ProBiS for drug development

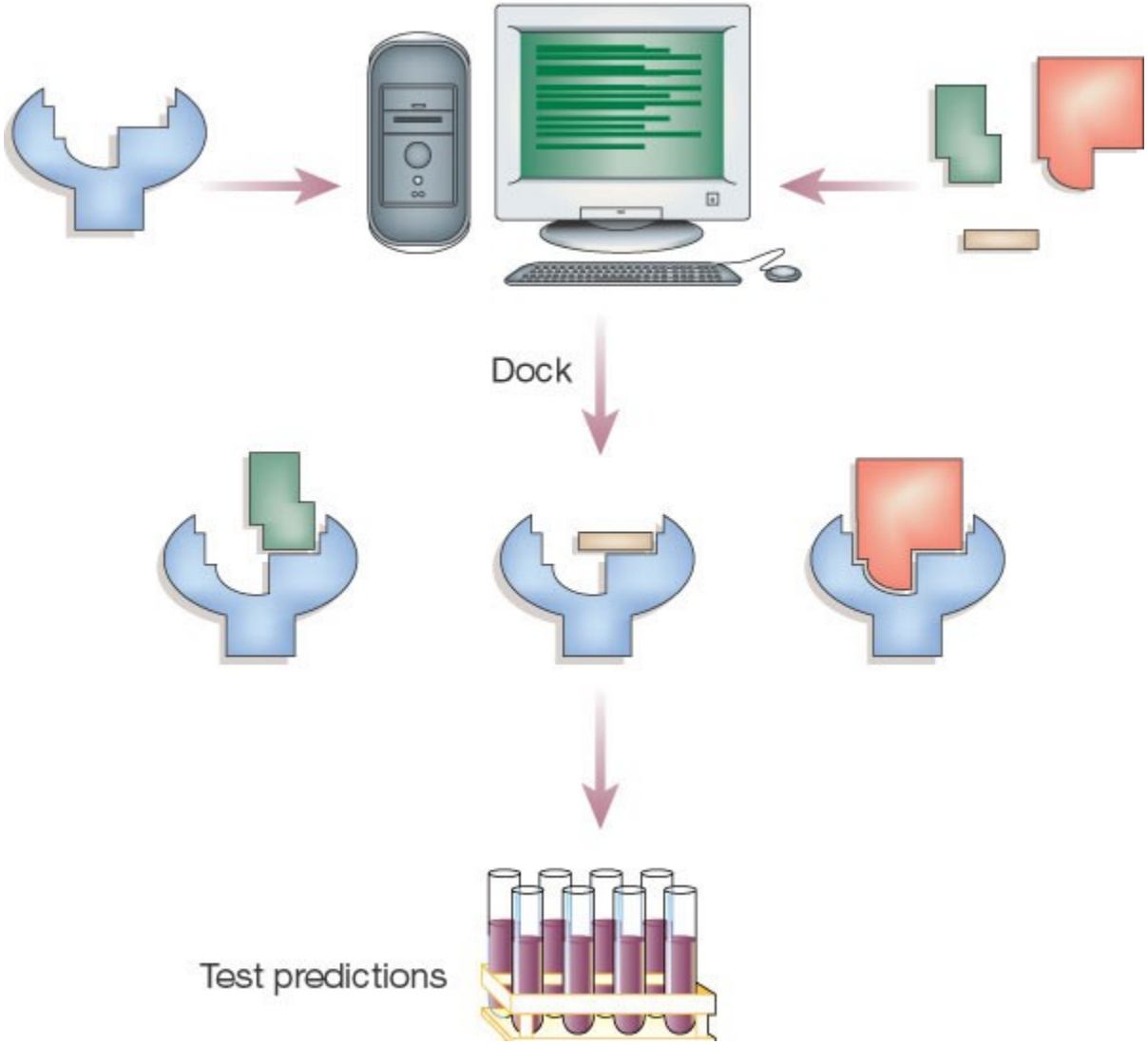
Janez Konc

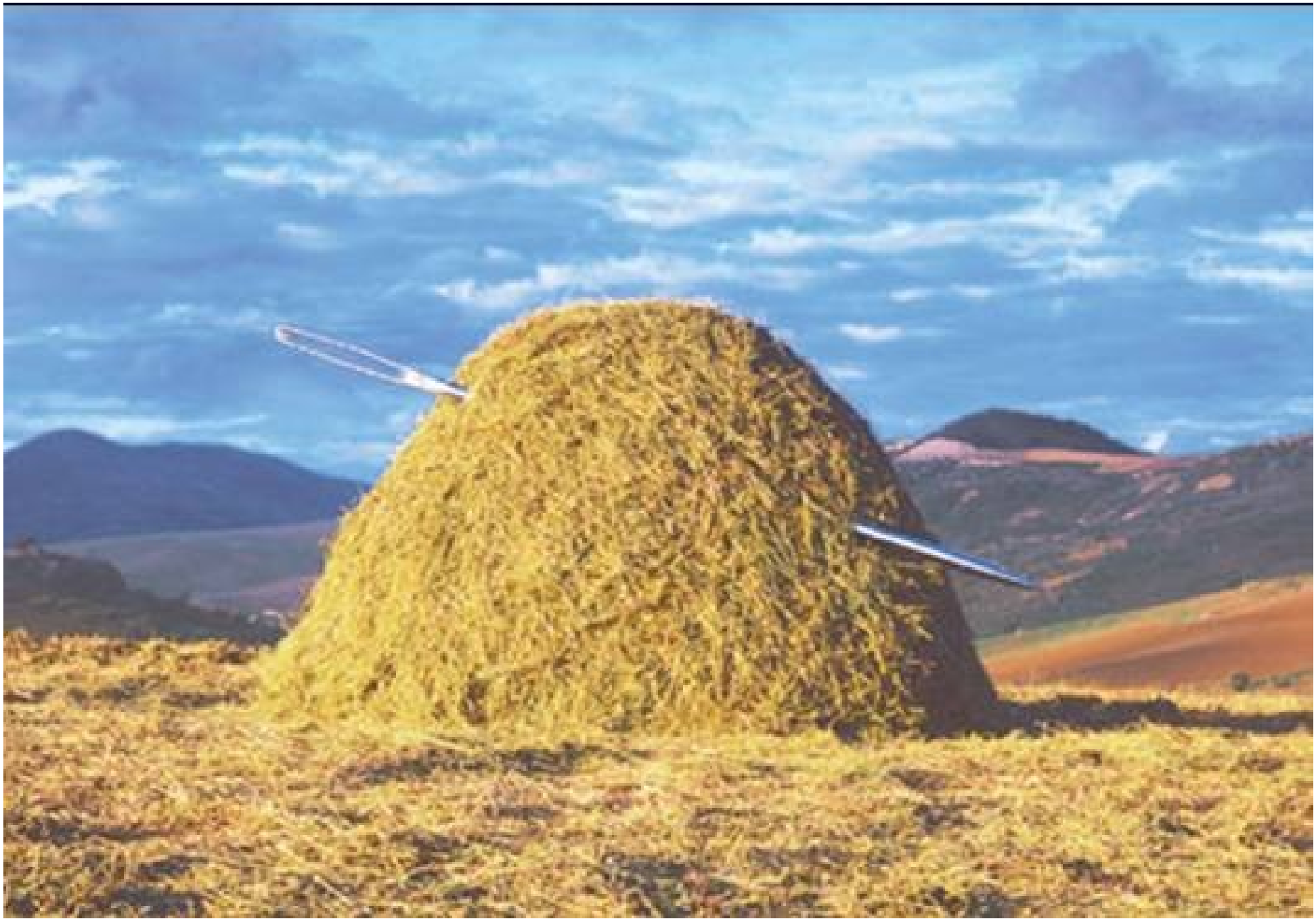
National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana

Ljubljana, 27/02/2024

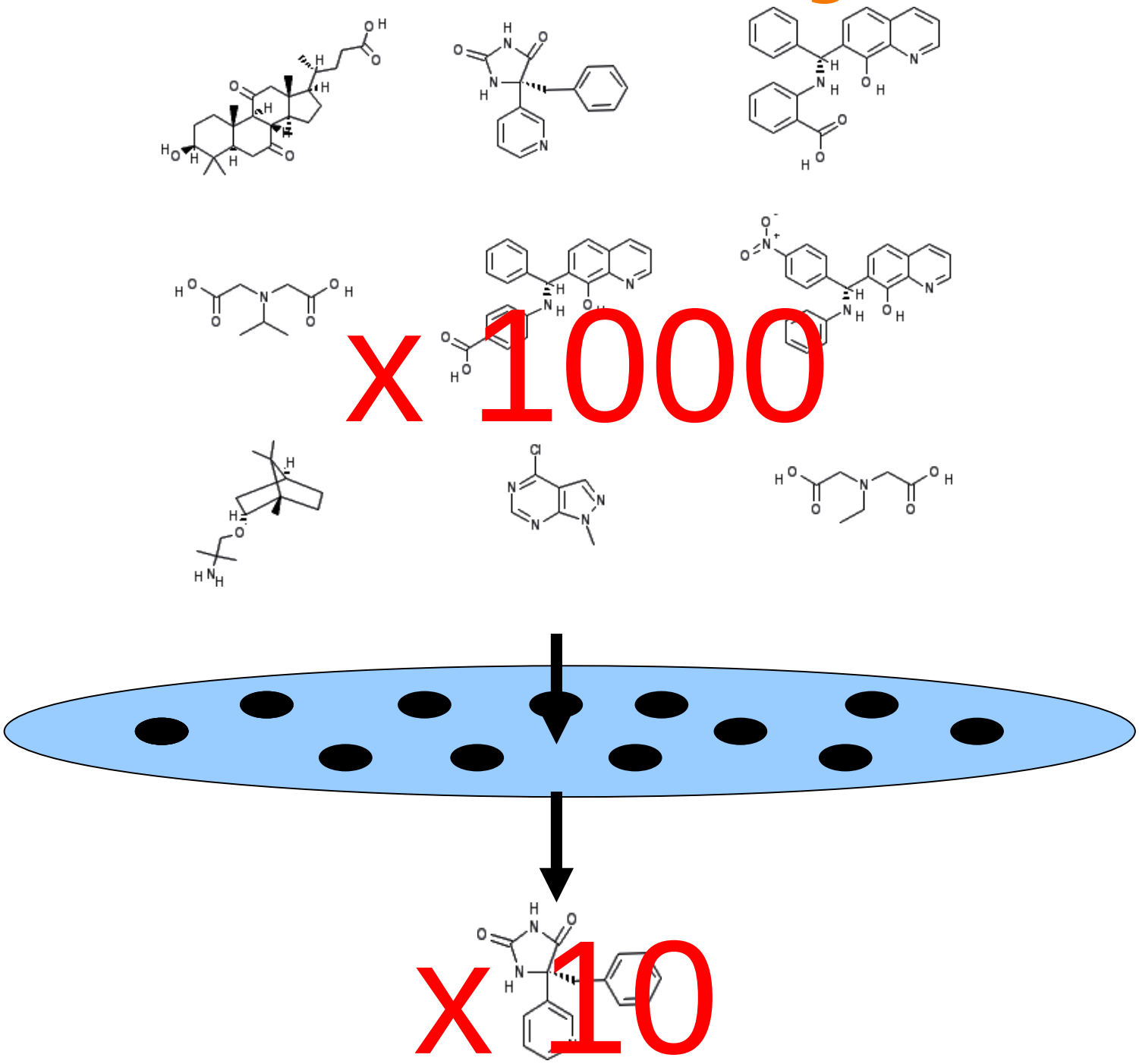
Why do we need a computer in drug development?

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

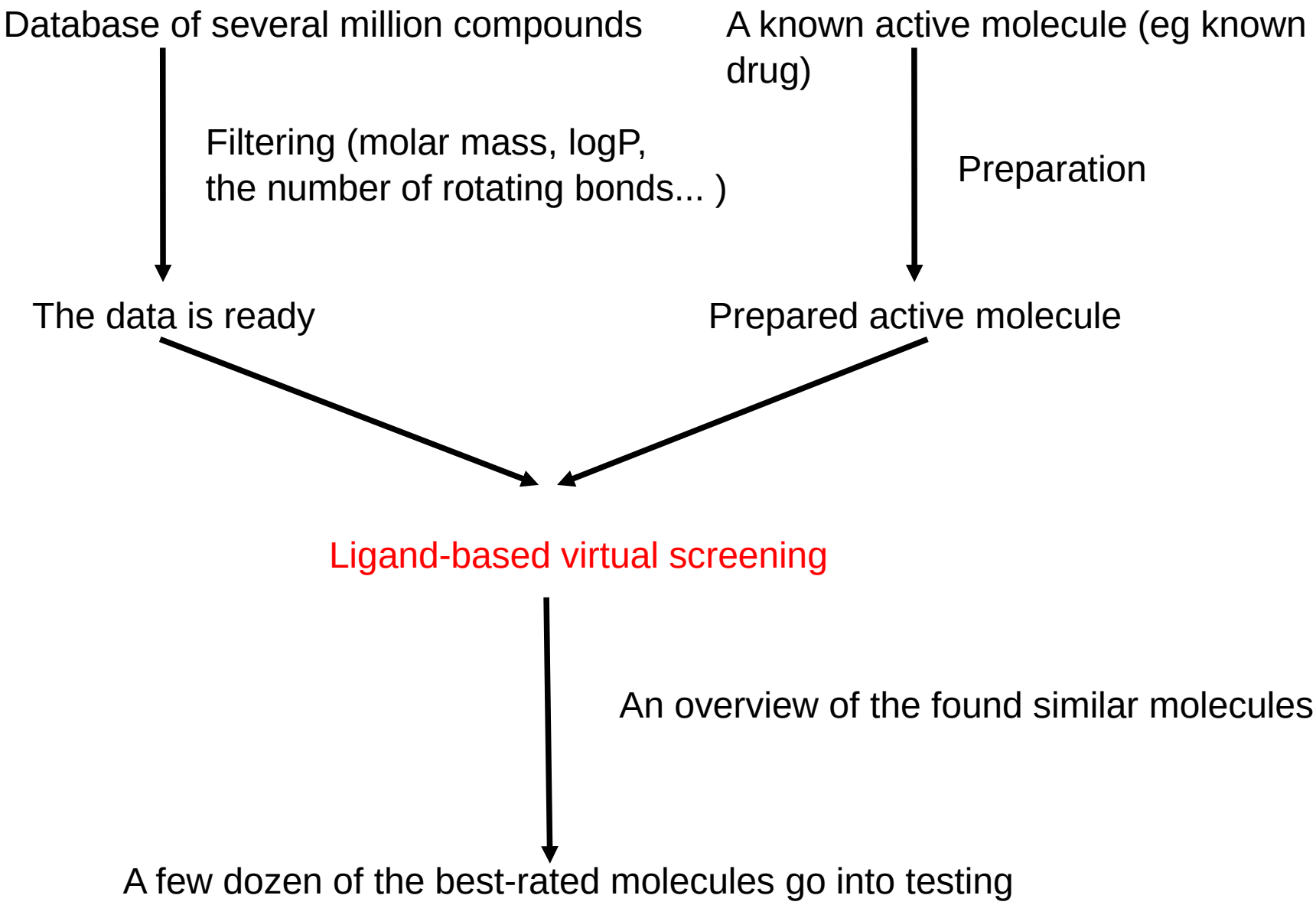




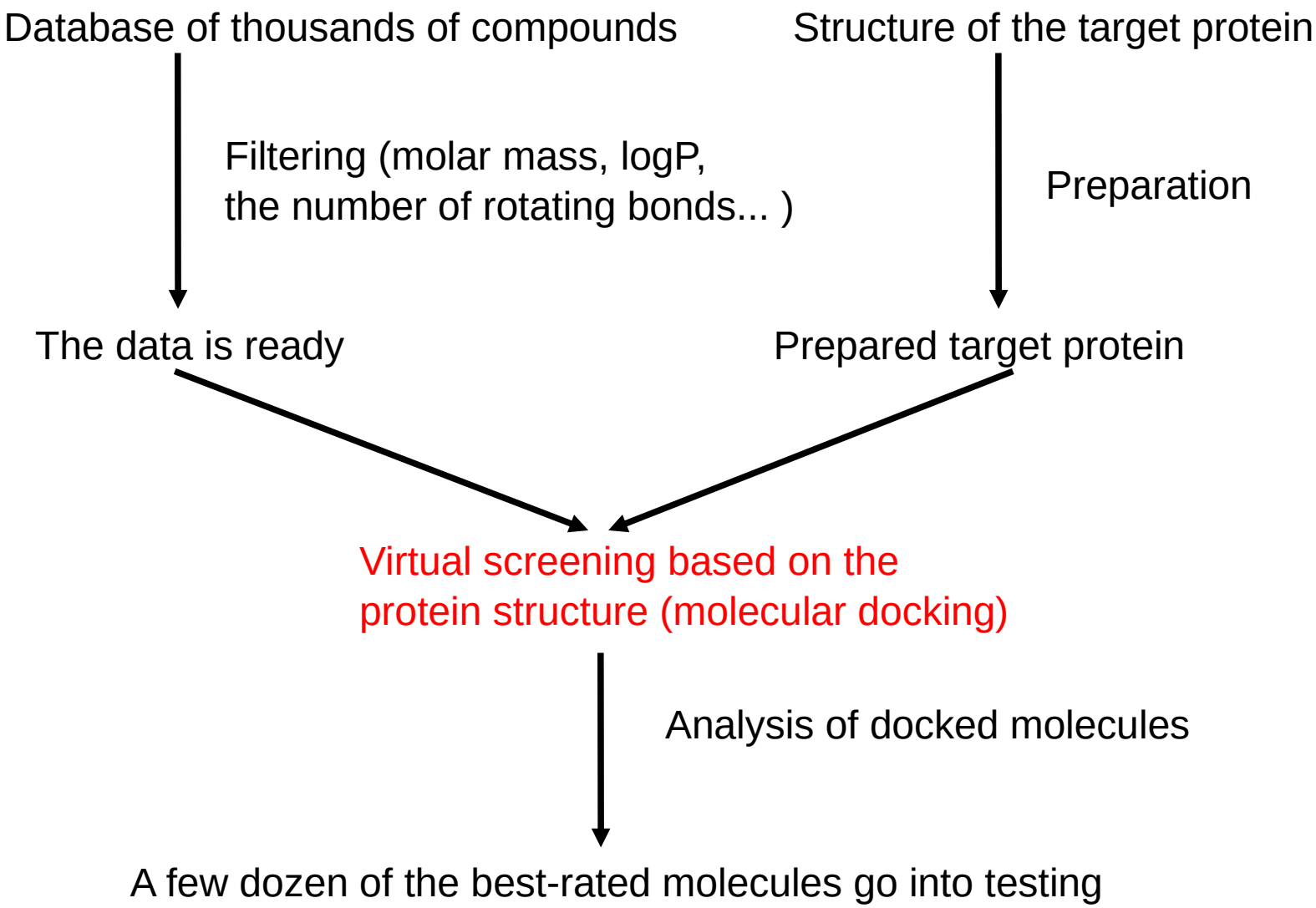
Virtual screening



General scheme I

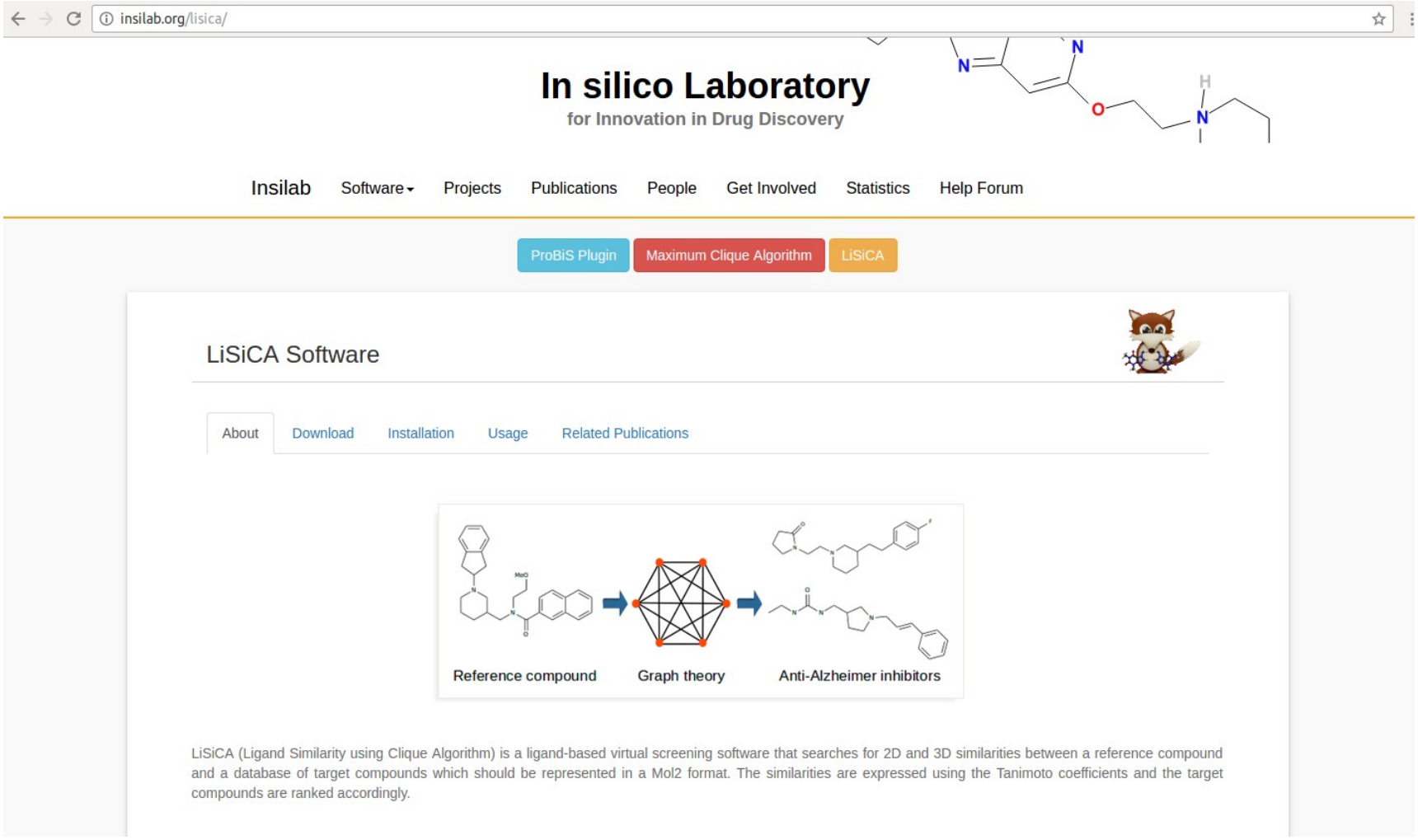


General scheme II



I. Ligand-based virtual screening

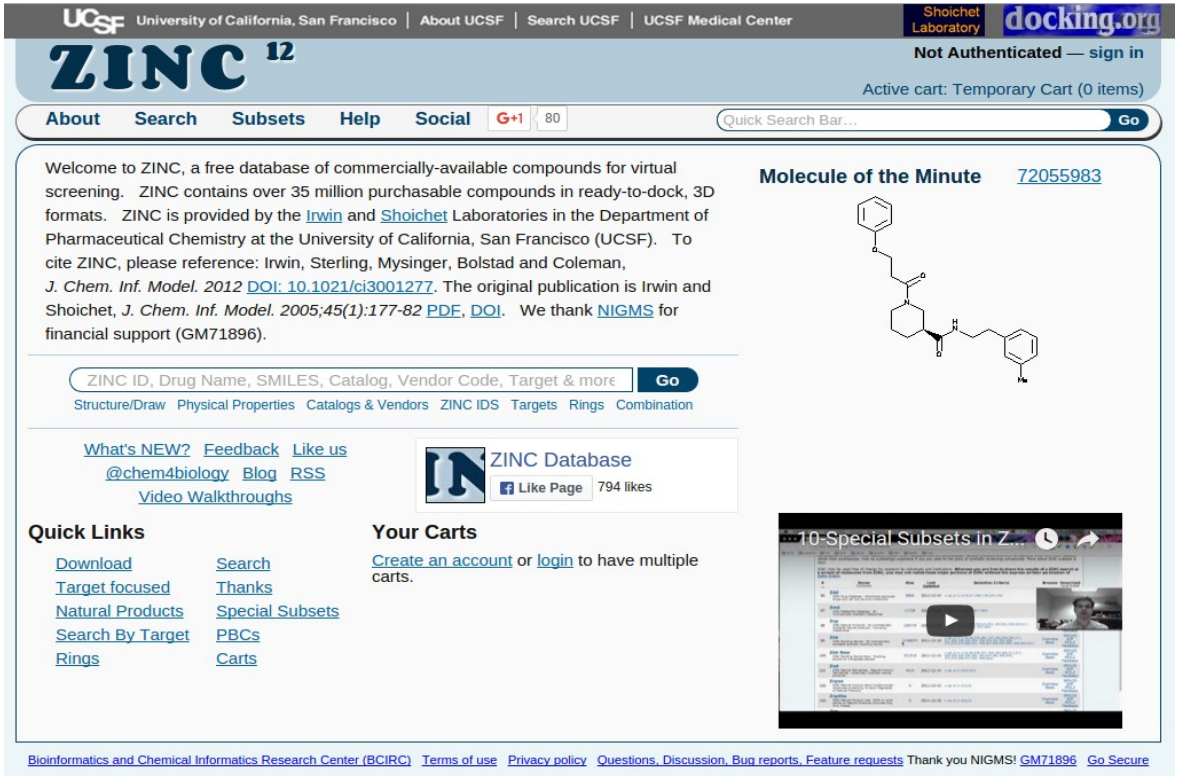
– Compound library
– Known active compound
– Solving algorithm (LiSiCA - Ligand Similarity using Clique Algorithm, http://insilab.org/lisica)



Compound libraries

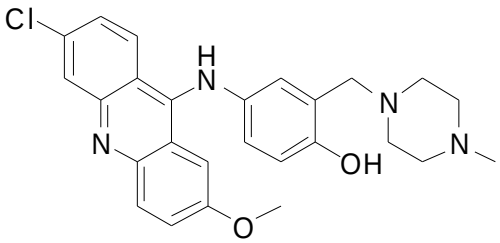
...are commercial and non-commercial databases of 2D and 3D structures of small synthetic and natural molecules.

- 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)

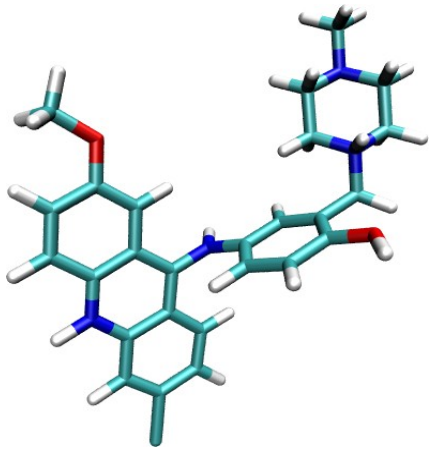


Preparation of small molecules

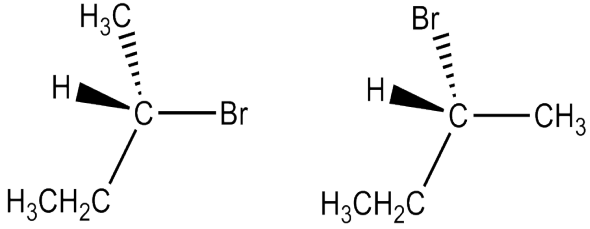
– Filter (MM, number of rotating bonds, functional groups – toxicity)
– 2D or 3D virtual screening?
– In 3D conversion from 2D to 3D structures
– Conformations
– Stereoisomers
– Ionization states



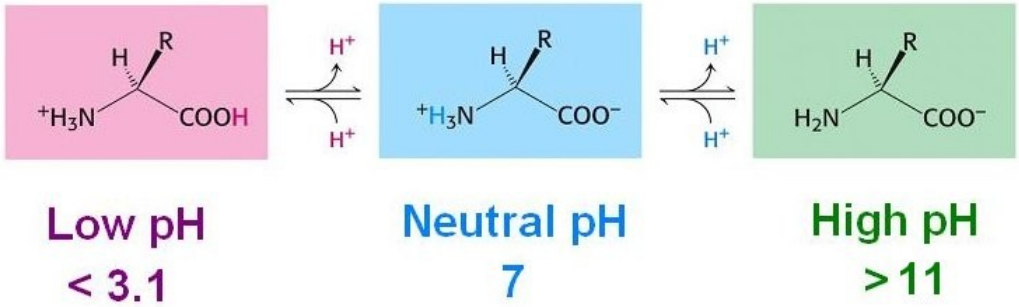
2D structure



3D structure

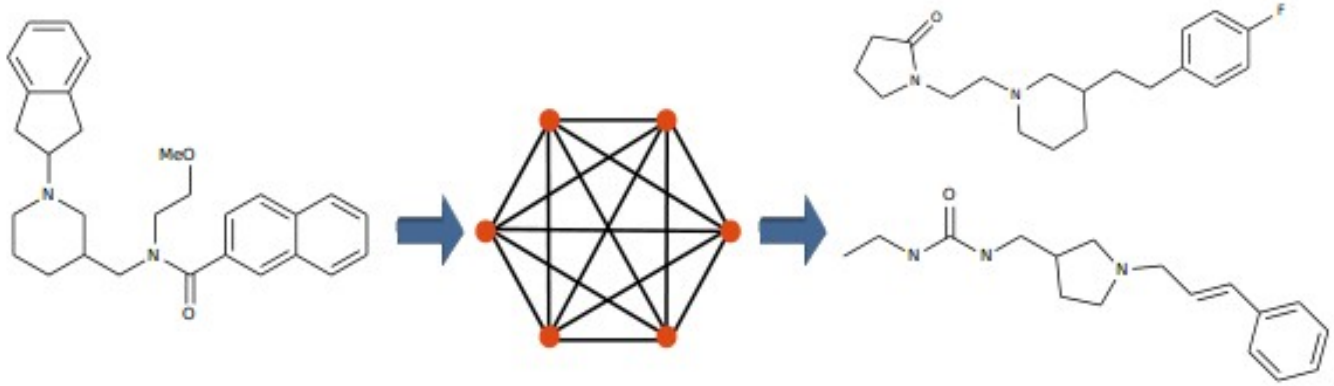


Stereoisomers



Ionization states

LiSiCA – program for ligand-based virtual screening

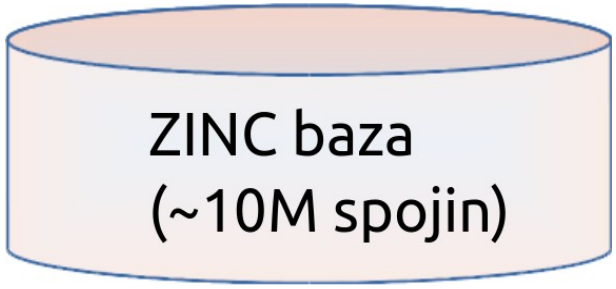


Reference compound (biologically active)

Graph theory

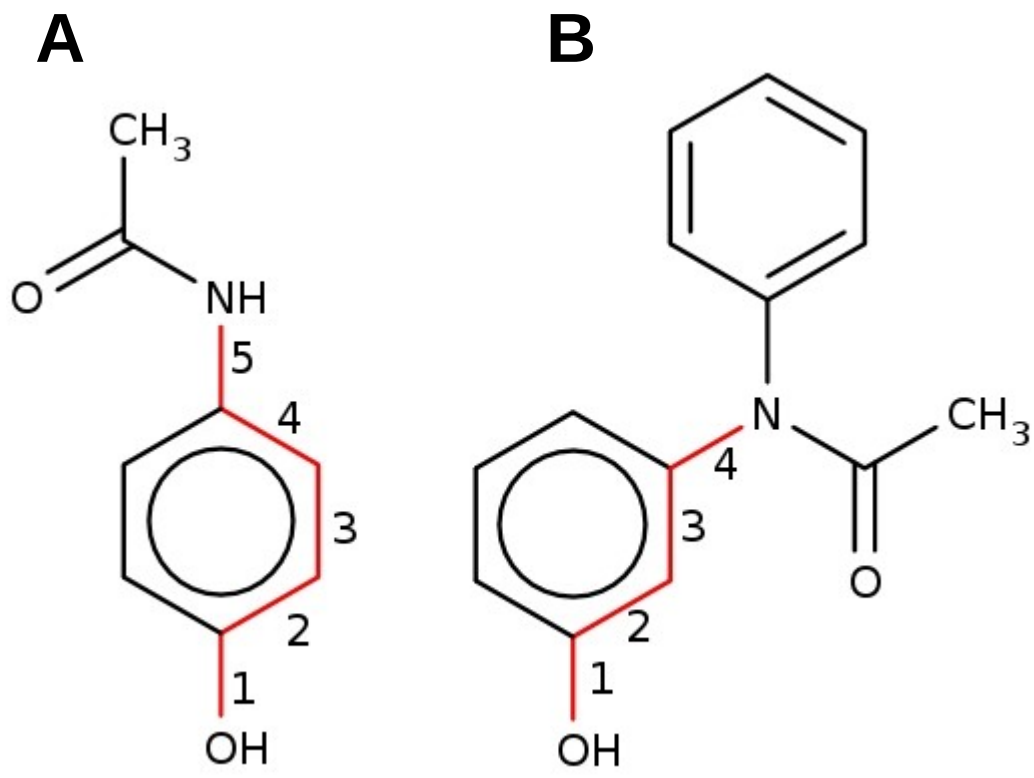
Similar compounds from database (also active)

+



J. Chem. Inf. Model., **2015** , *55* , 1521-1528.
J. Cheminform., **2016** , *8:46* .

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

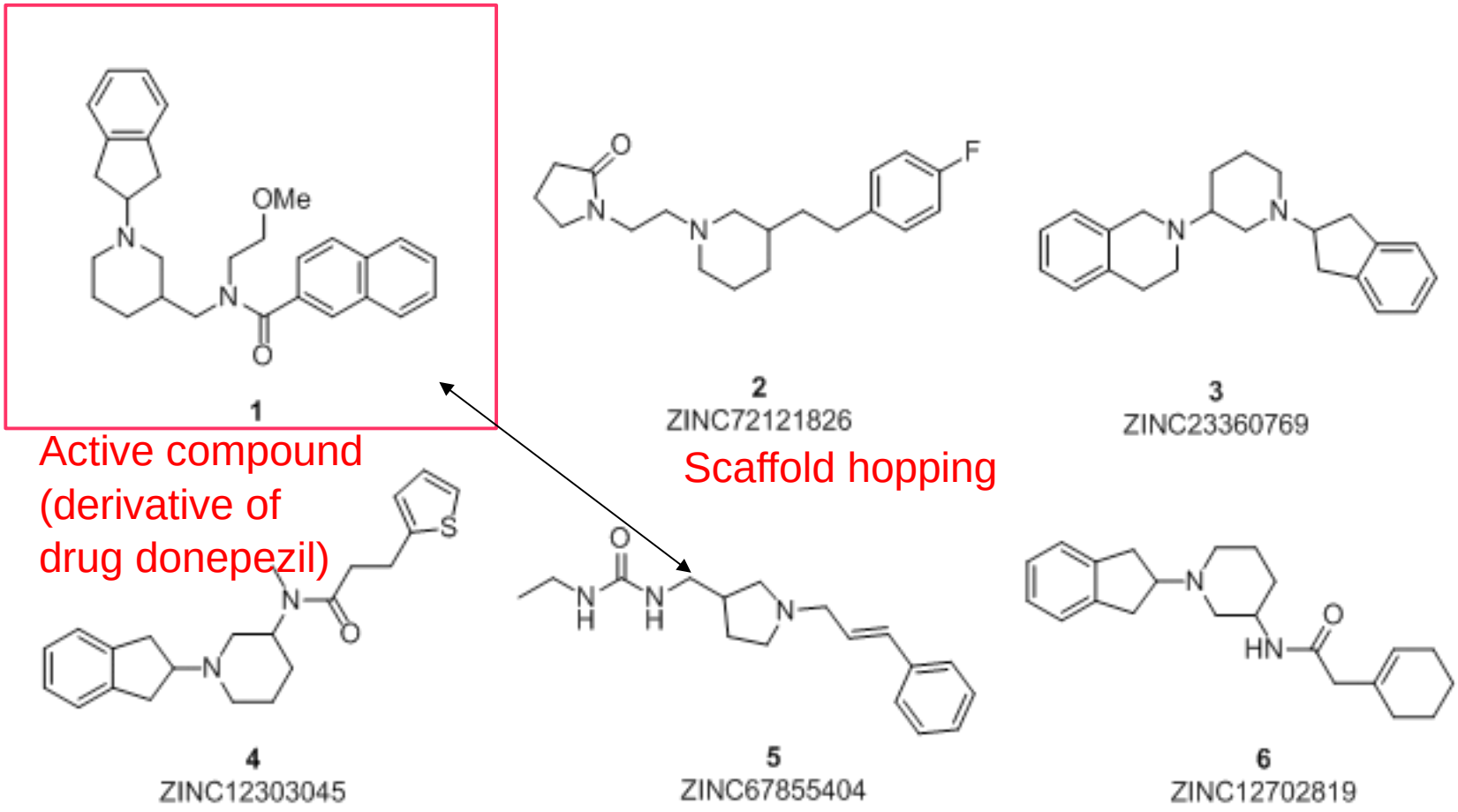


T(AB) = c / (a + b – c) = 11 / 11 + 17 – 11) = 11 / 17 = 0.65

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

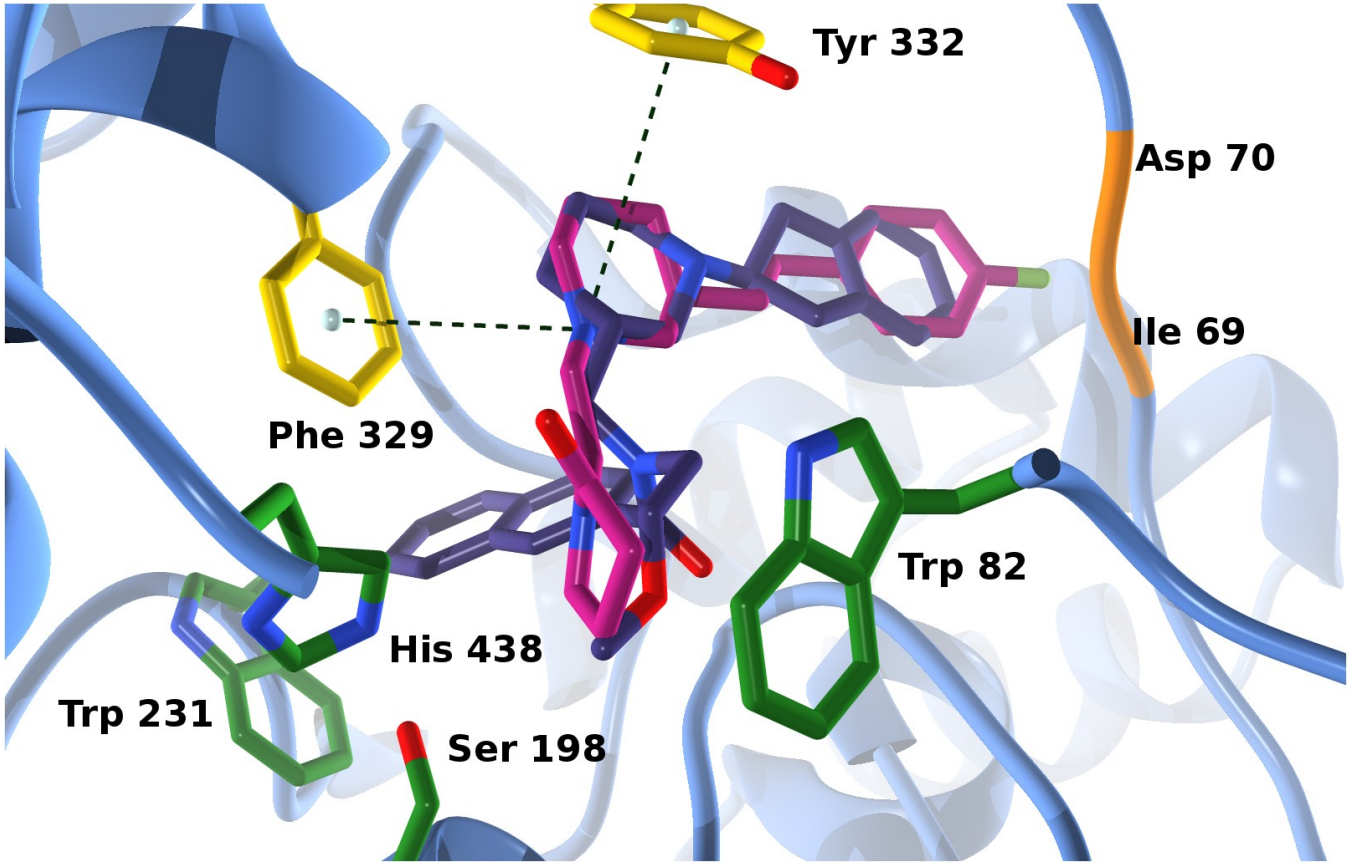
Alzheimer's - screening ~10M compounds

* 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



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

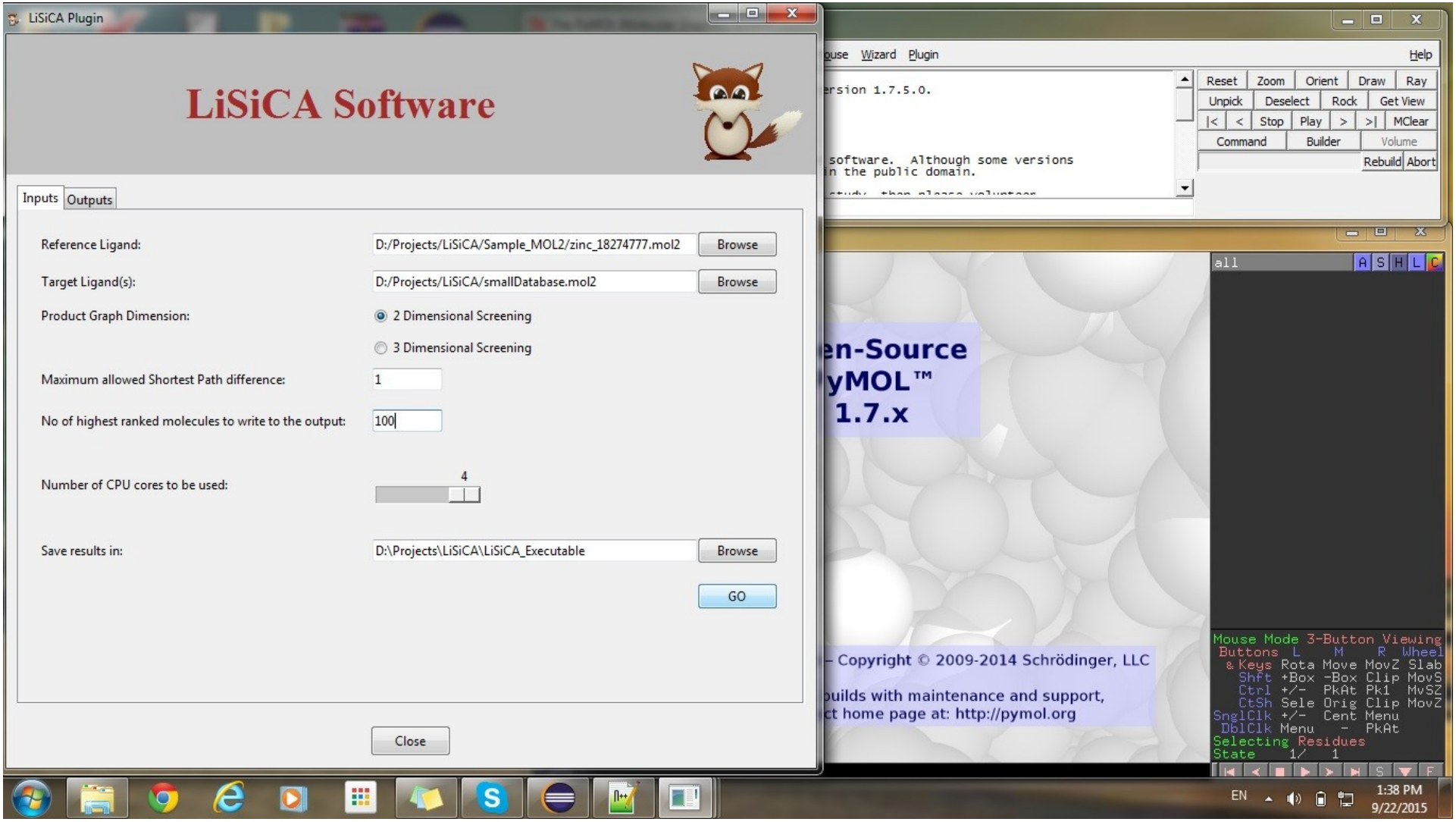
Comparison of known inhibitor: a new inhibitor



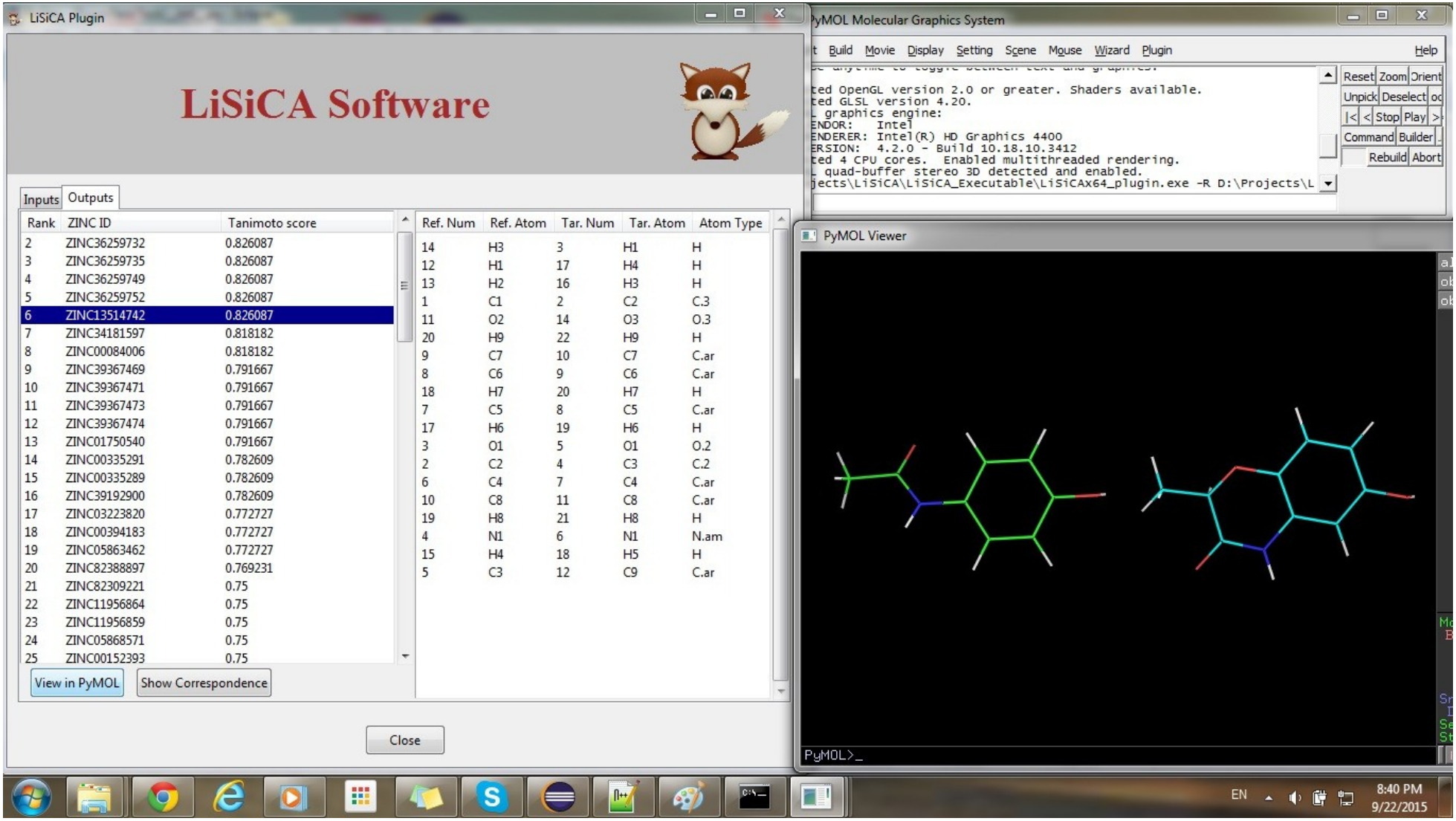
3D overlay:

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

LiSiCA - user interface



LiSiCA - user interface

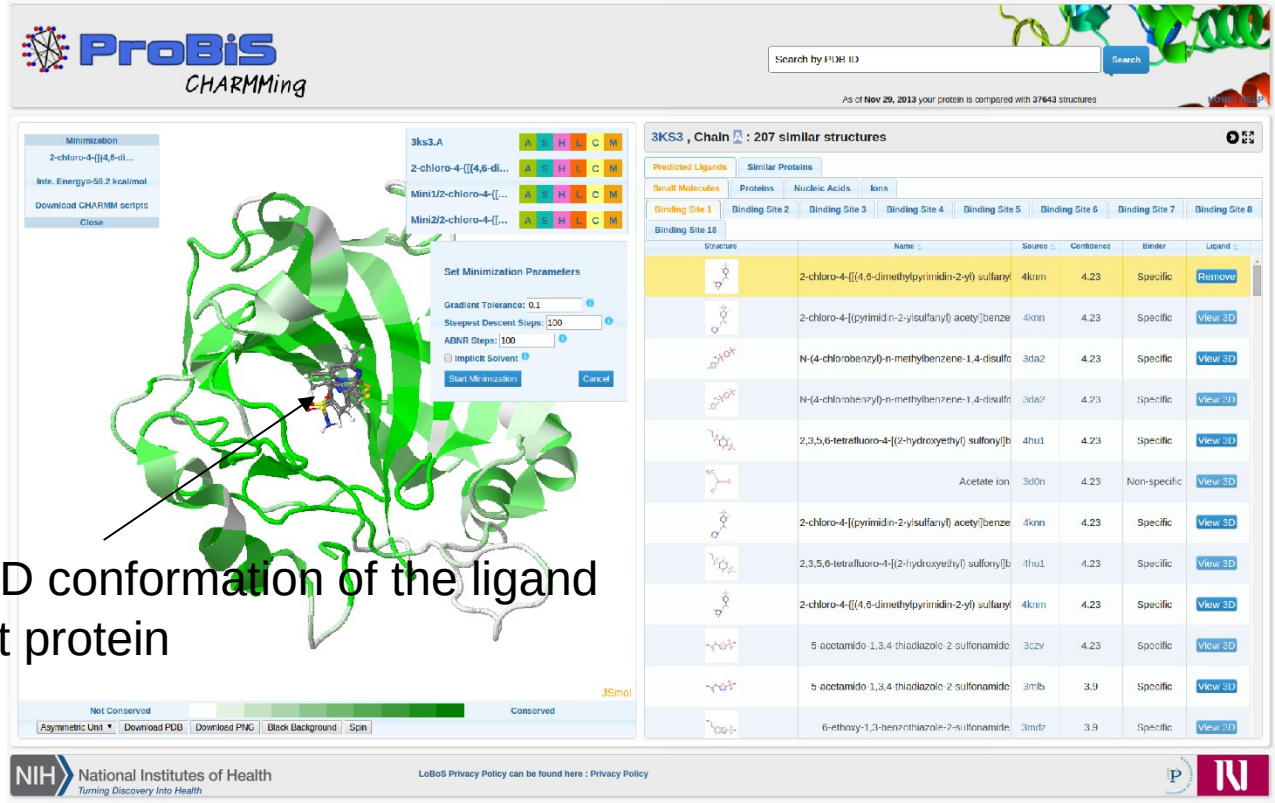


Rank	ZINC ID	Tanimoto score
2	ZINC36259732	0.826087
3	ZINC36259735	0.826087
4	ZINC36259749	0.826087
5	ZINC36259752	0.826087
6	ZINC13514742	0.826087
7	ZINC34181597	0.818182
8	ZINC00084006	0.818182
9	ZINC39367469	0.791667
10	ZINC39367471	0.791667
11	ZINC39367473	0.791667
12	ZINC39367474	0.791667
13	ZINC01750540	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

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

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



Predicted 3D conformation of the ligand in the target protein

J. Chem. Inf. Model., **2015** , *55* , 2308-2314.
Bioinformatics, **2010** , *26* , 1160–1168.

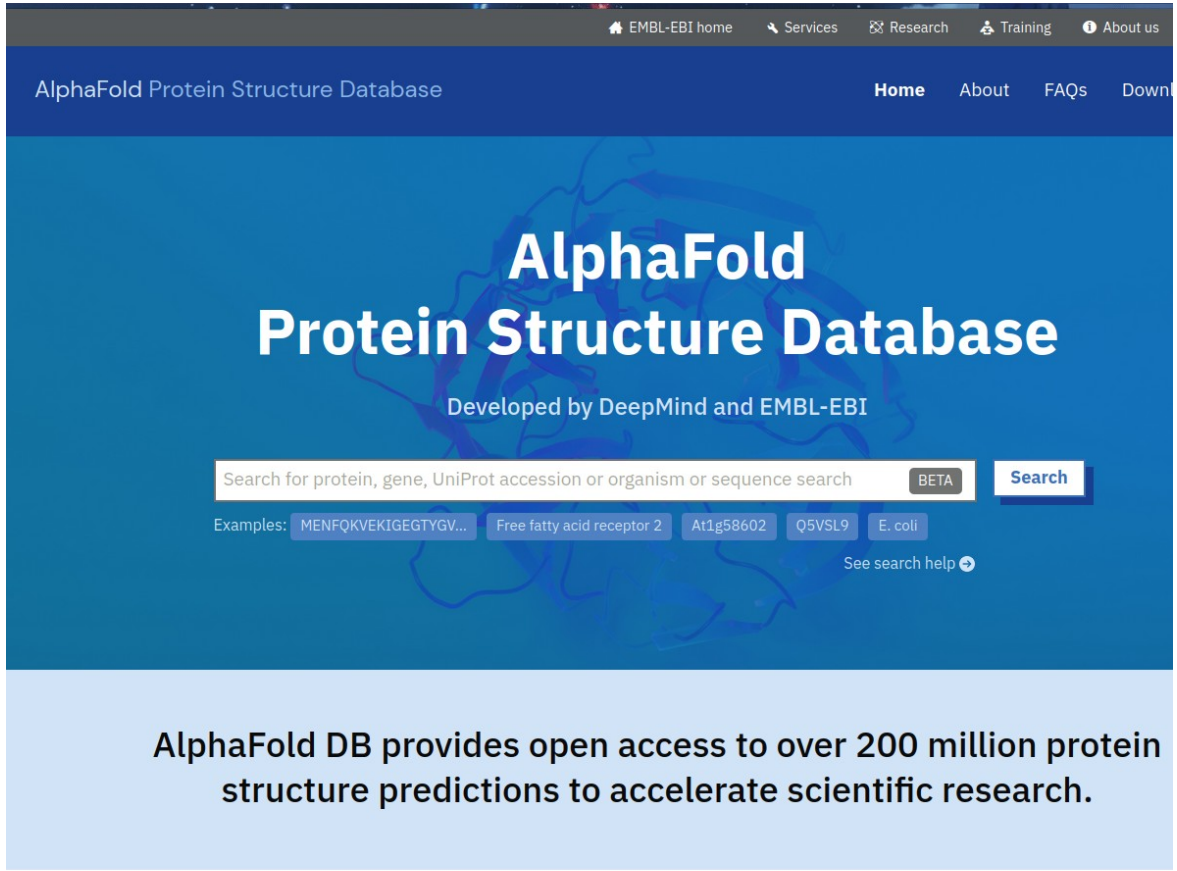
http://probis.cmm.ki.si and https://probis.nih.gov

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

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

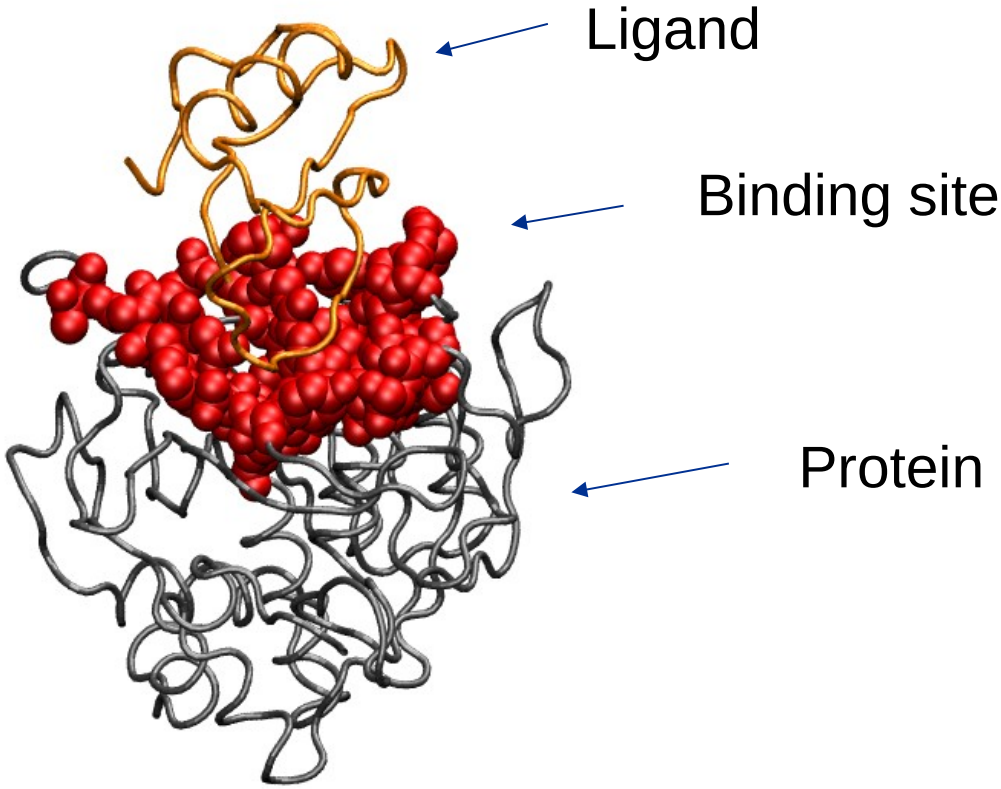


Background

Protein molecular dynamics

Proteins are dynamic: which conformation to use?

Determination of the binding site

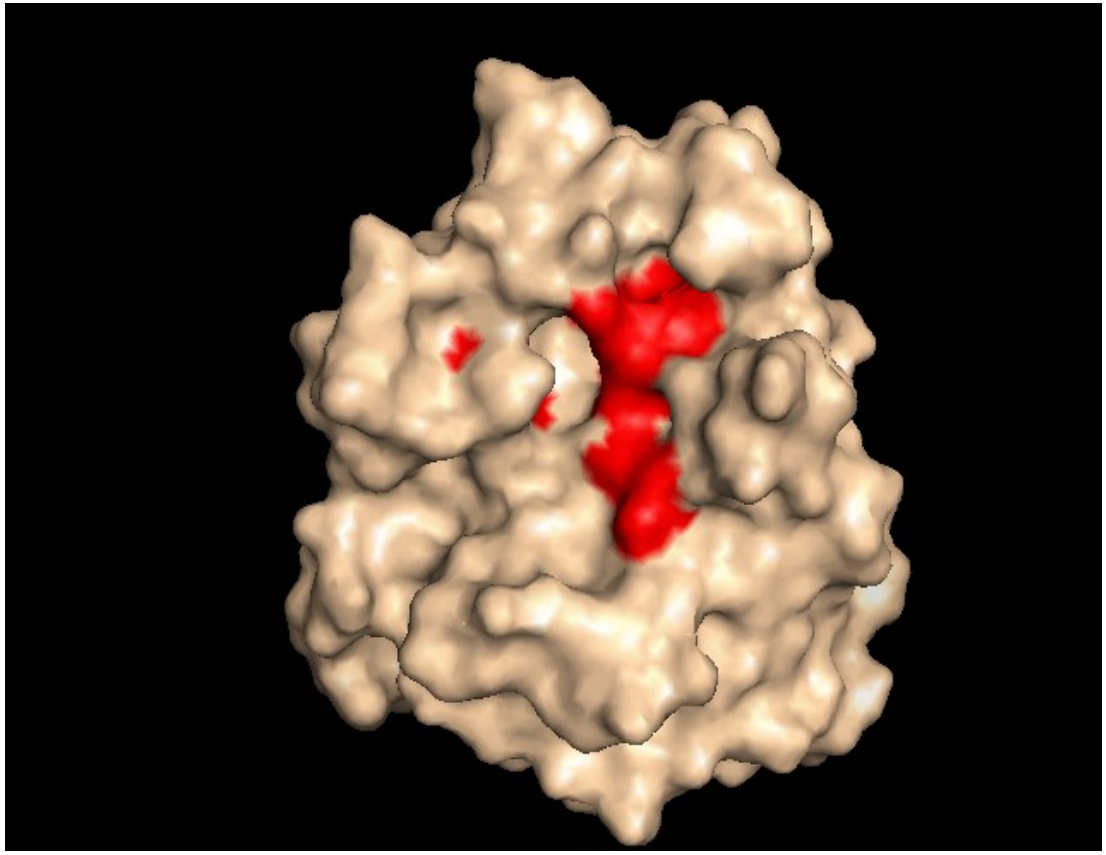


Ligands:

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

Binding site

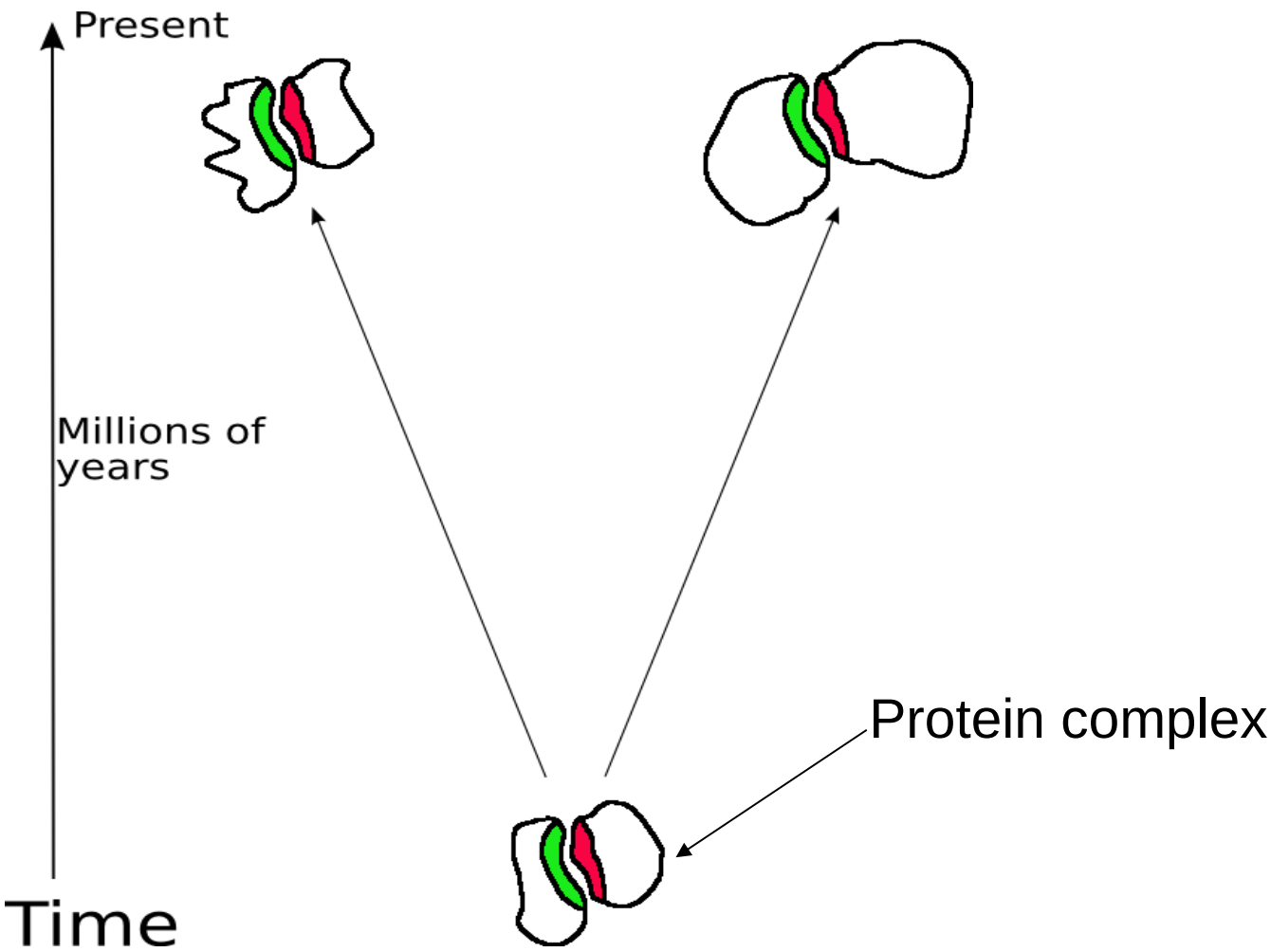
The number of potential target proteins is estimated to be ~6000, but pharma uses only ~200 for development of new drugs.



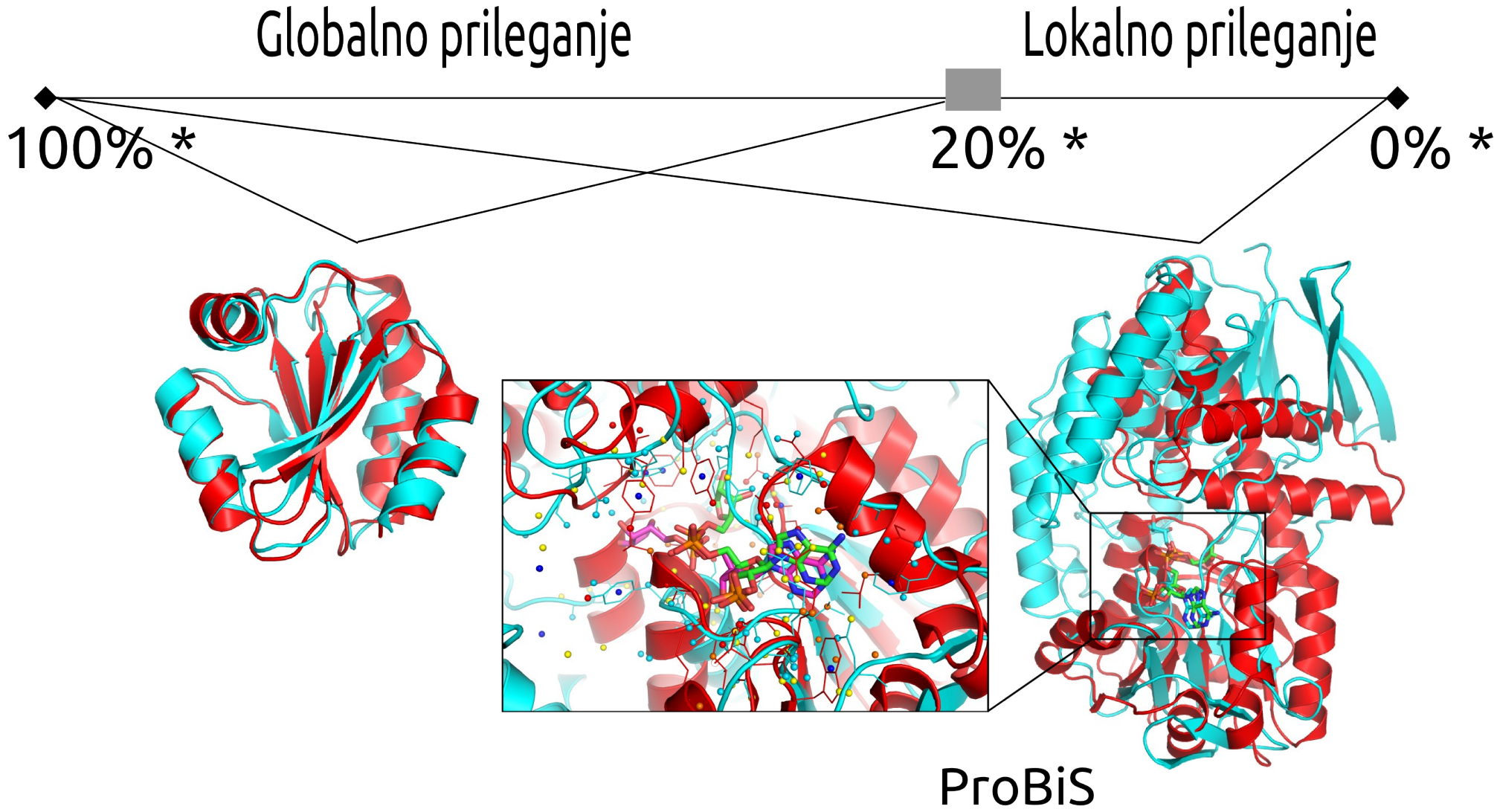
The binding sites for the active substances are in cavities in the surface of the protein.

Evolution of binding sites

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



ProBiS algorithm for structural protein fitting



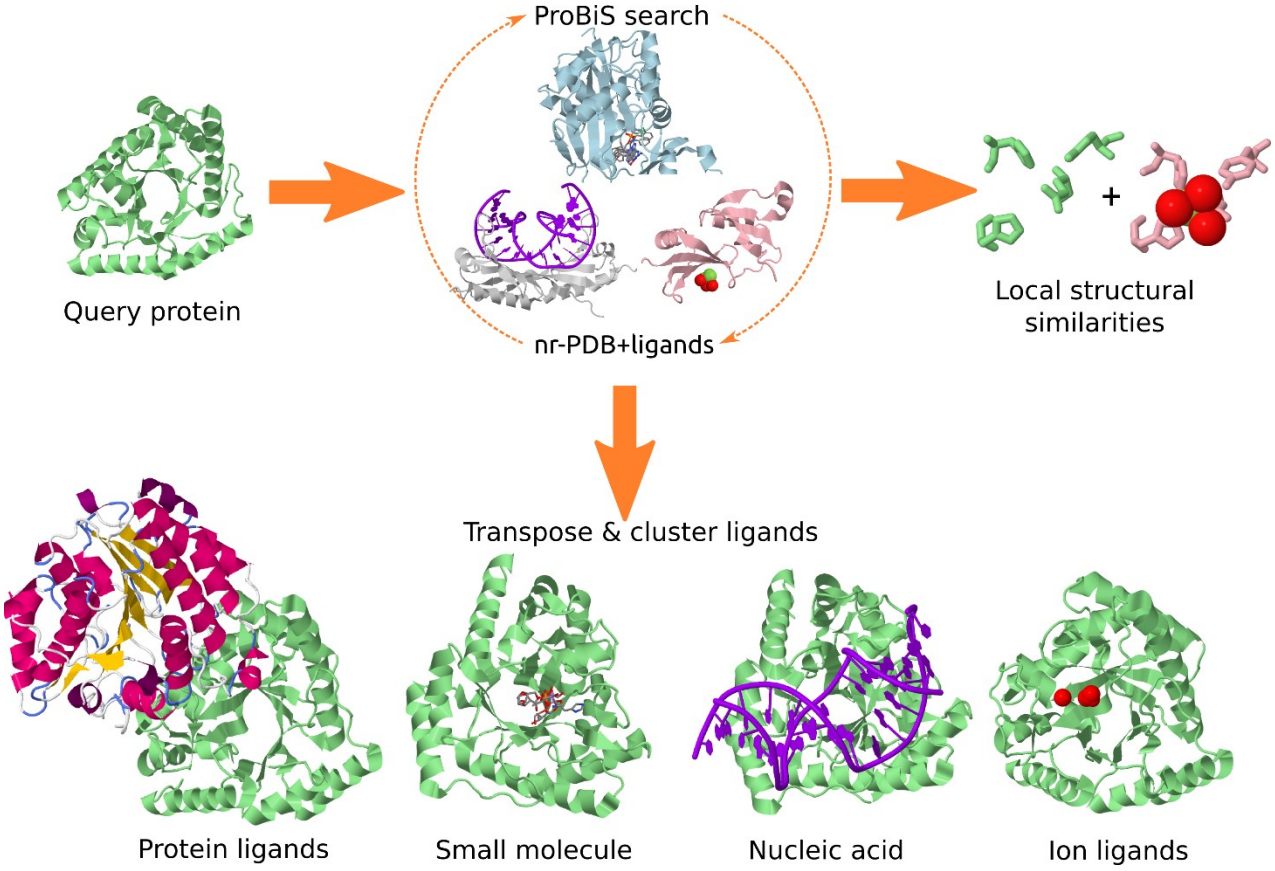
* sequence identity

Bioinformatics, **2010** , *26* , 1160–1168.

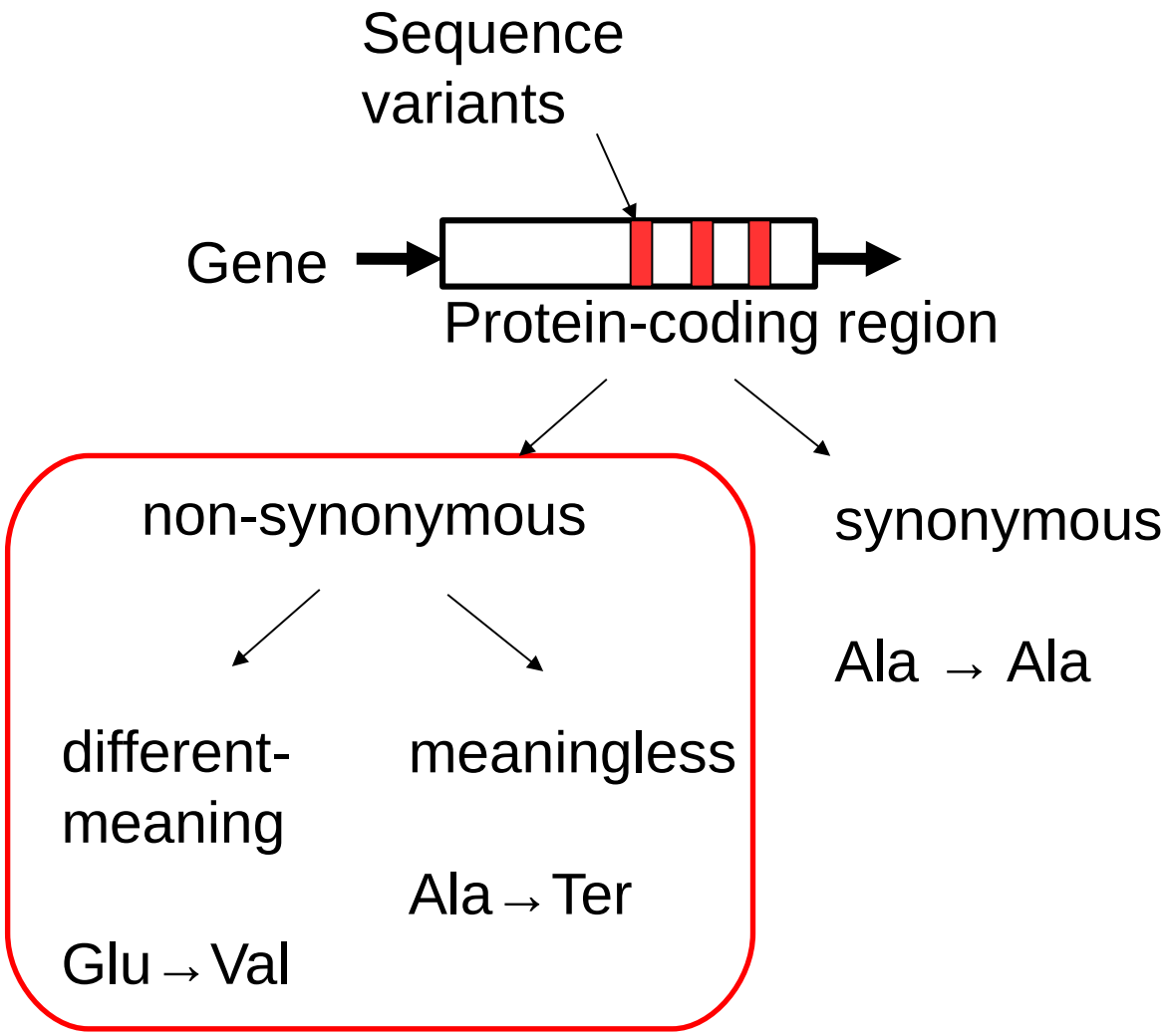
Virtual screening with the ProBiS approach

– 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



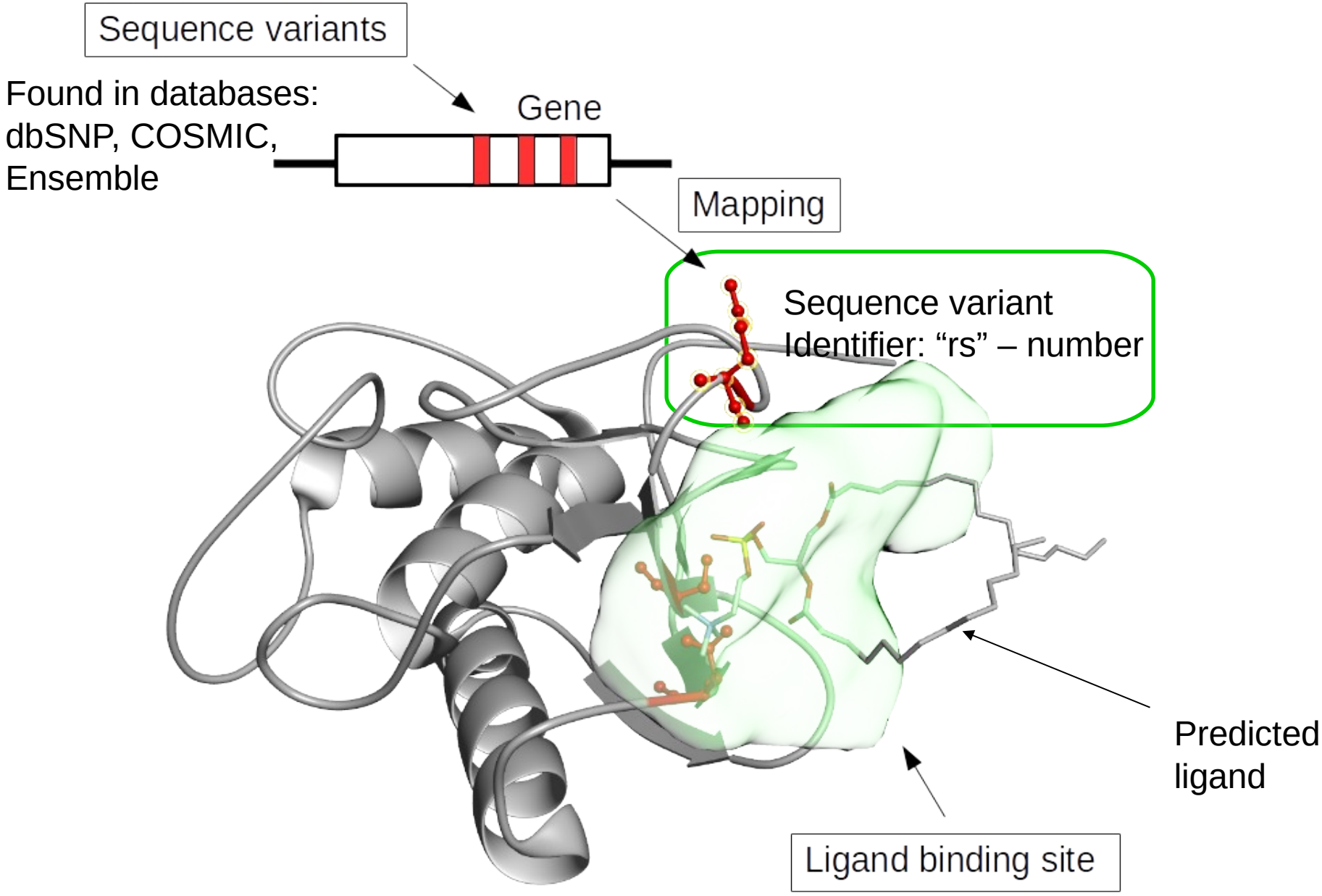
Sequence variants

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

Normal Red Blood Cell Sickle Cell

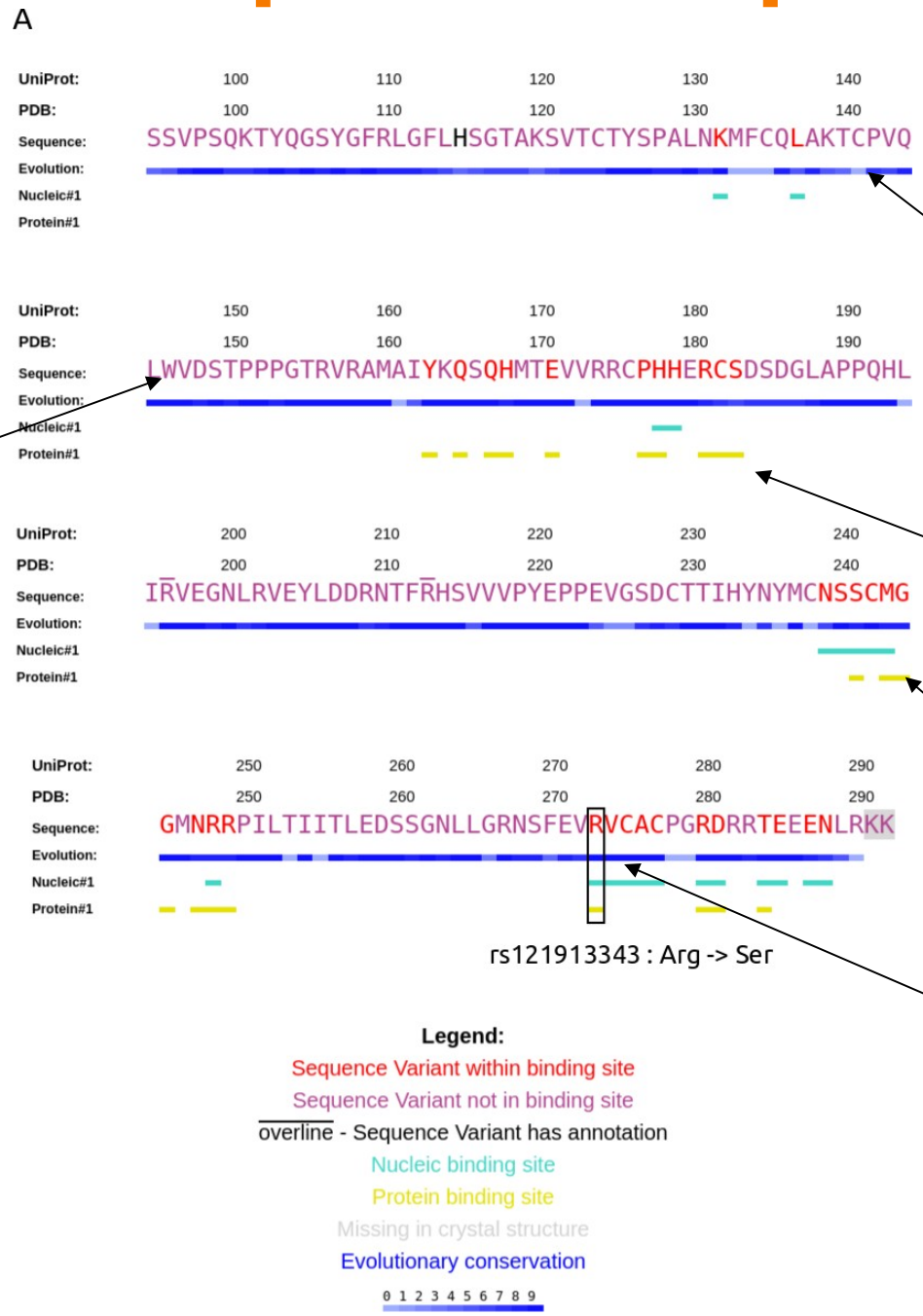
Sickle cell anemia is caused by a hemoglobin variant that has the amino acid valine instead of glutamic acid at position 6 in the sequence.

GenProBiS: mapping sequence variants to binding sites



Nucleic Acids Res., 2017, 45, W253-W259.

Variants and binding sites in the protein sequence



Evolutionary conservation

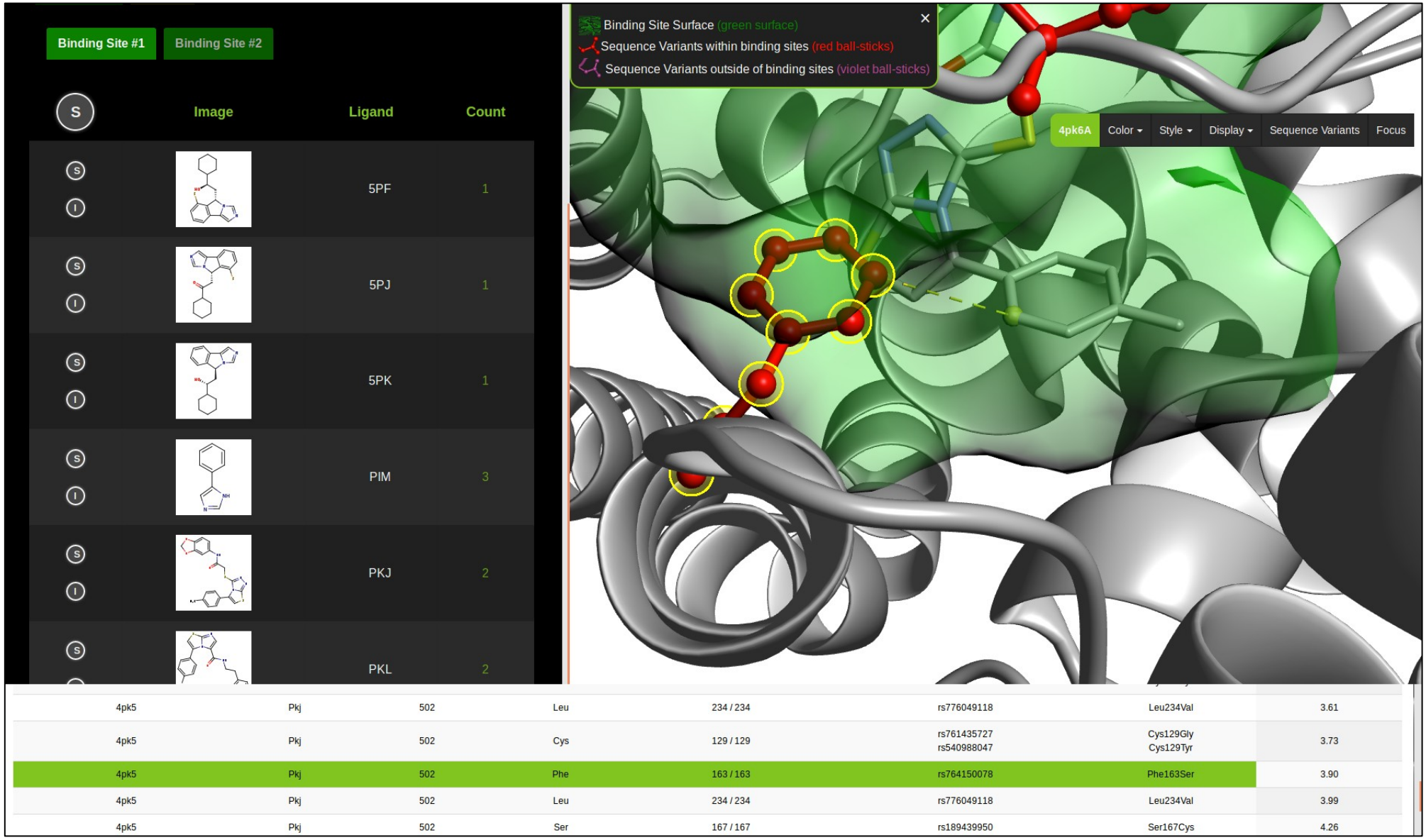
Amino acid sequence of a protein

Protein-protein binding site

Protein-DNA binding site

Variant in two binding sites

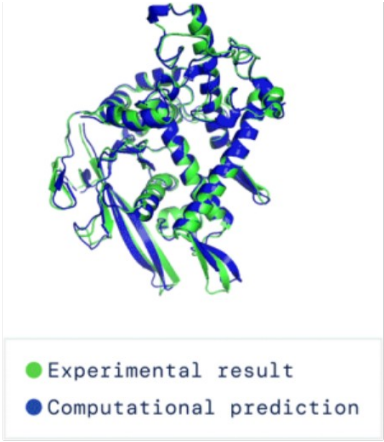
GenProBiS web server



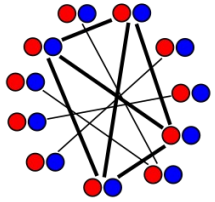
Nucleic Acids Res., 2017, 45, W253-W259.

Prediction of binding sites for AlphaFold structures

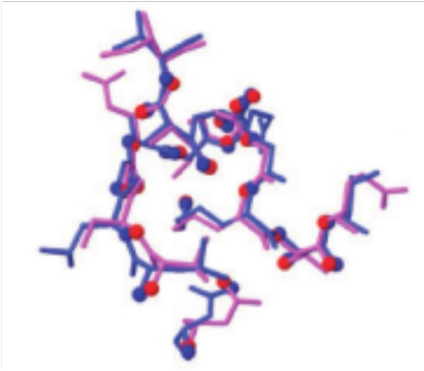
Input: AlphaFold protein structure model (human)



Compare against PDB (with ProBiS algorithm)



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

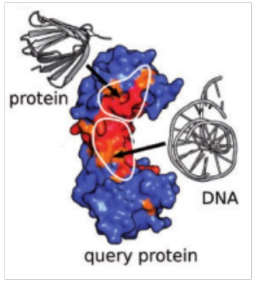
↓

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

↓

Cluster ligands of same type by their proximity

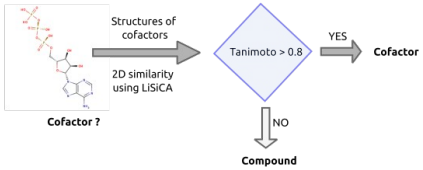
↓



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

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

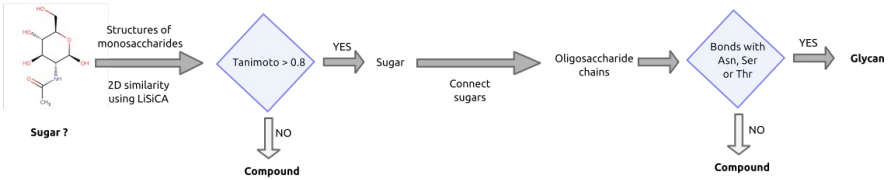
$$R_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

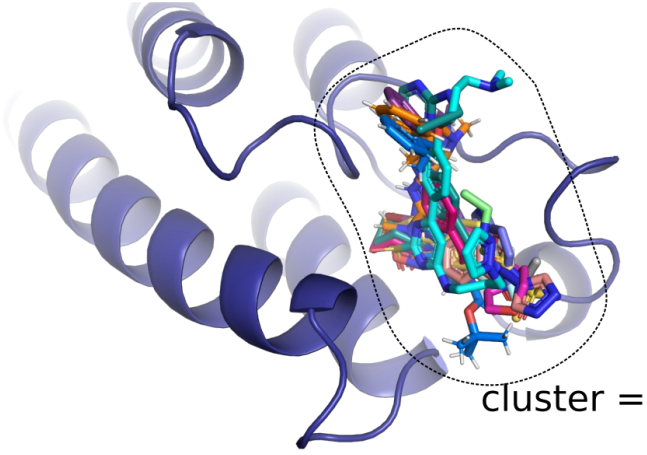


Protein: any amino acid chain
Peptide: < 20 aa.
Nucleic acid: DNA or RNA

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

Water: HOH residue in PDB file





cluster = binding site

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:

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

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

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

Druggability score:

$\mathbf{S_{bsite}} = \max_{1 \le i \le n\_ligands}(S_{cplx,i} + w \times O_{bsite})$

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

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

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

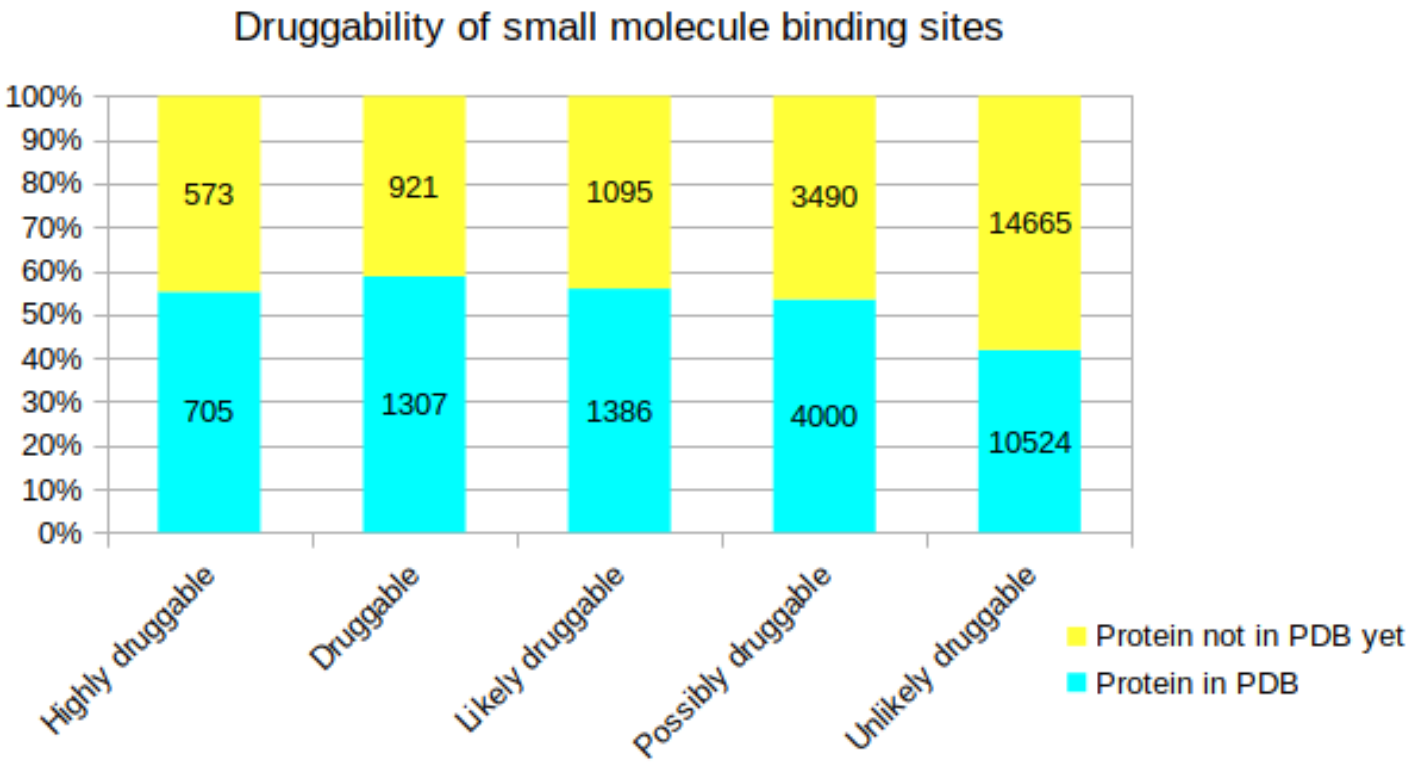
↓

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

Confidence score

$\mathbf{C_{bsite}} = \min_{1 \le i \le n\_bsite\_residues}(\text{AF model confidence}_{residue\_i})$

New binding sites useful for drug development



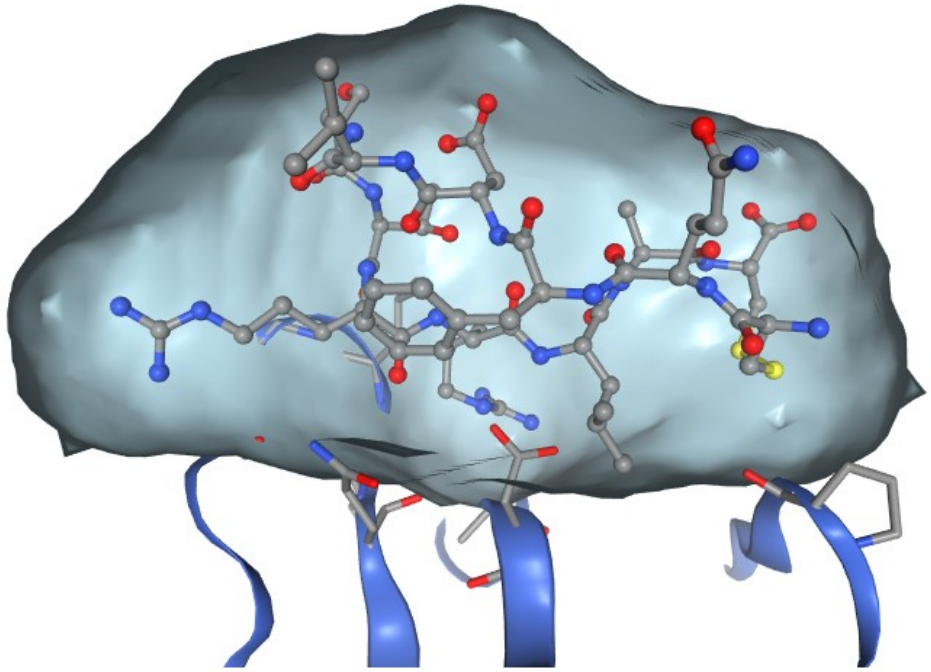
J. Chem. Inf. Model., **2022** , 62, 22, 5821–5829. http://probis-fold.insilab.org/

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

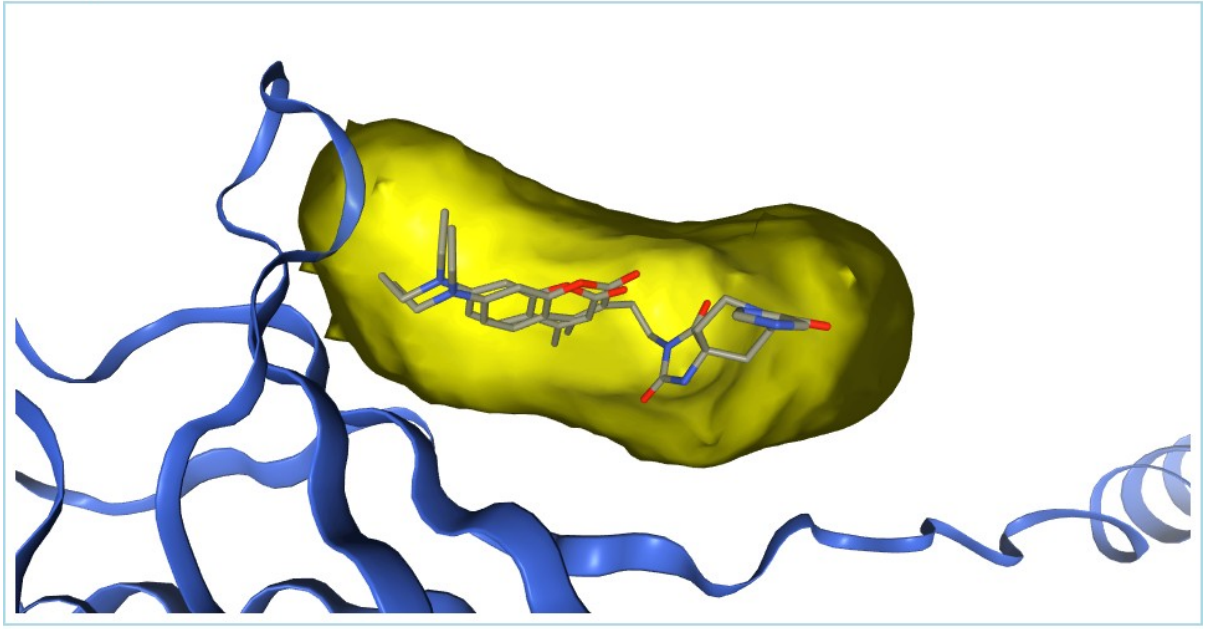
Binding site viewer 3D



J. Chem. Inf. Model., **2022** , 62, 22, 5821–5829. http://probis-fold.insilab.org/

Small molecule binding sites

Binding site viewer 3D

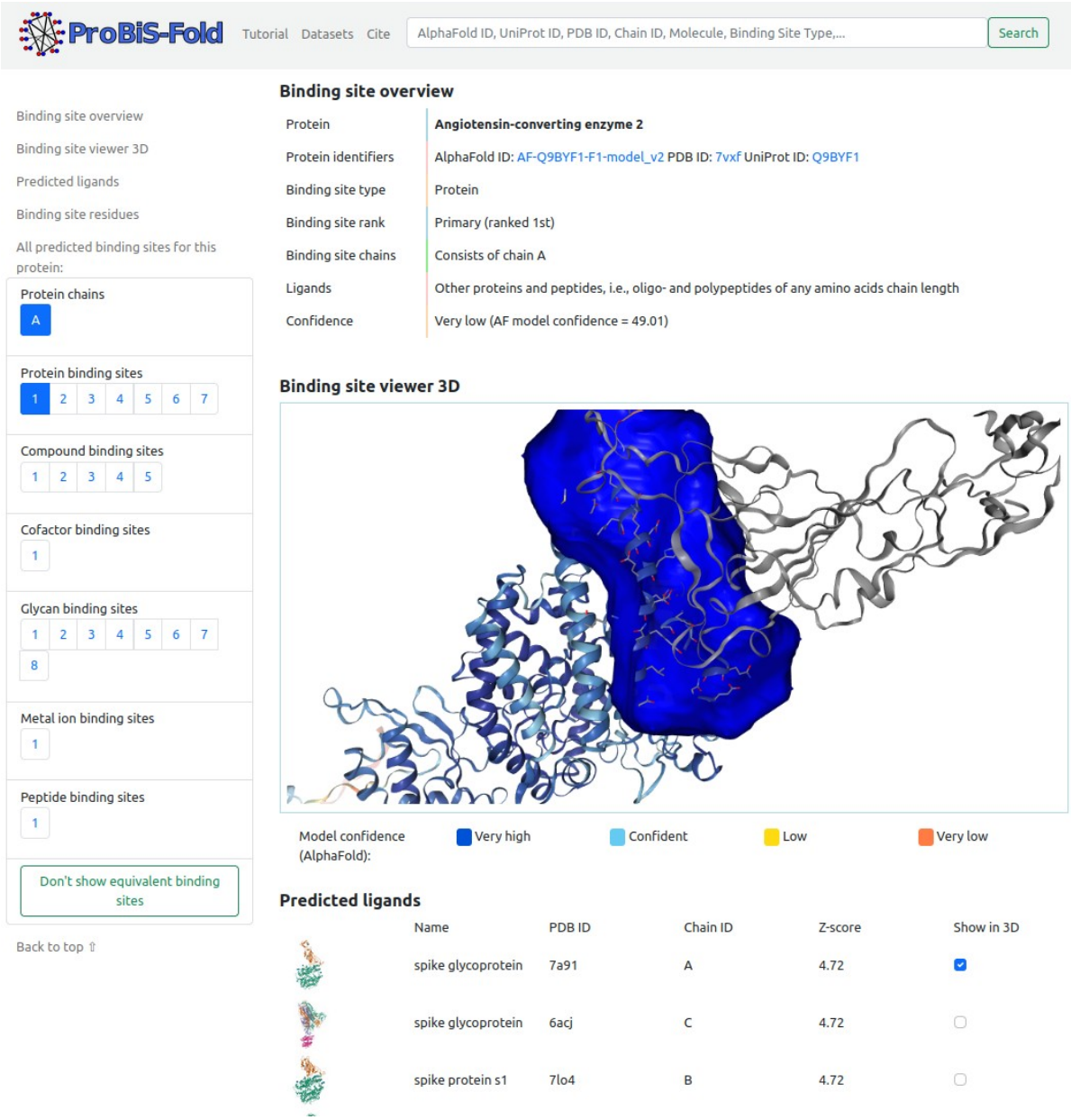


Predicted ligands

	Chem ID	Name	PDB ID	Z-score	Show in 3D
	9ZW	4-methyl-3-(morpholi...	7lmq	2.9	☐
	6ZW	7-(diethylamino)-4-m...	6mg5	2.9	☑
	9ZX	4-methyl-3-(morpholi...	7lmr	2.9	☐
	NY6	3-[2-[7-(diethylamin...	7lmp	2.9	☑

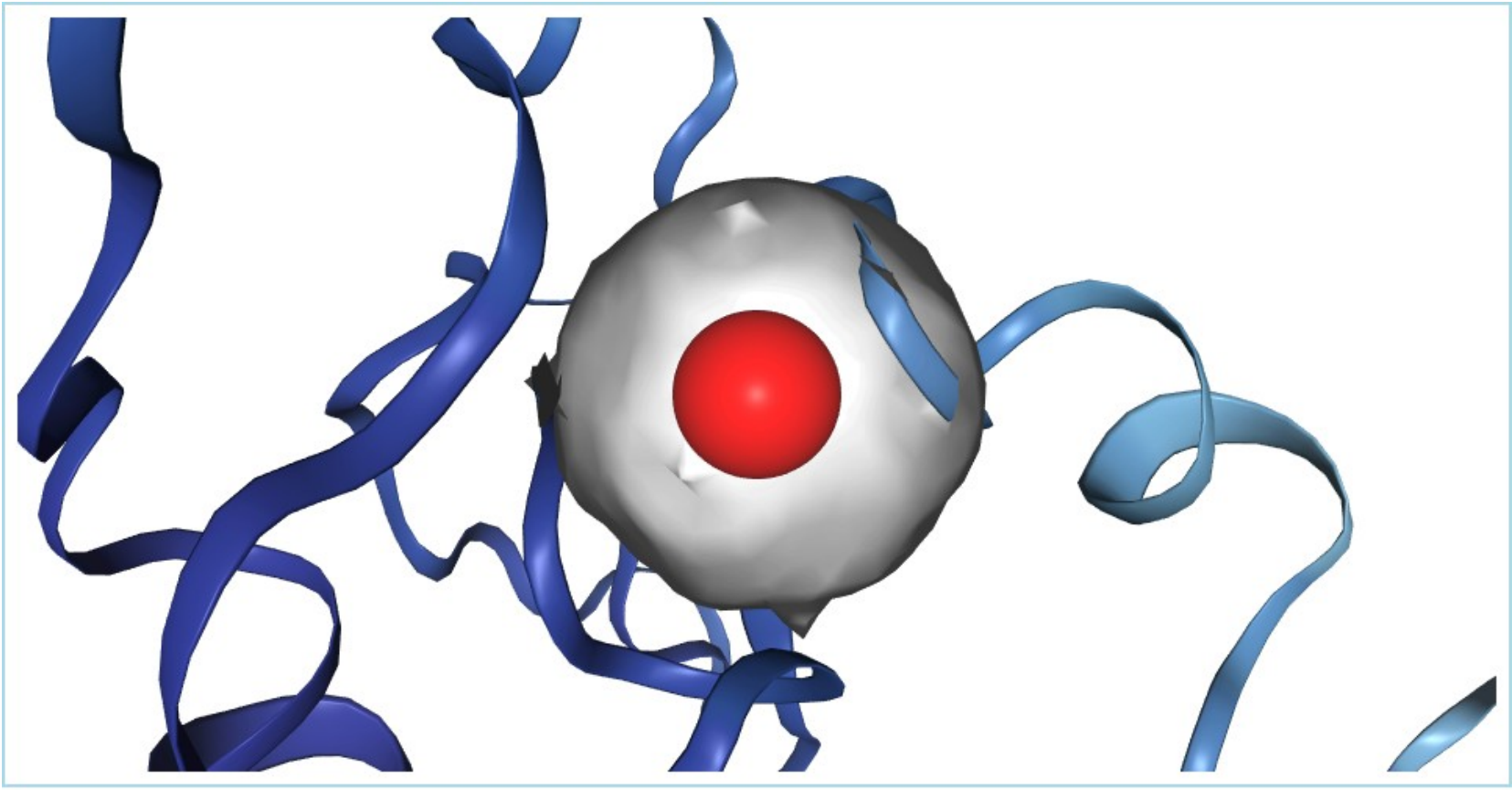
J. Chem. Inf. Model., **2022** , 62, 22, 5821–5829. http://probis-fold.insilab.org/

Protein binding sites



J. Chem. Inf. Model., **2022** , 62, 22, 5821–5829. http://probis-fold.insilab.org/

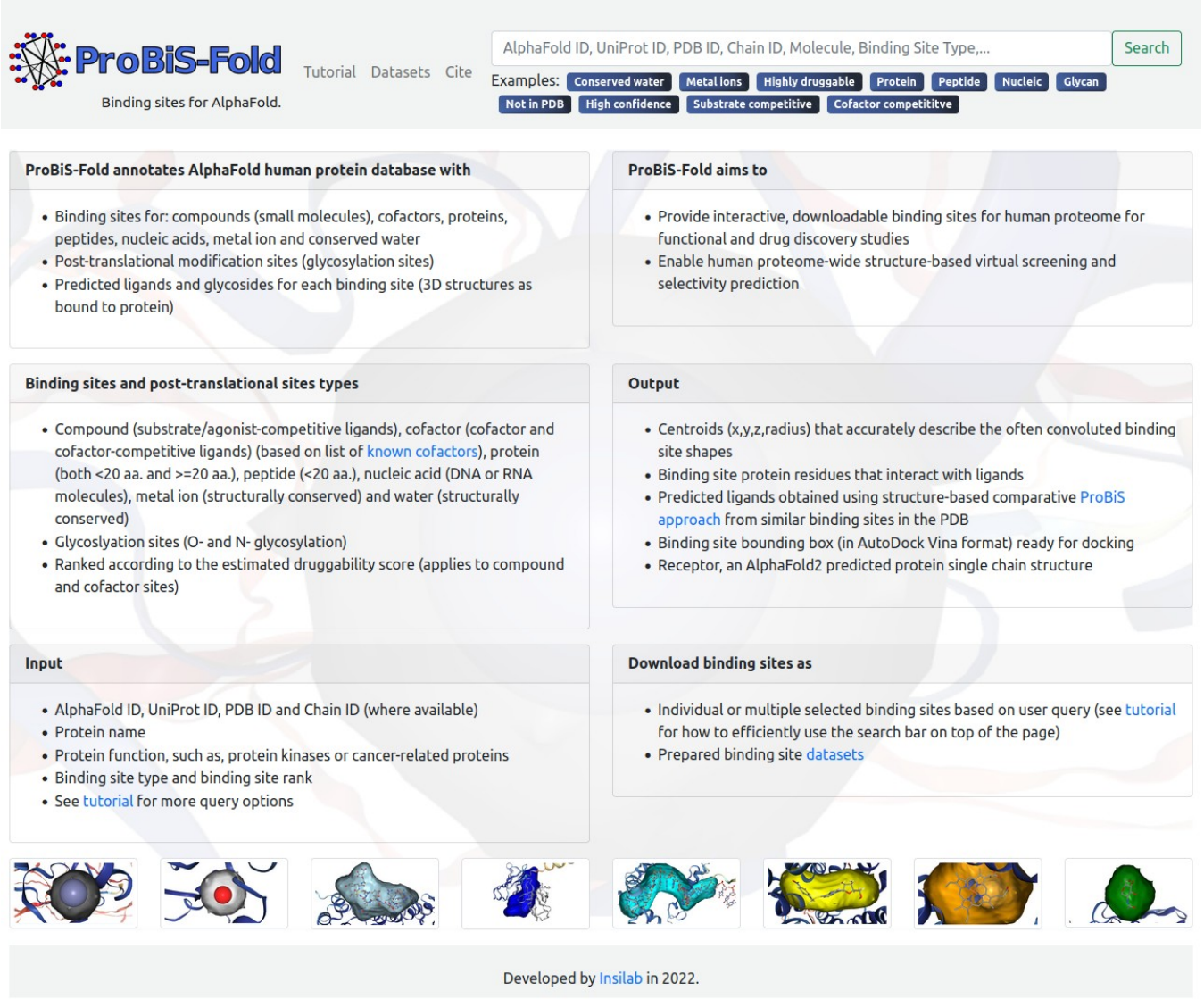
Binding sites for water and ions



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

J. Chem. Inf. Model., **2022** , 62, 22, 5821–5829. http://probis-fold.insilab.org/

ProBiS-Fold web server



J. Chem. Inf. Model., **2022** , 62, 22, 5821–5829. http://probis-fold.insilab.org/

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 competititve

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.

J. Chem. Inf. Model., **2022** , 62, 22, 5821–5829. http://probis-fold.insilab.org/

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