Janez Konc

Uporaba spletnega orodja ProBiS-Dock Database za napovedovanje PDB vezavnih mest primernih za razvoj zdravil

Učno gradivo za študente univerzitetnega študijskega programa 2. stopnje Matematika pri predmetu Izbrane teme iz računsko intenzivnih metod

Univerza na Primorskem, Fakulteta za matematiko, naravoslovje in informacijske tehnologije
Glagoljaška 8, SI-6000, Koper, Slovenija

Koper, 2023

Uporaba spletnega orodja ProBiS-Dock Database za napovedovanje PDB vezavnih mest primernih za razvoj zdravil; Učno gradivo pri predmetu: Izbrane teme iz računsko intenzivnih metod

Namenjeno študentom univerzitetnega študijskega programa 2. stopnje Matematika, Univerza na Primorskem, Glagoljaška 8, SI-6000, Koper, Slovenija

Avtor: Janez Konc, konc@cmm.ki.si

Izšlo: Koper, 2023

Table of Contents

Background

The relevance of docking as an approach to target prediction is continually growing due to the expanding availability of protein structures obtained through X-ray crystallography, nuclear magnetic resonance spectroscopy, and electron microscopy. This increase in structural data has facilitated the development of the ProBiS-Dock Database, which encompasses all protein binding sites inferred from the Protein Data Bank (PDB), currently containing over 200,000 protein structures as of 2023. This database serves as a valuable resource for identifying proteins with specific binding sites, thereby holding tremendous potential for drug design.

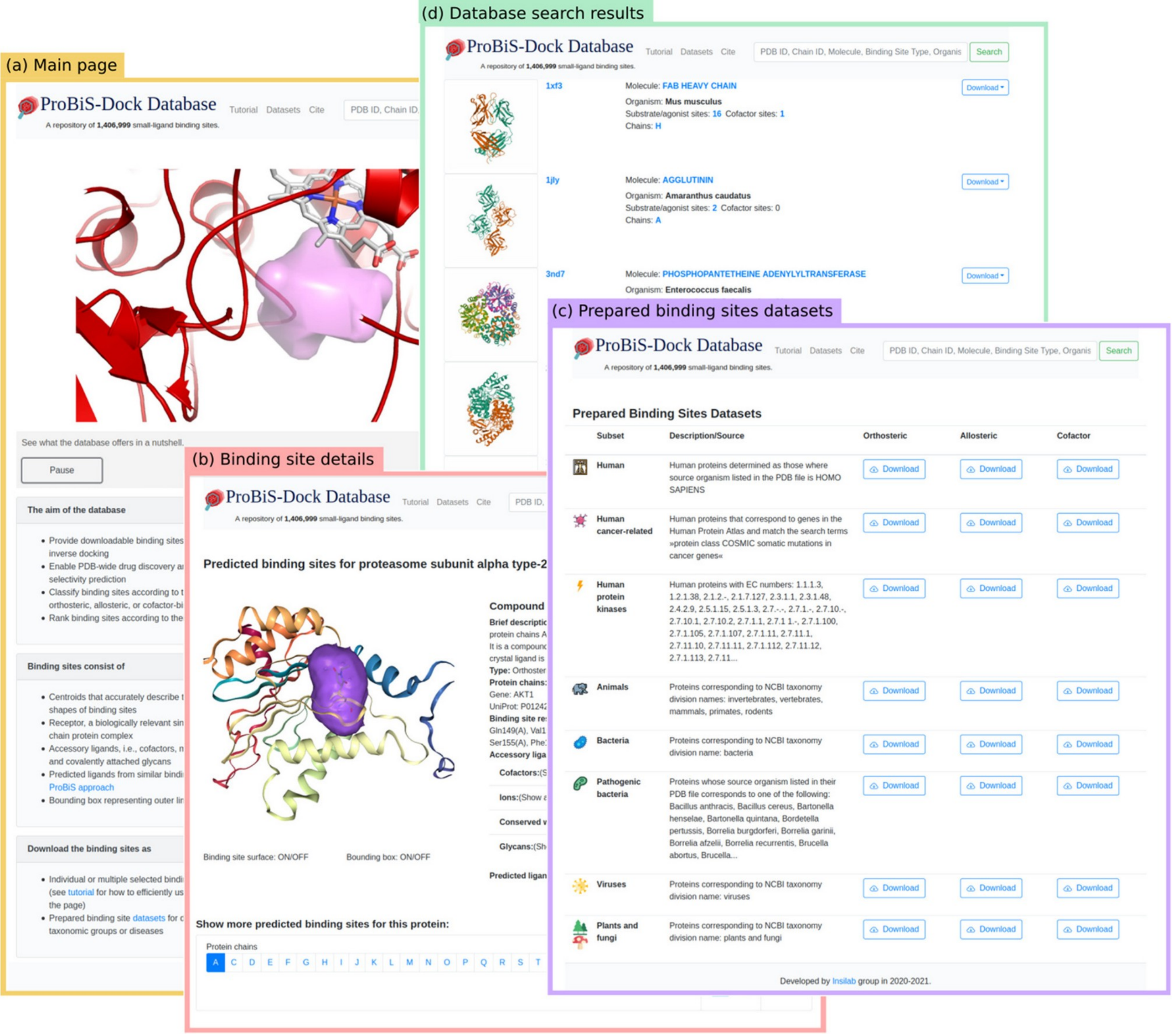
Advancements in homology modeling further contribute to the accumulation of structural information, enabling broader coverage of the proteome for various pharmacologically significant protein classes. Nevertheless, the challenge persists when it comes to docking techniques that consider the entire protein surface, particularly for large collections of proteins like the PDB. To address this issue, researchers have proposed narrowing down the search space for docking by focusing on binding sites. However, in many cases involving apoproteins with missing ligands, the locations of these binding sites may remain unknown.

ProBiS binding site prediction approaches

The ProBiS-ligand web server employs the ProBiS algorithm to accurately predict ligands and their respective poses for a given binding site by detecting similarities among different binding sites. This web server has undergone comprehensive validation to demonstrate its effectiveness in inferring similar binding sites even in the absence of closely related homologous proteins. Academic users can freely access this web server at http://probis.cmm.ki.si.

An extension of the ProBiS-ligand web server, called the ProBiS-CHARMMing web interface, is also available at http://probis.nih.gov. This interface not only predicts ligands but also facilitates their minimization and calculates the interaction energy of the predicted protein-ligand complex. Integration with the CHARMMing web server enables the utilization of the CHARMM force field to perform minimization and compute potential energies in various biomolecular systems. The ProBiS-CHARMMing web interface offers the capability to construct proteins with bound ligands based on their corresponding protein structures.

Additionally, the recently introduced GenProBiS web server (http://genprobis.insilab.org) maps sequence variants to protein structures and identifies their associations with protein-protein, protein-nucleic acid, protein-compound, and protein-metal ion binding sites. This server allows intuitive visual exploration of extensively mapped variants, such as somatic missense mutations associated with human cancer and nonsynonymous single nucleotide polymorphisms from 21 different species, within the predicted binding site regions of approximately 80,000 PDB protein structures. It serves as a valuable tool for discovering potentially deleterious sequence variants and generating new hypotheses in the field of drug discovery.



(a) Main page

ProBiS-Dock Database Tutorial Datasets Cite PDB ID, Chain ID

A repository of **1,406,999** small-ligand binding sites.

See what the database offers in a nutshell.

Pause

The aim of the database

- Provide downloadable binding sites inverse docking
- Enable PDB-wide drug discovery a selectivity prediction
- Classify binding sites according to t orthosteric, allosteric, or cofactor-bi
- Rank binding sites according to the

Binding sites consist of

- Centroids that accurately describe t shapes of binding sites
- Receptor, a biologically relevant sin chain protein complex
- Accessory ligands, i.e., cofactors, n and covalently attached glycans
- Predicted ligands from similar bindi ProBiS approach
- Bounding box representing outer lin

Download the binding sites as

- Individual or multiple selected bindi (see tutorial for how to efficiently us the page)
- Prepared binding site datasets for c taxonomic groups or diseases

(b) Binding site details

ProBiS-Dock Database Tutorial Datasets Cite PDB ID,

A repository of **1,406,999** small-ligand binding sites.

Predicted binding sites for proteasome subunit alpha type-2

Compound
Brief descriptic
protein chains A
It is a compounc
crystal ligand is
Type: Orthoster
Protein chains:
Gene: AKT1
UniProt: P01242
Binding site re:
Gln149(A), Val1
Ser155(A), Phe:
Accessory liga
Cofactors:(S
Ions:(Show a
Conserved v
Glycans:(Sh
Predicted ligan

Binding site surface: ON/OFF Bounding box: ON/OFF

Show more predicted binding sites for this protein:

Protein chains

A C D E F G H I J K L M N O P Q R S T

(c) Prepared binding sites datasets

ProBiS-Dock Database Tutorial Datasets Cite PDB ID, Chain ID, Molecule, Binding Site Type, Organis Search

A repository of **1,406,999** small-ligand binding sites.

Prepared Binding Sites Datasets

Subset	Description/Source	Orthosteric	Allosteric	Cofactor
Human	Human proteins determined as those where source organism listed in the PDB file is HOMO SAPIENS	Download	Download	Download
Human cancer-related	Human proteins that correspond to genes in the Human Protein Atlas and match the search terms »protein class COSMIC somatic mutations in cancer genes«	Download	Download	Download
Human protein kinases	Human proteins with EC numbers: 1.1.1.3, 1.2.1.38, 2.1.2.-, 2.1.7.127, 2.3.1.1, 2.3.1.48, 2.4.2.9, 2.5.1.15, 2.5.1.3, 2.7.-.-, 2.7.1.-, 2.7.10.-, 2.7.10.1, 2.7.10.2, 2.7.1.1, 2.7.1 1.-, 2.7.1.100, 2.7.1.105, 2.7.1.107, 2.7.1.11, 2.7.11.1, 2.7.11.10, 2.7.11.11, 2.7.1.112, 2.7.11.12, 2.7.1.113, 2.7.11...	Download	Download	Download
Animals	Proteins corresponding to NCBI taxonomy division names: invertebrates, vertebrates, mammals, primates, rodents	Download	Download	Download
Bacteria	Proteins corresponding to NCBI taxonomy division name: bacteria	Download	Download	Download
Pathogenic bacteria	Proteins whose source organism listed in their PDB file corresponds to one of the following: Bacillus anthracis, Bacillus cereus, Bartonella henselae, Bartonella quintana, Bordetella pertussis, Borrelia burgdorferi, Borrelia garinii, Borrelia afzelii, Borrelia recurrentis, Brucella abortus, Brucella...	Download	Download	Download
Viruses	Proteins corresponding to NCBI taxonomy division name: viruses	Download	Download	Download
Plants and fungi	Proteins corresponding to NCBI taxonomy division name: plants and fungi	Download	Download	Download

Developed by Insilab group in 2020-2021.

(d) Database search results

ProBiS-Dock Database Tutorial Datasets Cite PDB ID, Chain ID, Molecule, Binding Site Type, Organis Search

A repository of **1,406,999** small-ligand binding sites.

1xf3 Molecule: FAB HEAVY CHAIN Download ▾
Organism: **Mus musculus**
Substrate/agonist sites: 16 Cofactor sites: 1
Chains: H

1jly Molecule: AGGLUTININ Download ▾
Organism: **Amaranthus caudatus**
Substrate/agonist sites: 2 Cofactor sites: 0
Chains: A

3nd7 Molecule: PHOSPHOPANTETHEINE ADENYLYLTRANSFERASE Download ▾
Organism: **Enterococcus faecalis**

Figure 1. ProBiS-Dock Database web server.

ProBiS-Dock represents an expansion of the ProBiS-ligands and ProBiS-CHARMMing methodologies, integrating their capabilities. This extension facilitates the creation of the ProBiS-Dock Database (**Figure 1**), a comprehensive repository of small ligand-binding sites derived from various species in the Protein Data Bank (PDB). Each ligand type within the database is meticulously defined. Unlike the previous GenProBiS web server, ProBiS-Dock also predicts binding sites involving multiple protein chains.

An earlier version of the ProBiS-Dock Database was employed in conjunction with an inverse molecular docking approach to identify novel target proteins associated with natural products like resveratrol, curcumin, and antidiabetic drugs such as rosiglitazone and troglitazone. This application aimed to elucidate the known effects of these compounds and make predictions about potential new effects. Notably, in contrast to protein structure-based methods for binding site prediction (28,29), the ProBiS-Dock Database defines a binding site as the collective space occupied by all predicted ligands from structurally similar entries in the PDB, utilizing the ProBiS-ligand approach.

In conventional docking procedures, binding sites necessitate manual curation before initiating the docking process. In contrast, the described method automates the binding site curation procedure by automatically selecting the relevant bound ligands within each PDB file, distinguishing them based on whether they should or should not be present in the specific binding site during docking. These auxiliary ligands, categorized as cofactors, metal ions, glycans, or conserved waters, can significantly impact the binding of a drug to a protein.

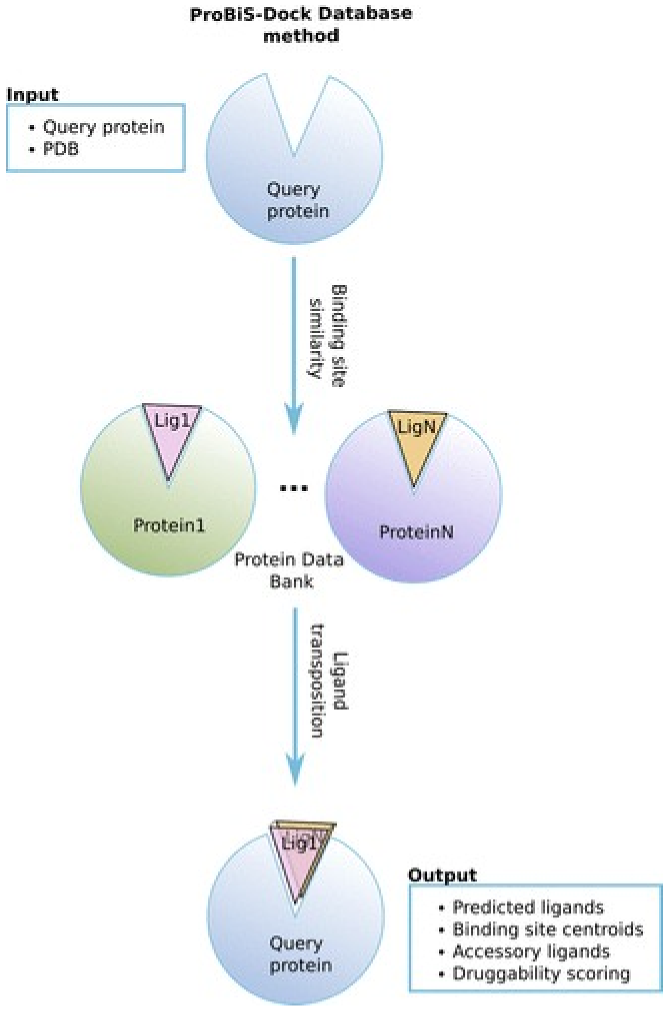


Figure 2. Construction of the binding site database.

Methodology

The ProBiS-Dock method is an extension of previous approaches called ProBiS-ligands and ProBiS-CHARMMing, and it aims to construct the ProBiS-Dock Database. This database contains small ligand-binding sites from different species in the Protein Data Bank (PDB), where each ligand type is defined. The method predicts binding sites between multiple protein chains, which is a new feature compared to the previous GenProBiS web server. The ProBiS-Dock Database has been used in conjunction with inverse molecular docking to discover new target proteins associated with natural products and drugs, explaining their effects and predicting new ones.

The construction of the ProBiS-Dock Database involves several steps (**Figure 2**). First, protein chains from the PDB are clustered based on sequence identity, resulting in representative protein chains for each cluster. Then, structural superimposition is performed within each cluster to align the protein chain structures. The rotational-translational matrices obtained from these superimpositions are saved for later use. Next, the representative protein chain structures are superimposed on each other across different clusters, resulting in a large number of protein local structural superimpositions. The rotational-translational matrices from these superimpositions are also saved.

In the prediction of ligands, biologically relevant ligands from similar proteins are transposed onto the query protein chains using the saved rotational-translational matrices. The similarity between binding sites is measured using a Z-score metric, and ligands are transferred if the similarity is sufficient. The ligand types are determined, including compounds, cofactors, glycans, metal ions, and conserved water molecules. Nonspecific binders and known crystallization artifacts are identified and filtered out. The ligand residues in each PDB entry are classified into different types based on comparisons with known cofactors and monosaccharides.

To generate biologically relevant protein complexes, the presumed biological assembly of each protein chain is determined, and all molecules in contact with the query protein chain are considered as possibly relevant ligands. Biologically relevant binding sites consisting of multiple protein chains are identified by combining the binding site predictions for the constituent single protein chains. The binding sites are represented by overlapping centroids (**Figue 3**) calculated at regular intervals in the space defined by the ligand atoms. Accessory ligands, such as cofactors, metal ions, conserved water molecules, and glycans, are determined based on their proximity to the binding site centroids.

Equivalent binding sites that are functionally equivalent are identified by comparing the predicted ligands. Binding sites are ranked according to their "druggability" based on the molecular complexity and number of constituent ligands. The ranking score takes into account the complexity of ligands and the occupancy of the binding site.

Overall, the ProBiS-Dock method constructs the ProBiS-Dock Database by predicting binding sites, determining ligand types, generating biologically relevant protein complexes, identifying binding sites consisting of multiple protein chains, and ranking the binding sites based on their druggability. This database can be useful for drug development and understanding ligand-protein interactions.

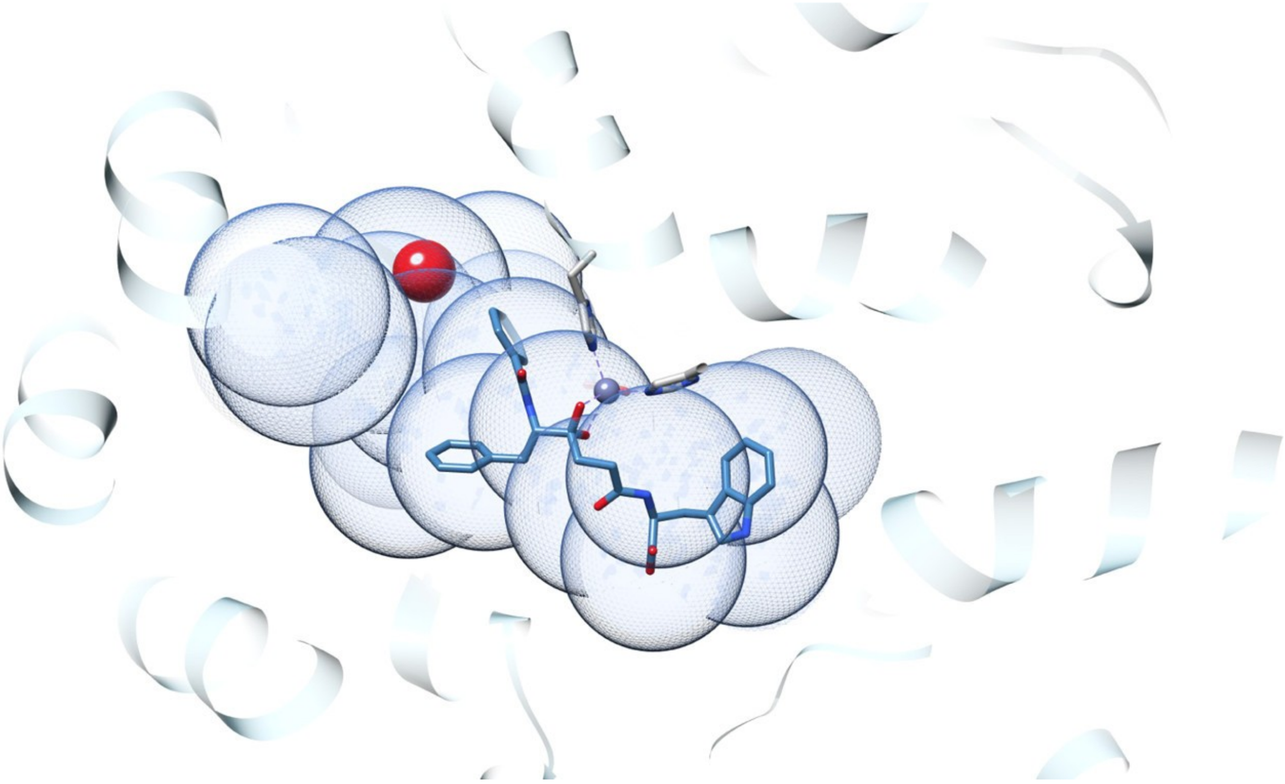


Figure 3. Overlapping centroids representing a binding site.

Using the search bar to retrieve binding sites

Basic usage

Any text can be entered in the search bar, which results in a free text search. A query term `kinase cancer`, for example, is translated to the search for the binding site entries that contain the (sub-)terms kinase and cancer anywhere in their descriptions. The white space character between the terms is interpreted as the logical AND operator.

Advanced usage

Alternatively, a query can be made specific by providing the *column name* in which to perform the search, and the value for this column. For example, `pdb_id=1got` retrieves exactly one protein from the database.

Valid column names are:

- **pdb_id** - Four letter PDB identifier, *e.g.*, 1all
- **chain_id** - Protein chain identifier, *e.g.*, A, B, ...
- **rest** - Binding site type [compound or cofactor]
- **bs_id** - Binding site number according to our druggability ranking

- **organism_scientific** - Scientific name of the organism from which the protein originates *e.g. Homo sapiens*
- **protein_class** - Protein class [kinase]
- **disease** - Disease relatedness [cancer]
- **ec** - Enzyme Commision number *e.g.*, 3.4.25.1
- **molecule** - Name of the protein *e.g.*, PROTEASOME SUBUNIT BETA TYPE-3
- **division_id** - NCBI taxonomy database division identifier [0=Bacteria, 1=Invertebrates, 2=Mammals, 3=Phages, 4=Plants and Fungi, 5=Primates, 6=Rodents, 7=Synthetic and Chimeric, 8=Unassigned, 9=Viruses, 10=Vertebrates, 11=Environmental samples]
- **tax_id** - NCBI taxonomy identifier, *e.g.*, 9606

The following examples show how to properly use the query expression language in the search bar to retrieve whatever binding sites you wish from the database:

- All binding sites: `leave empty`
- Primary binding sites (mostly orthosteric): `bs_id=1`
- Secondary binding sites (allosteric/orthosteric/other): `bs_id>1` (note: our method does not reliably determine if a site is allosteric)
- Compound binding sites (for substrates, agonists,...): `rest=compound`
- Cofactor binding sites: `rest=cofactor`
- Binding sites ranked No. 1 and 2: `bs_id=[1,2]`
- Human kinase binding sites: `organism_scientific="homo sapiens" protein_class=kinase` (note the use of double quotes in search terms composed of two or more words)
- Human cancer-related binding sites: `organism_scientific="homo sapiens" disease=cancer`
- Animal binding sites: `division_id=[1,10,2,5,6]`
- Bacterial binding sites: `division_id=0`
- Pathogenic bacterial binding sites: `organism_scientific=["BACILLUS ANTHRACIS","BACILLUS CEREUS","BARTONELLA HENSELAE",...]`

Predicted binding sites

Binding sites were predicted using the ProBiS-Dock Database method for the whole PDB. Here, we showcase a few examples of potential usage, advantages and properties of the predicted binding sites.

Binding site centroids

Binding site grid is then generated for each ligand cluster by sampling hexagonal close-packed points spaced 0.2 Å apart that fall within the radius of any of the ligands' atoms but do not overlap with any of the protein atoms. A binding site grid thus follows the contours of the molecular surface of the biological assembly and the space occupied by the predicted ligands up to 8 Å from the protein surface. Binding site centroids are then calculated as grid points sampled at approximately 3 Å intervals. Each centroid is assigned a radius of about 4 Å, and each binding site is represented by a set of overlapping centroids with radiuses that closely follow its contours.

Primary and secondary binding sites

Primary binding sites are those with rank equal to 1 and typically correspond to a main binding site in a protein. Secondary binding sites are those with rank>1 according to our binding sites prioritization score. Ligand binding to an secondary site that is also an allosteric site can lead to a conformational change within the orthosteric binding site, thus modulating the protein's activity.[5,6] As such, secondary sites are important in proteins as they often serve as natural control loops, such as feedback from downstream products of enzymes, while also being crucial in cell signaling. The secondary binding sites in our database can readily be used in the identification of previously unknown binding locations and subsequently the design of drugs acting on previously un-targeted binding sites, potentially resulting in drugs exhibiting novel and unique scaffolds, while still acting on the same target protein as the existing drugs that target orthosteric sites.

Accurate binding sites 3D shapes for docking

Most docking algorithms require that the search space in which the algorithm samples ligand poses, is defined prior to the start of the docking process. If this space is too large, the algorithm will require long computational times and may predict non-native poses; if the search space however, is too small, this could prevent larger potentially appropriate ligands to dock. It was determined that the optimal search space for the AutoDock Vina docking program [1] is a cube centered on the co-crystallized ligand with sides having 2.9× the radius of gyration of the currently docked ligand [2]. We show that for the human adenosine A2A receptor (PDB: 2ydo), this definition can lead to a suboptimal search space (**Figure 4**, orange box). Here, the cubic search space is defined as the box centered on the co-crystallized adenosine with sides 2.9× radius of gyration of a reported high-affinity antagonist (PDB: 3eml; ligand code: ZMA). It can be observed that that search space calculated using this definition does not include the fenolic oxygen of the ZMA ligand, which may result in poor docking of this compound. On the contrary, the ProBiS-Dock Database method defines this binding site using a set of centroids (Figure 4, green mesh spheres) following the shape of transposed ligands from similar binding sites, which are translated to a rectangular box shape. Here, both the adeonsine and the ZMA ligands are within the binding site search space (**Figure 1**, green box). This extended binding site determined by ProBiS-Dock Database should enable successful docking of larger ligands that could form additional interactions with the part of the binding site unoccupied by the co-crystallized ligands in this and other protein structures.

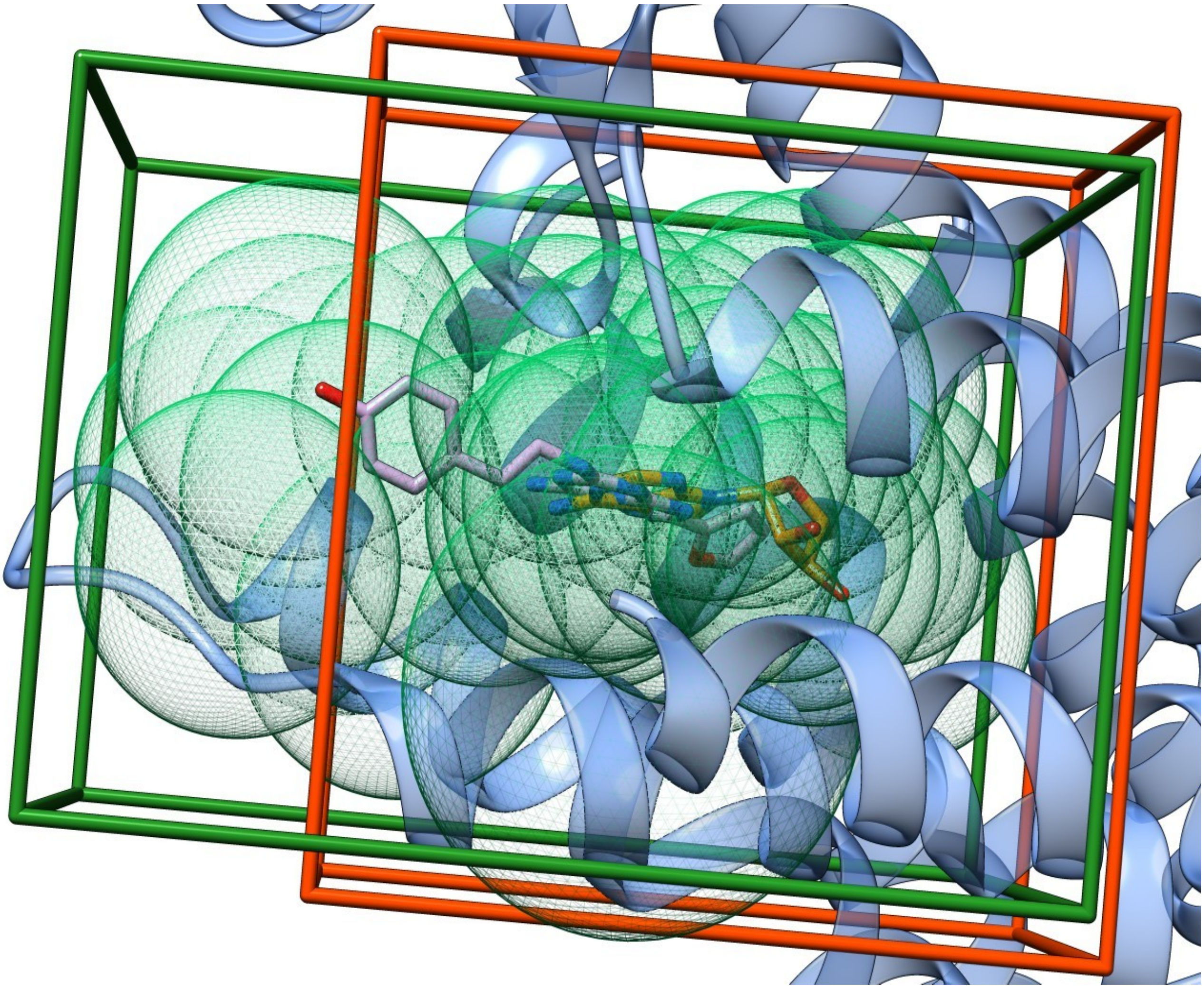


Figure 4. Alternative docking search spaces in human adenosine A2A receptor (PDB: 2ydo) (blue ribbons). The first search space (orange box) is defined by the procedure in Ref. [2] with adenosine (orange sticks) at its center. The second search space is defined by the ProBiS-Dock Database method (green box) as the smallest rectangular box that envelops the binding site centroids (green mesh spheres). The oxygen atom of the phenolic group of the ZMA ligand (pink sticks) is outside of the first search space, but within the search space defined by the green box.

Orthosteric and allosteric binding sites

Protein function can be regulated by binding of ligands to allosteric binding sites that are distinct from orthosteric sites [3]. Here, we present the ability of our method to detect allosteric binding sites in addition to orthosteric sites on the cathepsin K protein structure (PDB: 7pck). We used an apo-form of the cathepsin K structure, therefore the algorithm does not have prior knowledge about the ligands positions or binding sites. The second ranked binding site detected by our method (**Figure 5**, red spheres) corresponds to a known allosteric site as reported in the PDB structure 5j94, while the first ranked binding site (**Figure 5**, cyan spheres) corresponds to a known orthosteric site found in the structure 4dmx. The allosteric binding site is located on a flat part of the protein surface and therefore could not be detected by using, for example, cavity detection algorithms. On the contrary, in the ProBiS-Dock Database binding sites are defined based on their ligands, and thus the method can detect such sites as well.

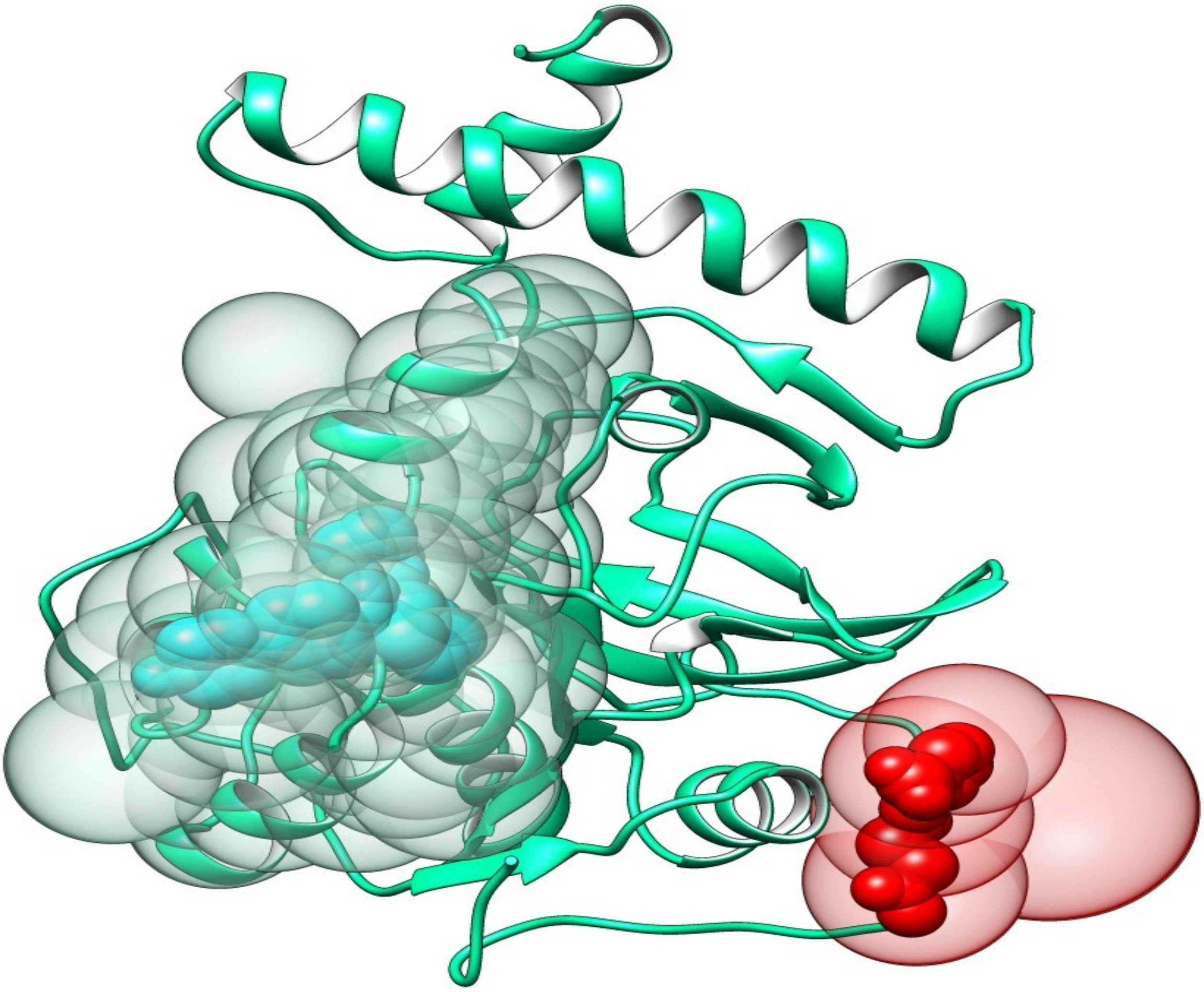


Figure 5. Allosteric binding site prediction on the structure of cathepsin K (PDB: 7pck). Allosteric binding site centroids are shown as red transparent spheres and orthosteric binding site is shown as cyan transparent spheres. Ligands within each binding site are nontransparent spheres.

Interface binding sites composed of multiple protein chains

Binding sites are often located at the interface of two or more protein chains. Using an asymmetric unit of such proteins instead of the biologically relevant assembly can lead to worse docking outcomes. Such binding sites are correctly identified in the ProBiS-Dock Database which is shown on the guanine nucleotide exchange factor (**Figure 6**). The ligand binding site in this protein is a cleft formed by the interface of two protein chains (A and E). With the omission of the chain E, a hydrogen bond that Tyr190 forms with the sp3-hybridized oxygen on the native ligand brefeldin A would be lost, in addition to two hydrophobic contacts with Ala157 and Thr197, which would be omitted. Taken together, considering only one protein chain would most likely reduce the accuracy of ligand docking to this binding site.

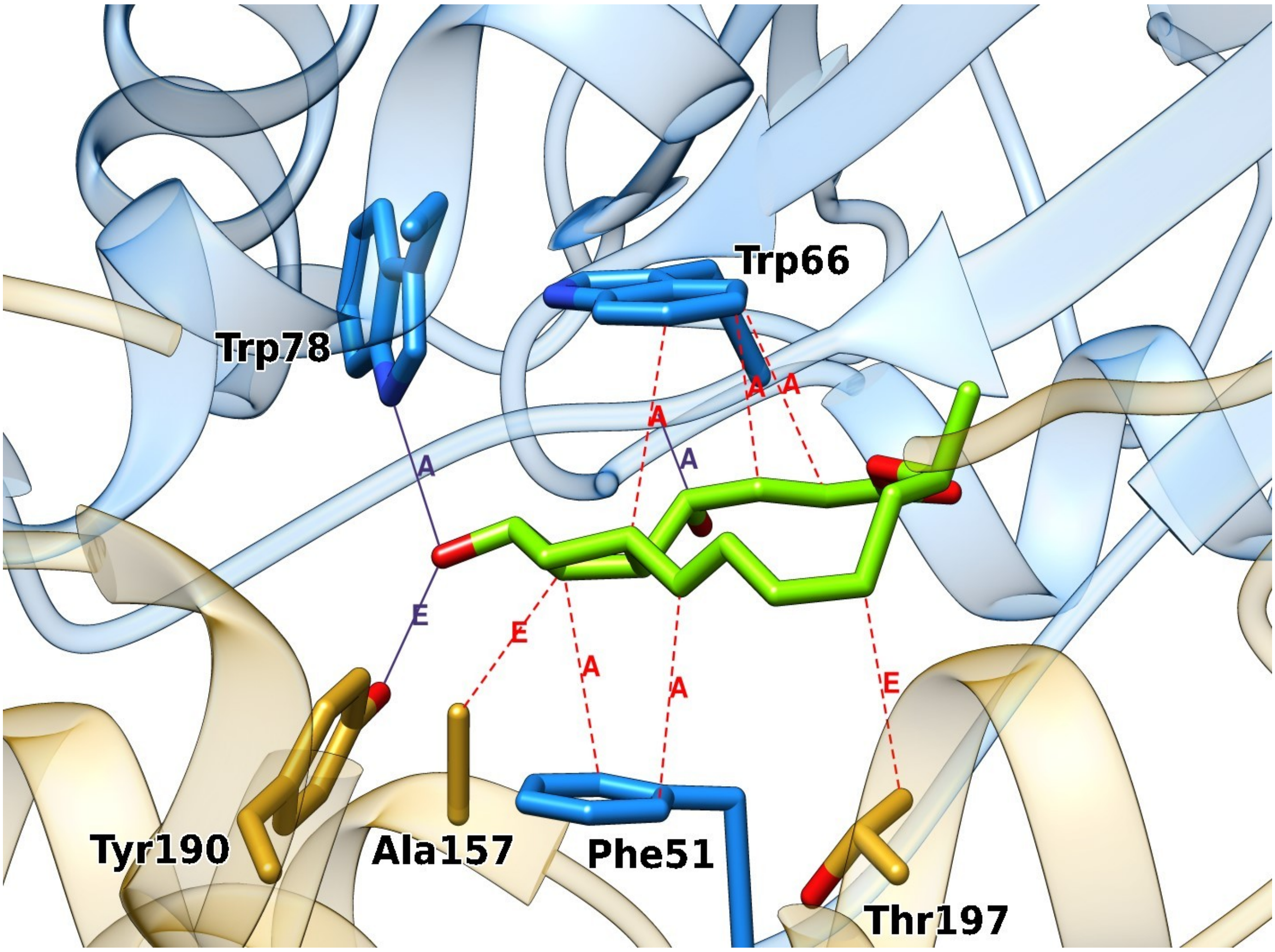


Figure 6. Interface binding site on the guanine nucleotide exchange factor (PDB: 1r8q). The ligand brefeldin A (green sticks) binds at the interface of chains A (blue ribbons and sticks) and E (gold ribbons and sticks). Hydrogen bonds are purple lines and hydrophobic interactions are red dotted lines, labeled with the chain identifiers.

Accessory ligands

Accessory ligands are ligands from the original PDB entry that may significantly influence binding of a docked ligand (eg. a drug candidate) to a protein (depending on the docking application) in the specific binding site during docking. They are determined by our method automatically. Accessory ligands are classified as cofactors, glycans, metal ions, and conserved waters.

Cofactor accessory ligands

Cofactor accessory ligands and cofactor binding sites are identified based on the list of known cofactors extended with a few more. The coordinate file for each cofactor are obtained from the PubChem database, and, basically, all PDB ligands that are very similar to the cofactors in this list, are considered cofactors themselves. The cofactors that we consider are the following:

- Adenosine-5'-diphosphate
- Adenosine monophosphate
- Guanosine-5'-triphosphate
- Phosphoadenosine phosphosulfate
- Pyrroloquinoline quinone
- Pyridoxal phosphate
- Thiamine diphosphate
- UDP-D-galactose dianion
- Coenzyme F420
- Glutathione
- Biotin
- Guanosine diphosphate mannose
- Coenzyme B
- Coenzyme M
- Tetrahydrobiopterin
- Heme B
- Molybdopterin
- Coenzyme Q10
- Vitamin K1 also Phytonadione
- N-Formylmethanofuran
- Tetrahydromethanopterin
- Cobalamine
- Triphosphopyridine nucleotide
- Adenosine-5'-triphosphate
- Cytidine-5'-triphosphate
- Methylcobalamin
- Flavin adenine dinucleotide
- Flavin mononucleotide
- Coenzyme A
- (R)-Lipoamide
- Tetrahydrofolic acid
- beta-Nicotinamide adenine dinucleotide
- S-Adenosyl methionine
- Ferroheme A also Heme A
- Vitamin C also Ascorbic acid

Glycan accessory ligands

This is the list of known monosaccharides that are used to determine if a PDB ligand is a part of a glycan.

- Pseudaminic acid
- L-Altrose
- 2-Keto-3-Deoxy-D-Mannooctanoic Acid
- D-Olivose

- N-Acetyl-D-Allosamine
- D-Ribose
- L-Idose
- D-Taluronic Acid
- D-Guluronic Acid
- N-Acetyl-D-Quinovosamine
- 6-Deoxy-L-Altose
- 6-Deoxy-D-gulose
- L-Colitose
- D-Tyvelose
- D-Paratose
- L-Apiose
- N-Acetyl-6-deoxy-L-altrosamine
- N-Acetyl-6-deoxy-D-talosamine
- Acinetaminic acid
- 4-Epilegionaminic acid
- N-Acetyl-D-Gulosamine
- N-Acetyl-L-Idosamine
- D-Xylose
- Keto-Deoxy-Nonulonic acid
- 3-Deoxy-D-Lyxo-Heptopyran-2-ularic Acid
- D-Abequose
- L-Fucose
- D-Mannose
- L-Glycero-D-Manno-Heptose
- D-Galactosamine
- L-Rhamnose
- D-Digitoxose
- D-Fructose
- N-Acetyl-D-Galactosamine
- N-Acetyl-D-Glucosamine
- L-Sorbose
- L-Arabinose
- N-Acetyl-Neuraminic Acid
- D-Glucosamine
- D-Galacturonic Acid
- D-Lyxose
- N-Acetyl-D-Mannosamine
- D-Tagatose
- D-Allose
- D-Mannuronic Acid
- D-Quinovose
- N-Glycolyl-Neuraminic acid
- D-Mannosamine

- D-Gulose
- D-Talose
- D-Psicose
- Neuraminic acid
- Muramic acid
- L-Iduronic Acid
- 6-Deoxy-D-Talose
- N-Acetyl-L-Altrosamine
- D-Glycero-D-Manno-Heptose
- D-Bacillosamine
- N-Acetyl-Muramic Acid
- L-Alturonic Acid
- N-Acetyl-D-Talosamine
- D-Glucose
- D-Galactose
- L-Altrosamine
- D-Alluronic Acid
- D-Allosamine
- Legionaminic acid
- N-Acetyl-L-Rhamnosamine
- N-Acetyl-L-Fucosamine
- N-Glycolyl-Muramic Acid
- L-Idosamine
- D-Glucuronic Acid
- D-Talosamine
- D-Gulosamine

Metal ions and conserved waters

Biologically relevant metal ions and conserved water molecules are determined by counting the number of times an ion is found in similar binding sites at the same or a similar position according to the methodology described in the ProBiS H2O approach. Biologically relevant ions are identified based on the candidate ions (see Figure 3 in the accompanying paper) and an additional filter which is used to determine that they belong to clusters of at least 10 members. Those ions that do not meet both criteria are considered artifacts and classified as buffer. Similarly, water is labeled as conserved water if it belongs to clusters with >10 members, otherwise it is considered to bind nonspecifically.

Predicted ligands

The criterion for assuming that a ligand of a protein can be transposed into a binding site on a query protein is the similarity between the binding site of the originating protein and the binding site of the query protein. Ligands are transposed from similar proteins if they have binding sites that are sufficiently similar to the binding site(s) on the query protein. Sufficiency for transposition is determined separately for each ligand type using a Z-score metric.

Prediction of ligands by transposition from similar binding sites

Z-scores are assigned by the ProBiS-ligands approach to each pairwise protein superimposition and measure the local structural similarity of the superimposed protein patches, where higher Z-scores indicate higher structural similarity of the compared binding sites.[4] For compounds, cofactors, glycans, and water molecules the superimpositions with Z-score ≥ 2.5 are used, while for metal ions this threshold is set to ≥ 2.0. Further, three different cases are distinguished for transposition: if a ligand originates from a non-representative protein within the same sequence cluster (Step 1, see our paper) as the query protein chain, then the rotational-translational matrix obtained in Step 2 is applied to the ligand's coordinates to transpose them into the coordinate frame of the query protein chain; if the ligand originates from a representative protein of another cluster, then the rotational-translational matrix obtained in Step 3 is used; finally, if the ligand is from a non-representative protein from another sequence cluster, then both the corresponding matrices from Step 2 and Step 3 are applied to the ligand's coordinates to transpose the ligand into the binding site of the query protein.

Nonspecific Binders

This is an updated and extended list of non-specific binders given as PDB Chemical IDs based on the list of non-specific binders available here.

12P, 144, 15P, 16D, 16P, 1BO, 1PE, 1PG, 1PS, ACA, ACE, ACN, ACT, ACY, AE3, AE4, AGC, AZI, B3P, B7G, BCN, BE7, BEN, BEQ, BEZ, BGC, BMA, BNG, BOG, BTB, BTC, BU1, BU2, BU3, C10, C15, C8E, CAC, CBM, CBX, CCN, CE1, CIT, CM, CM5, CN, CPS, CRY, CXE, CYN, CYS, D10, DDQ, DHD, DIA, DIO, DMF, DMS, DMU, DMX, DOD, DOX, DPR, DR6, DTT, DXE, DXG, EDO, EEE, EGL, EOH, EPE, ETE, ETF, FCL, FCY, FMT, FRU, GBL, GCD, GLC, GLO, GLY, GOL, GPX, HEZ, HTG, HTO, ICI, ICT, IDT, IOH, IPA, IPH, JEF, LAK, LAT, LBT, LDA, LMT, M2M, MA4, MAN, ME2, MES, MG8, MHA, MLI, MOH, MPD, MPO, MRD, MRY, MTL, MXE, N8E, NDG, NH4, NHE, NO3, O4B, OTE, P15, P33, P3G, P4C, P4G, P6G, PDO, PE3, PE4, PE5, PE7, PE8, PEG, PEU, PG0, PG4, PG5, PG6, PGE, PGF, PGO, PGQ, PGR, PIG, PIN, PO4, POL, SAL, SBT, SCN, SDS, SO4, SOR, SPD, SPK, SPM, SUC, SUL, SYL, TAR, TAU, TBU, TEP, TLA, TMA, TOE, TRE, TRS, TRT, UMQ, UNK, URE, VO4, XPE, XYP, AL, CS, BR, CL, F, IOD, PB, LI, HG, K, RB, AG, NA, SR, YT3, Y1, XE

References

1. Konc, J.; Lešnik, S.; Škrlj, B.; Janežič. D. ProBiS-Dock Database: A Web Server and Interactive Web Repository of Small Ligand Protein Binding Sites for Drug Design. *J. Chem. Inf. Model.*, **2021**, 61, 4097-4107.
2. Trott, O.; Olson, A. J. *AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading.* J. Comput. Chem. **2010**, 31 (2), 455–461.

3. Feinstein, W. P.; Brylinski, M. *Calculating an Optimal Box Size for Ligand Docking and Virtual Screening against Experimental and Predicted Binding Pockets.* J. Cheminf. **2015**, 7 (1), 18.
4. Huang, Z.; Zhu, L.; Cao, Y.; Wu, G.; Liu, X.; Chen, Y.; Wang, Q.; Shi, T.; Zhao, Y.; Wang, Y.; Li, W.; Li, Y.; Chen, H.; Chen, G.; Zhang, J. *ASD: A Comprehensive Database of Allosteric Proteins and Modulators.* Nucleic Acids Res. **2011**, 39 (suppl_1), D663–D669.
5. Konc, J.; Česnik, T.; Konc, J. T.; Penca, M.; Janežič, D. *ProBiS-Database: Precalculated Binding Site Similarities and Local Pairwise Alignments of PDB Structures.* J. Chem. Inf. Model. **2012**, 52 (2), 604–612.
6. Bu, Z.; Callaway, D. J. E. *Chapter 5 - Proteins MOVE! Protein Dynamics and Long-Range Allostery in Cell Signaling.* In Advances in Protein Chemistry and Structural Biology; Donev, R., Ed.; Protein Structure and Diseases; Academic Press, **2011**; Vol. 83, pp 163–221.
7. Kern, D.; Zuiderweg, E. R. *The Role of Dynamics in Allosteric Regulation.* Curr. Opin. Struct. Biol. **2003**, 13 (6), 748–757.10.008.