such an assembly, or even a fragment of a protein as long as it carries enough structural information (local structure conformation formed by atom coordinates) to reflect a structural shape of a putative binding site.

2.2 PDBspheres similarity searches

The fundamental identification method used by PDBspheres is structural alignment of a sphere with the query structure, and the assessment of the structural match. For general use, a comprehensive search – that is, a structural alignment-based search of the entire PDBspheres library (over 2 million sphere templates) – would be computationally expensive. One means to address this issue is to conduct an initial search for matches to a query protein structure on a subset of sphere templates. Then, if needed, the search can be expanded to additional templates from the library. In its current implementation for possibly faster processing, the PDBspheres searches can be restricted to a subset of template spheres selected based on (1) ligand similarity i.e., ligand(s) information, or (2) sequence similarity between the query protein and PDB proteins from which the PDBspheres library entries were derived. In the case of former approach, the sequence similarity searches are conducted using the Smith-Waterman algorithm against FASTA-format sequences of all proteins contributing to the PDBspheres library entries. In the current version of PDBspheres, the Smith-Waterman algorithm “ssearch36” [15] is used.

In the PDBspheres library each sphere template entry includes a number of characteristics such as the ligand name (PDB ligand ID), the number of heavy atoms in the ligand, and the number of residues in the protein fragment forming the protein-ligand "sphere" (that is, the size of the pocket template). Thus, the list of matching spheres can be screened using specific criteria related to the expected pocket size, exact ligand name or ligand similarity. A selection of the template spheres based on the estimated similarity between the expected ligand and the PDB ligands from the PDBspheres system can be quantified by the Tanimoto or Tversky similarity indexes [16]. In our currently implemented approach, the pocket identification is restricted to the template spheres derived from the similarity between the query ligand and the set of PDB ligands collected within the PDBspheres library.

After the identification of a subset of spheres, for example, those matching by sequence similarity and/or ligand similarity, each member of this subset is evaluated by assessing its structure similarity with a query protein. The evaluation is performed in two steps: (1) detection of the region resembling the binding site, and (2) an assessment of the similarity in residue conformations including a conformation in their side chain atoms. The primary search involves a calculation of structural alignment between template spheres and a query protein. The structural alignment is conducted using the LGA program [10] and is calculated using a single point representation of aligning residues (Calpha atoms, Cbeta atoms or any other point that can represent a residue position.) The structural similarity search returns similarity scores (LGA_S scores) and alignments between template spheres for various "pockets" and the query protein. The final evaluation of identified pockets is done by assessing similarities in the conformation of all atoms (including side chain atoms) using the GDC metric [11].

The PDB ligand from the sphere template is translated and rotated according to the transformation that results from aligning the sphere template atoms to the query protein atoms. The ligand atoms are not considered in this alignment, and the crystal ligand conformation is not altered. It means that potential steric clashes between protein residues and atoms of inserted ligands can be observed. The number of possible clashes is reported.

When for a given protein structure and selected set of sphere templates, PDBspheres searches are completed the reported binding site predictions are screened and evaluated by several criteria.
PDBspheres reports the following set of characteristics for all pockets detected within a given protein:

(a) PDB identifier for the template sphere protein matching the query protein pocket,
(b) PDB identifier for the template sphere ligand matching the query protein pocket,
(c) list of residues in the query protein in direct contact with the inserted ligand (residues within the distance <= 4.5 Ångstroms),
(d) coordinates of the centroid of the ligand inserted into the detected pocket in the query protein.

Because the pocket searches can match template spheres to different regions of the query protein all identified pocket candidates are organized into clusters. Within the PDBspheres approach the merging and clustering of detected pockets is performed to satisfy the following two criteria:

(1) within each cluster the pockets are grouped together based on sets of residues interacting with ligands that are in common (more than 80% of ligand contact residues are the same),
(2) pockets and the sets of residues identified as in contact with ligands are merged when the inserted ligands overlap (the distances between ligand centroids are not larger than 2.0 Ångstroms.)

This grouping of pockets and their corresponding inserted ligands allows to define sets of residues interacting with similar ligands, and helps identify overlapping parts of ligands that approximate the ligand's "core" conformation whose location is used to define representative (consensus) pockets within the protein.

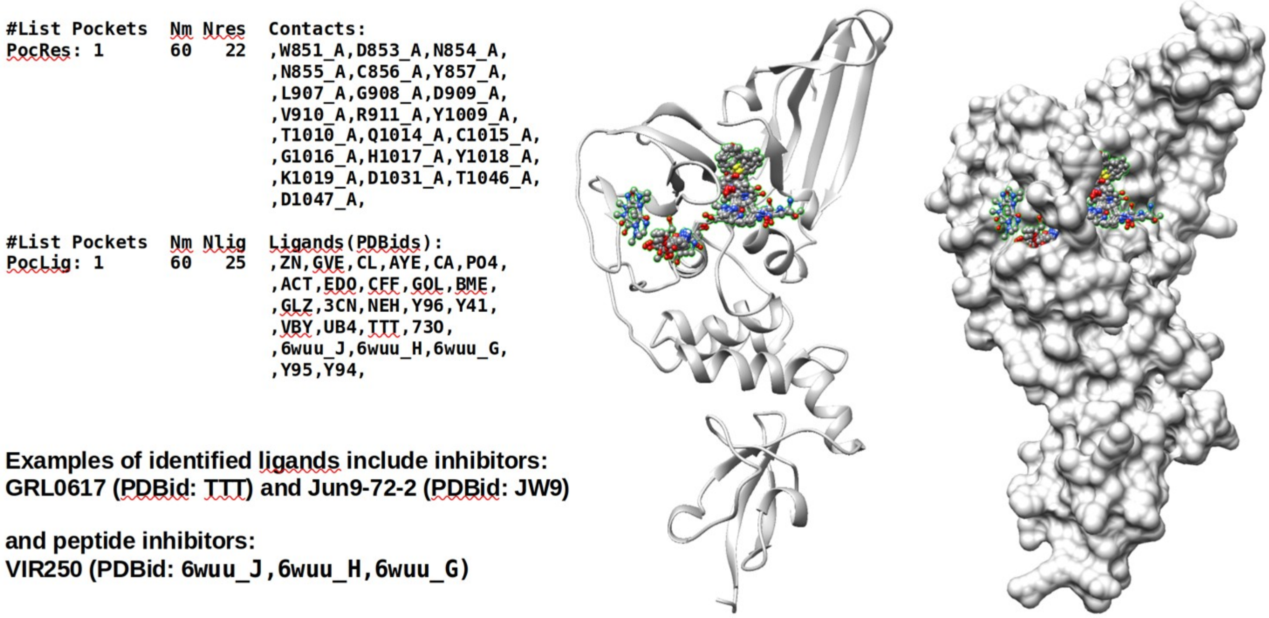
For a given protein the following set of measurements and scorings are provided to help assess the confidence level of the pocket prediction (see examples provided in Fig 4 and supplemental File 1):

- LIGAND – protein-ligand template sphere identifier.
- Ns - number of residues in the protein-ligand template sphere.
- RMSD – root mean square deviation calculated on superimposed Calpha or Cbeta atoms from sphere template and detected protein pocket.
- Nc – number of conserved, i.e., "tightly" superimposed residues between "sphere template" and detected pocket in evaluated protein.
- SeqID – sequence identity in structure aligned residues. Higher value indicates that protein forming a sphere template and our protein might be close homologs.
- LGA – structure similarity based on aligned by LGA program Calpha or Cbeta atoms.
- GDC – structure similarity calculated by LGA program assessing agreement in conformations of all atoms (i.e., including side chain atoms).

- N4 - the number of protein residues within 4.5 Å of the inserted ligand.
- cl - the number of query protein residues that may have possible steric clashes with inserted ligand's atoms.

Firstly, to be considered a predicted site candidate, there must be at least ten aligned residues (Nc>=10) that are conserved between the query and the sphere. Secondly, the sequence identity between query and sphere conserved residues must be at least 10% (SeqID>=10.0). Thirdly, the structural similarity measured by GDC (Global Distance Calculations) [11], which counts how many atoms (including side chain atoms) in the query protein and the sphere protein are in close superposition, must be at least 55% (GDC>=55.0). Fourthly, there must be at least one query protein binding site residue that have atoms within less than 4.5 Å of an atom of the aligned sphere PDB ligand atom (N4>=1). Finally, there are no more than two steric clashes (cl<=2) (distance less than 1.0 Å) allowed between query protein binding site residue atoms and any atom of the aligned sphere ligand, counted per residue (that is, multiple clashing atoms all within a single residue are allowed, but only for not more than two residues.) We may expect a higher confidence in predicted binding sites when: Nc>=25, GDC>=65, and cl<=1. However, these thresholds cannot be considered as absolute requirements. For example, possible clashes indicated by "cl>0" may suggest a need to correct the placement of the inserted ligand. Ideally, a score of "cl=0" would enhance the confidence in ligand-protein binding prediction, however, in many cases "cl>0" indicates that the ligand placement within the pocket may need additional adjustment or refinement of protein side-chain atoms. More specifically: "cl>0" indicates that the location of the inserted ligand needs some optimization (through docking or MD simulations), or that the side-chain and/or backbone conformations of the protein residues forming the pocket may need some relaxation to accommodate the ligand. It is also important to keep in mind that the query protein structure may not be in its "holo" conformation to properly accommodate the ligand (i.e., without clashes), so the conformation of binding site residues of the protein may require some optimization.

The example below illustrates the PDBspheres method applied to find binding sites in papain-like proteinase (PL2pro) from COVID-19. In Fig.2. we show results from detected pocket cluster #1 which was defined based on PL2pro pocket structure similarities with 60 pocket-ligand sphere templates (Nm=60 in Fig.2). The number of predicted different bound ligands is Nlig=25, and the combined total number of residues being in contact with bound ligands is Nres=22.



```
#List Pockets   Nm Nres   Contacts:
PocRes: 1       60   22   ,W851_A,D853_A,N854_A,
                          ,N855_A,C856_A,Y857_A,
                          ,L907_A,G908_A,D909_A,
                          ,V910_A,R911_A,Y1009_A,
                          ,T1010_A,Q1014_A,C1015_A,
                          ,G1016_A,H1017_A,Y1018_A,
                          ,K1019_A,D1031_A,T1046_A,
                          ,D1047_A,

#List Pockets   Nm Nlig   Ligands(PDBids):
PocLig: 1       60   25   ,ZN,GVE,CL,AYE,CA,PO4,
                          ,ACT,EDO,CFF,GOL,BME,
                          ,GLZ,3CN,NEH,Y96,Y41,
                          ,VBY,UB4,TTT,730,
                          ,6wuu_J,6wuu_H,6wuu_G,
                          ,Y95,Y94,
```

Examples of identified ligands include inhibitors:
GRL0617 (PDBid: TTT) and Jun9-72-2 (PDBid: JW9)

and peptide inhibitors:
VIR250 (PDBid: 6wuu_J,6wuu_H,6wuu_G)

Fig.2. Identified in PL2pro pocket cluster #1. A list of identified ligands includes three important inhibitors: TTT (GRL0617), JW9 (Jun9-72-2), and VIR250. Agreement in sets of contact residues is also observed (see Fig.4.).

In Fig.3. we show that the poses of three TTT, JW9, and VIR250 ligands overlap significantly. These ligand poses are brought from 11 different sphere templates listed in Fig.4: three for VIR250, five for TTT, and three for JW9.

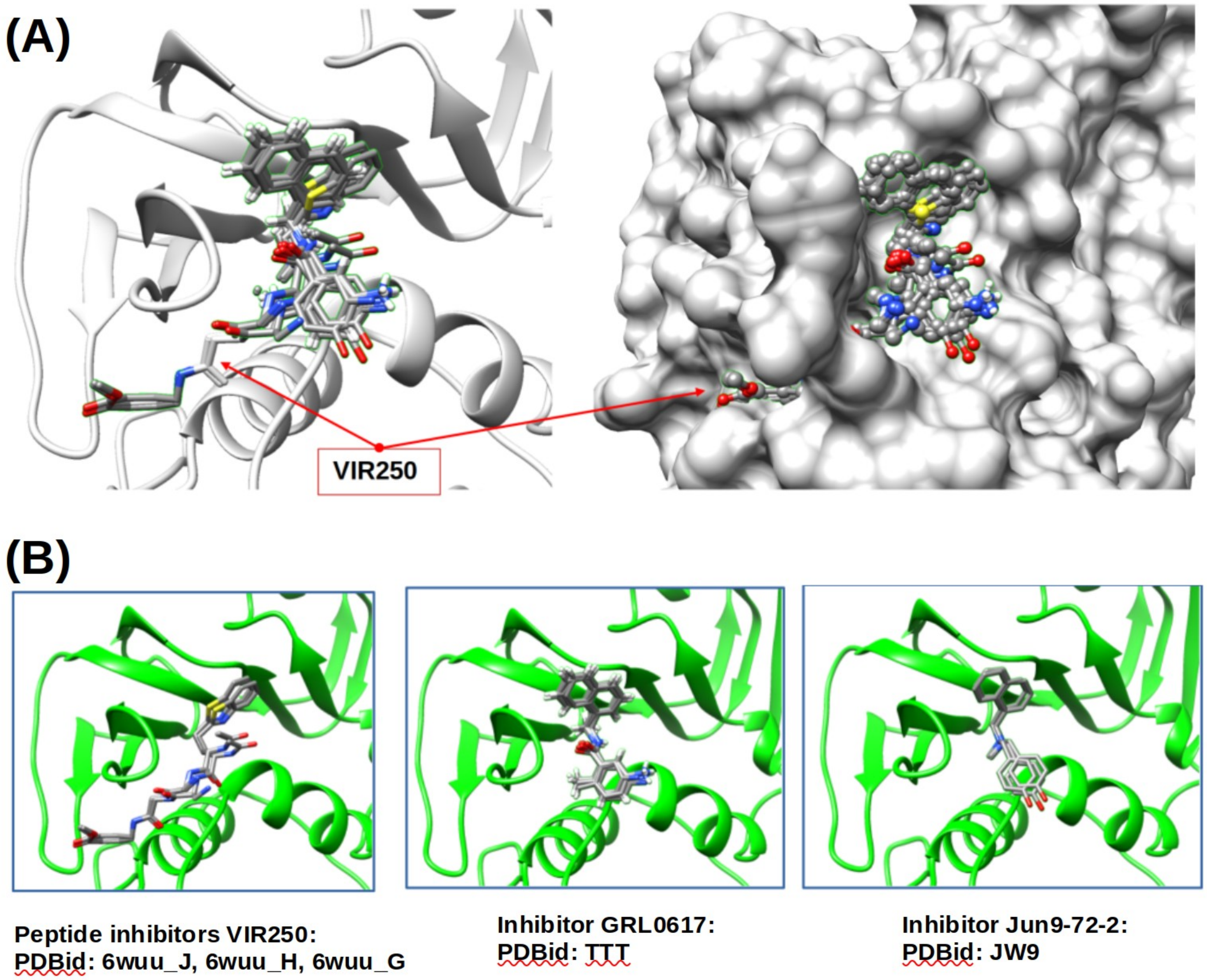
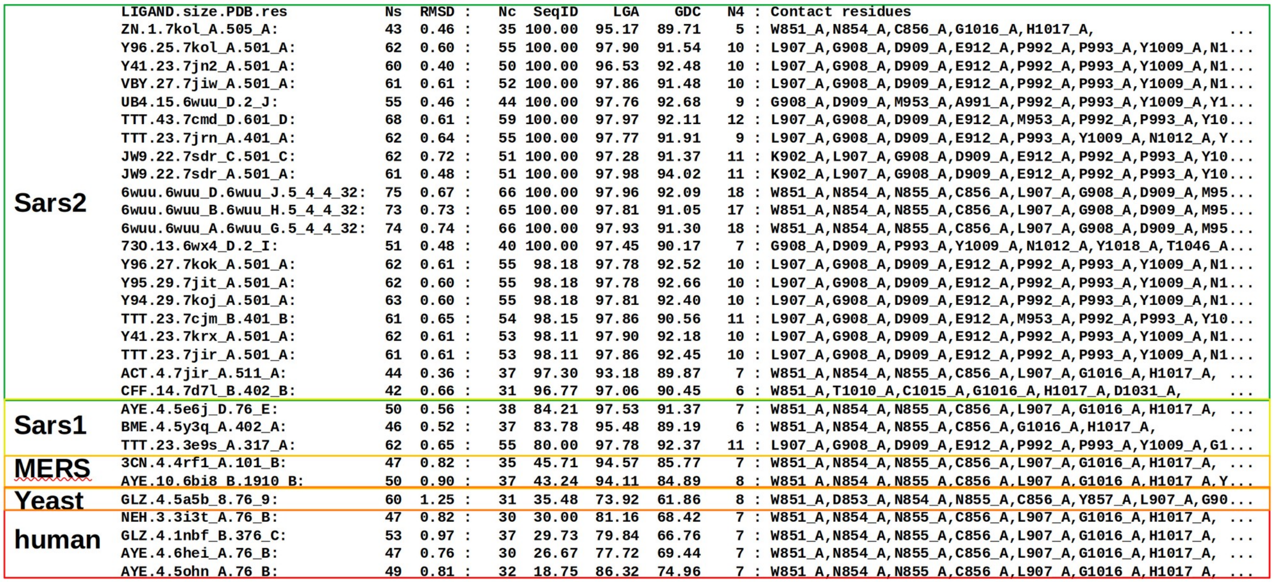


Fig.3. (A) Example of the peptide inhibitors VIR250 (6wuu_J,6wuu_H,6wuu_G) that pass through the “gorge”. Inhibitors TTT and JW9 are places on the one side (right part) of the cavity only. (B) Assessed by PDBspheres poses of all three ligands that come from different sphere templates show strong overlap.

In Fig.4. we present a snapshot from the summary table automatically created by the PDBspheres system. A complete summary table of predicted pocket-ligands for a structural model of papain-like proteinase (PL2pro model: nCoV_nsp3.6w9c_A.pdb) is provided in supplemental File1.


Sars2
Sars1
MERS
Yeast
human

```
LIGAND.size.PDB.res              Ns  RMSD :   Nc  SeqID    LGA    GDC    N4 : Contact residues
ZN.1.7kol_A.505_A:               43  0.46 :   35 100.00  95.17  89.71     5 : W851_A,N854_A,C856_A,G1016_A,H1017_A,               ...
Y96.25.7kol_A.501_A:             62  0.60 :   55 100.00  97.90  91.54    10 : L907_A,G908_A,D909_A,E912_A,P992_A,P993_A,Y1009_A,N1...
Y41.23.7jn2_A.501_A:             60  0.40 :   50 100.00  96.53  92.48    10 : L907_A,G908_A,D909_A,E912_A,P992_A,P993_A,Y1009_A,N1...
VBY.27.7jiw_A.501_A:             61  0.61 :   52 100.00  97.86  91.48    10 : L907_A,G908_A,D909_A,E912_A,P992_A,P993_A,Y1009_A,N1...
UB4.15.6wuu_D.2_J:               55  0.46 :   44 100.00  97.76  92.68     9 : G908_A,D909_A,M953_A,A991_A,P992_A,P993_A,Y1009_A,Y1...
TTT.43.7cmd_D.601_D:             68  0.61 :   59 100.00  97.97  92.11    12 : L907_A,G908_A,D909_A,E912_A,M953_A,P992_A,P993_A,Y10...
TTT.23.7jrn_A.401_A:             62  0.64 :   55 100.00  97.77  91.91     9 : L907_A,G908_A,D909_A,E912_A,P993_A,Y1009_A,N1012_A,Y...
JW9.22.7sdr_C.501_C:             62  0.72 :   51 100.00  97.28  91.37    11 : K902_A,L907_A,G908_A,D909_A,E912_A,P992_A,P993_A,Y10...
JW9.22.7sdr_A.501_A:             61  0.48 :   51 100.00  97.98  94.02    11 : K902_A,L907_A,G908_A,D909_A,E912_A,P992_A,P993_A,Y10...
6wuu.6wuu_D.6wuu_J.5_4_4_32:     75  0.67 :   66 100.00  97.96  92.09    18 : W851_A,N854_A,N855_A,C856_A,L907_A,G908_A,D909_A,M95...
6wuu.6wuu_B.6wuu_H.5_4_4_32:     73  0.73 :   65 100.00  97.81  91.05    17 : W851_A,N854_A,N855_A,C856_A,L907_A,G908_A,D909_A,M95...
6wuu.6wuu_A.6wuu_G.5_4_4_32:     74  0.74 :   66 100.00  97.93  91.30    18 : W851_A,N854_A,N855_A,C856_A,L907_A,G908_A,D909_A,M95...
73O.13.6wx4_D.2_I:               51  0.48 :   40 100.00  97.45  90.17     7 : G908_A,D909_A,P993_A,Y1009_A,N1012_A,Y1018_A,T1046_A...
Y96.27.7kok_A.501_A:             62  0.61 :   55  98.18  97.78  92.52    10 : L907_A,G908_A,D909_A,E912_A,P992_A,P993_A,Y1009_A,N1...
Y95.29.7jit_A.501_A:             62  0.60 :   55  98.18  97.78  92.66    10 : L907_A,G908_A,D909_A,E912_A,P992_A,P993_A,Y1009_A,N1...
Y94.29.7koj_A.501_A:             63  0.60 :   55  98.18  97.81  92.40    10 : L907_A,G908_A,D909_A,E912_A,P992_A,P993_A,Y1009_A,N1...
TTT.23.7cjm_B.401_B:             61  0.65 :   54  98.15  97.86  90.56    11 : L907_A,G908_A,D909_A,E912_A,M953_A,P992_A,P993_A,Y10...
Y41.23.7krx_A.501_A:             62  0.61 :   53  98.11  97.90  92.18    10 : L907_A,G908_A,D909_A,E912_A,P992_A,P993_A,Y1009_A,N1...
TTT.23.7jir_A.501_A:             61  0.61 :   53  98.11  97.86  92.45    10 : L907_A,G908_A,D909_A,E912_A,P992_A,P993_A,Y1009_A,N1...
ACT.4.7jir_A.511_A:              44  0.36 :   37  97.30  93.18  89.87     7 : W851_A,N854_A,N855_A,C856_A,L907_A,G1016_A,H1017_A, ...
CFF.14.7d7l_B.402_B:             42  0.66 :   31  96.77  97.06  90.45     6 : W851_A,T1010_A,C1015_A,G1016_A,H1017_A,D1031_A,      ...
AYE.4.5e6j_D.76_E:               50  0.56 :   38  84.21  97.53  91.37     7 : W851_A,N854_A,N855_A,C856_A,L907_A,G1016_A,H1017_A, ...
BME.4.5y3q_A.402_A:              46  0.52 :   37  83.78  95.48  89.19     6 : W851_A,N854_A,N855_A,C856_A,G1016_A,H1017_A,        ...
TTT.23.3e9s_A.317_A:             62  0.65 :   55  80.00  97.78  92.37    11 : L907_A,G908_A,D909_A,E912_A,P992_A,P993_A,Y1009_A,G1...
3CN.4.4rf1_A.101_B:              47  0.82 :   35  45.71  94.57  85.77     7 : W851_A,N854_A,N855_A,C856_A,L907_A,G1016_A,H1017_A, ...
AYE.10.6bi8_B.1910_B:            50  0.90 :   37  43.24  94.11  84.89     8 : W851_A,N854_A,N855_A,C856_A,L907_A,G1016_A,H1017_A,Y...
GLZ.4.5a5b_8.76_9:               60  1.25 :   31  35.48  73.92  61.86     8 : W851_A,D853_A,N854_A,N855_A,C856_A,Y857_A,L907_A,G90...
NEH.3.3i3t_A.76_B:               47  0.82 :   30  30.00  81.16  68.42     7 : W851_A,N854_A,N855_A,C856_A,L907_A,G1016_A,H1017_A, ...
GLZ.4.1nbf_B.376_C:              53  0.97 :   37  29.73  79.84  66.76     7 : W851_A,N854_A,N855_A,C856_A,L907_A,G1016_A,H1017_A, ...
AYE.4.6hei_A.76_B:               47  0.76 :   30  26.67  77.72  69.44     7 : W851_A,N854_A,N855_A,C856_A,L907_A,G1016_A,H1017_A, ...
AYE.4.5ohn_A.76_B:               49  0.81 :   32  18.75  86.32  74.96     7 : W851_A,N854_A,N855_A,C856_A,L907_A,G1016_A,H1017_A, ...
```

Fig.4. Fragment of the PDBspheres summary table reporting predicted PL2pro ligands assigned to pocket cluster #1. Results highlighted in green (when SeqID ">95") indicate that similar pockets are detected in Sars2 or variants of Sars2. When SeqID "<30" (results highlighted in red) examples of similar pockets detected in human proteins are shown. An agreement in sets of contact residues indicates that all listed ligands are predicted to bind the same pocket identified through clustering as "pocket cluster #1".

Recent studies of ligand binding site refinements show significant success in generating reliable "holo" (ligand-bound) protein structures from their "apo" (ligand-free) conformations [20]. Similarly, if N4 – the number of protein atoms within 4.5 A of the ligand – is very low, then this indicates that the current placement of the ligand does not show strong interactions with residues of the protein binding site. This may suggest that the pocket is too big (or some residues in the protein model are missing, or side-chain atoms are not in the right conformation). Of course, it may also indicate that the identified pocket is incorrect, or the location of the inserted ligand is wrong, but these conclusions can only be confirmed by more detailed inspection. In such cases it can be informative to check the overlap of the template pocket and the query protein pocket (i.e., Nc, the number of conserved superimposed pocket-forming residues). Higher overlaps (e.g., Nc>=25 or more) indicate greater size of the region forming the pocket and higher confidence in reported pocket similarities (more residues identified as conserved); however, such thresholds cannot be definitive because in cases of shallow cavities or interface sites on the surface of the protein the Nc number can be low.

2.3 Benchmark dataset and evaluation metrics

The benchmark dataset used for the comparative analysis of performance of PDBspheres with seven other methods is the "LBSp dataset" described in Clark et al. 2019 [9] and 2020 [1] papers. To assess accuracy of evaluated methods we followed the same metrics and evaluation procedures as described in the Clark et al. 2020 paper [1]. The authors defined a reference list called unified binding sites (UBS) as a union of all residues contacted by any bound ligand within a family. The scoring of each algorithm's site prediction is determined by agreement with the UBS reference using calculated numbers of true positives (TP), true negative (TN), false positive (FP), and false negative (FN) predicted contacts. Both Matthew's Correlation Coefficients (MCCs), and F scores as calculated by formulas (Eqs. 1-2) have been used as metrics to represent the predictive power of the various methods.

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

$$F = \frac{2 * TP}{2 * TP + FP + FN} \quad (2)$$

There are other metrics sometimes used to evaluate protein-ligand binding sites detection, e.g., DCC [13] and DVO [14]. DCC is the distance between the predicted and the actual center of the pocket. This metric evaluates the correct location of the pocket. DVO (discretized volume overlap) is defined as the ratio of the volume of the intersection of the predicted and the actual segmentations to the volume of their union. It assesses the correctness of the shape of predicted pockets. If DCC is below 4 Å, the pocket is considered as correctly located [13]. However, in our study we follow metrics described in the Clark et al. 2020 [1] paper as they focus on assessing accuracy of binding site prediction methods based on the correctness of identified pocket residues being in contact with ligands.

3. Results and discussion

3.1 Comparison of PDBspheres to other methods

The PDBspheres method was compared to Surfnet [2], Ghecom [3], LIGSITE [4], Fpocket [5], Depth [6], AutoSite [7] and Kalasanty [8], described in more detail in Clark et al. [1]. Five of these seven methods are considered to be geometry-based, while one is energy-based, and one is machine-learning-based. Evaluation results for these seven methods are taken from the publication [1], i.e., no new calculations/predictions have been performed for these other methods. The metrics used for the assessment of the prediction accuracy of PDBspheres are the same as used in [1]. Predictive power is assessed using two metrics: F scores and Matthew's Correlation Coefficients (MCCs). The F scores and MCCs provide a good description of the relative success of evaluated algorithms assessing whether or not a method produced a predicted binding site containing residues in common with the definition of Unified Binding Sites (UBS) [1]. They assign high scores not just when at least one residue in a given site is identified correctly, but rather reward methods with more correct and less false predictions of contact residues implying which of the algorithm pocket predictions are close to the "correct" location on the binding surface of the protein. The results from evaluation different methods are provided in Table 1, where the first seven result rows are taken from [1].

Method	Apo		Holo		Apo		Holo	
	F	F IQR	F	F IQR	MCC	MCC IQR	MCC	MCC IQR
Surfnet	0.23	0.23	0.23	0.24	0.22	0.26	0.23	0.28
Ghecom	0.48	0.5	0.54	0.55	0.5	0.54	0.53	0.62
LIGSITE	0.48	0.46	0.52	0.48	0.46	0.52	0.5	0.56
Fpocket	0.42	0.57	0.53	0.56	0.43	0.61	0.51	0.62
Depth	0.4	0.29	0.42	0.27	0.38	0.29	0.4	0.27
AutoSite	0.35	0.59	0.45	0.6	0.34	0.67	0.42	0.67
Kalasanty	0.49	0.51	0.51	0.43	0.48	0.56	0.54	0.48
PDBspheres(100)	0.81	0.14	0.82	0.15	0.80	0.15	0.82	0.15

	0.80		0.82		0.79		0.81	
PDBspheres(90)	0.77 0.73	0.18	0.79 0.74	0.16	0.76 0.72	0.18	0.77 0.73	0.17
PDBspheres(80)	0.76 0.72	0.17	0.78 0.73	0.18	0.75 0.71	0.18	0.77 0.72	0.18
PDBspheres(70)	0.76 0.71	0.19	0.78 0.73	0.17	0.75 0.70	0.19	0.77 0.72	0.17
PDBspheres(60)	0.76 0.71	0.19	0.78 0.72	0.17	0.75 0.70	0.19	0.77 0.72	0.17
PDBspheres(50)	0.75 0.70	0.20	0.77 0.71	0.17	0.74 0.69	0.20	0.75 0.70	0.19

Table 1. Median of family median F scores and MCCs for apo and holo datasets for all seven LBS-prediction methods and PDBspheres (for which mean values are also given, colored in blue). IQR (interquartile range) describes the difference between maximum and minimum scores within the middle 50% of values when ordered from lowest to highest. IQR indicates how close the middle 50% of family F and MCC values are to their respective medians. F and MCC scores are described in Section 2.3 {Benchmark dataset and evaluation metrics}.

The final six data rows in Table 1 show results from the evaluation of PDBspheres when the pocket detection was performed using the complete PDBspheres library (100%), and when the PDBspheres library was restricted to templates with no more than 90%, 80%, 70%, 60% and 50% sequence identity with query proteins, respectively. Results illustrate that in contrast to the large diversity of possible protein sequences the number of structurally distinct pockets is limited, therefore proteins that by sequence are very different may still share almost identical structural conformations in local regions (e.g., binding sites) that can perform similar functions. This observation served as a basis for the development of our PDBspheres structure template-based binding site detection method. The only limitation in its ability to identify correct pockets for a given protein might be an underrepresentation of particular pocket's conformation in currently experimentally solved protein structures deposited in the PDB. For example, in case when a non-restricted library is used the binding sites for all 304 families ("apo" and "holo" protein conformations) were predicted. However, in the case of the restricted library (which excludes template spheres derived from proteins showing more than 90% sequence identity to the query proteins) the binding sites for 5 families of "apo" versions and 7 families of "holo" versions were not predicted (see Table 2). In Table 2 we show even more results indicating that the protein binding sites are highly structurally conserved and can be successfully detected using structure-based template spheres taken from proteins sharing low sequence identity with targeted proteins. The PDBspheres method leverages this quality efficiently. Indeed, thanks to the richness in diversity of protein structures deposited in PDB the system is able to identify location of 97% binding sites from the LBPS dataset using structural templates that share as low as 50% sequence identity with targeted proteins. And even with such significant restriction to the library of template spheres the accuracy in identified sets of residues interacting with ligands is still high ~75% as measured by F and MCC metrics (see Table 1).

Restricted sequence identity	Number of detected pockets			Number of pocket's families		
	All	Apo	Holo	All	Apo	Holo
100%	2528	1082	1456	304	304	304
90%	2481	1056	1435	299	299	297

80%	2469	1051	1428	298	296	297
70%	2469	1051	1428	298	296	297
60%	2464	1048	1426	297	295	296
50%	2457	1046	1421	295	294	294
	Percent of detected pockets			Percent of pocket's families		
50%	97.2%	96.7%	97.6%	97.0%	96.7%	96.7%

Table 2. Number of detected pockets and pocket's families when the PDBspheres libraries are restricted to templates with sequence identity to proteins from the LBPs dataset not exceeding introduced cutoffs.

Based on MCC and F scores shown in Table 1, the PDBspheres method identifies similar pockets better than the other seven evaluated methods. We should emphasize that PDBspheres is an exclusively structure-based method, not exactly template-based (i.e., PDBspheres does not utilize any prior information from libraries of sequences/residues forming binding sites). Additionally, PDBspheres does not use any prior information about the location of searched pockets in proteins from the same family. We treat all structures equally and independently in finding structural similarities between the query protein and template spheres. Of course, we can find such cavities more easily in the "holo" conformations, but as the results show, we can also find adequate structure similarities in "apo" versions of the query protein; again, strictly based on similarities in structure without any sequence-based knowledge of residues forming the pocket (residue information that could be transferred from some databases of "holo" structures.) All results reported here are based on PDBspheres superpositions calculated on C-alpha atoms. Results based on superposition of C-beta atoms (results not shown) are virtually identical.

3.2 Structural similarity of binding sites vs. sequence identity

Here, we discuss how close in sequence identity two proteins need to be to have structurally similar binding sites, addressing these two questions:

- How low the level of sequence identity between two proteins can be and still share similar pockets and perform similar functions?
- Do such proteins have enough similarity in their functional sites to bind ligands in a similar manner?

Fig. 5 illustrates results from the LBS database pocket identification calculations, which were summarized in Table 1. Fig.5. shows that the correct detection of binding sites using strictly structure-based conformation similarity criteria does not require high sequence identity between targeted protein and the binding site templates. The analysis of the PDBspheres predictions of pockets from the LBS database indicates that the binding sites can be correctly predicted based on template spheres derived from proteins that share significantly lower overall sequence identity with the query protein. Let us note that the main difference between results shown in the Table 2 and Fig 5 is that here we discuss the correctness of the prediction (i.e., as accurate as possible identification of residues involved in protein-ligand interaction) while in Table 2 we report results from just detection of the pocket location in the protein (i.e., without any assessment of the accuracy and completeness of predicted interacting residues).

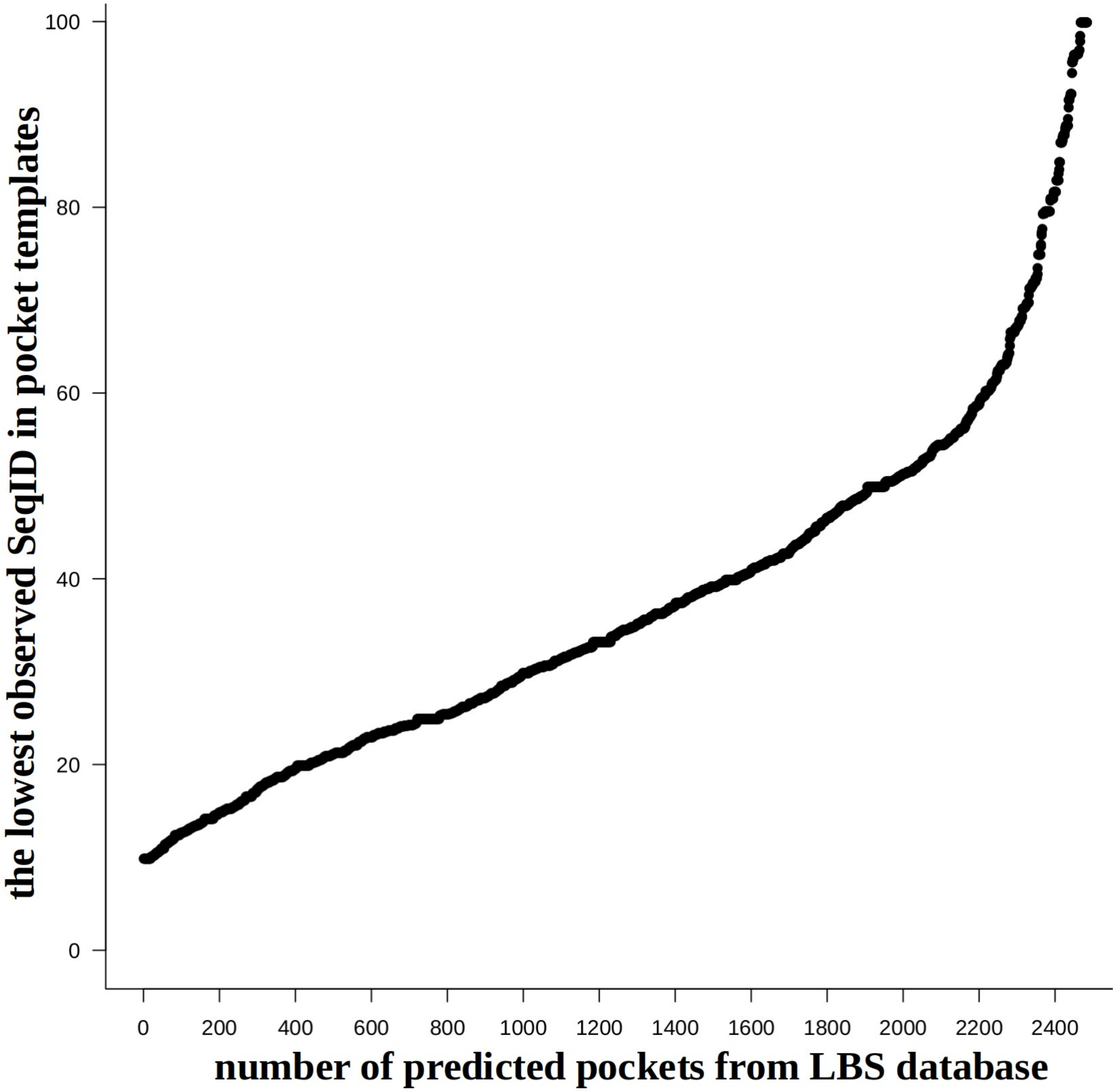
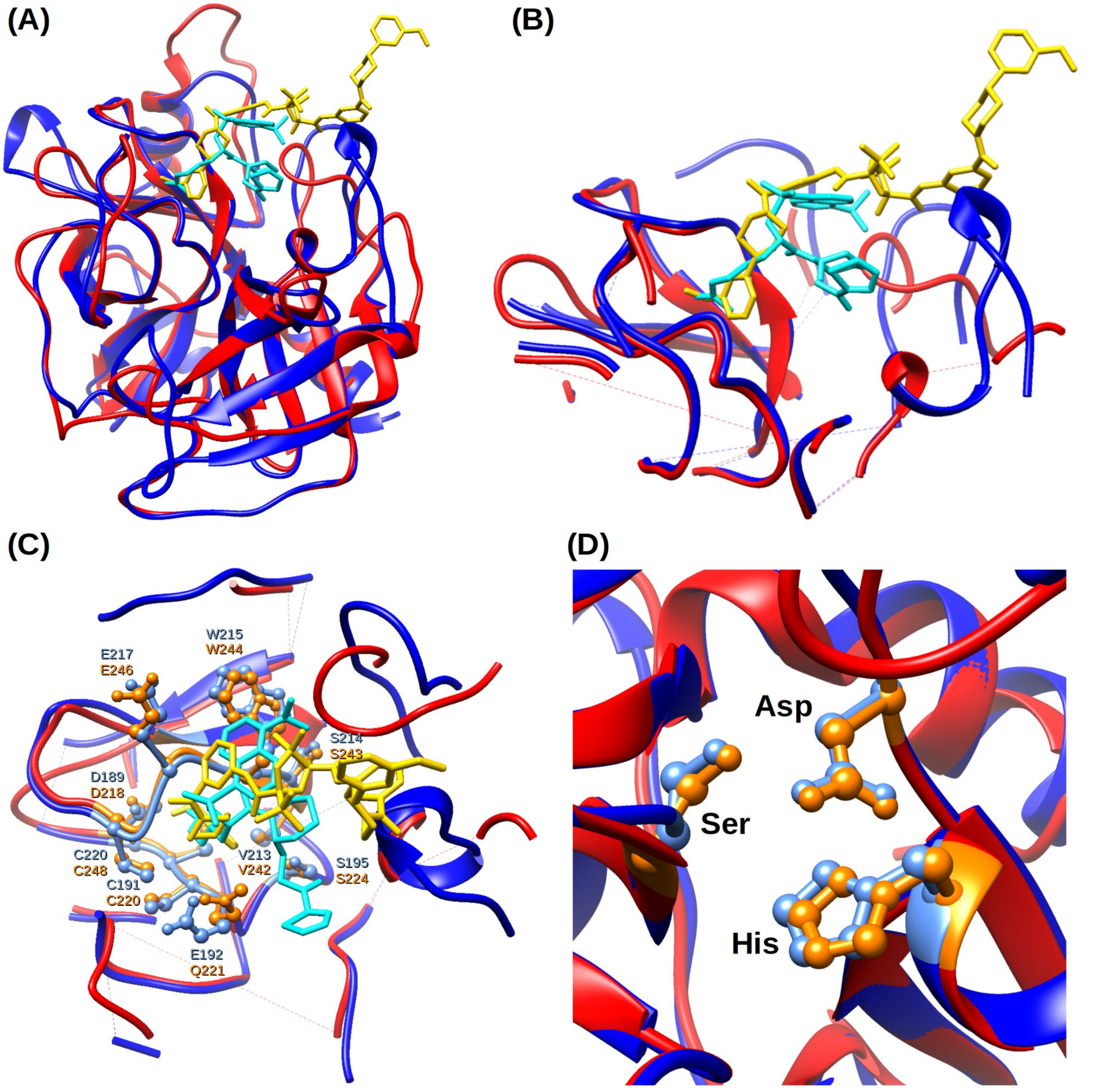


Fig.5. The plot illustrates how many pockets from the LBS database proteins (all 2,528 pockets; "apo" and "holo" combined) can be predicted based on binding site templates derived from proteins with lowest possible sequence identity to the query proteins. For example, at least 400 pockets (see left bottom part of the plot) can be predicted based on template binding sites derived from proteins that share as low as 20% sequence identity (SeqID) with a query protein. On the other hand, more than 2,400 pockets (~95%) can be correctly predicted based on templates sharing no more than 80% of SeqID with query proteins.

In Fig.6. we show an example of correct pocket detection by PDBspheres using different pocket-ligand sphere templates derived from proteins that share low sequence identity. Two compared proteins (serine proteases) have less than 38% of sequence identity, but they are structurally similar at the level of over 86% by the LGA assessment of similarity on the C-alpha atoms level and 77% by GDC (all atoms level). Of the residues that are in contact with corresponding ligands (distance below 4.5 Å) in the identified similar pockets, 11 out of 17 are identical (65%). The pocket template spheres derive from two PDB complexes that bind inhibitor compounds having PDB ligand identifiers 2A4 and QWE, respectively. These ligands are significantly different in size and the distance between ligand centroids calculated from the complexes of the same orientations are very different, i.e., the distance between centroids of ligands when inserted in any of these pockets is about 6.0 Å. However, the portions of the ligands inserted in pockets have a similar overlapping region and are in contact with similar amino acids from the pockets. In each of these two protein-ligand complexes the core parts of the ligands are in contact with 13 residues of which 11 are identical (85%) (see Fig.6 (C) and (E)). These results illustrate how PDBspheres can be used to detect conservation in local structural conformations and to assess conservation of critical contact residues; both can assist in inferring protein function.



(E)

1arw_H:	Y60A	P60B	P60C	W60D	D60E	-	-	-	-	D189	A190	C191	E192	S195	V213	S214	W215	G216	E217	G219	C220	G226
4mpv_A:	-	-	-	-	-	H74	T115	A116	Q117	D218	S219	C220	Q221	S224	V242	S243	W244	G245	E246	G247	C248	G255

Fig.6. Example of two structures of serine proteases: Tryptase (in red PDB chain: 4mpv_A) and Thrombin (in blue PDB chain: 1a4w_H) that share high structure similarity in their binding sites (over 77% by GDC) while the level of sequence identity between them is no higher than 38%. (A) Overall structure superposition of two protein-ligand complexes showing location of bound inhibitors. (B) PDBspheres-based local superposition of corresponding protein spheres surrounding ligands. (C) Structurally superimposed spheres of 4mpv and 1a4w show significant similarity in side-chain conformation of residues interacting with corresponding ligands 2A4 and QWE. Residues interacting with ligands 2A4 and QWE are highlighted in orange and light blue, respectively. (D) Local superposition indicates a perfect agreement in the nearby catalytic triad residue conformations (His, Asp, Ser). (E) Structural alignment of residues from close distance (4.5 Å) from the corresponding ligands shows that 11 out of 13 (85%) of residues that are in contact with similar core parts of the ligands are identical.

3.3 Clustering binding sites from PDBbind database

In this section, we describe how PDBspheres can be used to perform structure-based clustering and structure similarity analysis of binding sites from the PDBbind dataset [17] (ver. 2019). Some of these results were leveraged in our previous work generating rigorous training and validation datasets for machine learning of ligand-protein interactions [24]. Here we want to address the following questions:

- To what extent can structurally-similar binding pockets having similar ligand placement allow inference of binding affinity from one pocket-ligand pair to another pocket-ligand pair?
- To what extent can clustering of detected pockets and calculated structure similarities among clustered pockets from different proteins provide functional information for protein annotation?

In our analysis, we focus on the "refined" 2019 dataset (4,852 structures) which we expanded by adding 24 structures from the previous PDBbind release that are not present in the 2019 version. Hence, in total the dataset of evaluated binding sites consists of 4,876 structures. Since we are interested in the assessment of similarities between specific pockets listed in proteins from PDBbind (the PDBbind database reports only one pocket for each protein regardless of how many different pockets a given protein may have or how many alternative locations of a given pocket in a multichain protein complex can be observed) we restricted our structure similarity searches and evaluations to only those regions in PDBbind proteins that encompass targeted pockets (many protein structures in the PDBbind refined dataset are multidomain complexes with total sizes of more than 2000 residues, so they can have multiple binding sites in addition to the targeted ones). Therefore, in our approach to evaluate similarities and cluster pockets in PDBbind each of its protein structures was reduced to the region in the close vicinity of a reported ligand (residues having any atom within 16 Å of any ligand atom), and each ligand binding site reported in PDBbind was associated with its corresponding PDBspheres template sphere (residues having an atom within 12 Å of a ligand atom). We performed an all-against-all PDBspheres detection and pocket similarity evaluation using the 16 Å protein region representation of each PDBbind protein. Results from the pair-wise pocket similarity evaluation are provided in supplemental File5. Structure similarity results allowed grouping of the 4,876 PDBbind protein pockets into 760 clusters. Cluster details are provided in supplemental File2, File3, and File4. Fig. 7 illustrates the clusters and it is a snapshot taken from the HTML supplemental File6 (interactive overview of predicted clusters created using “plotly” R graphic library). Each axis of Fig. 7 indexes a sample of 4,876 protein pockets, where sampled proteins are reported to help identification of corresponding pockets or clusters on the plot. A “zoom-in” option in “plotly” graph expands each rectangle to show a list of individual memebers of a selected cluster. It allows for each of the sampled proteins to be labeled with its exact location within the cluster (see the example in Fig. 8; supplemental File6 contains further illustrations). Each rectangle in Fig. 7 represents a cluster and is composed of small markers – one for each sampled protein-pocket pair within the cluster – colored by the GDC all-atom similarity score between each member of the cluster, where colors closer to red indicate a higher degree of similarity between members. For example, the first three large clusters from the bottom left of Fig. 7 represent predicted clusters, i.e., cluster #22 (318 members), #8 (351 members), and #5 (322 members), respectively. Evaluation of predicted clusters (see supplemental File2) shows that all members of the cluster #22 are assigned with EC subclass 4.2.1.1, members of the cluster #8 - EC 3.4.21, and cluster #5 - EC 3.4.23. All clusters along the diagonal in Fig. 7 have high within-cluster similarity, and are separated from other clusters according to the applied “exclusive” clustering approach. Some of the defined clusters are formed after grouping together proteins from several “finer” subclusters. For example, at the top right there is a large cluster (#25 with 382 members assigned with several subclasses of the broad-spectrum transferases - EC class 2.7 (“Transferring phosphorus-containing groups”), with varying degrees of similarity among its members as they belong to different and more specific subclasses which are still similar enough (according to the selected thresholds) to form one distinct cluster (see supplemental File2 and File6).

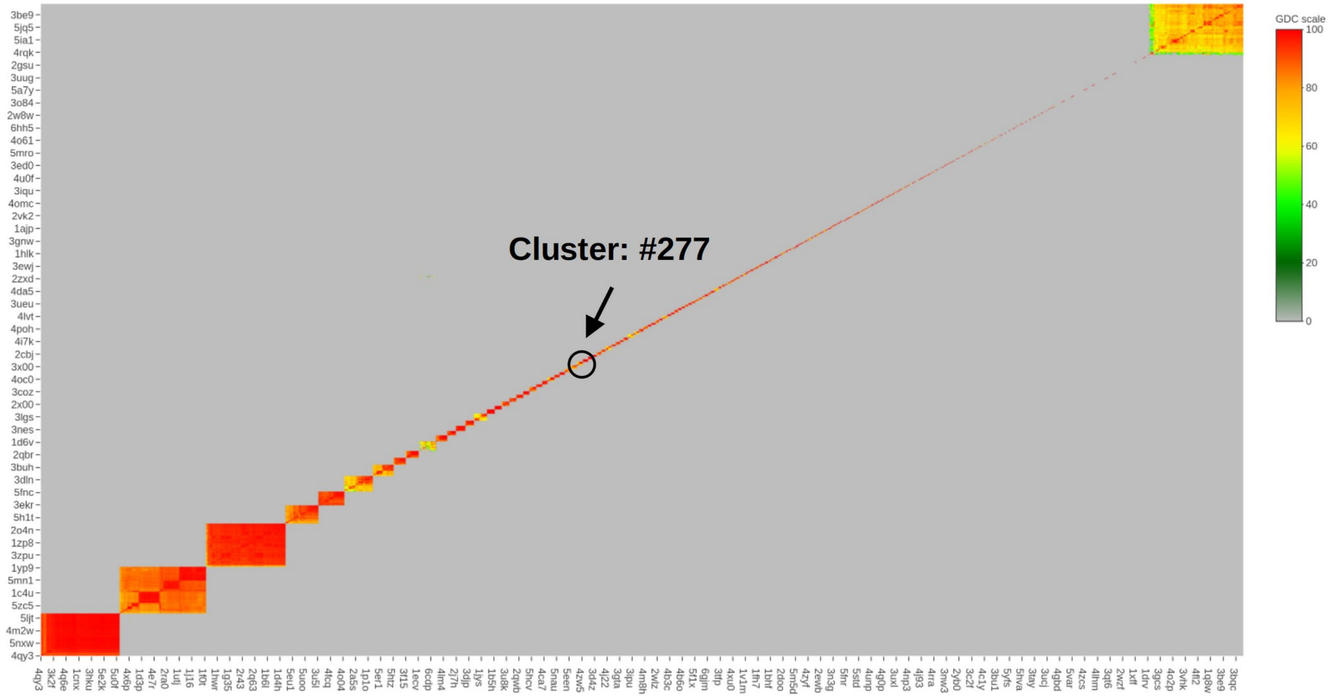


Fig.7. Clustering of the refined set of 4,876 pockets from the PDBbind based on their structure similarity. PDBspheres-based pocket detection and similarity evaluation resulted in 760 constructed clusters. HTML file allowing interactive overview of predicted clusters is provided in supplemental File6. Fig. 8 shows a “zoom-in” to cluster #277, which contains 24 protein-ligand pairs.

PDBspheres can assist in making predictions related to protein functional annotation. Similar pockets that are clustered together by PDBspheres share similar functions as indicated in the supplemental File2, File3, and File4, where the members from each cluster are checked for their agreement in assigned EC, SCOP, and GO annotations. An example in Fig. 8 illustrates such agreement showing individual protein-pocket-pair GDC similarity values for a particular cluster (cluster #277). Each of the proteins grouped into this cluster share the same EC subclass and GO annotation (:0006508:0008237:0008270: see cluster #277 in supplemental File2 and File4 for details) identified in PDB, which is an indication that the PDBspheres-based clustering can be functionally meaningful.

4zw7	:3.4.11.-:
4zx4	:3.4.11.-:
4zw8	:3.4.11.-:
3ebi	:3.4.11.-:
4zw5	:3.4.11.-:
3ebh	:3.4.11.-:3.4.11.2:
4zw6	:3.4.11.-:
4r5t	:3.4.11.-:
2zxg	:3.4.11.2:
4k5p	:3.4.11.-:
4zx3	:3.4.11.-:
6eed	:3.4.11.-:3.4.11.2:
3t8v	:3.4.11.-:
3q44	:3.4.11.-:
4kxb	:3.4.11.7:
4kx8	:3.4.11.1:3.4.11.2:3.4
4fys	:3.4.14.-:3.4.11.2:
4ou3	:3.4.11.2:
2vj8	:3.3.2.6:3.4.11.4:
2r59	:3.3.2.6:3.4.11.4:
3fh7	:3.3.2.6:3.4.11.4:
3b7r	:3.3.2.6:3.4.11.4:
3b7u	:3.3.2.6:3.4.11.4:
4l2l	:3.3.2.6:3.4.11.4:

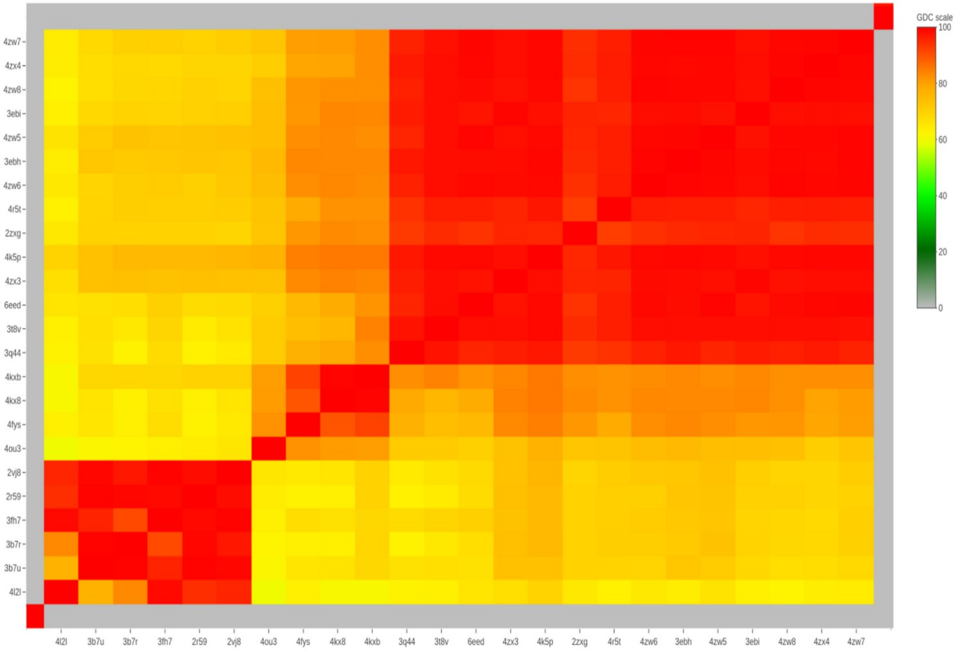


Fig.8. Plot showing an example of a cluster #277 identified by PDBspheres (see Fig. 7 and supplemental File2.) Pocket-based clustering groups proteins with the same function. All 24 proteins grouped together within the cluster #277 belong to the same enzyme subclass 3.4. - hydrolases that act on peptide bonds (EC 3.4.11 are those hydrolases that cleave off the amino-terminal amino acid from a polypeptide). For each of the listed enzymes (PDB id in the first column which corresponds to the Y-axis in the plot) the assigned alternative EC numbers (second column) are separated by colon ':'. In addition to the general functional clustering of proteins, the PDBspheres clustering approach provides finer subclustering of proteins within predicted clusters. The bottom 6 proteins from the cluster #277 form a clear subcluster as they share additional EC 3.3.2.6 - bifunctional zinc metalloprotease activities.

Another important question is: can we transfer binding affinity scores from one ligand binding site to another if the pockets and the ligand placements within the pockets are similar? Interestingly, pockets from PDBbind that share high structure similarity and that have similar ligand placement (distances between the centroids of inserted ligands within the aligned pockets no greater than 0.5 Å) show similarities in reported protein-ligand binding affinities. For PDBbind entries having Kd values, Fig. 9 (A) compares the Kd values of each pair of protein-ligand complexes for the set of pairs that have GDC structural similarity greater than 95% and aligned ligand centroids within 0.5 Å. The R^2 and Spearman values of 0.5 and 0.7, respectively, indicate the strength of the relation – similar protein-ligand pairs tend to have similar Kd values. Similarly, Fig 9 (B) shows R^2 and Spearman values of 0.44 and 0.574 for pairs that have Ki values and meet the similarity criteria.

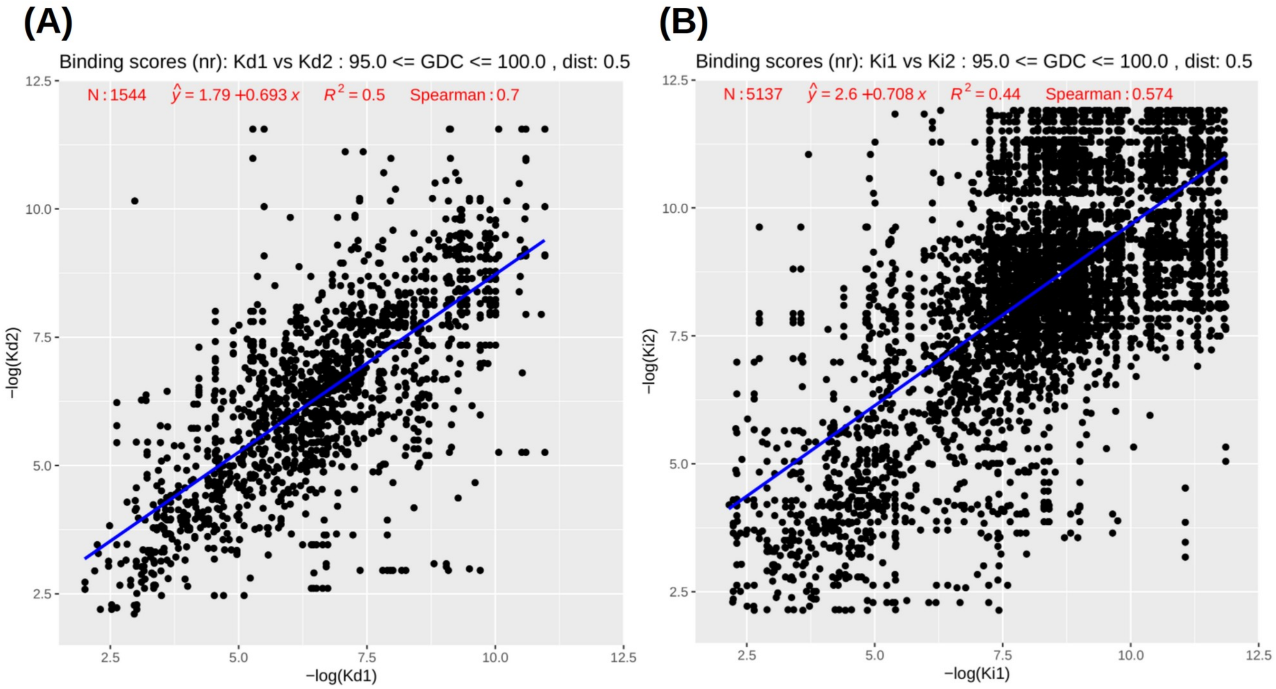


Fig.9. (A) Scatter plot of Kd values from 1,544 pocket pairs, and (B) Ki values from 5,137 pocket pairs. Redundant pairs – self-comparisons and symmetry duplicates – are not included.

These similarities in binding affinities are even higher between pockets from different proteins when they bind the same ligand. The corresponding R^2/Spearman scores for Kd and Ki are 0.5/0.738 (194 pairs) and 0.69/0.772 (411 pairs), respectively. If we relax the criteria – a ligand placement centroid value cutoff of 1.0 Å and GDC as low as 90% – then of the resulting PDBbind Kd subset of similar complexes, 287 pairs have the "same ligand" in the pocket, and there are 631 "same-ligand" pairs in the Ki subset. Figs. 10 (A) and (B) compare the Kd and Ki values, respectively, with respective R^2/Spearman values of 0.46/0.703 (287 pairs) and 0.64/0.752 (631 pairs).

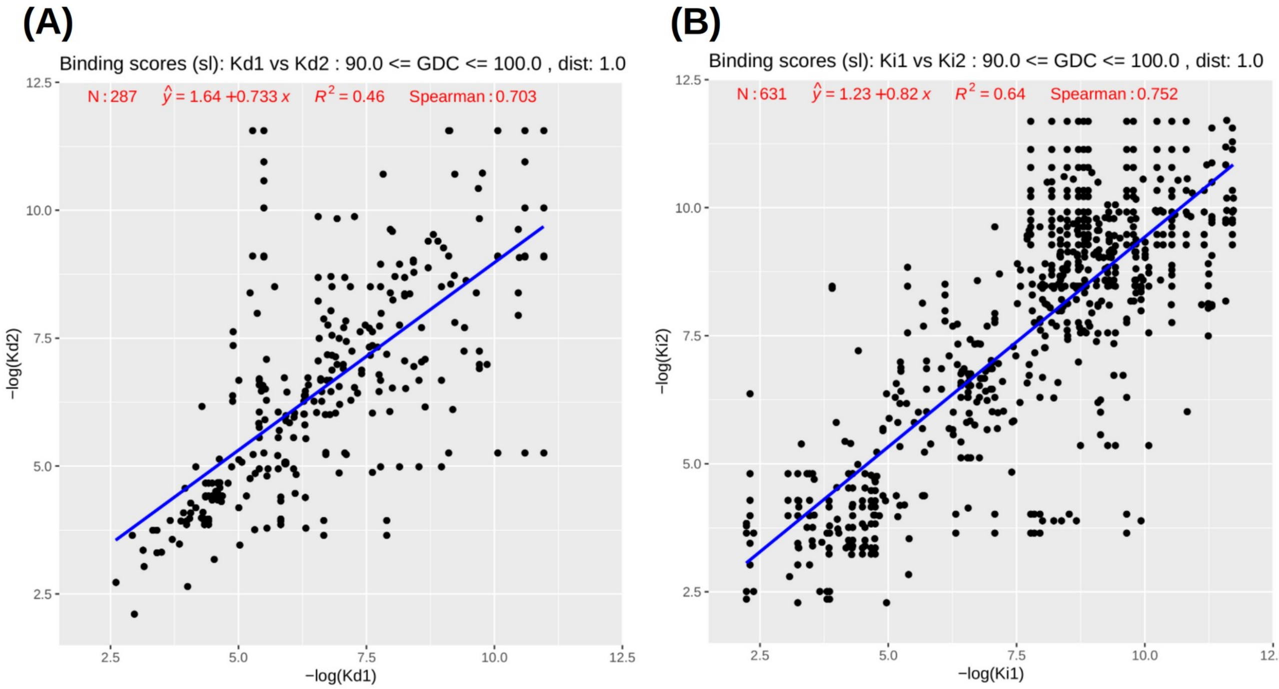


Fig.10. (A) Scatter plot of Kd values from 287 “same ligand” pocket pairs, and (B) Ki values from 631 “same ligand” pocket pairs. Redundant pairs – self-comparisons and symmetry duplicates – are removed from calculations.

Results from the PDBspheres clustering of the PDBbind dataset suggest that the structure similarity between pockets/ligand placements is a significant characteristic that can allow prediction of similar binding affinity values. In future work, we anticipate enhancing current measurements with additional information about the specific atom location of protein residues interacting with the ligand, which may improve predictions. Complete detailed results from PDBspheres analysis of pair-wise similarities in binding sites between proteins from PDBbind are provided in the supplemental File5.

4. Conclusions

While developing PDBspheres we focused on two goals: (1) binding pocket detection; and (2) identification of characteristics and scores to assess similarities between pockets to help further protein functional characterization and clustering.

In particular, with regard to binding pocket detection we find that PDBspheres’ strictly structure-based approach can correctly predict binding site regions in protein structures known to be in a “holo” (i.e., ligand-binding) conformation as well as in protein structures in “apo” (without a ligand present) conformation. Since local regions in functionally similar proteins are remarkably conserved in their structural conformations, the method allows detection of similar binding sites even in proteins that share very low sequence similarity. In comparisons with other binding site prediction methods the PDBspheres’ strictly structure-based approach allows a very high accuracy in identifying protein-ligand contact residues.

With regard to characterizing and evaluating binding pocket similarities among proteins we find that a high level of sequence similarity between different proteins is not essential to identify structurally similar binding sites that may perform similar functions. In the structure-based detection of binding sites the similarity assessed based on calculated structural alignment using C-alpha atom positions is sufficient, and the use of other residues (e.g., C-beta atoms, or other points representing residue) in the similarity assessment does not yield better results. Structurally similar binding pockets having similar ligand placements allow inference of binding affinity from one pocket-ligand pair to another pocket-ligand pair. PDBspheres-based clustering of detected pockets and calculated structure similarities among pockets from different proteins provide information that can significantly help protein function annotation efforts.

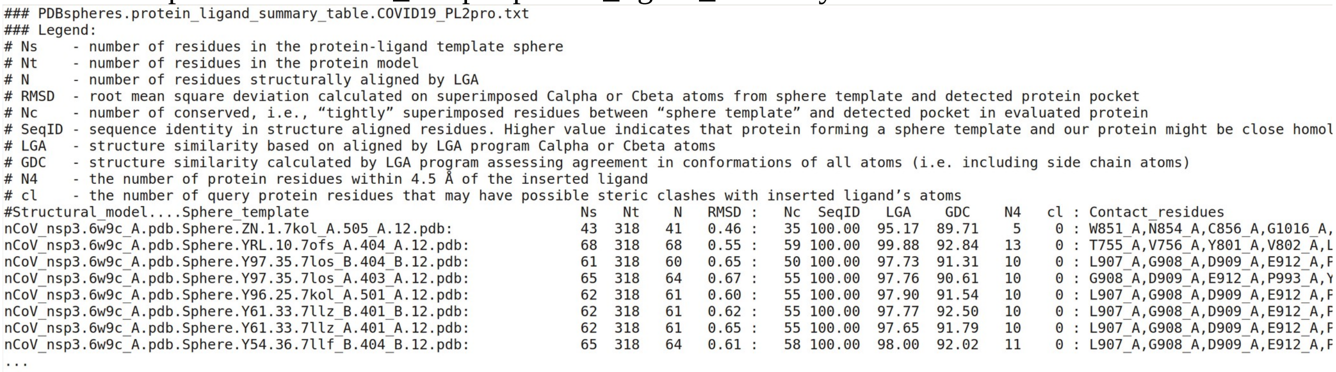
Acknowledgments

This research was supported by the American Heart Association under CRADA TC02274, by Laboratory Directed Research and Development (LDRD 20-ERD-062) at Lawrence Livermore National Laboratory (LLNL), and by DTRA under award HDTRA1036045. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Supplemental data:

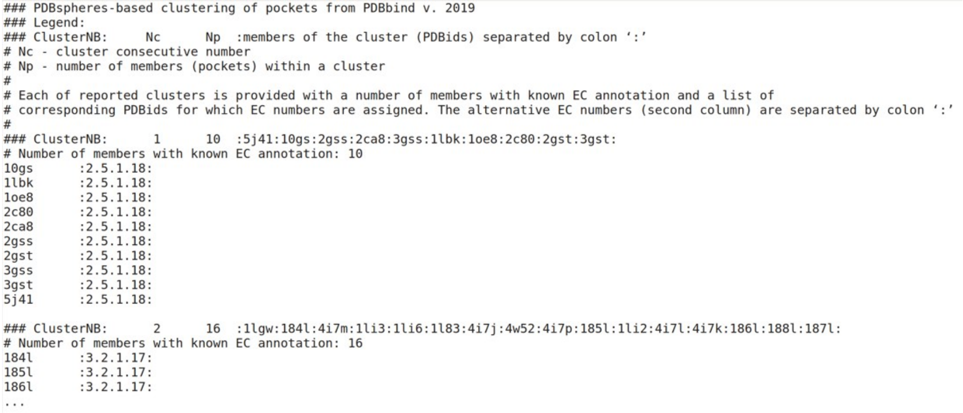
The supplemental data is available for download at https://proteinmodel.org/AS2TS/PDBspheres

File1: PDBspheres.COVID19_PL2pro.protein_ligand_summary.txt



```
### PDBspheres.protein_ligand_summary_table.COVID19_PL2pro.txt
### Legend:
# Ns    - number of residues in the protein-ligand template sphere
# Nt    - number of residues in the protein model
# N     - number of residues structurally aligned by LGA
# RMSD  - root mean square deviation calculated on superimposed Calpha or Cbeta atoms from sphere template and detected protein pocket
# Nc    - number of conserved, i.e., “tightly” superimposed residues between “sphere template” and detected pocket in evaluated protein
# SeqID - sequence identity in structure aligned residues. Higher value indicates that protein forming a sphere template and our protein might be close homol
# LGA   - structure similarity based on aligned by LGA program Calpha or Cbeta atoms
# GDC   - structure similarity calculated by LGA program assessing agreement in conformations of all atoms (i.e. including side chain atoms)
# N4    - the number of protein residues within 4.5 Å of the inserted ligand
# cl    - the number of query protein residues that may have possible steric clashes with inserted ligand’s atoms
#Structural_model....Sphere_template                               Ns   Nt    N   RMSD :   Nc  SeqID   LGA    GDC    N4   cl : Contact_residues
nCoV_nsp3.6w9c_A.pdb.Sphere.ZN.1.7kol_A.505_A.12.pdb:              43  318   41   0.46 :   35 100.00  95.17  89.71    5    0 : W851_A,N854_A,C856_A,G1016_A,
nCoV_nsp3.6w9c_A.pdb.Sphere.YRL.10.7ofs_A.404_A.12.pdb:            68  318   68   0.55 :   59 100.00  99.88  92.84   13    0 : T755_A,V756_A,Y801_A,V802_A,L
nCoV_nsp3.6w9c_A.pdb.Sphere.Y97.35.7los_B.404_B.12.pdb:            61  318   60   0.65 :   50 100.00  97.73  91.31   10    0 : L907_A,G908_A,D909_A,E912_A,F
nCoV_nsp3.6w9c_A.pdb.Sphere.Y97.35.7los_A.403_A.12.pdb:            65  318   64   0.67 :   55 100.00  97.76  90.61   10    0 : G908_A,D909_A,E912_A,P993_A,Y
nCoV_nsp3.6w9c_A.pdb.Sphere.Y96.25.7kol_A.501_A.12.pdb:            62  318   61   0.60 :   55 100.00  97.90  91.54   10    0 : L907_A,G908_A,D909_A,E912_A,F
nCoV_nsp3.6w9c_A.pdb.Sphere.Y61.33.7llz_B.401_B.12.pdb:            62  318   61   0.62 :   55 100.00  97.77  92.50   10    0 : L907_A,G908_A,D909_A,E912_A,F
nCoV_nsp3.6w9c_A.pdb.Sphere.Y61.33.7llz_A.401_A.12.pdb:            62  318   61   0.65 :   55 100.00  97.65  91.79   10    0 : L907_A,G908_A,D909_A,E912_A,F
nCoV_nsp3.6w9c_A.pdb.Sphere.Y54.36.7llf_B.404_B.12.pdb:            65  318   64   0.61 :   58 100.00  98.00  92.02   11    0 : L907_A,G908_A,D909_A,E912_A,F
...
```

File2: PDBspheres.PDBbind_Clusters.EC_included.txt



```
### PDBspheres-based clustering of pockets from PDBbind v. 2019
### Legend:
### ClusterNB:     Nc      Np  :members of the cluster (PDBids) separated by colon ‘:’
# Nc - cluster consecutive number
# Np - number of members (pockets) within a cluster
#
# Each of reported clusters is provided with a number of members with known EC annotation and a list of
# corresponding PDBids for which EC numbers are assigned. The alternative EC numbers (second column) are separated by colon ‘:’
#
### ClusterNB:      1      10  :5j41:10gs:2gss:2ca8:3gss:1lbk:1oe8:2c80:2gst:3gst:
# Number of members with known EC annotation: 10
10gs      :2.5.1.18:
1lbk      :2.5.1.18:
1oe8      :2.5.1.18:
2c80      :2.5.1.18:
2ca8      :2.5.1.18:
2gss      :2.5.1.18:
2gst      :2.5.1.18:
3gss      :2.5.1.18:
3gst      :2.5.1.18:
5j41      :2.5.1.18:

### ClusterNB:      2      16  :1lgw:184l:4i7m:1li3:1li6:1l83:4i7j:4w52:4i7p:185l:1li2:4i7l:4i7k:186l:188l:187l:
# Number of members with known EC annotation: 16
184l      :3.2.1.17:
185l      :3.2.1.17:
186l      :3.2.1.17:
...
```

File3: PDBspheres.PDBbind_Clusters.SCOP_included.txt

(the same format as File2, but reporting SCOP annotation)

File4: PDBspheres.PDBbind_Clusters.GO_included.txt
(the same format as File2, but reporting GO annotation)

File5: PDBspheres.PDBbind_Binding_sites_similarities.GDC_and_affinities.txt

```
### PDBspheres-based analysis of similarities in binding sites between proteins from PDBbind v. 2019
### Legend:
# pdb2   - sphere (expanded by 4 angstroms) surrounding pocket from protein pdb2
# pdb1   - pocket from protein pdb1
# lig2   - ligand name (PDBid code) of protein pdb2 used in PDB-spheres
# lig1   - ligand name (PDBid code) of protein pdb1 used in PDB-spheres
# sz2    - size (number of heavy atoms) of ligand from protein pdb2
# sz1    - size (number of heavy atoms) of ligand from protein pdb2
# i22    - number of residues interacting with lig2 in protein pdb2
# i11    - number of residues interacting with lig1 from protein pdb1
# i12    - number of residues interacting with lig1 inserted in protein pdb2
# RMS12  - rmsd calculated on superimposed pockets pdb1 and pdb2
# Sid12  - sequence identity calculated on superimposed pockets pdb1 and pdb2
# LGA12  - LGA structure similarity (Calpha only) calculated on superimposed pockets pdb1 and pdb2
# GDC12  - GDC structure similarity (all atoms including sidechains) calculated on superimposed pockets pdb1 and pdb2
# dist12 - distance between ligand centroids: lig1 inserted in protein pdb2 and original lig2 in protein pdb2
# pblig2 - ligand name of lig2 used in PDBbind
# baf2   - binding affinity "-logKd/Ki" of lig2 reported in PDBbind
# KdKi2  - measurements of Kd/Ki of lig2 reported in PDBbind
# pblig1 - ligand name of lig1 used in PDBbind
# baf1   - binding affinity "-logKd/Ki" of lig1 reported in PDBbind
# KdKi1  - measurements of Kd/Ki of lig1 reported in PDBbind
pdb2   pdb1   lig2   lig1   sz2 sz1 i22 i11 i12  RMS12  Sid12  LGA12  GDC12 : dist12 : pblig2         baf2  KdKi2             pblig1         baf1  KdKi1
10gs   3f37   VWW    2MY     33   9  15   4  10   0.64  13.33  47.25  41.18 :  16.85 : (VWW)          6.40  Ki=0.4uM          (2MY)          2.62  Kd=2.4mM
10gs   5j41   VWW    3LF     33  17  15   8   7   0.25 100.00 100.00  97.96 :   3.94 : (VWW)          6.40  Ki=0.4uM          (2-mer)        3.70  Ki=199uM
10gs   2vo4   VWW    4NM     33  11  15   8   4   1.21  17.39  52.32  43.68 :  16.47 : (VWW)          6.40  Ki=0.4uM          (4NM)          4.94  Kd=11.6uM
10gs   3f8f   VWW    DM1     33  38  15  11  14   0.89  11.11  49.06  42.66 :   5.68 : (VWW)          6.40  Ki=0.4uM          (DM1)          6.63  Kd=236nM
10gs   2gss   VWW    EAA     33  19  15   8   8   0.25 100.00 100.00  99.21 :   3.72 : (VWW)          6.40  Ki=0.4uM          (EAA)          4.94  Ki=11.5uM
10gs   3gx0   VWW    GDS     33  40  15  16  14   1.75  13.64  75.70  63.16 :   3.64 : (VWW)          6.40  Ki=0.4uM          (GDS)          3.48  Kd=330uM
10gs   4f0c   VWW    GDS     33  40  15  16  14   1.90  11.54  72.28  58.79 :   3.75 : (VWW)          6.40  Ki=0.4uM          (GDS)          6.27  Kd=0.54uM
10gs   4zb6   VWW    GDS     33  40  15  19  16   2.09  11.36  63.94  57.27 :   4.02 : (VWW)          6.40  Ki=0.4uM          (GDS)          6.28  Kd=0.53uM
10gs   4zb8   VWW    GDS     33  40  15  14  13   2.00  10.64  64.16  58.34 :   4.13 : (VWW)          6.40  Ki=0.4uM          (GDS)          6.03  Kd=0.94uM
...
```

File6: PDBspheres.PDBbind_Clusters.interactive_plot.html
(interactive overview of predicted clusters shown on Fig.7.)

References

1. Clark, J.J. et al. (2020). Predicting binding sites from unbound versus bound protein structures. Nature Scientific Reports. 10:15856.

2. Laskowski, R.A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J. Mol. Graph 13(323–330), 307–328.

3. Kawabata, T. (2010). Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins 78, 1195–1211. https://doi.org/10.1002/prot.22639.

4. Huang, B., Schroeder, M. (2006). LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct. Biol. 6, 19. https://doi.org/10.1186/1472-6807-6-19.

5. Le Guilloux, V., Schmidtke, P., Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. BMC Bioinform. 10, 168. https://doi.org/10.1186/1471-2105-10-168.

6. Tan, K.P., Varadarajan, R., Madhusudhan, M.S. (2011). DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. Nucleic Acids Res. 39, W242-248. https://doi.org/10.1093/nar/gkr356.

7. Ravindranath, P.A., Sanner, M.F. (2016). AutoSite: an automated approach for pseudo-ligands prediction-from ligand-binding sites identification to predicting key ligand atoms. Bioinformatics 32, 3142–3149. https://doi.org/10.1093/bioinformatics/btw36 7.

8. Stepniewska-Dziubinska, M.M., Zielenkiewicz, P., Siedlecki, P. (2019). Detection of protein-ligand binding sites with 3D segmentation. arXiv e-prints. https://ui.adsas.harvard.edu/abs/2019arXiv190406517S.

9. Clark, J.J. et al. (2019). Inherent versus induced protein flexibility: Comparisons within and between apo and holo structures. PLoS Comput. Biol. 15, e1006705. https://doi.org/10.1371/journal.pcbi.1006705.

10. Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Research, 31(13), 3370–3374.

11. Keedy, D.A. et al. (2009). The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. Proteins, 77 Suppl 9, 29–49.

12. Yoon, S. et al. (2007). Clustering protein environments for function prediction: finding PROSITE motifs in 3D. BMC Bioinformatics, 8.

13. Chen, K. et al. (2011). A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. Structure, 19, 613–621.

14. Jiménez, J. et al. (2017). DeepSite: protein-binding site predictor using 3D-convolutional neural networks, Bioinformatics, Volume 33, Issue 19.

15. Pearson, W.R. (2016). Finding Protein and Nucleotide Similarities with FASTA. Current protocols in bioinformatics, 53, 3.9.1–3.9.25. https://doi.org/10.1002/0471250953.bi0309s53

16. Maggiora, G., Vogt, M., Stumpfe, D., Bajorath, J. (2014). Molecular Similarity in Medicinal Chemistry. J. Med. Chem., 57, 8, 3186–3204, https://doi.org/10.1021/jm401411z

17. Wang, R., Fang, X., Lu, Y., Wang, S. (2004). The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J. Med. Chem. 47 (12): 2977–80. doi:10.1021/jm030580l. PMID 15163179.

18. Gao, M., Skolnick, J. (2013). A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. PLoS Comput Biol 9(10): e1003302. doi:10.1371/journal.pcbi.1003302.

19. Brylinski, M., Skolnick, J. (2008). A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proceedings of the National Academy of Sciences., 105 (1) 129-134; https://doi.org/10.1073/pnas.0707684105.

20. Guterres, H., Park, S., Jiang, W., Im, W. (2021). Ligand-Binding-Site Refinement to Generate Reliable Holo Protein Structure Conformations from Apo Structures. J. Chem. Inf. Model., 61, 1, 535–546, https://doi.org/10.1021/acs.jcim.0c01354.

21. Berman, H.M. et al. (2000). The Protein Data Bank. Nucleic Acids Research, Volume 28, Issue 1, Pages 235–242, https://doi.org/10.1093/nar/28.1.235

22. Zou, K.H. et al. (2007) Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation. 115(5):654-7. doi: 10.1161/CIRCULATIONAHA. 105.594929. PMID: 17283280, https://pubmed.ncbi.nlm.nih.gov/17283280/

23. Powell, J., Mota, F., Steadman, D., Soudy, C., Miyauchi, J.T., Crosby, S., Jarvis, A., Reisinger, T., Winfield, N., Evans, G., Finniear, A., Yelland, T., Chou, Y.T., Chan, A.W.E., O'Leary, A., Cheng, L., Liu, D., Fotinou, C., Milagre, C., Martin, J.F., Jia, H., Frankel, P., Djordjevic, S., Tsirka, S.E., Zachary, I.C., Selwood, D.L. (2018) Small Molecule Neuropilin-1 Antagonists Combine Antiangiogenic and Antitumor Activity with Immune Modulation through Reduction of Transforming Growth Factor Beta (TGF beta) Production in Regulatory T-Cells. J Med Chem 61: 4135-4154, http://dx.doi.org/10.1021/acs.jmedchem.8b00210

24. D. Jones, H. Kim, X. Zhang, A. Zemla, G. Stevenson, W. F. Bennett, D. Kirshner, S. E. Wong, F. C. Lightstone, J. E. Allen (2021). Improved Protein-Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference, Journal of Chemical Information and Modeling, 61, 4, 1583-1592, https://doi.org/10.1021/acs.jcim.0c01306, PMID: 33754707.

25. Degac, J. et al. (2015). Graph-Based Clustering of Predicted Ligand-Binding Pockets on Protein Surfaces. Journal of Chemical Information and Modeling, 55(9), 1944–1952.

26. Ehrt, C. et al. (2018). A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). PLOS Computational Biology, 14(11).

27. Govindaraj, R.G. et al. (2018). Comparative assessment of strategies to identify similar ligand-binding pockets in proteins. BMC Bioinformatics, 19.

28. Kahraman, A. et al. (2007). Shape variation in protein binding pockets and their ligands. Journal of Molecular Biology, 368(1), 283–301.

29. Kahraman, A. et al. (2010). On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. Proteins. 78(5), 1120–36.

30. Hoffmann, B. et al. (2010). A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. BMC Bioinformatics, 11.

31. Konc, J. et al. (2010). ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics. 26(9), 1160–8.

32. Dias, S. et al. (2019). CavBench: A benchmark for protein cavity detection methods. PLoS ONE 14(10).

33. Glaser, F. et al. (2006). A Method for Localizing Ligand Binding Pockets in Protein Structures. Proteins, 62, 479–488.

34. Spitzer, R. et al. (2011). Surface-Based Protein Binding Pocket Similarity. Proteins. 79(9): 2746–2763.

35. Stepniewska-Dziubinska, M.M., Zielenkiewicz, P., Siedlecki, P. (2020). Improving detection of protein-ligand binding sites with 3D segmentation. Nature Scientific Reports, 10:5035, https://doi.org/10.1038/s41598-020-61860-z