

In silico laboratory: new and updated tools for protein-centered drug discovery

Dušanka Janežič¹ and Janez Konc^{1,2,3,*}

¹Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska,
Glagoljaška ulica 8, SI-6000 Koper, Slovenia.

²National Institute of Chemistry, Hajdrihova 19, SI-1000 Ljubljana, Slovenia.

³Faculty of Pharmacy, University of Ljubljana, Aškerčeva 7, SI-1000 Ljubljana, Slovenia.

Corresponding author's contact address: konc@cmm.ki.si

Running Head: In silico laboratory for *drug discovery*

Abstract

Computational tools for protein-ligand binding predictions are essential in modern drug discovery. This chapter provides an updated overview of the ProBiS tools, which identify binding sites, predict ligand interactions, and analyze conserved water molecules. Using an efficient maximum clique algorithm, these tools enable large-scale mining of protein structure databases, including the Protein Data Bank (PDB) and AlphaFoldDB, for drug repurposing and functional annotation. Recent advances include the ProBiS-Dock Database, which provides precomputed binding sites for proteome-wide ligand binding prediction, and the ProBiS-Dock algorithm, a hybrid flexible docking method. Additionally, ProBiS-H2O and ProBiS-H2O MD identify conserved water clusters crucial for ligand binding, while the ProBiS-Fold web server predicts binding sites on AI-modeled protein structures. We highlight applications of these tools in therapeutic research. ProBiS-H2O has identified conserved hydration sites in enzymes, ProBiS-H2O MD has revealed allosteric water networks in kinases, and ProBiS-Dock has facilitated virtual screening for cancer drug targets. ProBiS-Fold has enhanced the annotation of AlphaFold2-modeled human proteome structures. By integrating these tools into computational workflows, researchers can accelerate the discovery of new therapeutics. This chapter provides protocols for their effective use in structural biology and drug discovery.

Key Words

Protein-ligand binding, drug discovery, ProBiS tools, binding site prediction, ligand docking, conserved water molecules, AlphaFoldDB, ProBiS-H2O, ProBiS-H2O MD, ProBiS-Dock Database, ProBiS-Dock algorithm, ProBiS-Fold

1 Introduction

Computational tools for protein-ligand binding predictions have advanced significantly, enabling researchers to explore molecular interactions with high accuracy, which is essential in the early stages of drug discovery. Several tools have been developed to facilitate these analyses, including the ProBiS suite of tools [1], which

provides a range of methods for detecting binding sites, predicting ligand interactions, and analyzing conserved water molecules that can be used individually or in combination to enhance the drug discovery pipeline.

The ProBiS tools leverage a highly efficient maximum clique algorithm [2–6], which enables the rapid detection of similarities between protein structures represented by molecular descriptors. This capability facilitates large-scale database mining, including the Protein Data Bank (PDB) [7] and the artificial intelligence (AI) protein model structures, such as the AlphaFoldDB [8]. Applications include identifying novel protein functions, discovering new binding sites, analyzing the effects of genetic variations on ligand binding, and exploring drug repurposing opportunities.

This chapter, which is an update of our earlier chapter in this book series [9], describes the latest ProBiS tools and their applications in structural biology and drug discovery. Our suite includes recent additions, like the ProBiS-Dock Database [10], which provides precomputed protein binding sites ready for docking for a vast collection of protein structures spanning the entire PDB, streamlining the process of proteome-wide ligand binding prediction. ProBiS-Dock Database is complemented by the ProBiS-Dock algorithm [11], a hybrid multitemplate homology flexible docking algorithm that enables accurate predictions of protein-ligand interactions by leveraging protein binding site comparisons.

We also introduce ProBiS-H₂O [12], a tool designed to identify conserved water molecules in protein structures, which play a crucial role in ligand binding and drug design. This tool is further extended by ProBiS-H₂O MD [13], which integrates molecular dynamics (MD) simulations to identify conserved water networks across multiple protein structures available in an MD trajectory.

Additionally, we present ProBiS-Fold [14], a web server that predicts binding sites on AI-modeled protein structures from the AlphaFoldDB based on template-based modeling and binding site conservation, enabling the discovery of structurally and functionally similar proteins even in the absence of sequence homology.

With the integration of these tools, researchers can efficiently explore protein-ligand interactions, gain deeper insights into molecular recognition, and accelerate the development of new therapeutic molecules.

In the following, we give specific examples of usage of our ProBiS tools summarized from various studies. Although not exhaustive, these examples serve as a guide in what way and with what purpose our tools can be used.

ProBiS-H2O has been successfully applied to study conserved water sites in proteins. It was used in a study on Bovine Pancreatic Ribonuclease A, where it identified conserved water molecules in the enzyme's active site by analyzing high-resolution structures [15]. The study revealed key water clusters that interact with catalytic residues, including His-12 and His-119, which are crucial for RNA cleavage. These findings provided insights into the role of water-mediated hydrogen bonding in the enzyme's catalytic mechanism, reinforcing the importance of conserved water sites in drug design.

Our ProBiS-H2O MD was employed in a study of WNK kinases, where it helped uncover two conserved water networks (CWN1 and CWN2) that regulate autophosphorylation and chloride sensitivity [16]. The study demonstrated that these water clusters function as allosteric regulators, influencing the enzymatic activity of WNK kinases.

The ProBiS-Dock Database was used in a study on colorectal cancer, where it assisted in identifying binding sites based on the evolutionary conservation of amino acids [17]. Our approach reduced docking complexity and improved the accuracy of ligand-protein interaction predictions. By analyzing potential drug targets, such as KEAP1 and COL1A1, ProBiS-Dock Database helped researchers explore novel small-molecule inhibitors and their impact on cancer therapy.

Additionally, ProBiS-Dock algorithm was applied in a virtual screening study targeting CC chemokine receptor 7 [18]. Using allosteric binding site comparisons and flexible docking, researchers screened a library of commercially available compounds, leading to a set of experimentally validated decoys that can aid future drug discovery efforts.

Our newest tool, ProBiS-Fold, has already gained recognition in the scientific community, as highlighted in a recent review on protein-ligand binding databases [19]. This review described ProBiS-Fold as the counterpart to ProBiS-Dock, focusing on the prediction of binding sites in AlphaFold2-modeled structures for the entire human proteome. Its integration of the ProBiS algorithm [20] to predict ligands and glycosylations further emphasizes its value in structural biology research.

Together, our ProBiS tools have proven invaluable in analyzing protein structures, mapping functional water networks, and predicting ligand interactions, making them essential components of modern computational structural biology.

2 ProBiS tools

2.1 ProBiS-H2O

ProBiS H2O is a computational tool for identifying conserved water sites in proteins, crucial for molecular docking and stability predictions [12]. Instead of costly simulations, it leverages experimentally determined water molecules from the PDB and a local structure alignment algorithm to predict conserved waters. By analyzing similar proteins and clustering transposed water molecules, it reveals key conserved hydration sites in binding pockets and interfaces. ProBiS H2O has been successfully validated in multiple studies, demonstrating its utility in drug design. The tool is freely available at <http://insilab.org/probis-h2o/>.

2.2 ProBiS-H2O MD

ProBiS H2O MD is an advanced tool for identifying conserved water sites in macromolecular systems, crucial for structure-based drug design [13]. Building on ProBiS H2O, it integrates molecular dynamics data to analyze less-characterized protein systems, for example, proteins without a structure in the PDB yet. By processing trajectory snapshots, performing local superpositions, and clustering water positions, it reveals key hydration sites. This fast, software-independent method has been successfully validated in reported protein systems. ProBiS H2O MD enables researchers to track water conservation trends and is freely available as a PyMOL plugin at <http://insilab.org/probis-h2o-md>.

2.3 ProBiS-Dock Database

The ProBiS-Dock Database is a comprehensive repository of small-molecule binding sites across different organisms in the PDB, enabling advanced structure-based drug discovery [10]. Built upon the ProBiS algorithm, it predicts binding sites based on ligand conservation in structurally similar PDB entries, including binding sites at protein-protein interfaces. The database automates binding site curation (in contrast to docking approaches where binding sites are often manually prepared) by identifying relevant accessory molecules like cofactors, glycans, and conserved waters as needed. The database introduces several key innovations, including the generation of multiple centroids to define docking search spaces, consideration of binding sites spanning multiple protein chains, and ranking of sites by druggability, which is assessed based on ligand complexity and frequency in known drug-target interactions. The ProBiS-Dock Database, freely

available to academic users at <http://probis-dock-database.insilab.org>, is particularly useful for multi target drug design, addressing off-target effects, and improving drug selectivity.

2.4 ProBiS-Dock algorithm

ProBiS-Dock is a flexible molecular docking algorithm designed to complement the ProBiS-Dock Database, leveraging existing protein-ligand structures from the PDB to improve docking predictions [11]. Unlike traditional docking methods, ProBiS-Dock treats both proteins and small molecules as fully flexible, allowing for conformational adjustments upon ligand binding. It employs a novel scoring function, combining a binding site-specific score (ProBiS-Score) with a general statistical score for enhanced accuracy. Validated against standard benchmarks, ProBiS-Dock has proven effective for identifying new active ligands and was successfully applied to discover novel inhibitors of human indoleamine 2,3-dioxygenase 1 (IDO1), a key enzyme in cancer-related tryptophan metabolism. Freely available to academic users at <http://insilab.org/probisdock>, ProBiS-Dock enables efficient virtual screening for drug discovery and development.

2.5 ProBiS-Fold web server

ProBiS-Fold is a web server that annotates AlphaFold-predicted human protein structures with druggable binding sites, enabling large-scale structure-based drug discovery [14]. It utilizes the ProBiS algorithm to compare AlphaFold-predicted structures with experimentally determined structures in the PDB, identifying previously unknown binding sites by transferring known ligands from similar PDB structures. The identified binding sites are assigned druggability scores based on ligand similarity to known drugs, ranking them for drug development potential.

ProBiS-Fold provides interactive, downloadable binding sites for the entire human proteome, covering small molecules, cofactors, peptides, nucleic acids, metal ions, conserved water molecules, and glycosylation sites. It supports structure-based virtual screening and enables protein function studies. The web server, freely available to academic users at <http://probis-fold.insilab.org>, offers a powerful resource for pharmaceutical research and functional proteomics.

3 Protocols

Here, we provide protocols for using ProBiS tools for protein-centered drug discovery, covering both stand-alone software and web servers. The ProBiS-H2O and ProBiS-H2O MD tools are PyMOL plugins and require a Linux or Windows operating system; the stand-alone tool ProBiS-Dock algorithm requires a Linux operating system, while the web servers, i.e., ProBiS-Fold and ProBiS-Dock Database, are compatible with all major browsers, including Firefox and Chrome, on any operating system. For stand-alone tools, commands provided in the following sections can be copied and pasted into a Linux terminal for execution.

3.1 Use of the ProBiS-H2O graphical user interface

The graphical user interface for ProBiS-H2O is available as a plugin for the widely used PyMOL molecular visualization software. Currently, we support PyMOL version 2.3 or above [21]. The computational workflow implemented in the ProBiS-H2O plugin is designed to identify conserved water sites in macromolecular structures (**Fig. 1**). It integrates structural alignment of proteins from the PDB using the ProBiS algorithm, water extraction from aligned PDB structures, and visualization to analyze water molecule conservation within protein-ligand and protein-protein interfaces as well as on the non-binding surface of a given query protein.

Below is a detailed step-by-step guide on how to use ProBiS-H2O. This step-by-step protocol provides a structured approach to using ProBiS-H2O for identifying conserved water sites in macromolecular structures. By following the outlined steps, users can efficiently analyze water molecule conservation and gain insights into their biological significance.

1. Installation of dependencies

The ProBiS-H2O plugin should automatically install its dependencies during first start in PyMOL. However, if that fails for some reason, here is the information on how to do this manually. Ensure the following software and libraries are installed before using ProBiS-H2O plugin:

- Python >3.0
- NumPy
- Scikit-learn
- PyMOL

- ProBiS algorithm binary (installed automatically during ProBiS-H2O setup) [20]

On Ubuntu, install the required Python libraries by running:

```
sudo apt-get install python3-numpy python3-sklearn
```

Alternatively, install scikit-learn using pip or conda:

```
pip3 install -U scikit-learn
```

```
conda install scikit-learn
```

2. Installing ProBiS-H2O in PyMOL molecular visualization program

1. Download the ProBiS-H2O plugin from <http://insilab.org/probis-h2o>.
2. Open PyMOL.
3. Navigate to Plugin > Plugin Manager > Install New Plugin.
4. Select ProBiS-H2O plugin file on your computer or, alternatively, provide an URL to where the plugin is located.
5. Click Install button.
6. Restart PyMOL to finalize the installation.

3. First-time setup

1. Open PyMOL and launch the ProBiS-H2O plugin.
2. Choose database directory in the *Settings tab*
3. Click the SETUP DATABASE button in the *Settings tab*.
 - This will create a /Probis_H2O/ folder in the default PyMOL working directory.
 - It will download protein sequence clustering data from the RCSB PDB.
 - It will also download the ProBiS algorithm binary file.
4. The ProBiS binary should be executable by default; if it is not, grant it executable rights:

```
sudo chmod +x ~/Probis_H2O/.pro/probis
```

4. Running conserved water analysis

To run the analysis, switch to the *Define System* tab. Then:

1. Input the chosen PDB ID or click the *Custom* button to choose a custom .pdb file [NOTE] If choosing a custom PDB file, the file should be named PDB_ID.pdb, where PDB_ID is a 4-letter code, and .pdb is the suffix.
2. Check boxes to define parameters.
3. Choose **Blastclust** value (cutoff for homologous protein chains).
4. Click **Find** button (finds homologous protein chains for the query protein in the PDB).
5. Click **Download** button to download the identified structures of homologous protein chains from the PDB (this process may take a while depending on the number of protein chains found – the plugin may stop responding during this process).
6. Click **Identify** to identify binding sites (based on bound ligands) on the query protein.
7. Pick a binding site to run the calculation from the list.
8. Optionally, switch to the *Settings tab* and set **Maximum distance** between two samples for one to be considered a neighbor of the other (default 0.9).
9. Switch back to the *Define System* tab, then click **GO** to start the calculation (may take a while and repeatedly open and close terminal on Windows).

5. Displaying conserved water clusters

Once the calculation is complete, the user will be automatically taken to the *Cluster Analysis* tab, and the results will be listed and highlighted in red. Select a conserved water cluster from the cluster list.

1. Click **Fetch/Reset** and wait for your protein to display in the PyMOL console. (*Required for all following functions*)
2. Click **Chain Box** to draw a red box around your chosen binding site.
3. Click **Display** to show selected water clusters as red dots in the PyMOL console.
4. Click **Contacts** to show distances between the center of the cluster of conserved water and the analyzed protein's binding site.
5. Click **Debye-Waller correction** to apply isotropic displacement correction for low-resolution structures.
6. Click **Keep selected display** to visualize conservation trends with a color gradient (white to red for increasing conservation).

7. Click **B-Site** to display amino acids making up the chosen binding site.

This workflow ensures a systematic approach to analyzing water molecule conservation in protein binding sites using ProBiS-H2O.

6. Logging and report generation

ProBiS-H2O generates a report file for each query structure in the /Probis_H2O/ folder.

- The report file is named: ./report_\$(pdb_query).txt
- It contains information about the analyzed structure, sequence identity clusters, and identified conserved water molecules.

3.2 Use of the ProBiS-H2O MD graphical user interface

The ProBiS-H2O MD plugin for PyMOL enables the identification and visualization of conserved water molecules from MD trajectories of proteins in water (**Fig. 2**). It allows for detailed analysis of conserved water positions within a protein structure based on MD simulation data, providing valuable insights into hydration patterns and protein-water interactions. The plugin is compatible with Linux, Windows, and Mac (from version 1.0.1 onwards) and requires PyMOL version 2.3 or higher for proper functionality.

1. Installation

Download the plugin file:

1. Download ProBiS_H2O_MD.py from <http://insilab.org/probis-h2o-md> (Download tab) for direct installation in PyMOL.
2. Alternatively, download the full package (ProBiS-H2O MD-plugin.zip) for manual installation.

Install the plugin in PyMOL:

1. Open PyMOL and navigate to **Plugin Manager > Install New Plugin > Choose File**.
2. Select the downloaded plugin file.
3. Click **Install**. The plugin will appear in the **Plugins** dropdown menu.
4. Click on **ProBiS-H2O MD** in the dropdown menu to launch the plugin.

Verify dependencies:

1. On first launch, the plugin checks for the pip3 module and installs additional dependencies (scikit-learn and MDAnalysis) automatically.
2. If the automatic installation fails, install the dependencies manually:

```
apt-get install python3-pip
```

```
pip3 install MDAnalysis scikit-learn
```

2. Running ProBiS-H2O MD water cluster analysis on an MD trajectory

The following steps pertain to the *Define System* tab:

1. Input the chosen topology file (PDB) and trajectory file (DCD).
2. Set analysis parameters by selecting the relevant checkboxes.
3. Identify sites by clicking *Identify* button to detect binding sites, water sites, or chains on the input protein based on the selected settings.
4. Select region for analysis on the query protein by choosing a binding site, water site, or chain from the identified list.
5. Optionally, switch to the *Settings* tab to configure neighbor distance by setting the maximum distance between two samples for one to be considered a neighbor of the other (default: 3.0).
6. Start the Calculation by clicking *GO* button in the *Define System* tab to begin the analysis. The process may take some time and might cause the terminal to open and close repeatedly on Windows—this is expected behavior.

Once completed, the results will appear in the *Cluster Analysis* tab, with identified clusters marked in red.

3. Displaying identified conserved water clusters on an MD trajectory

1. Load analysis results by selecting one or more clusters from the cluster list.
2. Click **Fetch/Reset** button to display the protein in the PyMOL console (required for subsequent steps).
3. Visualize trajectory by clicking *add MD* to load the trajectory file into PyMOL.
4. Use *mp lay* and *mstop* commands in PyMOL to control the trajectory playback.

5. Highlight binding site location by clicking **Chain Box** button to draw a red box around the selected binding site.
6. Click **Display** to visualize selected clusters as red dots in the PyMOL console.
7. Click **Contacts** to show distances between the conserved water cluster center and the analyzed protein's binding site.
8. Click **B-Site** to display amino acids constituting the selected binding site.

This structured workflow enables an intuitive approach to analyzing conserved water molecules in MD simulations, facilitating a deeper understanding of their role in protein-ligand interactions.

3.3 Use of the ProBiS-Dock Database web server

The ProBiS-Dock Database web server is a tool for structure-based drug discovery, enabling comprehensive binding site analysis and ligand prediction. It serves as a resource for providing downloadable binding sites ready for docking and inverse docking. It enables PDB-wide drug discovery and predictions of intra- and cross-organism selectivity, classifying binding sites as primary (orthosteric in most cases) and secondary (allosteric/orthosteric/other), ranking binding sites based on their druggability, as well as detecting binding sites at the interface of multiple protein chains (**Fig. 3**).

1. Using the search bar to retrieve binding sites

The search bar allows for free-text searches, meaning any entered text is matched against the database. For instance, entering the query `kinase cancer` will return binding site entries that contain both `kinase` and `cancer` anywhere in their descriptions. The whitespace between terms functions as a logical **AND** operator.

For more precise searches, users can specify both the **column name** and **value** in the query. For example, searching for `pdb_id=1got` retrieves the specific protein associated with that PDB identifier.

Valid column names are:

- **pdb_id** – Four-letter PDB identifier (e.g., `1a1l`)
- **chain_id** – Protein chain identifier (e.g., `A`, `B`, ...)
- **rest** – Binding site type (compound or cofactor)
- **bs_id** – Binding site number based on druggability ranking

- **organism_scientific** – Scientific name of the protein's originating organism (e.g., Homo sapiens)
- **protein_class** – Protein classification (e.g., kinase)
- **disease** – Disease association (e.g., cancer)
- **ec** – Enzyme Commission (EC) number (e.g., 3.4.25.1)
- **molecule** – Protein name (e.g., PROTEASOME SUBUNIT BETA TYPE-3)
- **division_id** – NCBI taxonomy 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 for Homo sapiens)

We give some example queries:

- **All binding sites:** *(leave empty)*
- **Primary binding sites (mostly orthosteric):** bs_id=1
- **Secondary binding sites (allosteric/orthosteric/other):** bs_id>1 *(Note: the method does not reliably determine if a site is allosteric.)*
- **Compound binding sites (substrates, agonists, etc.):** 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 *(Double quotes are required for multi-word search terms.)*

- **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", ...]

This structured approach enables precise filtering and retrieval of binding site information from the ProBiS-Dock Database.

2. Detected protein binding sites format

The detected protein binding sites are structured into components (or files). Each binding site entry consists of:

- **Centroids:** Spheres, each described by coordinates (x,y,z) and radius r, accurately describing the often complex shapes of binding sites.
- **Receptor:** A biologically relevant single-chain protein or multi-chain protein complex.
- **Accessory Ligands:** Including cofactors, metal ions, conserved waters, and covalently attached glycans that are kept in the binding site due to their biological relevance.
- **Predicted Ligands:** Identified using the ProBiS approach from similar binding sites.
- **Bounding Box:** Defining the outer limits of a binding site.

Binding sites are categorized as:

- **Compound Sites:** Containing substrate/agonist-competitive ligands.
- **Cofactor Sites:** Containing known cofactors and cofactor-competitive ligands.
- **Primary Sites (Rank 1):** The most reliable binding sites, often corresponding to orthosteric binding site.
- **Secondary Sites (Rank >1):** Additional binding sites of interest, likely allosteric binding sites.
- **Druggable Sites:** Druggability score, measuring drug-likeness of the predicted ligands, expresses utility of a binding site for drug discovery.

- **Interface Binding Sites:** Defining binding sites at protein-protein interfaces for improved docking accuracy.

Users can download binding sites as:

- Individual or multiple selected sites based on search queries.
- Pre-prepared datasets for different organisms, taxonomic groups, or diseases.

The database includes accessory ligands that influence docking outcomes, classified as:

- **Cofactors:** Essential small molecules such as ATP, NADH, and heme.
- **Glycans:** Monosaccharides and polysaccharides that impact binding.
- **Metal Ions and Conserved Waters:** Key structural and functional components of binding sites.

Ligands are predicted using structural similarity analysis, offering insights into:

- **Ligand Transposition:** Identifying potential binders based on similar sites.
- **Nonspecific Binders:** Filtering ligands for docking applications.

3.4 Use of the ProBiS-Dock algorithm for molecular docking

The ProBiS-Dock algorithm is a flexible docking algorithm that complements the ProBiS-Dock Database.

This section provides a structured overview of ProBiS-Dock, covering its functionality, usage, and availability for research applications. In the following examples we show three different use cases for the ProBiS-Dock algorithm:

1. Docking a single ligand to a single receptor

To dock a ligand to a receptor, download and install ProBiS-Dock algorithm and example file from <http://insilab.org/probisdock/>. The input files for the receptor, centroids, and template ligands can be obtained from the ProBiS-Dock Database. Ensure the ProBiS-Dock executable is in your system path before execution.

```
unzip example1.zip
```

```
cd example1
```

```
chmod +x PATH_TO_PROBISDOCK/probisdock # Only required once
```

```
probisdock --ligands_file crystal_ligand.mol2
```

```
--receptor_file receptor_3max_A_compound_1.pdb.gz
```

```
--template_ligands_file predicted_ligands_3max_A_compound_1.pdb.gz
```

```
--centro_file probisdock_centro_3max_A_compound_1.cen.gz
```

After successful completion, the docking summary will be stored in the `summary` file, while the 3D coordinates of the docked ligand poses will be in the `docked_complexes` file with scores included in the `TITLE` lines.

2. Docking multiple ligands to a single receptor

To screen multiple ligands against a single receptor use the following command:

```
unzip example2.zip
```

```
cd example2
```

```
chmod +x PATH_TO_PROBISDOCK/probisdock # Only required once
```

```
probisdock --ligands_file actives-hdac2-chunk0.mol2.gz
```

```
--receptor_file receptor_3max_A_compound_1.pdb.gz
```

```
--template_ligands_file predicted_ligands_3max_A_compound_1.pdb.gz
```

```
--centro_file probisdock_centro_3max_A_compound_1.cen.gz
```

As before, results will be stored in `summary` and `docked_complexes` files.

3. Docking multiple ligands to multiple receptors

For more complex scenarios where multiple ligands are docked to different binding sites on multiple receptors, for example, when predicting protein targets for a set of compounds such as in the inverse molecular docking, execute:

```
unzip example3.zip
```

```
cd example3
```

```
chmod +x PATH_TO_PROBISDOCK/probisdock # Only required once
```

```
probisdock --ligands_file actives-hdac2-chunk0.mol2.gz \
```

```
--probisdock_db_dir ProBiS-DB-3max
```

The results, including docking scores and ligand poses, will be available in `summary` and `docked_complexes` files.

3.5 Use of the ProBiS-Fold web server

ProBiS-Fold annotates the AlphaFold human protein database by identifying binding sites, post-translational modifications, and predicted ligands with 3D structures (**Fig. 4**). It provides interactive, downloadable data for functional and drug discovery research while enabling structure-based virtual screening and selectivity predictions across the human proteome. It is available in all major browsers and enables free and fast download of pre-prepared binding site datasets.

1. Retrieving binding sites using the search bar

The search bar allows users to perform a free-text search by entering any text string. For instance, the query **kinase cancer** retrieves binding site entries containing the terms *kinase* and *cancer* anywhere in their descriptions. The whitespace between the terms acts as a logical **AND** operator, ensuring that both terms appear in the search results.

To refine searches, users can specify the column name and the corresponding value to search within a particular dataset field. For example, `pdb_id=1got`, retrieves the binding site entry associated with the protein **1got** from the database.

Valid column names are the following:

- **alphafold_id** – AlphaFold DB identifier, e.g., `alphafold_id=AF-075616-F1-model_v2`.
- **min_beta_factor** – AlphaFold model confidence score (0-100, highest confidence), e.g., `min_beta_factor>90` to filter sites with a minimum confidence level of 90.
- **compound_score** – Druggability score (Sbsite score, applicable to substrate- and cofactor-competitive binding sites):
 - `compound_score>90` (highly druggable)
 - `compound_score>50` (druggable)
 - `compound_score>30` (likely druggable)
 - `compound_score>15` (possibly druggable)
 - `compound_score>0` (unlikely druggable)
- **uniprot_id** – UniProt identifier, e.g., `uniprot_id=075616`.

- **pdb_id** – PDB four-letter identifier, e.g., `pdb_id=1all` or `pdb_id=null` (for models with no equivalent structure in PDB).
- **chain_id** – Protein chain identifier, e.g., `A`, `B`, ...
- **bsite_rest** – Binding site type:
 - `small.substrate.competitive` (for substrate-competitive binding sites)
 - `small.cofactor.competitive` (for cofactor-competitive binding sites)
- **bs_id** – Binding site ranking number based on druggability.
- **ec** – Enzyme Commission number, e.g., `ec=3.4.25.1`.
- **molecule** – Protein name, e.g., `molecule=PROTEASOME SUBUNIT BETA TYPE-3`.
- **protein_class** – Protein class, e.g., `protein_class=kinase`.
- **disease** – Disease association, e.g., `disease=cancer`.
- **cons** – Conservation score (range: 0 to 1), defined as the number of similar binding sites with the same ligand divided by the total number of similar binding sites found, e.g., `cons>0.9` for highly conserved ligands.
- **num_occ** – Number of occurrences of a ligand in similar binding sites, e.g., `num_occ>10`.

Users can retrieve specific binding sites using structured queries in the search bar:

- **All binding sites:** Leave the search bar empty.
- **Primary binding sites (mostly orthosteric):** `bs_id=1`
- **Secondary binding sites (allosteric/orthosteric/other):** `bs_id>1`
 - *Note: The method does not reliably differentiate allosteric sites.*
- **Substrate-competitive binding sites:**
`bsite_rest=small.substrate.competitive`
- **Cofactor-competitive binding sites:**
`bsite_rest=small.cofactor.competitive`
- **Top-ranked binding sites (e.g., No. 1 and 2):** `bs_id=[1, 2]`
- **Highly confident AlphaFold models (confidence > 90):** `min_beta_factor>90`
- **Highly druggable binding sites:** `compound_score>90`

- **Druggable binding sites (Sbsite scores 51-90):** `compound_score>50`
`compound_score<91`
 - *Note: A space character between terms is interpreted as logical AND.*
- **Highly conserved water molecules binding to a similar residue motif in at least 10 other PDB binding sites:**

`bsite_rest=water cons>0.9 num_occ>10`

- *Note: These water molecules may be critical for protein function or structural stability.*

By leveraging these search functionalities, users can efficiently identify relevant binding sites and explore structural and functional aspects of proteins using the ProBiS-Fold web server.

2. Output of the ProBiS-Fold web server

The ProBiS-Fold web server provides several key outputs that aid in structural and functional analysis of protein binding sites:

- **Centroids** – 3D coordinates (x, y, z) and radius values that accurately describe the often convoluted shapes of binding sites.
- **Binding site protein residues** – A list of protein residues that interact with ligands, providing insight into binding interactions.
- **Predicted ligands** – Ligands obtained through the structure-based comparative ProBiS approach from similar binding sites in the PDB, allowing for functional annotation and potential drug discovery.
- **Binding site bounding box** – A docking-ready bounding box in **AutoDock Vina** format, enabling molecular docking studies.
- **Receptor structure** – A single-chain protein structure predicted using **AlphaFold2**, providing a reliable structural framework for further computational analysis.

3. Annotation of AlphaFold human protein database by ProBiS-Fold

The ProBiS-Fold web server annotates the AlphaFold human protein database with detailed structural and functional insights, including binding site annotations:

- **Small molecule compounds** (substrates, agonists, inhibitors)
- **Cofactors** (cofactor and cofactor-competitive ligands, based on a curated list of known cofactors)
- **Proteins and peptides** (categorized as shorter than 20 amino acids or longer than 20 amino acids)
- **Nucleic acids** (DNA or RNA molecules)
- **Metal ions** (structurally conserved metal-binding sites)
- **Water molecules** (highly conserved, structurally relevant water molecules)

as well as post-translational modification site annotations:

- **Glycosylation sites** – Includes **O-glycosylation** and **N-glycosylation** predictions.

Binding sites are ranked according to their estimated druggability score, which applies specifically to **compound** and **cofactor** binding sites. The ranking helps prioritize sites most suitable for drug discovery efforts.

By integrating these annotations, ProBiS-Fold enhances the utility of AlphaFold-predicted protein structures, making them more accessible for functional analysis, drug discovery, and computational modeling.

Acknowledgment

Financial support through Slovenian Research and Innovation Agency grants N1-0142, N1-0209 in J1-4414 are gratefully acknowledged.

References

1. Konc J, Janežič D (2022) Protein binding sites for drug design. *Biophys Rev* 14:1413–1421. <https://doi.org/10.1007/s12551-022-01028-3>
2. Rozman K, Ghysels A, Janežič D, Konc J (2024) An exact algorithm to find a maximum weight clique in a weighted undirected graph. *Scientific reports* 14:9118
3. Rozman K, Ghysels A, Zavalnij B, Kunej T, Bren U, Janežič D, Konc J (2024) Enhanced molecular docking: Novel algorithm for identifying highest weight k-cliques in weighted general and protein-ligand graphs. *Journal of Molecular Structure* 1304:137639. <https://doi.org/10.1016/j.molstruc.2024.137639>

4. Konc J, Janezic D (2007) An improved branch and bound algorithm for the maximum clique problem. *MATCH Commun Math Comput Chem* 58:569–590
5. Depolli M, Konc J, Rozman K, Trobec R, Janežič D (2013) Exact Parallel Maximum Clique Algorithm for General and Protein Graphs. *J Chem Inf Model* 53:2217–2228. <https://doi.org/10.1021/ci4002525>
6. Reba K, Guid M, Rozman K, Janežič D, Konc J (2022) Exact Maximum Clique Algorithm for Different Graph Types Using Machine Learning. *Mathematics* 10:97. <https://doi.org/10.3390/math10010097>
7. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Mol Biol* 10:980–980. <https://doi.org/10.1038/nsb1203-980>
8. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Židek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 50:D439–D444. <https://doi.org/10.1093/nar/gkab1061>
9. Lešnik S, Konc J (2020) In Silico Laboratory: Tools for Similarity-Based Drug Discovery. In: Labrou NE (ed) *Targeting Enzymes for Pharmaceutical Development: Methods and Protocols*. Springer US, New York, NY, pp 1–28
10. Konc J, Lešnik S, Škrlić B, Janežič D (2021) ProBiS-Dock Database: A Web Server and Interactive Web Repository of Small Ligand–Protein Binding Sites for Drug Design. *J Chem Inf Model* 61:4097–4107. <https://doi.org/10.1021/acs.jcim.1c00454>
11. Konc J, Lešnik S, Škrlić B, Sova M, Proj M, Knez D, Gobec S, Janežič D (2022) ProBiS-Dock: A Hybrid Multitemplate Homology Flexible Docking Algorithm Enabled by Protein Binding Site Comparison. *J Chem Inf Model* 62:1573–1584. <https://doi.org/10.1021/acs.jcim.1c01176>
12. Jukić M, Konc J, Gobec S, Janezic D (2017) Identification of Conserved Water Sites in Protein Structures for Drug Design. *J Chem Inf Model* 57:3094–3103
13. Jukić M, Konc J, Janežič D, Bren U (2020) ProBiS H₂O MD Approach for Identification of Conserved Water Sites in Protein Structures for Drug Design. *ACS Med Chem Lett* 11:877–882. <https://doi.org/10.1021/acsmchemlett.9b00651>
14. Konc J, Janežič D (2022) ProBiS-Fold Approach for Annotation of Human Structures from the AlphaFold Database with No Corresponding Structure in the PDB to Discover New Druggable Binding Sites. *J Chem Inf Model* 62:5821–5829. <https://doi.org/10.1021/acs.jcim.2c00947>
15. Lisgarten DR, Palmer RA, Cooper JB, Naylor CE, Talbert RC, Howlin BJ, Lisgarten JN, Konc J, Najmudin S, Lobley CMC (2023) Ultra-high resolution X-ray structure of orthorhombic bovine pancreatic Ribonuclease A at 100K. *BMC Chemistry* 17:91. <https://doi.org/10.1186/s13065-023-00959-6>
16. Teixeira LR, Akella R, Humphreys JM, He H, Goldsmith EJ (2024) Water and chloride as allosteric inhibitors in WNK kinase osmosensing. *eLife* 12:RP88224. <https://doi.org/10.7554/eLife.88224>
17. Wang G, Zhu Z-M, Wang K (2024) Identification of ROS and KEAP1-related genes and verified targets of α -hederin induce cell death for CRC. *Drug Development Research* 85:e22200. <https://doi.org/10.1002/ddr.22200>
18. Proj M, De Jonghe S, Van Loy T, Jukić M, Meden A, Ciber L, Podlipnik Č, Grošelj U, Konc J, Schols D, Gobec S (2022) A Set of Experimentally Validated Decoys for the Human CC Chemokine Receptor

7 (CCR7) Obtained by Virtual Screening. *Front Pharmacol* 13:.
<https://doi.org/10.3389/fphar.2022.855653>

19. Carpenter KA, Altman RB (2024) Databases of ligand-binding pockets and protein-ligand interactions. *Computational and Structural Biotechnology Journal* 23:1320–1338.
<https://doi.org/10.1016/j.csbj.2024.03.015>
20. PyMOL | pymol.org. <https://www.pymol.org/>. Accessed 9 Mar 2025
21. Konc J, Janežič D (2010) ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* 26:1160–1168. <https://doi.org/10.1093/bioinformatics/btq100>

Figure Captions

Fig. 1 ProBiS-H2O detects conserved water molecules in proteins. A conserved water molecule (red sphere) found in the protein (white cartoon model; PDB ID: 4duh) with the native ligand shown as blue sticks model is shown on the left side. List of detected water clusters on this protein sorted by their conservation scores from highly conserved (bright red) to less conserved is on the right side.

Fig. 2 Output of the ProBiS-H2O MD: conserved waters identified from an MD trajectory. Two conserved waters are shown as red spheres on an antibody structure (green cartoon) in a water box used in the MD trajectory where waters are shown as wire frame models.

Fig. 3 ProBiS-Dock Database detects 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.

Fig. 4 ProBiS-Fold web server output. A predicted ligand is shown as sticks model in the predicted binding site (yellow surface) on an AlphaFold model protein structure (blue cartoon model).