

ProBiS for drug development

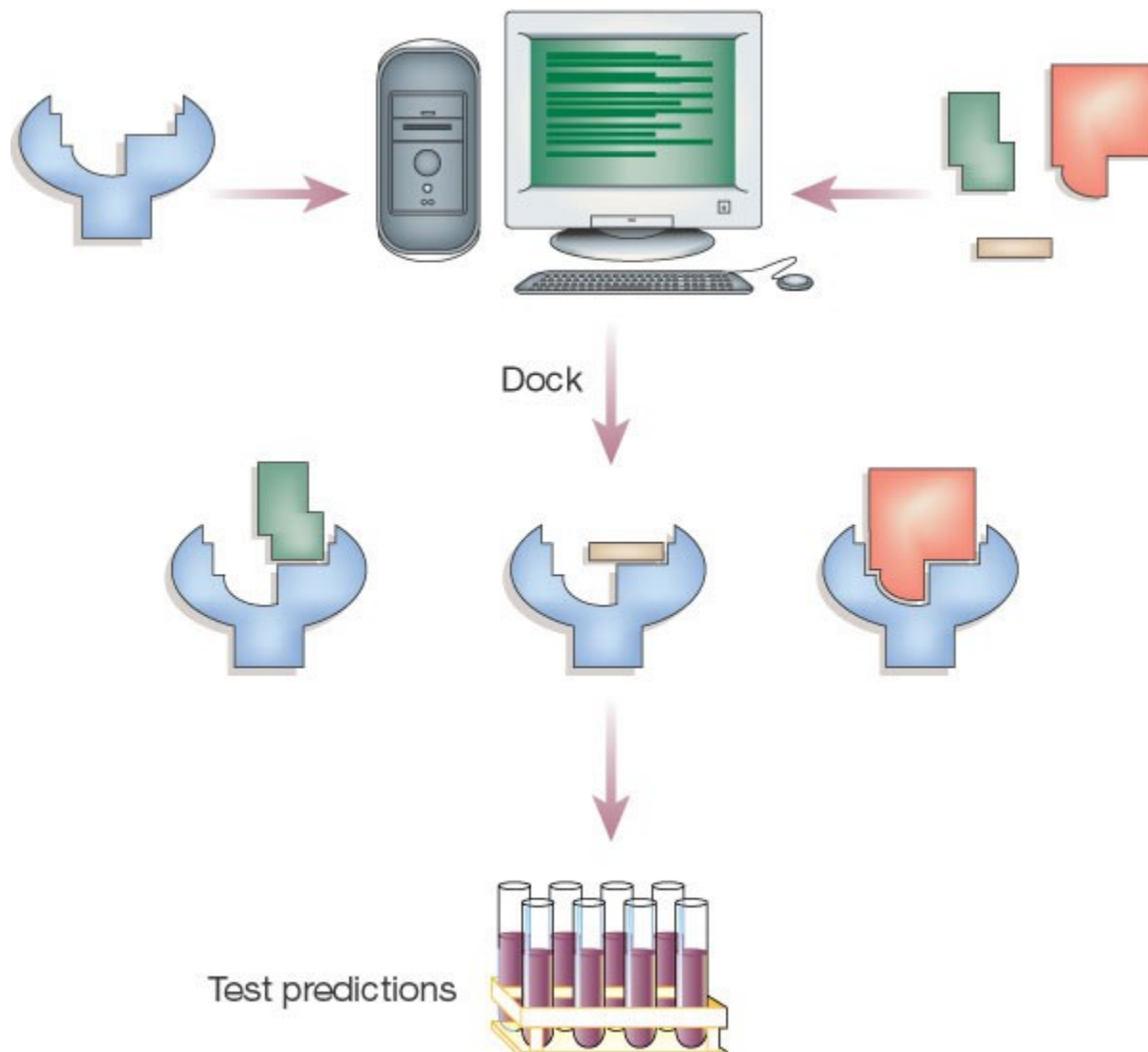
Janez Konc

National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana

Ljubljana, 27/02/2024

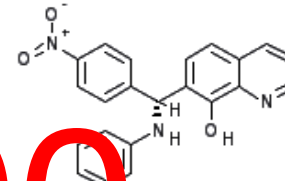
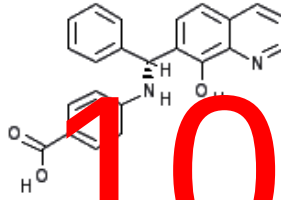
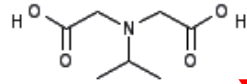
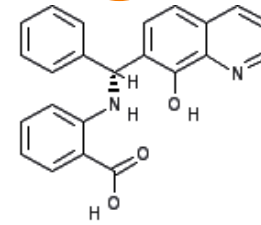
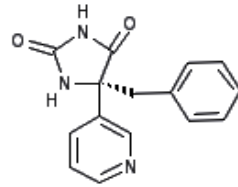
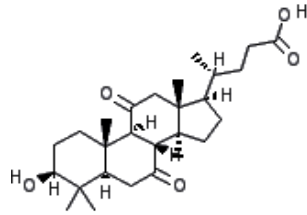
Why do we need a computer in drug development?

Let's speed up the discovery of new active compounds.

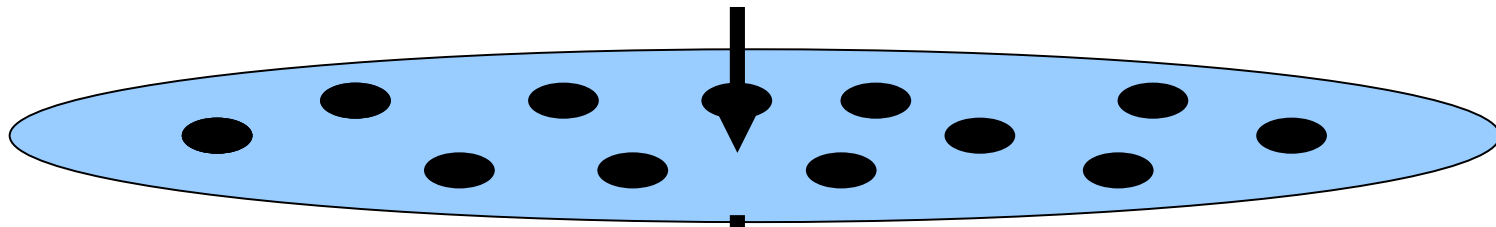
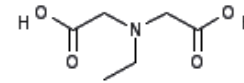
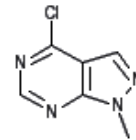
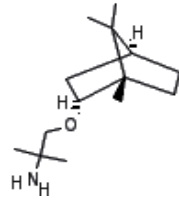




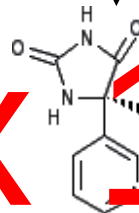
Virtual screening



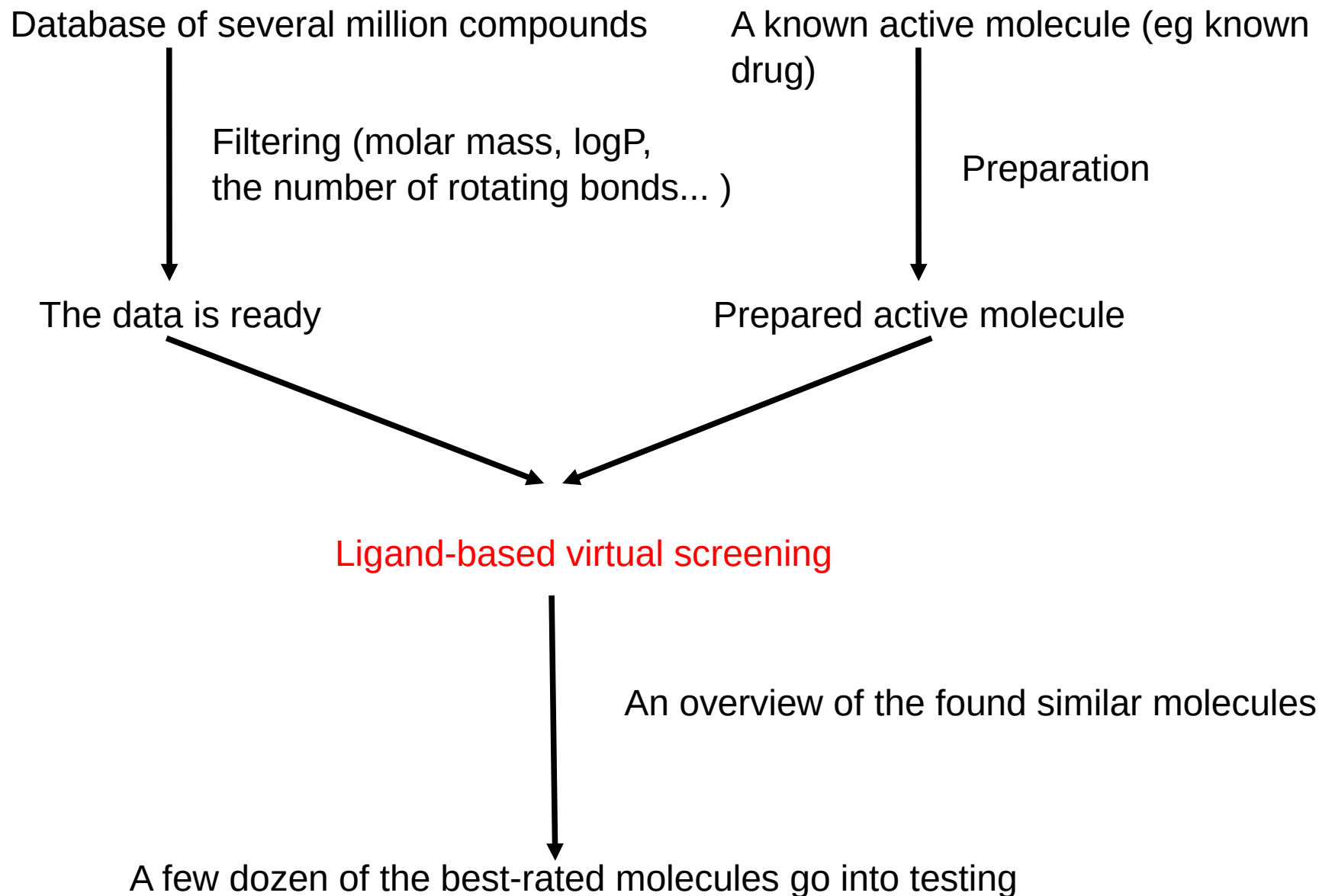
x 1000



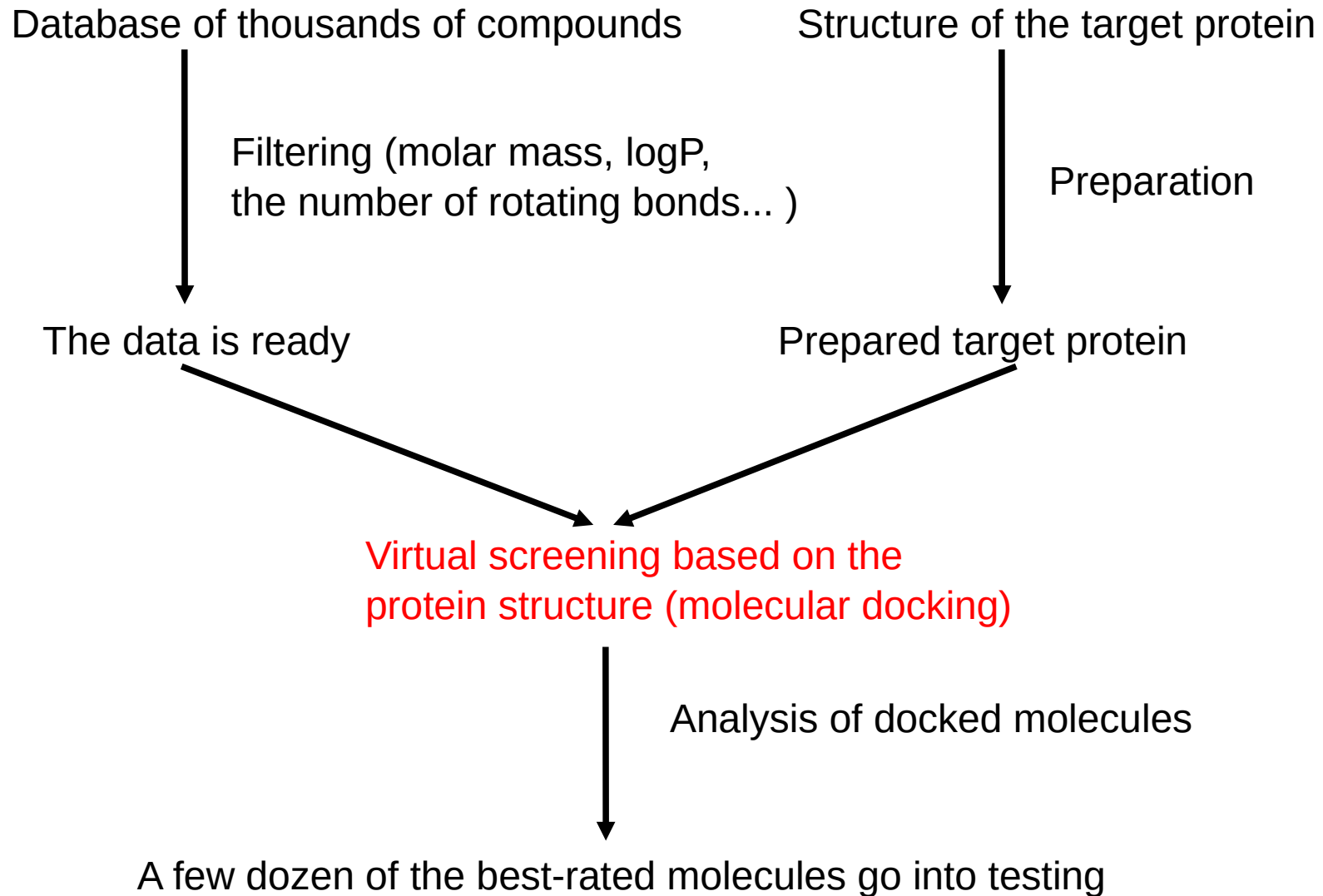
x 10



General scheme I

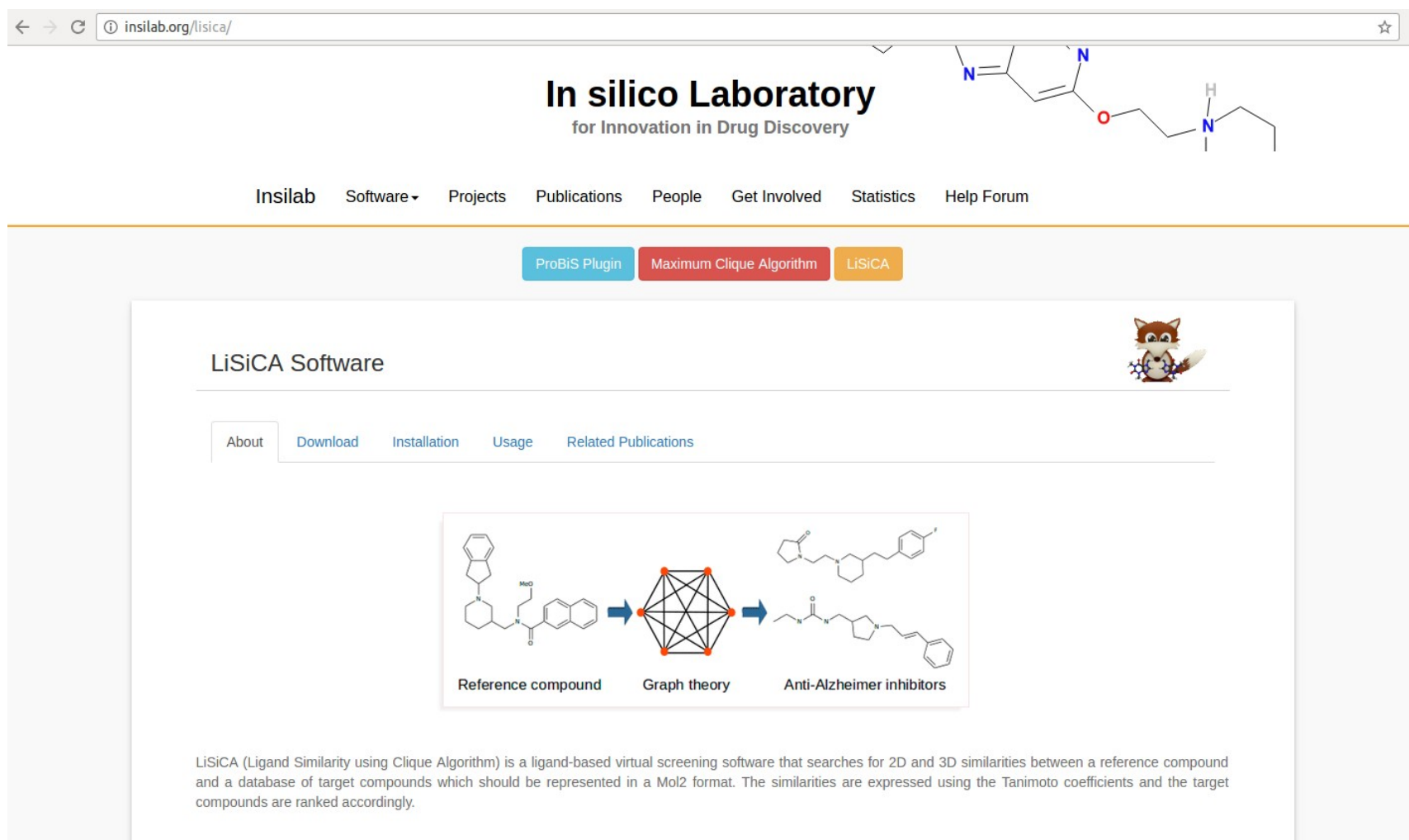


General scheme II



I. Ligand-based virtual screening

- Compound library
- Known active compound
- Solving algorithm (LiSiCA - Ligand Similarity using Clique Algorithm, <http://insilab.org/lisica/>)



The screenshot displays the Insilab website interface. At the top, the browser address bar shows insilab.org/lisica/. The main header features the text "In silico Laboratory for Innovation in Drug Discovery" and a chemical structure of a ligand. Below the header is a navigation menu with links: Insilab, Software, Projects, Publications, People, Get Involved, Statistics, and Help Forum. A secondary navigation bar contains buttons for "ProBiS Plugin", "Maximum Clique Algorithm", and "LiSiCA". The main content area is titled "LiSiCA Software" and includes a sub-menu with "About", "Download", "Installation", "Usage", and "Related Publications". A central diagram illustrates the workflow: a "Reference compound" (a complex organic molecule) is processed through "Graph theory" (represented by a network graph) to identify "Anti-Alzheimer inhibitors" (a set of related molecules). Below the diagram, a paragraph explains that LiSiCA is a ligand-based virtual screening software that searches for 2D and 3D similarities between a reference compound and a database of target compounds, using Tanimoto coefficients for ranking.

LiSiCA (Ligand Similarity using Clique Algorithm) is a ligand-based virtual screening software that searches for 2D and 3D similarities between a reference compound and a database of target compounds which should be represented in a Mol2 format. The similarities are expressed using the Tanimoto coefficients and the target compounds are ranked accordingly.

Compound libraries

...are commercial and non-commercial databases of 2D and 3D structures of small synthetic and natural molecules.

- PubChem (93M structures, <https://pubchem.ncbi.nlm.nih.gov>)
- ZINC (35M structures; several subsets of compounds, <http://zinc.docking.org>)
- Cambridge Structural Database (crystal structures of small molecules)
- NCI (140K structures; cancer research)

The screenshot shows the ZINC 12 website interface. At the top, there is a navigation bar with the UCSF logo and text: "University of California, San Francisco | About UCSF | Search UCSF | UCSF Medical Center". On the right, it says "Shoichet Laboratory" and "docking.org". Below this, it indicates "Not Authenticated — sign in" and "Active cart: Temporary Cart (0 items)".

The main header features the "ZINC 12" logo and a search bar with "Quick Search Bar..." and a "Go" button. Navigation links include "About", "Search", "Subsets", "Help", "Social", and a "G+1 80" badge.

The main content area is divided into two columns. The left column contains a welcome message: "Welcome to ZINC, a free database of commercially-available compounds for virtual screening. ZINC contains over 35 million purchasable compounds in ready-to-dock, 3D formats. ZINC is provided by the Irwin and Shoichet Laboratories in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF). To cite ZINC, please reference: Irwin, Sterling, Mysinger, Bolstad and Coleman, *J. Chem. Inf. Model.* 2012 DOI: 10.1021/ci3001277. The original publication is Irwin and Shoichet, *J. Chem. Inf. Model.* 2005,45(1):177-82 PDE, DOI. We thank NIGMS for financial support (GM71896)." Below this is a search bar with fields for "ZINC ID, Drug Name, SMILES, Catalog, Vendor Code, Target & more" and a "Go" button. Further down are links for "Structure/Draw", "Physical Properties", "Catalogs & Vendors", "ZINC IDS", "Targets", "Rings", and "Combination".

The right column features a "Molecule of the Minute" section with the ID "72055983" and a chemical structure of a complex organic molecule. Below this is a video player titled "10-Special Subsets in Z..." with a play button and a thumbnail image.

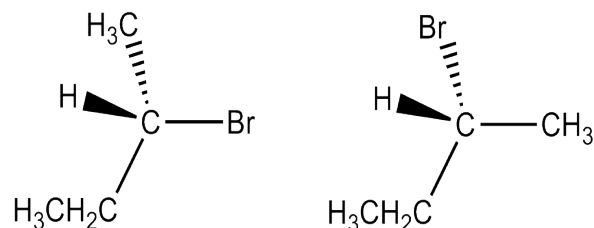
At the bottom left, there are "Quick Links" for "Download", "Search", "Target focused", "Thanks", "Natural Products", "Special Subsets", "Search By Target", "PBCs", "Rings", and "Carts". There is also a "Your Carts" section with the text "Create an account or login to have multiple carts."

At the bottom right, there is a social media section for the "ZINC Database" with a Facebook "Like Page" button showing "794 likes".

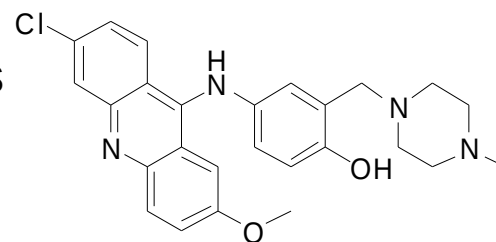
The footer contains the text: "Bioinformatics and Chemical Informatics Research Center (BCIRC) Terms of use Privacy policy Questions, Discussion, Bug reports, Feature requests Thank you NIGMS! GM71896 Go Secure".

Preparation of small molecules

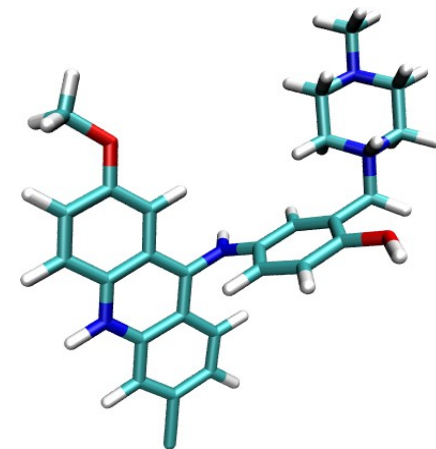
- Filter (MM, number of rotating bonds, functional groups – toxicity)
- 2D or 3D virtual screening?
- In 3D conversion from 2D to 3D structures
- Conformations
- Stereoisomers
- Ionization states



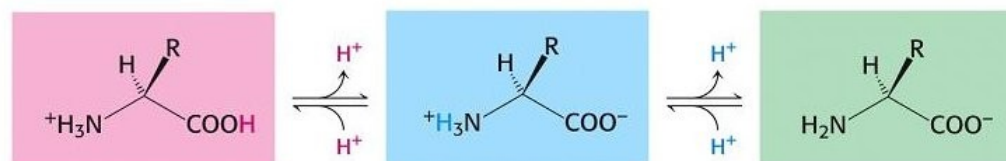
Stereoisomers



2D structure



3D structure



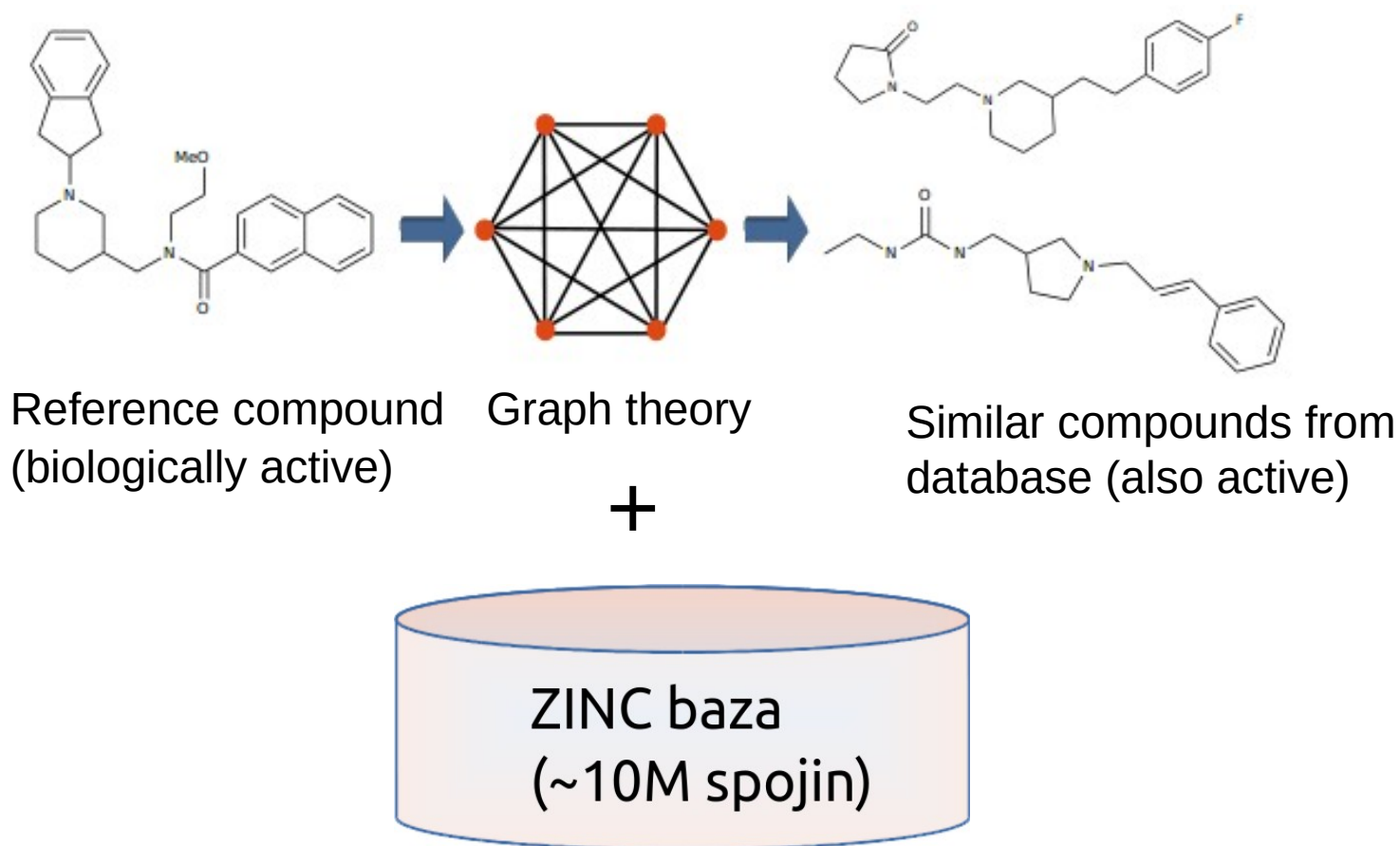
Low pH
< 3.1

Neutral pH
7

High pH
> 11

Ionization states

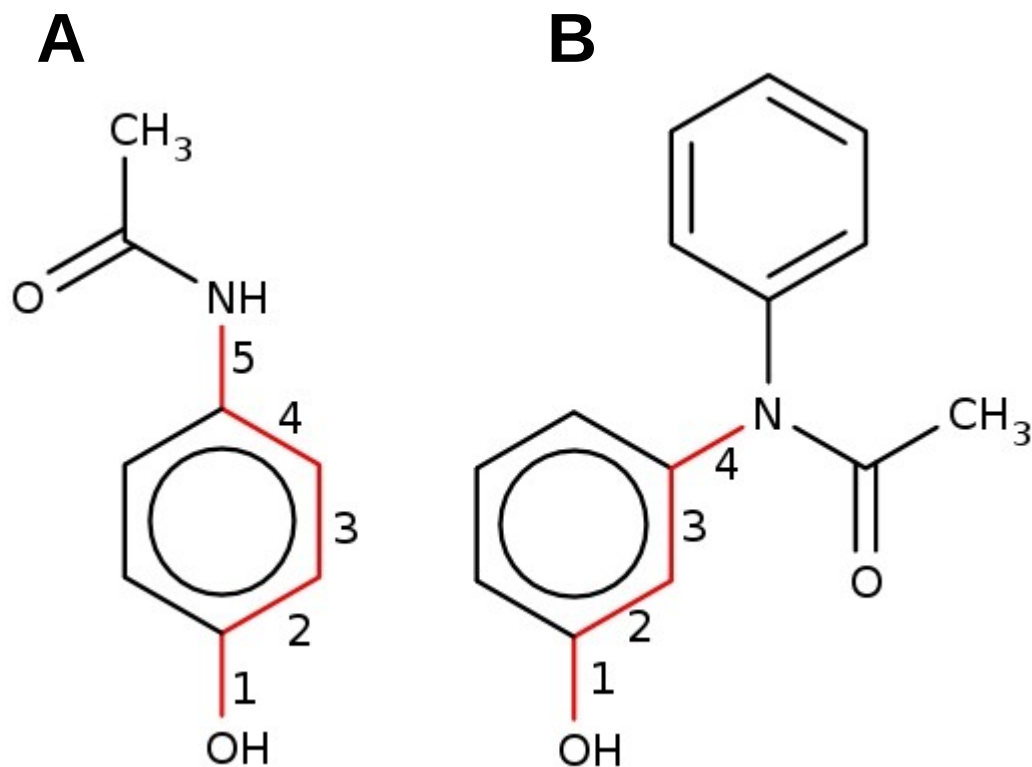
LiSiCA - program for ligand-based virtual screening



J. Chem. Inf. Model., **2015** , 55 , 1521-1528.

J. Cheminform., **2016** , 8:46 .

Tanimoto coefficient (T) - a measure of the similarity of molecules

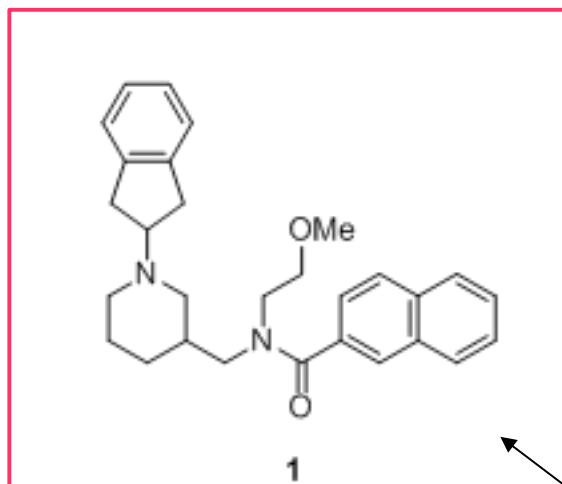


$$T(AB) = c / (a + b - c) = 11 / (11 + 17 - 11) = 11 / 17 = 0.65$$

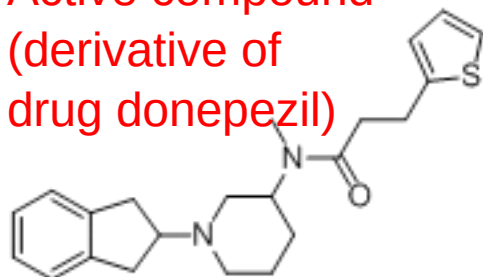
- c number of total atoms
- a, b number of atoms of molecule A and B

Alzheimer's - screening ~10M compounds

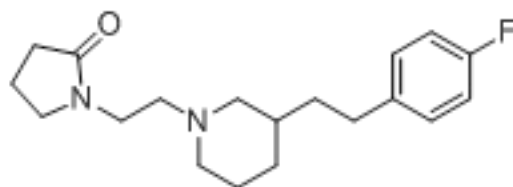
- * Degeneration of neurons -> decrease in acetylcholine concentration
- * Increased expression of the butyrylcholinesterase enzyme in Alzheimer's disease
- * Based on known butyrylcholinesterase inhibitors, "make" new ones



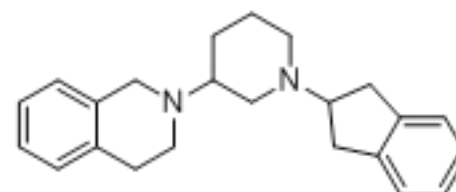
Active compound
(derivative of
drug donepezil)



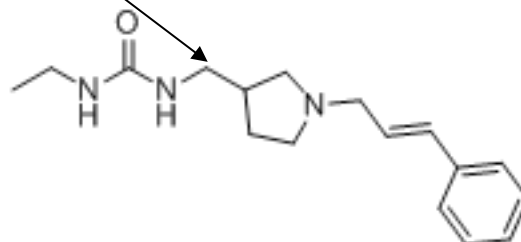
ZINC12303045



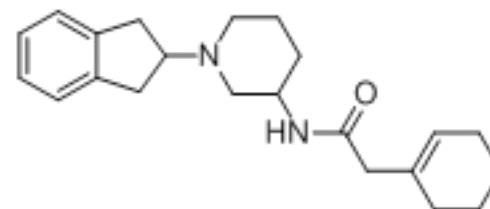
ZINC72121826



ZINC23360769



ZINC67855404

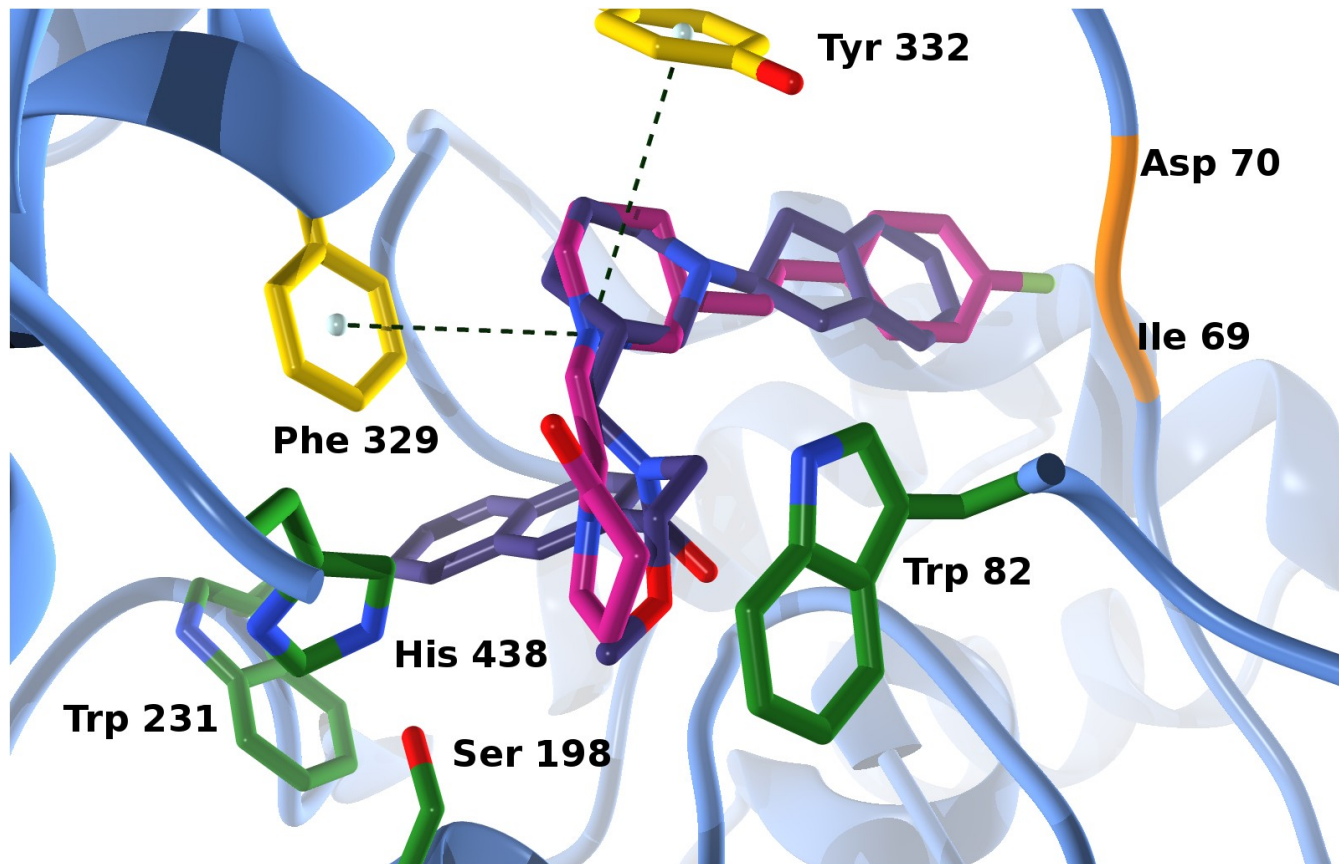


ZINC12702819

Scaffold hopping

Compound 1: known inhibitor of butyrylcholinesterase
Compounds 2-6: active compounds found by LiSiCA
(measured IC₅₀ values of new inhibitors: 80 nM – 840 nM)

Comparison of known inhibitor: a new inhibitor



3D overlay:

- known inhibitor (dark purple)
- new inhibitor (IC₅₀: 80 nM) (pink)

LiSiCA - user interface

The image displays the LiSiCA user interface, which is a graphical user interface for a software plugin. The main window is titled "LiSiCA Plugin" and features a logo of a fox and the text "LiSiCA Software". The interface is divided into two main sections: "Inputs" and "Outputs".

The "Inputs" section contains the following fields and controls:

- Reference Ligand:
- Target Ligand(s):
- Product Graph Dimension: 2 Dimensional Screening 3 Dimensional Screening
- Maximum allowed Shortest Path difference:
- No of highest ranked molecules to write to the output:
- Number of CPU cores to be used:
- Save results in:

A "GO" button is located at the bottom right of the "Inputs" section, and a "Close" button is at the bottom center.

The "Outputs" section is currently empty.

In the background, a PyMOL window is visible, showing a 3D visualization of a molecular structure. The window title is "PyMOL Plugin" and it displays version 1.7.5.0. The interface includes a toolbar with buttons for "Reset", "Zoom", "Orient", "Draw", "Ray", "Unpick", "Deselect", "Rock", "Get View", "Command", "Builder", "Volume", "Rebuild", and "Abort". The main view area shows a molecular structure with a blue box containing the text "Open-Source PyMOL™ 1.7.x" and "Copyright © 2009-2014 Schrödinger, LLC". A terminal window at the bottom right displays the following text:

```
Mouse Mode 3-Button Viewing
Buttons L M R Wheel
& Keys Rota Move MovZ Slab
Shft +Box -Box Clip MovS
Ctrl +/- PkAt PkI MwSZ
CtSh Sele Orig Clip MovZ
SnglClk +/- Cent Menu
DbiClk Menu - PkAt
Selecting Residues
State 1/ 1
```

The system tray at the bottom of the screen shows the time as 1:38 PM and the date as 9/22/2015.

LiSiCA - user interface

The screenshot displays the LiSiCA Software interface, which includes a list of ZINC IDs and their corresponding Tanimoto scores. The interface also features a PyMOL Molecular Graphics System window showing a 3D molecular model.

LiSiCA Software

Inputs

Rank	ZINC ID	Tanimoto score
2	ZINC36259732	0.826087
3	ZINC36259735	0.826087
4	ZINC36259749	0.826087
5	ZINC36259752	0.826087
6	ZINC13514742	0.826087
7	ZINC34181597	0.818182
8	ZINC00084006	0.818182
9	ZINC39367469	0.791667
10	ZINC39367471	0.791667
11	ZINC39367473	0.791667
12	ZINC39367474	0.791667
13	ZINC01750540	0.791667
14	ZINC00335291	0.782609
15	ZINC00335289	0.782609
16	ZINC39192900	0.782609
17	ZINC03223820	0.772727
18	ZINC00394183	0.772727
19	ZINC05863462	0.772727
20	ZINC82388897	0.769231
21	ZINC82309221	0.75
22	ZINC11956864	0.75
23	ZINC11956859	0.75
24	ZINC05868571	0.75
25	ZINC00152393	0.75

Outputs

Ref. Num	Ref. Atom	Tar. Num	Tar. Atom	Atom Type
14	H3	3	H1	H
12	H1	17	H4	H
13	H2	16	H3	H
1	C1	2	C2	C.3
11	O2	14	O3	O.3
20	H9	22	H9	H
9	C7	10	C7	C.ar
8	C6	9	C6	C.ar
18	H7	20	H7	H
7	C5	8	C5	C.ar
17	H6	19	H6	H
3	O1	5	O1	O.2
2	C2	4	C3	C.2
6	C4	7	C4	C.ar
10	C8	11	C8	C.ar
19	H8	21	H8	H
4	N1	6	N1	N.am
15	H4	18	H5	H
5	C3	12	C9	C.ar

PyMOL Molecular Graphics System

Build Movie Display Setting Scene Mouse Wizard Plugin Help

Reset Zoom Orient
Unpick Deselect
|< < Stop Play >|
Command Builder
Rebuild Abort

PyMOL Viewer

PyMOL>

II. Virtual screening based on the protein structure (molecular docking)

- Structure of the target protein
- Compound library
- Binding sites with the ProBiS (Protein Binding Sites) approach
- Virtual screening with the GenProBiS and ProBiS-Fold web server
- Determination of ligand interactions with sequence variants (sequence variants) in the binding site

The screenshot displays the ProBiS web interface. On the left, a 3D ribbon representation of a protein structure is shown in green, with a predicted ligand (a small molecule) docked in the binding site. The interface includes a search bar at the top right, a table of predicted ligands, and a table of similar protein structures. The table lists various ligands with their names, sources, confidence scores, and binding site information.

Structure	Name	Source	Confidence	Binder	Ligand
	2-chloro-4-[[4,6-di...	4krn	4.23	Specific	Remove
	2-chloro-4-[[pyrimidin-2-yl)sulfanyl] acetyl]benze	4krn	4.23	Specific	View 3D
	N-(4-chlorobenzyl)-n-methylbenzene-1,4-disulfon	3da2	4.23	Specific	View 3D
	N-(4-chlorobenzyl)-n-methylbenzene-1,4-disulfon	3da2	4.23	Specific	View 3D
	2,3,5,6-tetrafluoro-4-[(2-hydroxyethyl) sulfanyl]b	4hu1	4.23	Specific	View 3D
	Acetate ion	3u0n	4.23	Non-specific	View 3D
	2-chloro-4-[[pyrimidin-2-yl)sulfanyl] acetyl]benze	4krn	4.23	Specific	View 3D
	2,3,5,6-tetrafluoro-4-[(2-hydroxyethyl) sulfanyl]b	4hu1	4.23	Specific	View 3D
	2-chloro-4-[[4,6-dimethylpyrimidin-2-yl) sulfanyl]	4krn	4.23	Specific	View 3D
	5-acetamido-1,3,4-thiazolo[2-sulfonamido]	3czv	4.23	Specific	View 3D
	5-acetamido-1,3,4-thiazolo[2-sulfonamido]	3mb	3.9	Specific	View 3D
	6-ethoxy-1,3-benzothiazole-2-sulfonamide	3mfp	3.9	Specific	View 3D

Predicted 3D conformation of the ligand in the target protein

Preparation of target protein I

- The Protein Data Bank (PDB) is a Database of Protein Structures (<http://www.rcsb.org>)
- Determined by X-ray, electronically microscopy or nuclear magnetic resonance
- The protein is uniquely identified by four letters PDB code and single letter chain code e.g. PDB ID: 4BQP
Chain ID: A

The screenshot shows the RCSB PDB website homepage. At the top, there is a navigation bar with links for Deposit, Search, Visualize, Analyze, Download, Learn, and More, along with a MyPDB Login button. Below the navigation bar is the PDB logo and the text "An Information Portal to 114217 Biological Macromolecular Structures". A search bar is located on the right side of the page, with a "Go" button. Below the search bar are links for "Advanced Search" and "Browse by Annotations". The main content area is divided into several sections. On the left, there is a "Welcome" section with a sidebar containing links for Deposit, Search, Visualize, Analyze, Download, and Learn. The main content area features a "A Structural View of Biology" section, a "December Molecule of the Month" section featuring Vancomycin, and a "Take an Interactive Tour of the PDB" section. The "December Molecule of the Month" section shows a 3D model of Vancomycin with two D-Ala residues highlighted in red.

The screenshot shows the RCSB PDB website with three main sections: "Latest Entries", "New Features", and "News". The "Latest Entries" section features a 3D model of the Crystal Structure of the Human Factor VIII C2 Domain in Complex with Murine 3E6 Inhibitory Antibody (PDB ID: 4XZU). The "New Features" section lists updates from October 2015 and September 2015. The "News" section includes articles about the Tour of Ligand Deposition, Validate Before Depositing to Save Time, Advanced Search: Multiple ID Search, Comparison Tool for Exploring Sequence and Structure Alignments, and the Phased PDB Release Process.

Latest Entries

As of *Tuesday, Dec 08*

4XZU PDB Entry

Crystal Structure of the Human Factor VIII C2 Domain in Complex with Murine 3E6 Inhibitory Antibody

[View in 3D](#)

New Features

October 2015 Release

- Redesigned Structure Summary Page**
New Organization. Improved Layout. Clean. Usable. Simple.
- Improved Literature Tab
- Better Support for Mobile Browsing
- Redesigned Ligand Summary Page

September 2015 Release

- Validation Track on Protein**

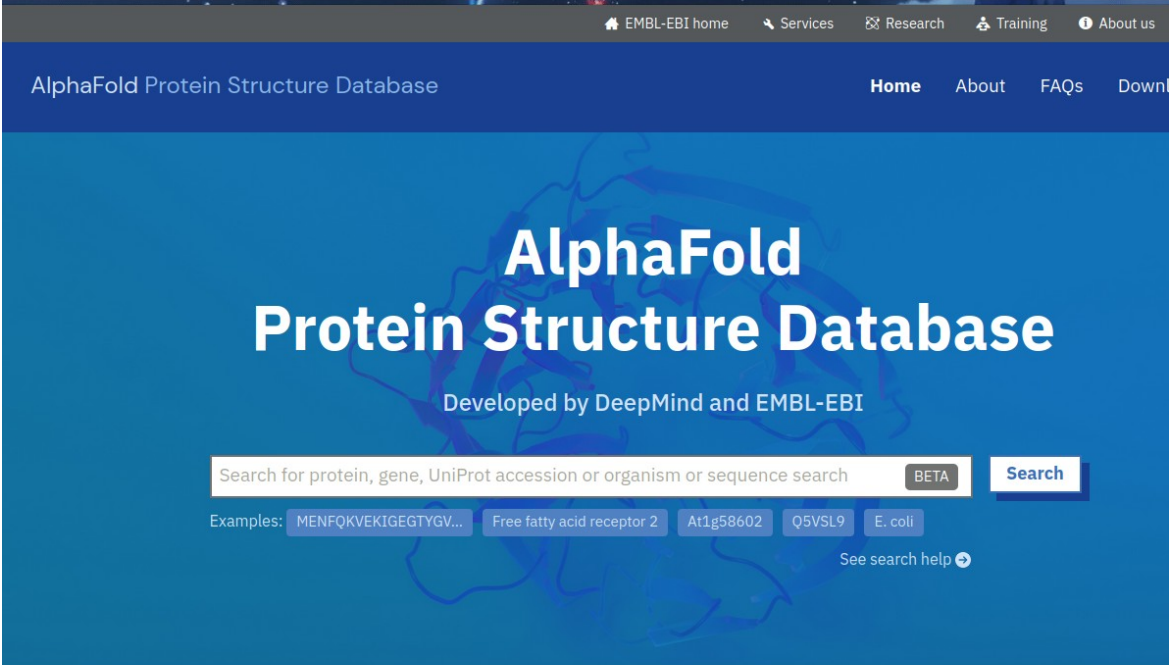
News

Publications ▾

- Tour of Ligand Deposition**
Watch how to review and submit ligands using the wwPDB Deposition Tool.
» 12/08/15
- Validate Before Depositing to Save Time » 12/01/15
- Advanced Search: Multiple ID Search » 11/24/15
- Comparison Tool for Exploring Sequence and Structure Alignments » 11/17/15
- Phased PDB Release Process**
» 08/24/15

Preparation of target protein II

- AlphaFold database of protein structures
(<https://alphafold.ebi.ac.uk/>)
- predicted protein structures using machine learning approach
- 48 organisms, more than 200 million structures



AlphaFold Protein Structure Database

Home About FAQs Down

AlphaFold Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism or sequence search BETA Search

Examples: MENFQKVEKIGEGTYGV... Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli

[See search help](#)

AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

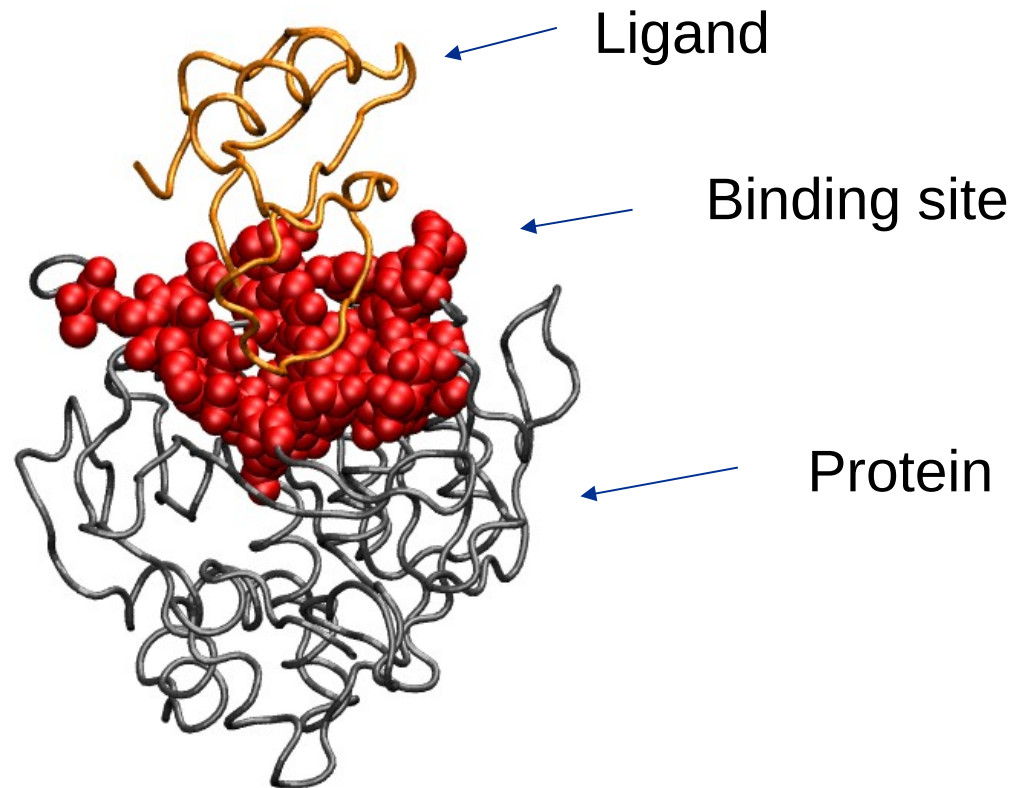
Background

Protein molecular dynamics



Proteins are dynamic: which conformation to use?

Determination of the binding site

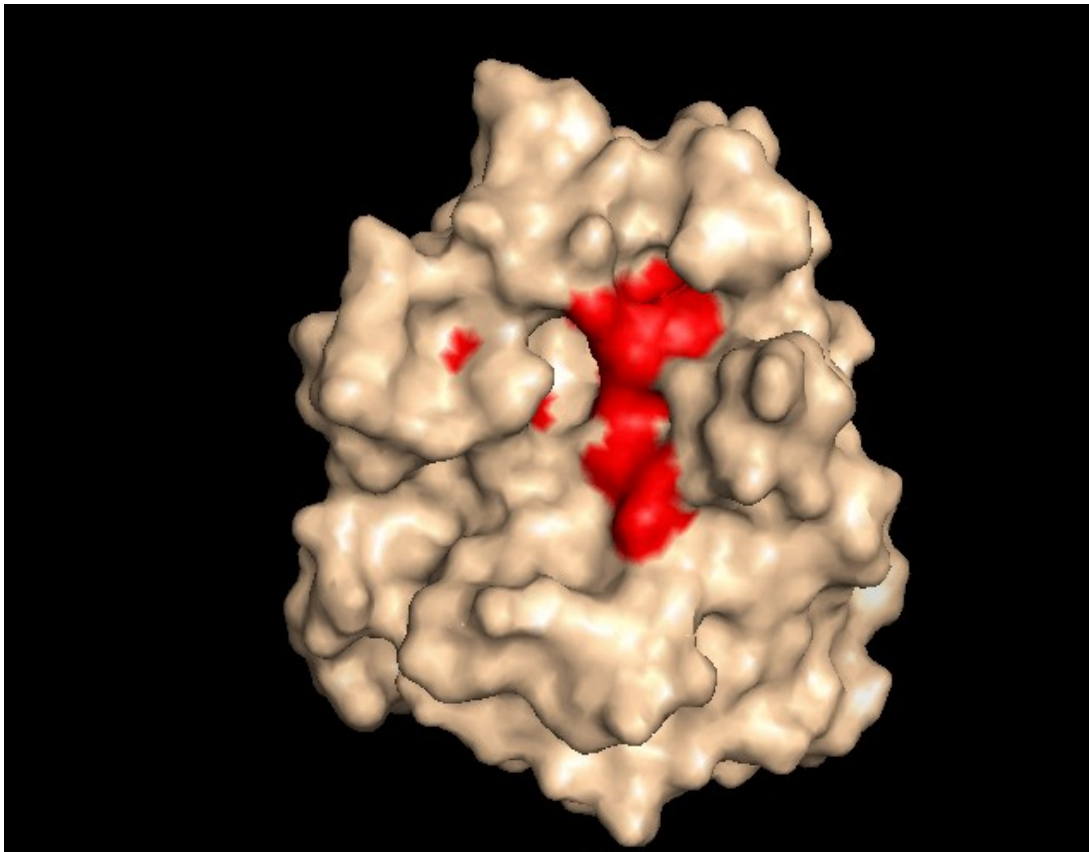


Ligands:

- proteins
- nucleic acids
- synthetic and natural compounds
- ions
- waters

Binding site

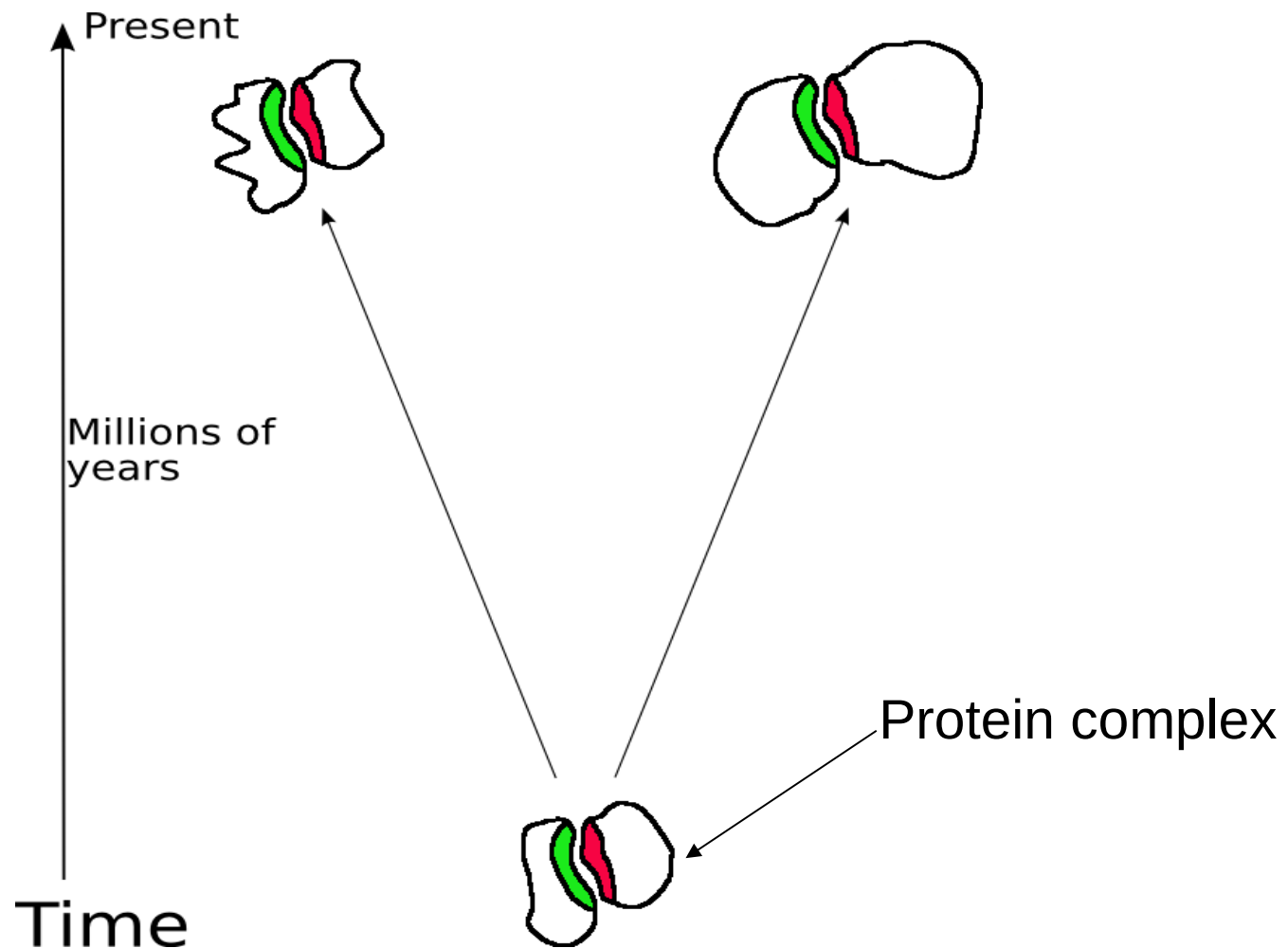
The number of potential target proteins is estimated to be ~6000, but pharma uses only ~200 for development of new drugs.



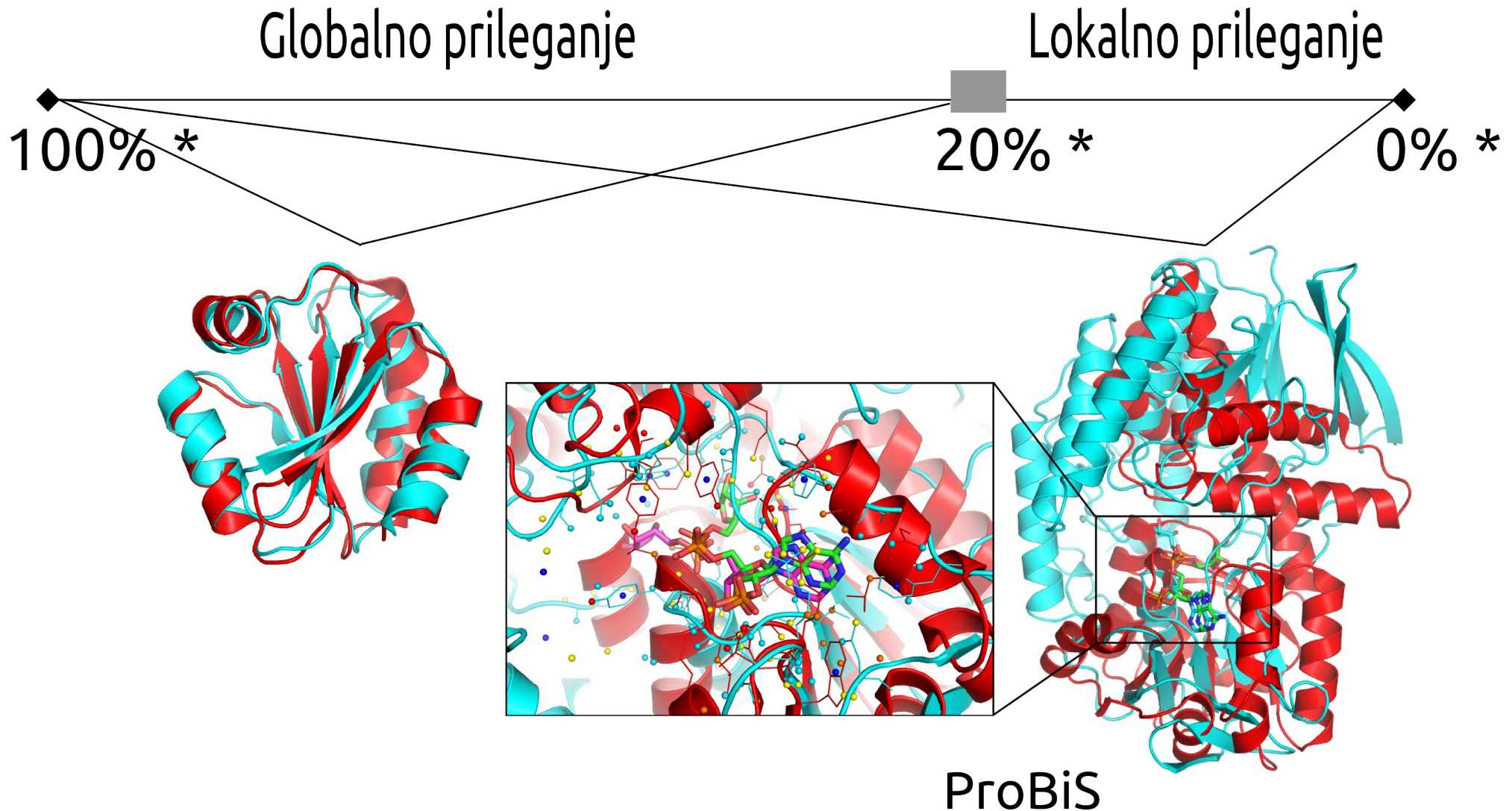
The binding sites for the active substances are in cavities in the surface of the protein.

Evolution of binding sites

Binding sites change more slowly than the rest of the protein through evolution.



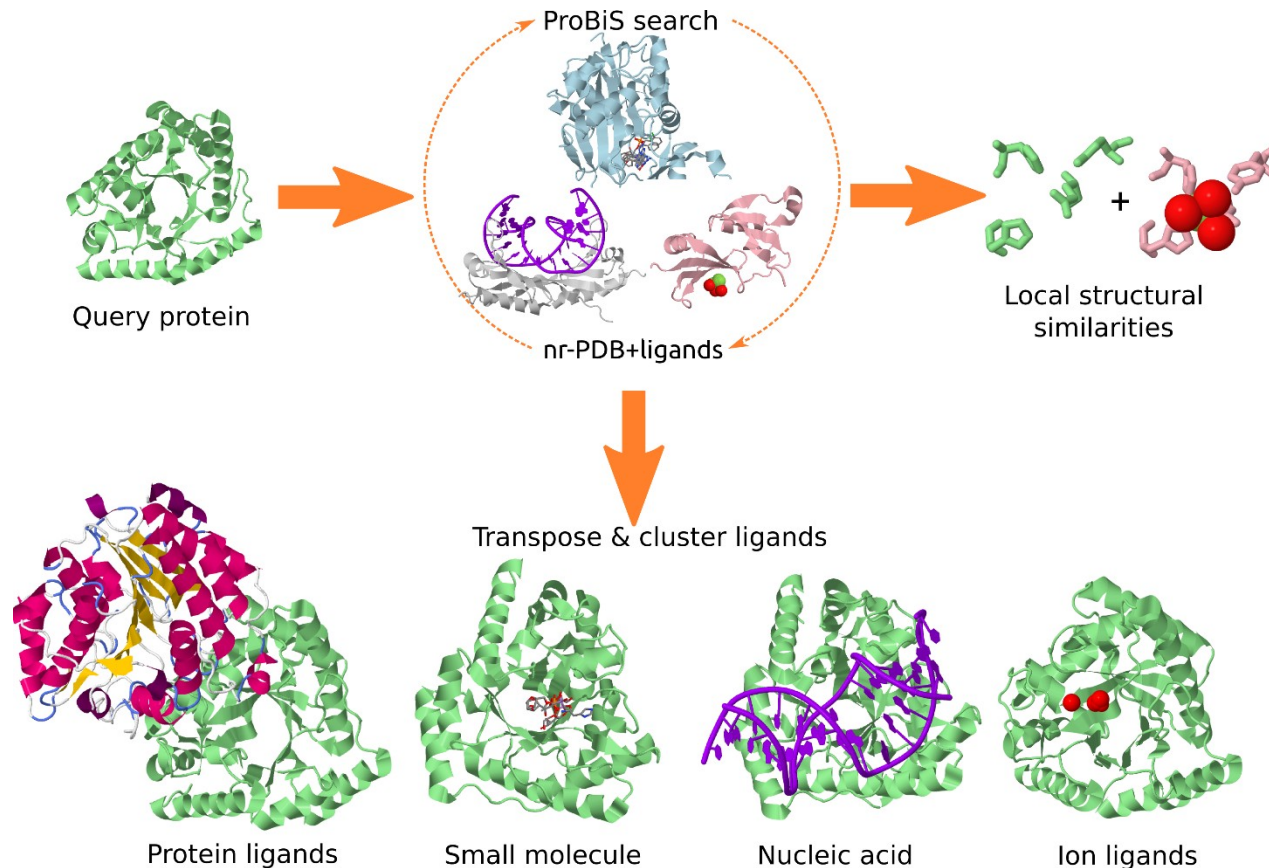
ProBiS algorithm for structural protein fitting



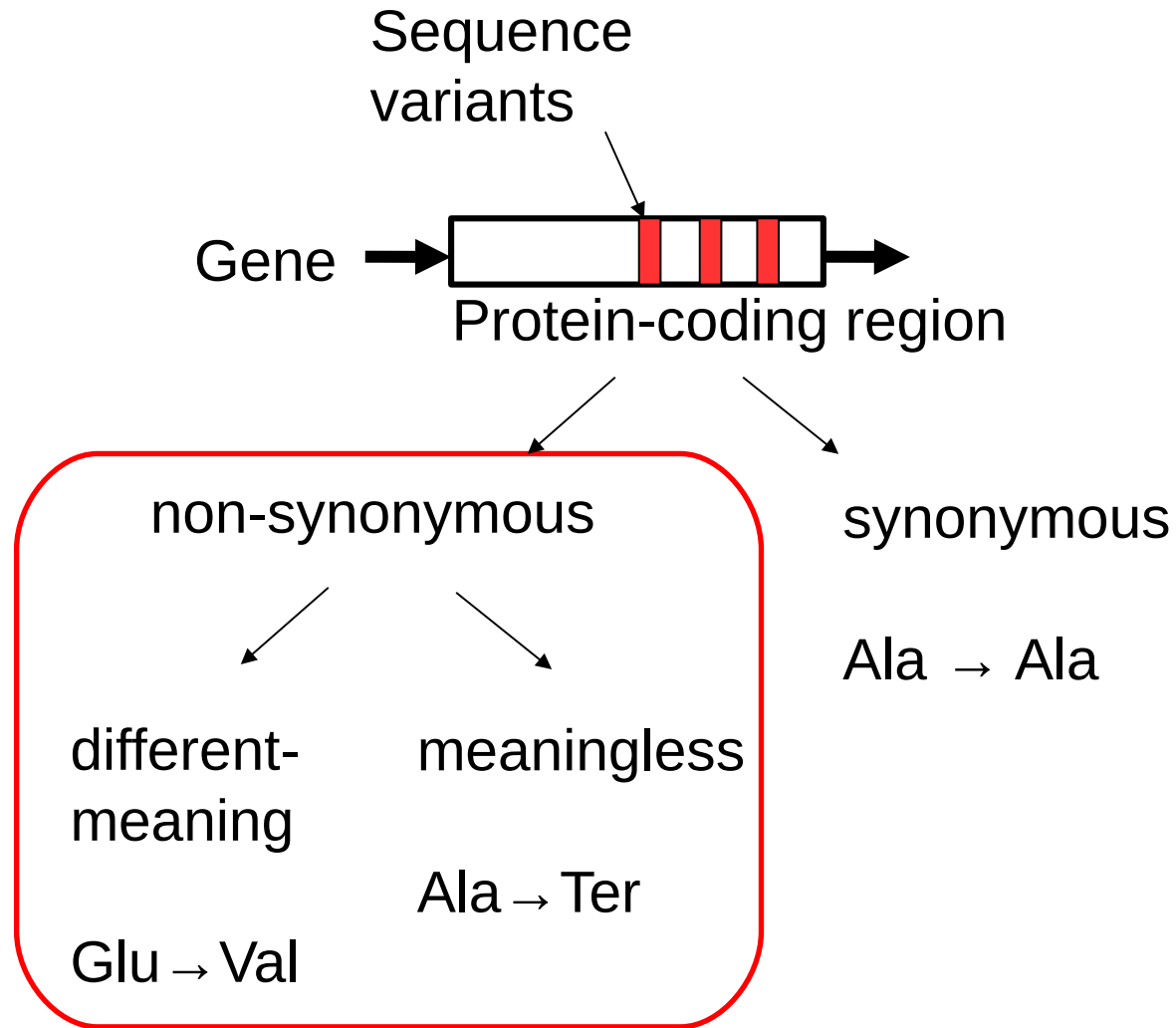
* sequence identity

Virtual screening with the ProBiS approach

- If the two binding sites are similar, similar ligands bind to them
- Ligands from the first binding site can be "moved" to the second, binding site provided that the two binding sites are similar

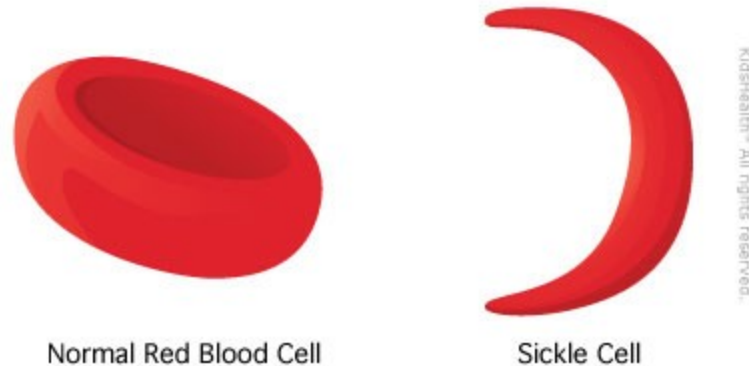


Sequence variants



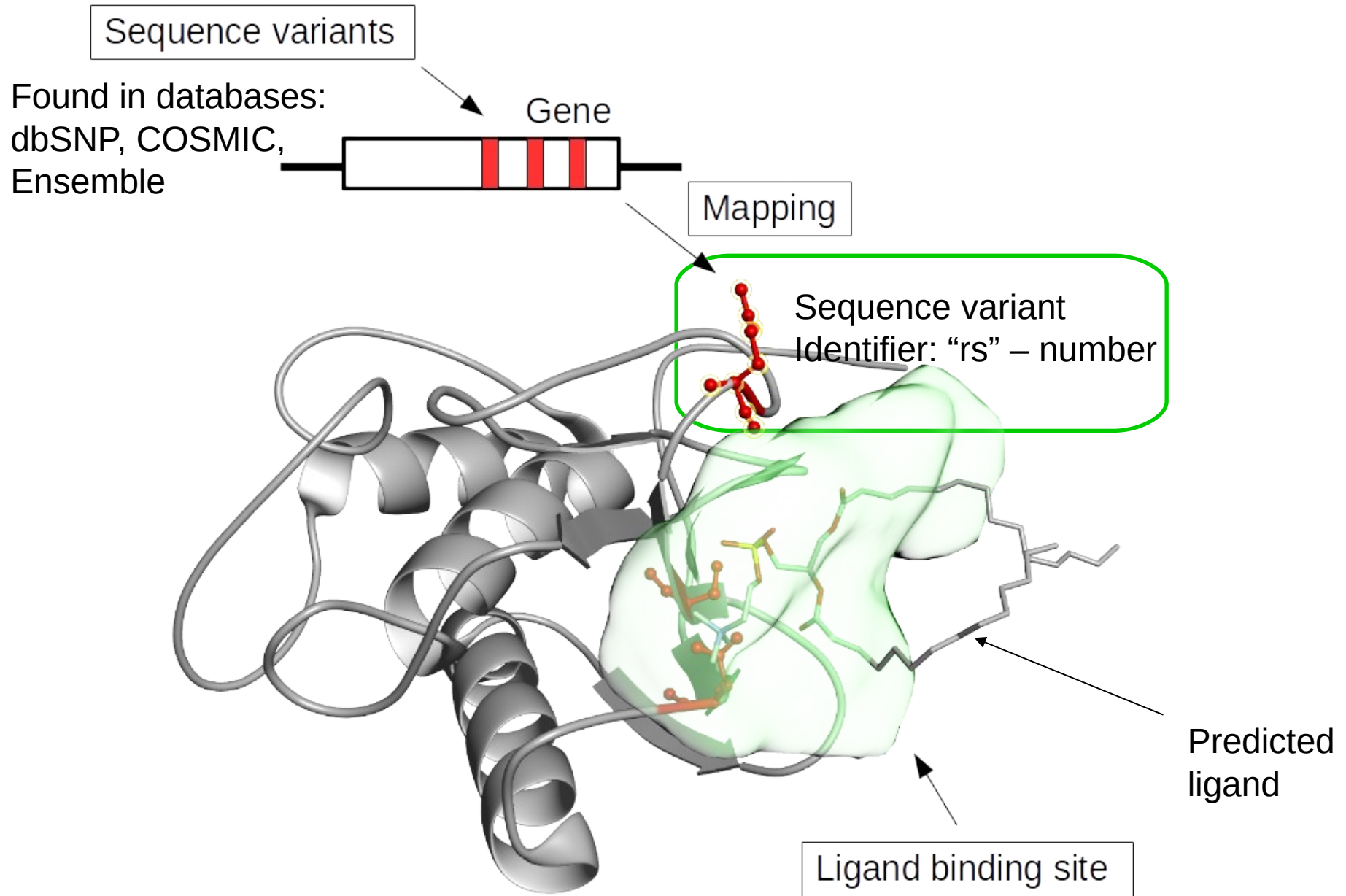
Sequence variants

- Responsible for the emergence of various diseases
- Responsible for different response to drugs
- Involved in the formation of cancer






Sickle cell anemia is caused by a hemoglobin variant that has the amino acid valine instead of glutamic acid at position 6 in the sequence.




GenProBiS: mapping sequence variants to binding sites





Variants and binding sites in the protein sequence

A

UniProt: 100 110 120 130 140
 PDB: 100 110 120 130 140
 Sequence: SSVPSQKTYQGSYGFRLGLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQ
 Evolution: 
 Nucleic#1 
 Protein#1 

UniProt: 150 160 170 180 190
 PDB: 150 160 170 180 190
 Sequence: LWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHL
 Evolution: 
 Nucleic#1 
 Protein#1 

UniProt: 200 210 220 230 240
 PDB: 200 210 220 230 240
 Sequence: IRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMG
 Evolution: 
 Nucleic#1 
 Protein#1 

UniProt: 250 260 270 280 290
 PDB: 250 260 270 280 290
 Sequence: GMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGDRDRTEENLRKK
 Evolution: 
 Nucleic#1 
 Protein#1 

rs121913343 : Arg -> Ser

Amino acid sequence of a protein

Evolutionary conservation

Protein-protein binding site

Protein-DNA binding site

Variant in two binding sites

Legend:

Sequence Variant within binding site

Sequence Variant not in binding site

overline - Sequence Variant has annotation

Nucleic binding site

Protein binding site

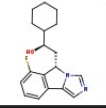
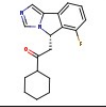
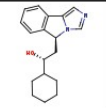
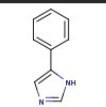
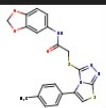
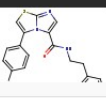
Missing in crystal structure


Evolutionary conservation

0 1 2 3 4 5 6 7 8 9

GenProBiS web server

Binding Site #1
Binding Site #2

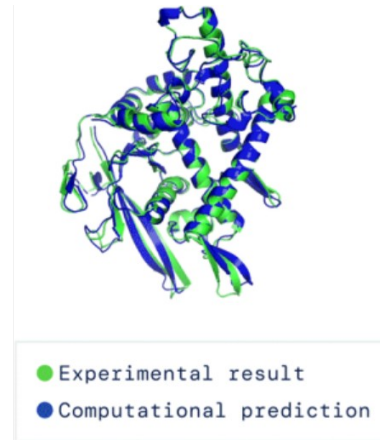
	S	Image	Ligand	Count
	<input type="radio"/>		5PF	1
	<input type="radio"/>		5PJ	1
	<input type="radio"/>		5PK	1
	<input type="radio"/>		PIM	3
	<input type="radio"/>		PKJ	2
	<input type="radio"/>		PKL	2



4pk5	Pkj	502	Leu	234 / 234	rs776049118	Leu234Val	3.61
4pk5	Pkj	502	Cys	129 / 129	rs761435727 rs540988047	Cys129Gly Cys129Tyr	3.73
4pk5	Pkj	502	Phe	163 / 163	rs764150078	Phe163Ser	3.90
4pk5	Pkj	502	Leu	234 / 234	rs776049118	Leu234Val	3.99
4pk5	Pkj	502	Ser	167 / 167	rs189439950	Ser167Cys	4.26

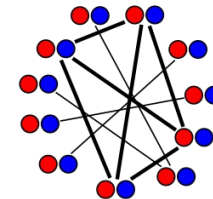
Prediction of binding sites for AlphaFold structures

Input: AlphaFold protein structure model (human)

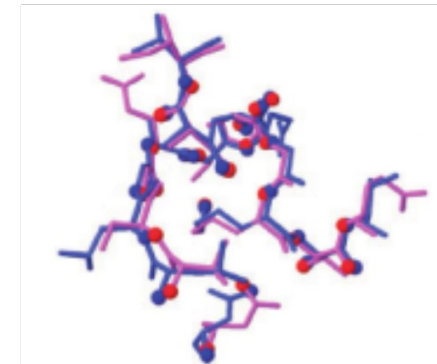


Compare against PDB
(with ProBiS algorithm)

RCSB **PDB**
PROTEIN DATA BANK



Gather pairwise local similarities between
AlphaFold structure and PDB structures

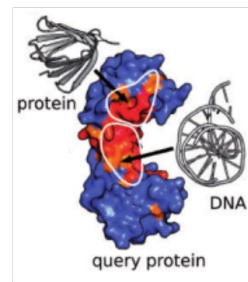


Prediction of binding sites for AlphaFold structures

Transpose the ligands from PDB to AlphaFold structure

Determine ligand types: protein, peptide, nucleic acid, small molecule (cofactors and compounds), metal ion, water, glycan

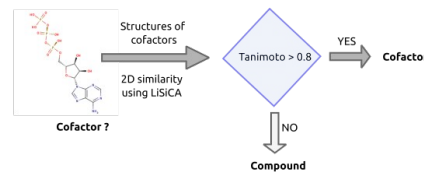
Cluster ligands of same type by their proximity



$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{bmatrix}$$

$$R_y(\theta) = \begin{bmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{bmatrix}$$

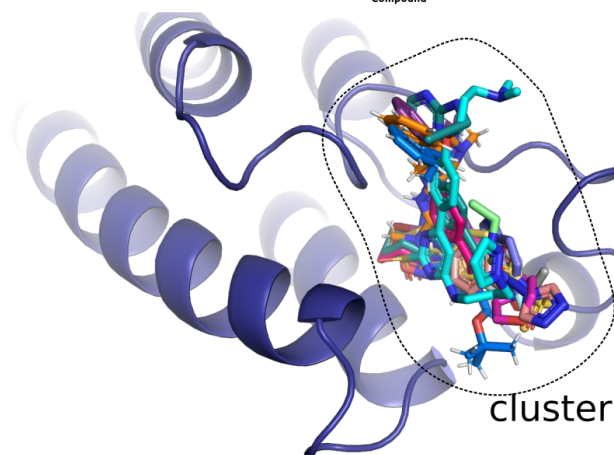
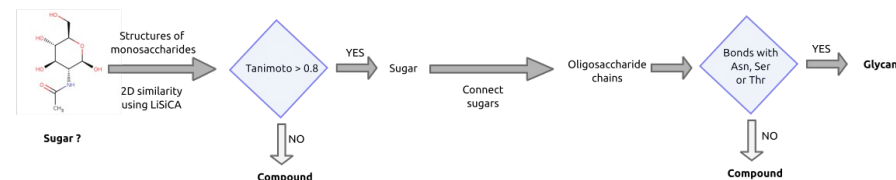
$$R_z(\theta) = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



Protein: any amino acid chain
Peptide: < 20 aa.
Nucleic acid: DNA or RNA

Metal ions: Ca^{2+} , $\text{Cu}^{(1+,2+)}$, Co^{2+} , Na^+ , Mg^{2+} , Cs^+ , Zn^{2+} , Ba^{2+} , Mn^{2+} , Ni^{2+} and $\text{Fe}^{(2+,3+)}$

Water: HOH residue in PDB file



Prediction of binding sites for AlphaFold structures

Refine small molecule binding site types
(with docking applications in mind)

Substrate-competitive: small molecule drugs, agonists, substrates, substrate-competitive inhibitors but not cofactors

Cofactor-competitive: Cofactors and ligands that overlap with cofactors, e.g., cofactor-competitive inhibitors

Calculate scores: druggability score for substrate- and cofactor-competitive binding sites, conservation score for water and metal ion binding sites

Water & ion conservation score:

$$O_{\text{bsite}} = n_{\text{ligand}} / n_{\text{super}} \quad (\text{occupancy})$$

n_{ligand} = number of liganded superimposed sites
 n_{super} = total number of superimposed sites

Hard-coded limits: $O_{\text{water}} > 0.6$, $n_{\text{water}} > 10$

Druggability score:

$$S_{\text{bsite}} = \max_{1 \leq i \leq n_{\text{ligands}}} (S_{\text{cplx},i} + w \times O_{\text{bsite}})$$

$$S_{\text{cplx}} = (n_{\text{rings}} + 1) \times n_{\text{elements}} \quad (\text{complex score})$$

n_{rings} = number of ring systems
 n_{elements} = number of chemical elements

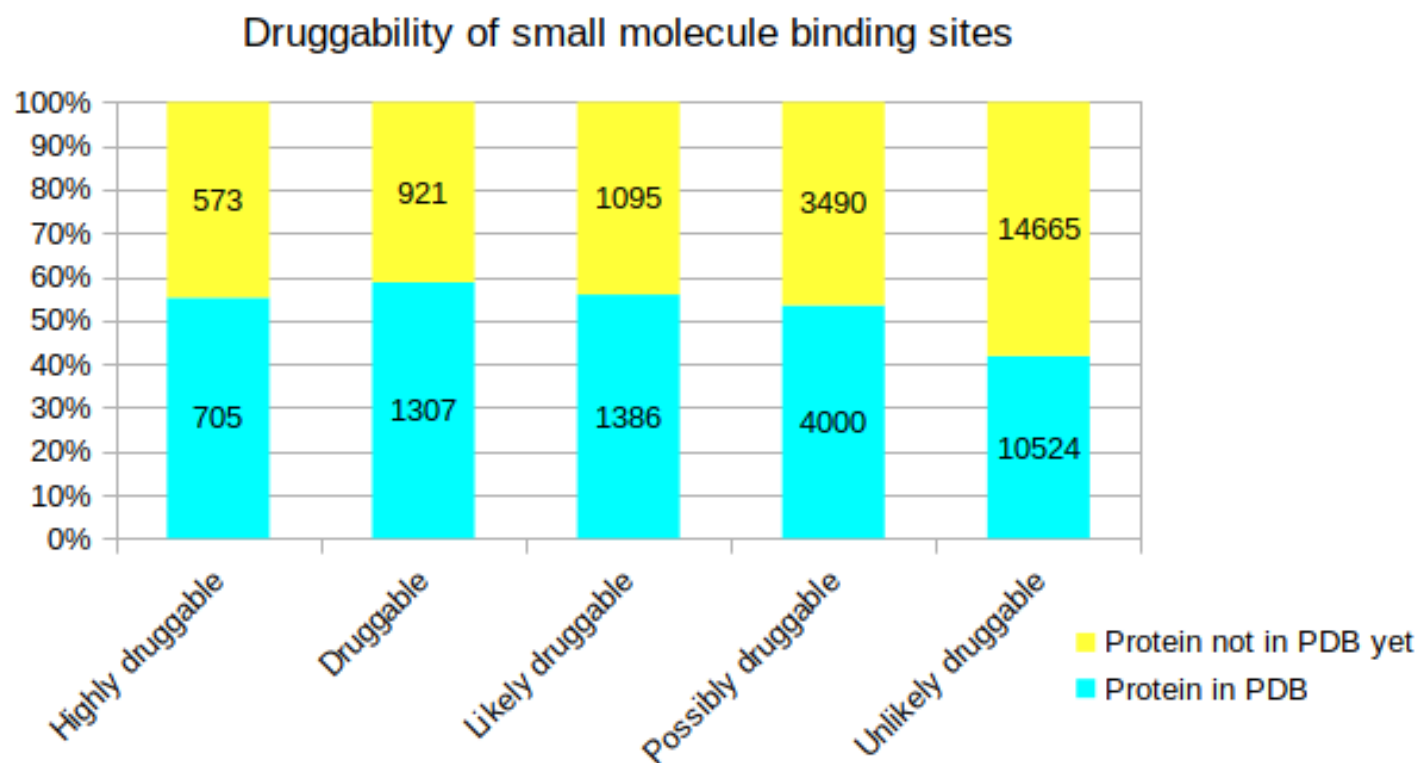
$w = w = 0$ if $S_{\text{cplx},i} < 12$, otherwise $w = 100$

Result: predicted binding sites and predicted ligands for each binding site in human proteome

Confidence score

$$C_{\text{bsite}} = \min_{1 \leq i \leq n_{\text{bsite_residues}}} (\text{AF model confidence}_{\text{residue}_i})$$

New binding sites useful for drug development

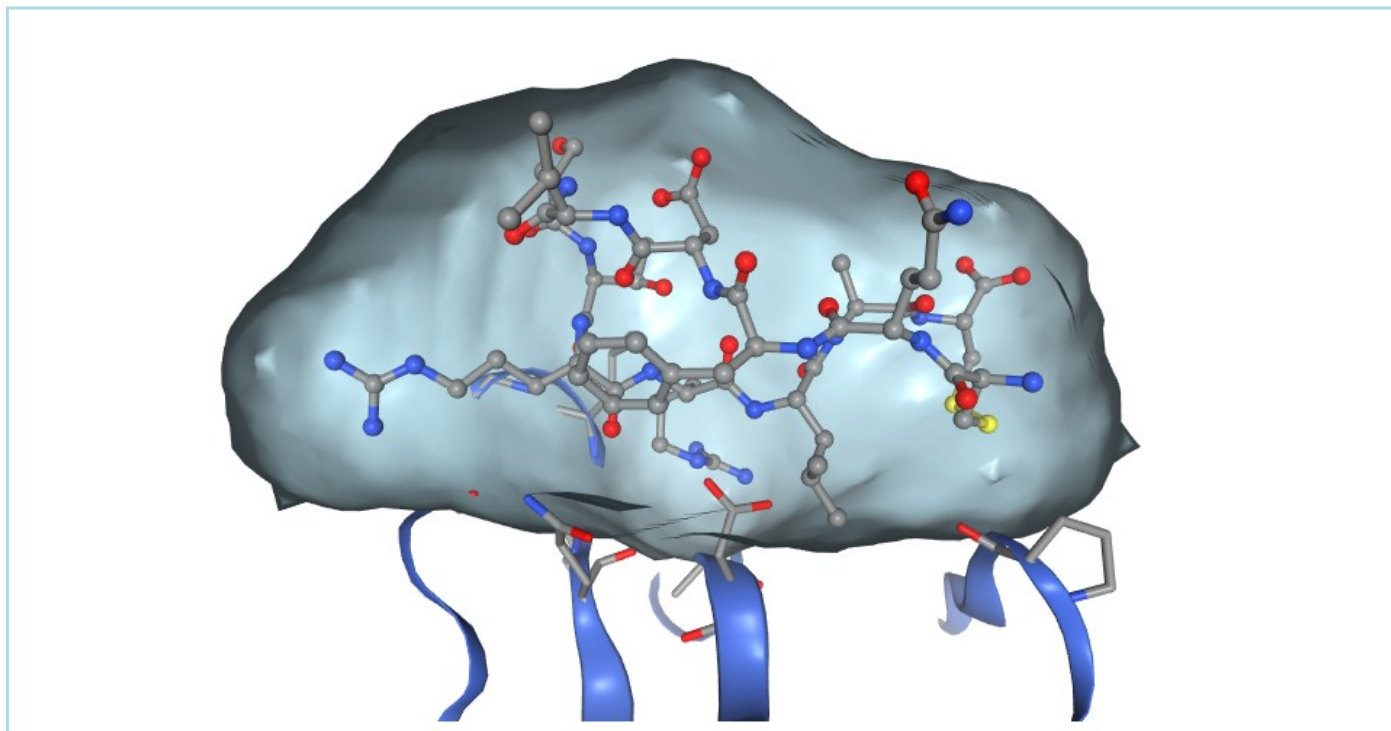


Peptide binding sites

Binding site overview

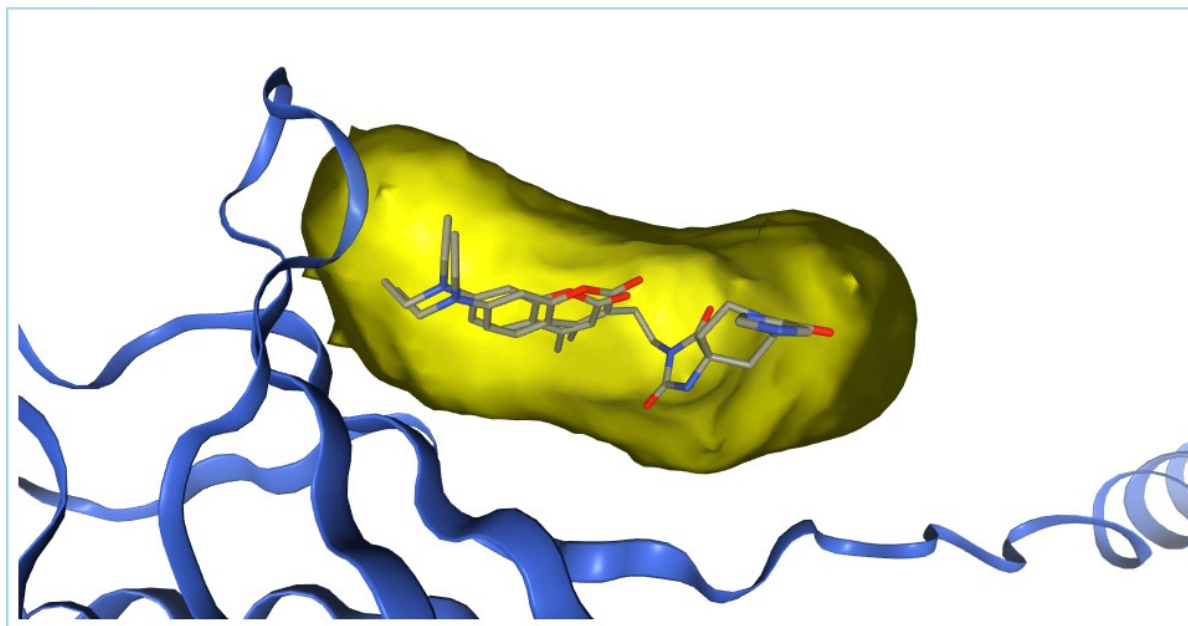
Protein	Probable non-functional immunoglobulin lambda variable
Protein identifiers	AlphaFold ID: AF-A0A075B6I3-F1-model_v2 PDB ID: None UniProt ID: A0A075B6I3
Binding site type	Protein
Binding site rank	Secondary (ranked 2nd)
Binding site chains	Consists of chain A
Ligands	Peptides, i.e., peptide chains with less than 20 amino acids in length

Binding site viewer 3D







Small molecule binding sites


Binding site viewer 3D



Predicted ligands

	Chem ID	Name	PDB ID	Z-score	Show in 3D
	9ZW	4-methyl-3-(morpholi...	7lmq	2.9	<input type="checkbox"/>
	6ZW	7-(diethylamino)-4-m...	6mg5	2.9	<input checked="" type="checkbox"/>
	9ZX	4-methyl-3-(morpholi...	7lmr	2.9	<input type="checkbox"/>
	NY6	3-[2-[7-(diethylamin...	7lmp	2.9	<input checked="" type="checkbox"/>

Protein binding sites

 **ProBiS-Fold** Tutorial Datasets Cite

Binding site overview

Binding site viewer 3D

Predicted ligands

Binding site residues

All predicted binding sites for this protein:

Protein chains

A

Protein binding sites

1 2 3 4 5 6 7

Compound binding sites

1 2 3 4 5

Cofactor binding sites

1

Glycan binding sites

1 2 3 4 5 6 7

8

Metal ion binding sites

1

Peptide binding sites

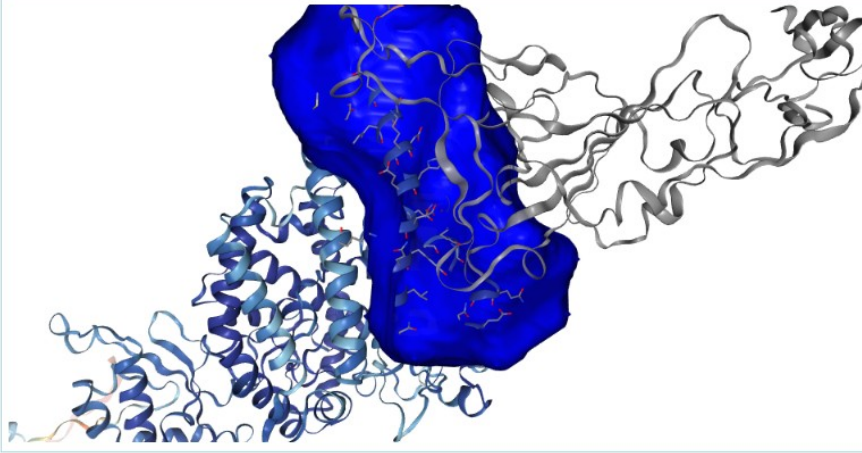
1

[Back to top ↑](#)

Binding site overview




Protein	Angiotensin-converting enzyme 2
Protein identifiers	AlphaFold ID: AF-Q9BYF1-F1-model_v2 PDB ID: 7vxf UniProt ID: Q9BYF1
Binding site type	Protein
Binding site rank	Primary (ranked 1st)
Binding site chains	Consists of chain A
Ligands	Other proteins and peptides, i.e., oligo- and polypeptides of any amino acids chain length
Confidence	Very low (AF model confidence = 49.01)

Binding site viewer 3D

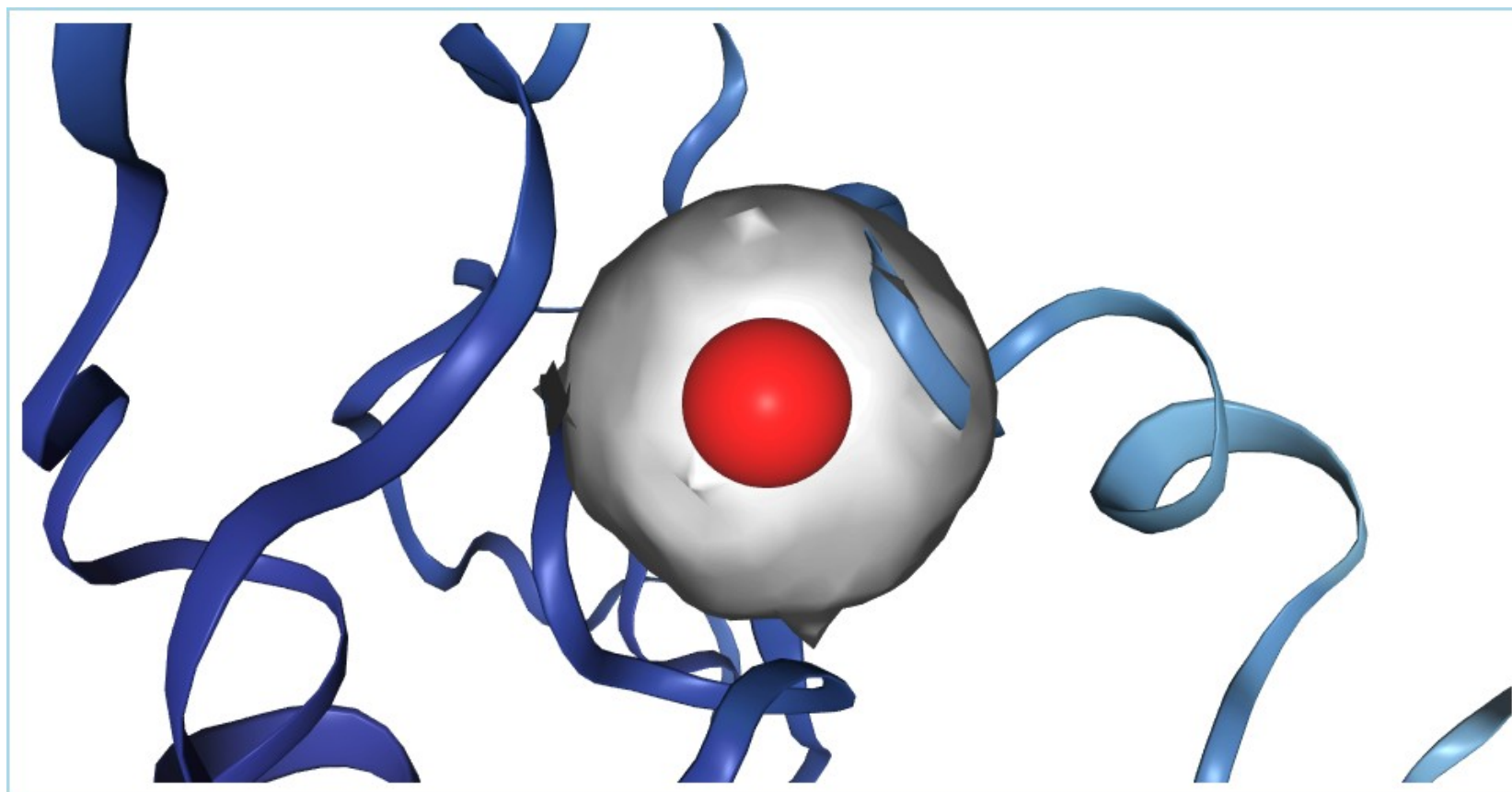


Model confidence (AlphaFold): Very high Confident Low Very low

Predicted ligands

	Name	PDB ID	Chain ID	Z-score	Show in 3D
	spike glycoprotein	7a91	A	4.72	<input checked="" type="checkbox"/>
	spike glycoprotein	6acj	C	4.72	<input type="checkbox"/>
	spike protein s1	7lo4	B	4.72	<input type="checkbox"/>


Binding sites for water and ions



Model confidence (AlphaFold):



ProBiS-Fold web server

**ProBiS-Fold**
Binding sites for AlphaFold.

Tutorial Datasets Cite

AlphaFold ID, UniProt ID, PDB ID, Chain ID, Molecule, Binding Site Type,...

Search

Examples: [Conserved water](#) [Metal ions](#) [Highly druggable](#) [Protein](#) [Peptide](#) [Nucleic](#) [Glycan](#)
[Not in PDB](#) [High confidence](#) [Substrate competitive](#) [Cofactor competitive](#)

ProBiS-Fold annotates AlphaFold human protein database with

- Binding sites for: compounds (small molecules), cofactors, proteins, peptides, nucleic acids, metal ion and conserved water
- Post-translational modification sites (glycosylation sites)
- Predicted ligands and glycosides for each binding site (3D structures as bound to protein)

ProBiS-Fold aims to

- Provide interactive, downloadable binding sites for human proteome for functional and drug discovery studies
- Enable human proteome-wide structure-based virtual screening and selectivity prediction

Binding sites and post-translational sites types

- Compound (substrate/agonist-competitive ligands), cofactor (cofactor and cofactor-competitive ligands) (based on list of [known cofactors](#)), protein (both <20 aa. and >=20 aa.), peptide (<20 aa.), nucleic acid (DNA or RNA molecules), metal ion (structurally conserved) and water (structurally conserved)
- Glycosylation sites (O- and N- glycosylation)
- Ranked according to the estimated druggability score (applies to compound and cofactor sites)

Input

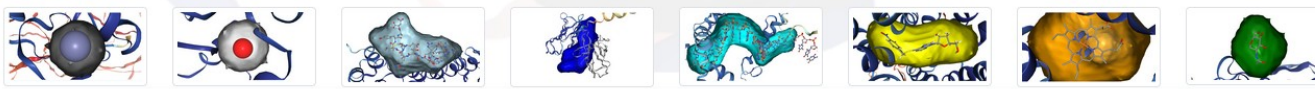
- AlphaFold ID, UniProt ID, PDB ID and Chain ID (where available)
- Protein name
- Protein function, such as, protein kinases or cancer-related proteins
- Binding site type and binding site rank
- See [tutorial](#) for more query options

Output

- Centroids (x,y,z,radius) that accurately describe the often convoluted binding site shapes
- Binding site protein residues that interact with ligands
- Predicted ligands obtained using structure-based comparative [ProBiS approach](#) from similar binding sites in the PDB
- Binding site bounding box (in AutoDock Vina format) ready for docking
- Receptor, an AlphaFold2 predicted protein single chain structure

Download binding sites as

- Individual or multiple selected binding sites based on user query (see [tutorial](#) for how to efficiently use the search bar on top of the page)
- Prepared binding site [datasets](#)



Developed by [Insilab](#) in 2022.

ProBiS-Fold database



ProBiS-Fold

Binding sites for AlphaFold.









[Tutorial](#) [Datasets](#) [Cite](#)

AlphaFold ID, UniProt ID, PDB ID, Chain ID, Molecule, Binding Site Type,...

[Search](#)

Examples: [Conserved water](#) [Metal ions](#) [Highly druggable](#) [Protein](#) [Peptide](#) [Nucleic](#) [Glycan](#)
[Not in PDB](#) [High confidence](#) [Substrate competitive](#) [Cofactor competitive](#)

Download binding sites datasets

Binding site type	Ligands	Dataset
 Protein	Other proteins and peptides, i.e., oligo- and polypeptides of any amino acids chain length	Download
 Nucleic	Nucleic acids (DNA or RNA)	Download
 Peptide	Peptides, i.e., oligopeptides with less than 20 amino acids in length	Download
 Compound	Small molecule drugs, agonists, substrates, substrate-competitive inhibitors but not cofactors	Download
 Cofactor	Cofactors and ligands that overlap with cofactors, e.g., cofactor-competitive inhibitors	Download
 Glycan	Covalently attached O- and N-glycans	Download
 Conserved water	Conserved water molecules, i.e., those found in more than 10 PDB structures (num_occ > 10) at the same location and having high conservation score (cons > 0.6)	Download
 Metal ion	Biologically relevant metal ions, i.e., those found in more than 10 PDB structures at the same location	Download

Developed by [Insilab](#) in 2022.

Exercise

- Exercise 1: LiSiCA – virtual screening based on a known active compound – medicine
- Exercise 2: GenProBiS – screening based on the protein structure

Exercise 1: LiSiCA - virtual screening based on the ligand

- Objective: To predict drug candidates based on known ligands (drugs)
- At <http://insilab.org/lisica>, under the "Download" tab, choose one of the drugs **acyclovir, aspirin, dopamine, paxlovid, remdesivir**
- Download the structure in MOL2 format
- Run the PyMOL program (already installed), then select "LiSiCA" in the "Plugin" menu
- Comparison of the molecule with the "database.mol2" database (approx. 14,000 molecules), you can find it at <http://insilab.org/lisica> under the "Download" tab
- 2D and 3D virtual solving with LiSiCA
- **Which compound among the first 100 has a different scaffold than the reference (scaffold hop) ?**

Exercise 2: GenProBiS - virtual screening based on the structure of the protein

Objective: To predict drug candidates based on the structure of the target protein

in GenProBiS (<http://genprobis.insilab.org>)

**1) What are the types of binding sites on ACE2?
("Table of Ligands" tab)**

**2) Find one known drug among the predicted "Compound" ligands,
which binds to ACE2! What medicine is this?**

3) What disease is it used to treat?

**4) Find the binding site (or sites) for the SARS-CoV "spike protein"! Which ones
amino acids make it up (list at least 3)?**

**5) Which sequence variants on ACE2 can affect SARS-CoV binding
"protein spike"?**

(clicking on "I" in the ligand table opens sequence variants interacting with the ligand)

More at:
<http://insilab.org>