Binding site identification

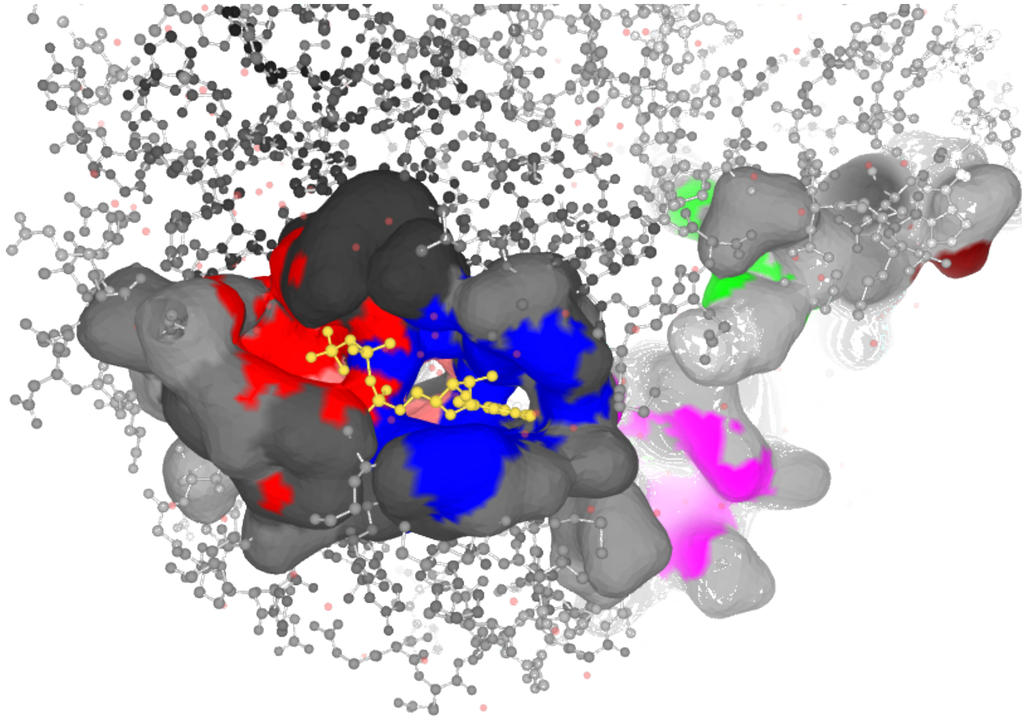
The Pocket Art

7ADD

Marian Novotný

Binding site

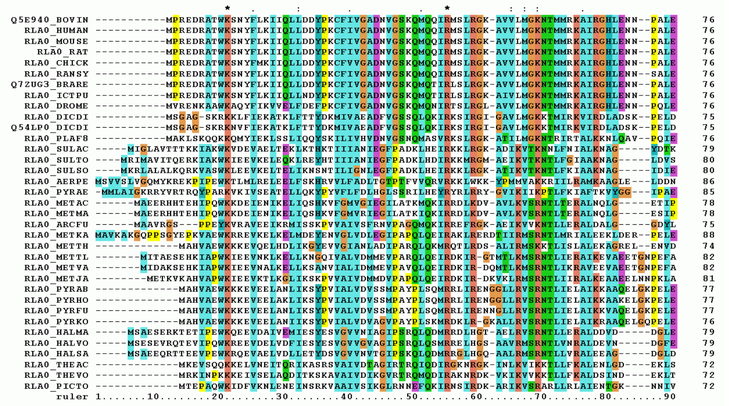
- Where a macromolecule interacts (forms contact) with other molecules
  - Structure-based drug discovery
  - Function annotation
  - Variation effect prediction
  - …

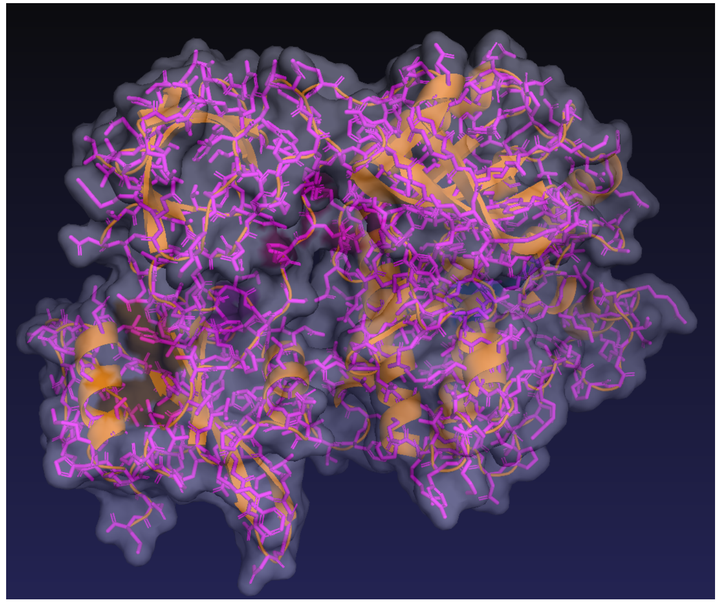


….FPWFGMDIGGTLVKLSYFEPIDITAEEEQEEVES….

Ligand binding sites prediction approaches

- **Automatic approaches**
  - Template-based
  - Template-free
    - Spatial
    - Energy-based
    - Knowledge-based (statistical)
    - Machine learning

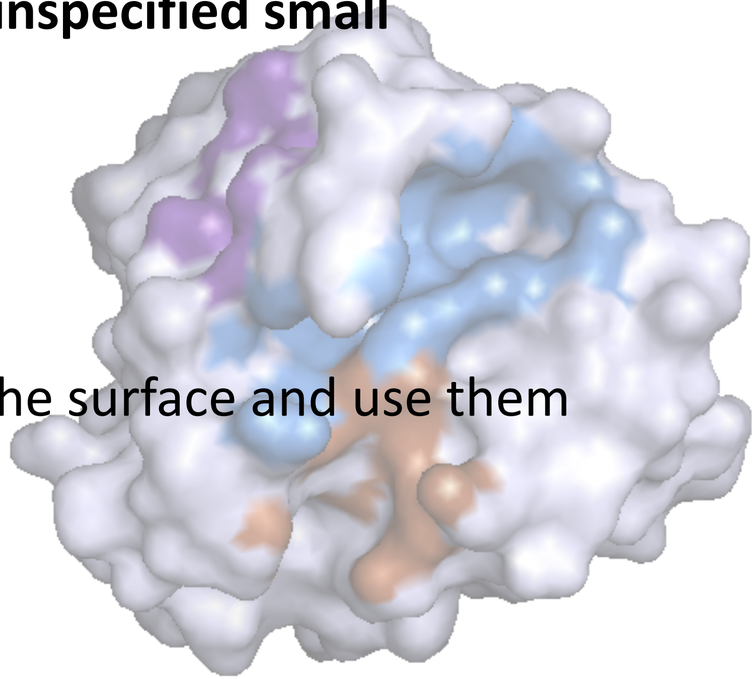




Over 50 methods developed, the comparison difficult!

P2RANK

- Method to **identify surface regions** which are **capable of binding unspecified small molecule**
- **Input**: a protein structure
- **Method:** build a supervised ML model from features of points on the surface and use them for prediction
- **Output**: a list of surface regions probably capable of ligand binding



Krivak, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J Cheminform* (2018)
Jakubec, D., Skoda, P., Krivak, R., Novotny, M. & Hoksza, D. PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures. *Nucleic Acids Res* (2022)

Model construction

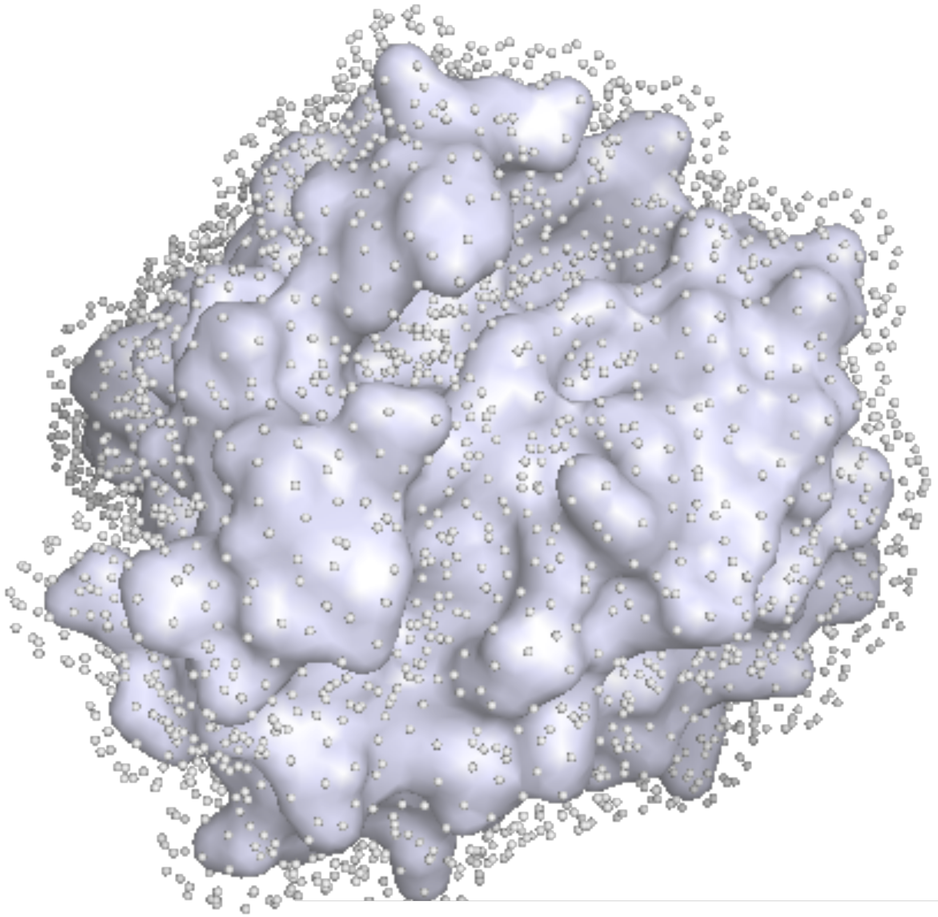
1. Obtaining known protein-ligand complexes

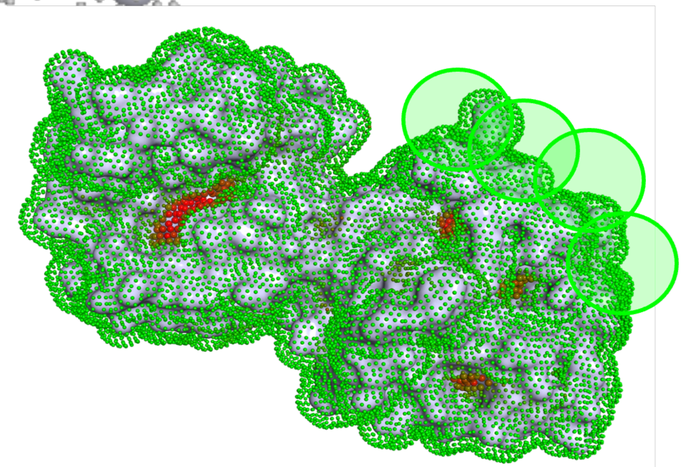
2. Cover the surface with a **mesh of points** (solvent accessible surface – SAS points)

3. **Label points** as binding/nonbinding

4. **Extract a vector of physico-chemical and structural features** for each SAS point of each protein

5. **Build a model,** which is able for given point (vector) decide with what probability is that point part of a pocket

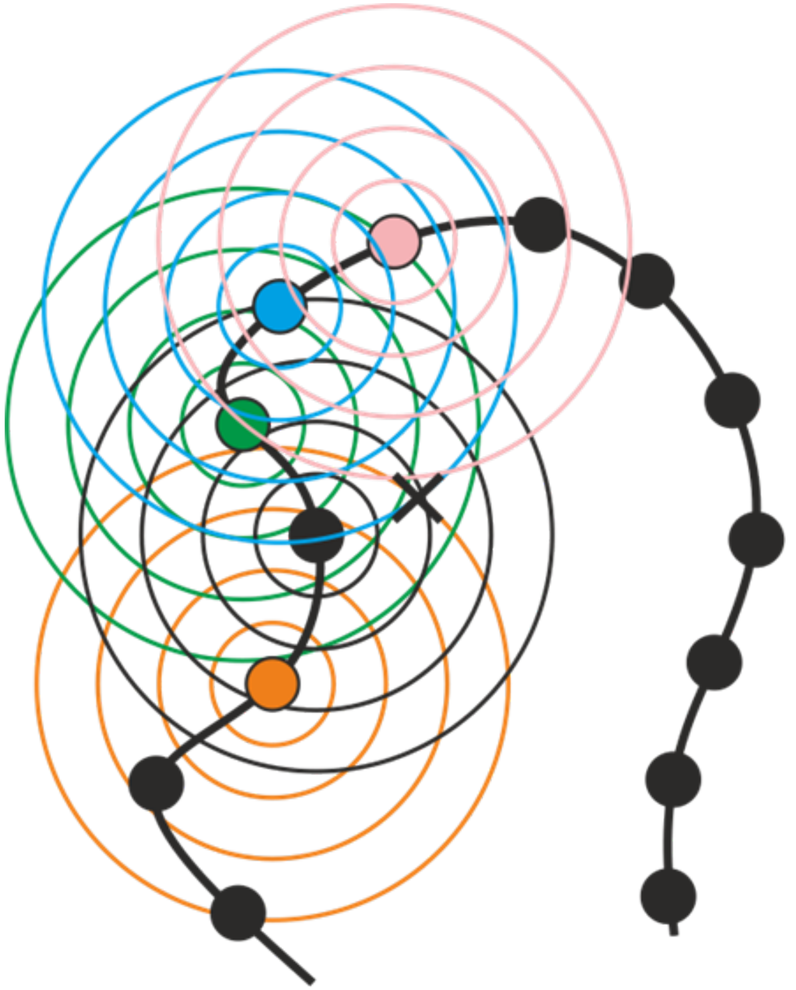




Features extraction

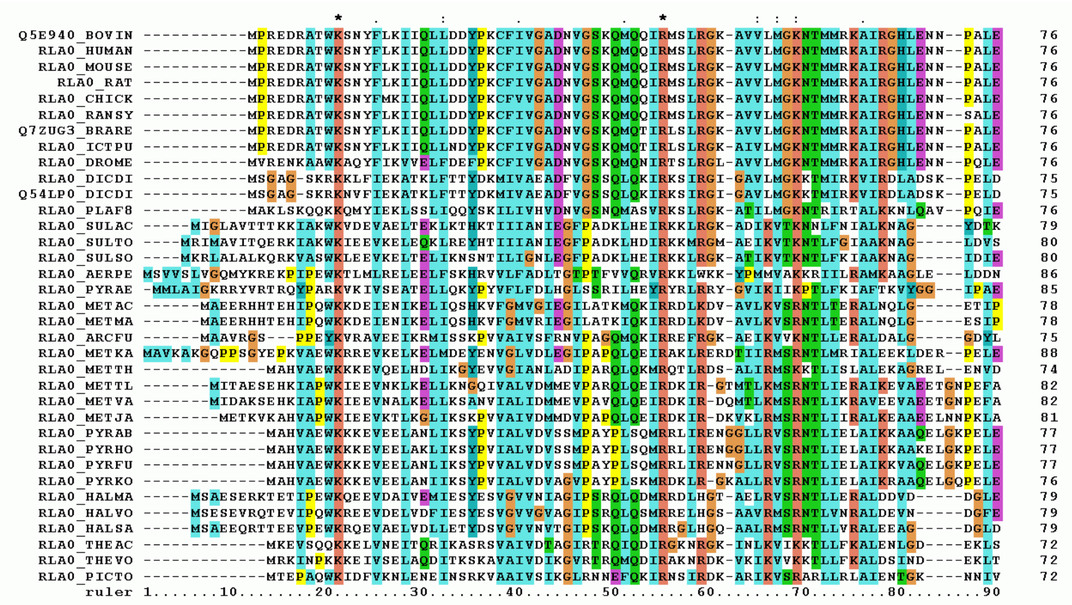
- **More than 30 attributes** describing physical-chemical properties of amino acids within the local neighborhood of given point

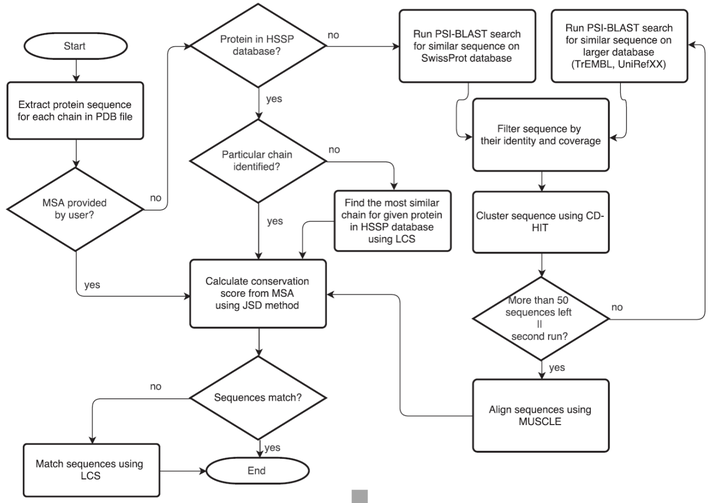
$$\mathrm{IFV}(P) = \sum_{\mathrm{A}_i \in\ \mathrm{A}(P)} \mathrm{AFV}\left(\mathrm{A}_i\right) \cdot w(\mathrm{dist}(P, \mathrm{A}_i)) \quad || \quad \mathrm{FV}(P)$$



Feature name	T*	source**	description
`hydrophobic`	a	AA tab.	binary attribute, 1 for hydrophobic residues
`hydrophilic`	a	AA tab.	binary attribute, 1 for hydrophilic residues
`hydrophatyIndex`	a	AA tab.	side-chain hydropathy index with values in range $\langle -4.5, 4.5 \rangle$ [5]
`aliphatic`	a	AA tab.	binary attribute, 1 for aliphatic residues
`aromatic`	a	AA tab.	binary attribute, 1 for aromatic residues
`sulfur`	a	AA tab.	binary attribute, 1 for residues containing sulfur
`hydroxyl`	a	AA tab.	binary attribute, 1 for hydroxyl group containing residues
`basic`	a	AA tab.	binary attribute, 1 for basic residues
`acidic`	a	AA tab.	binary attribute, 1 for acidic residues
`amide`	a	AA tab.	binary attribute, 1 for amide group containing residues
`posCharge`	a	AA tab.	binary attribute, 1 for positively charged residues
`negCharge`	a	AA tab.	binary attribute, 1 for negatively charged residues
`hBondDonor`	a	AA tab.	binary attribute, 1 for H-bond donor containing residues
`hBondAcceptor`	a	AA tab.	binary attribute, 1 for H-bond acceptor containing residues
`hBondDonorAcceptor`	a	AA tab.	binary attribute, 1 for residues that have H-bond donor AND acceptor
`polar`	a	AA tab.	binary attribute, 1 for polar residues
`ionizable`	a	AA tab.	binary attribute, 1 for ionizable residues
`vsAromatic`	a	AT tab.	VolSite atomic level features [1]
`vsCation`	a	AT tab.	
`vsAnion`	a	AT tab.	
`vsHydrophobic`	a	AT tab.	
`vsAcceptor`	a	AT tab.	
`vsDonor`	a	AT tab.	
`atomicHydrophobicity`	a	AT tab.	Atom type hydrophobicity scale [3]
`apRawValids`	a	AT tab.	Ligand binding propensity for biologically valid ligands [4]
`apRawInvalids`	a	AT tab.	Ligand binding propensity for biologically invalid ligands [4]
`bfactor`	a	given	B-factor number of the atom from pdb file
`atoms`	p	calc.	absolute number of protein exposed atoms in the neighbourhood (within 6 Å radius of the point)
`atomDensity`	p	calc.	number of protein exposed atoms weighted by distance
`atomC`	p	calc.	number of carbon atoms in the neighbourhood
`atomO`	p	calc.	number of oxygen atoms in the neighbourhood
`atomN`	p	calc.	number of nitrogen atoms in the neighbourhood
`hDonorAtoms`	p	calc.	number of H-bond donor atoms in the neighbourhood
`hAcceptorAtoms`	p	calc.	number of H-bond acceptor atoms in the neighbourhood
`protrusion`	p	calc.	Protein surface protrusion inspired by [6] calculated simply as number of all protein atoms (not just exposed) within 10 Å radius of the point

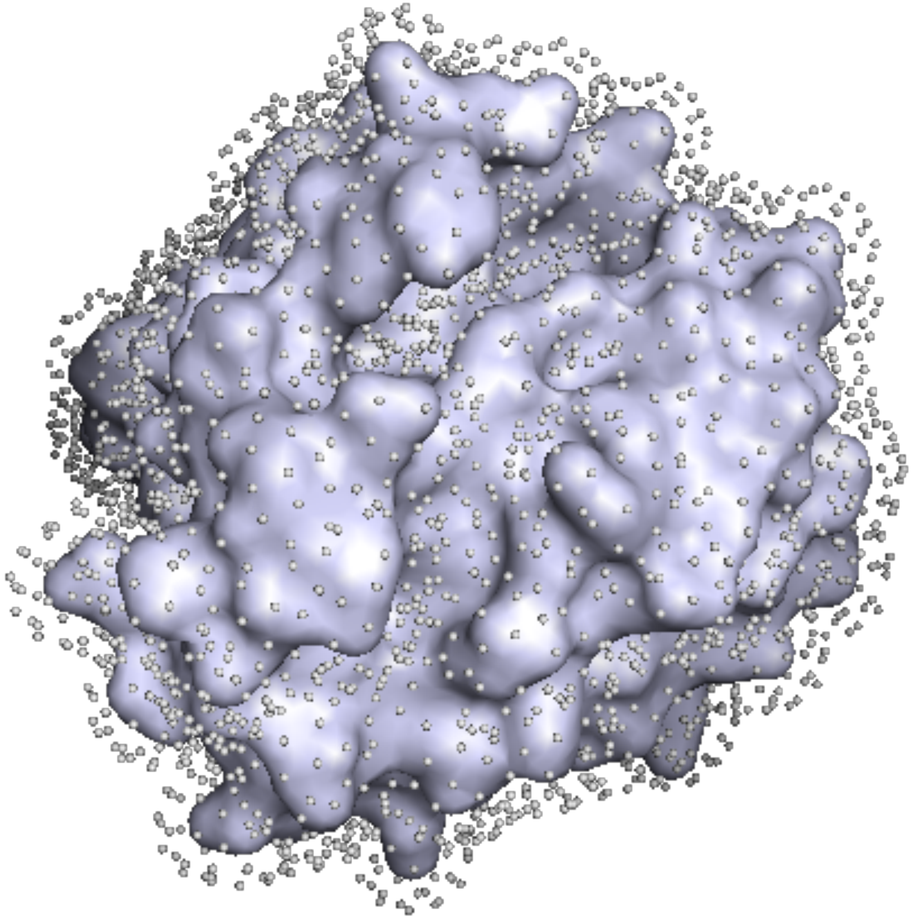
Features extraction - conservation

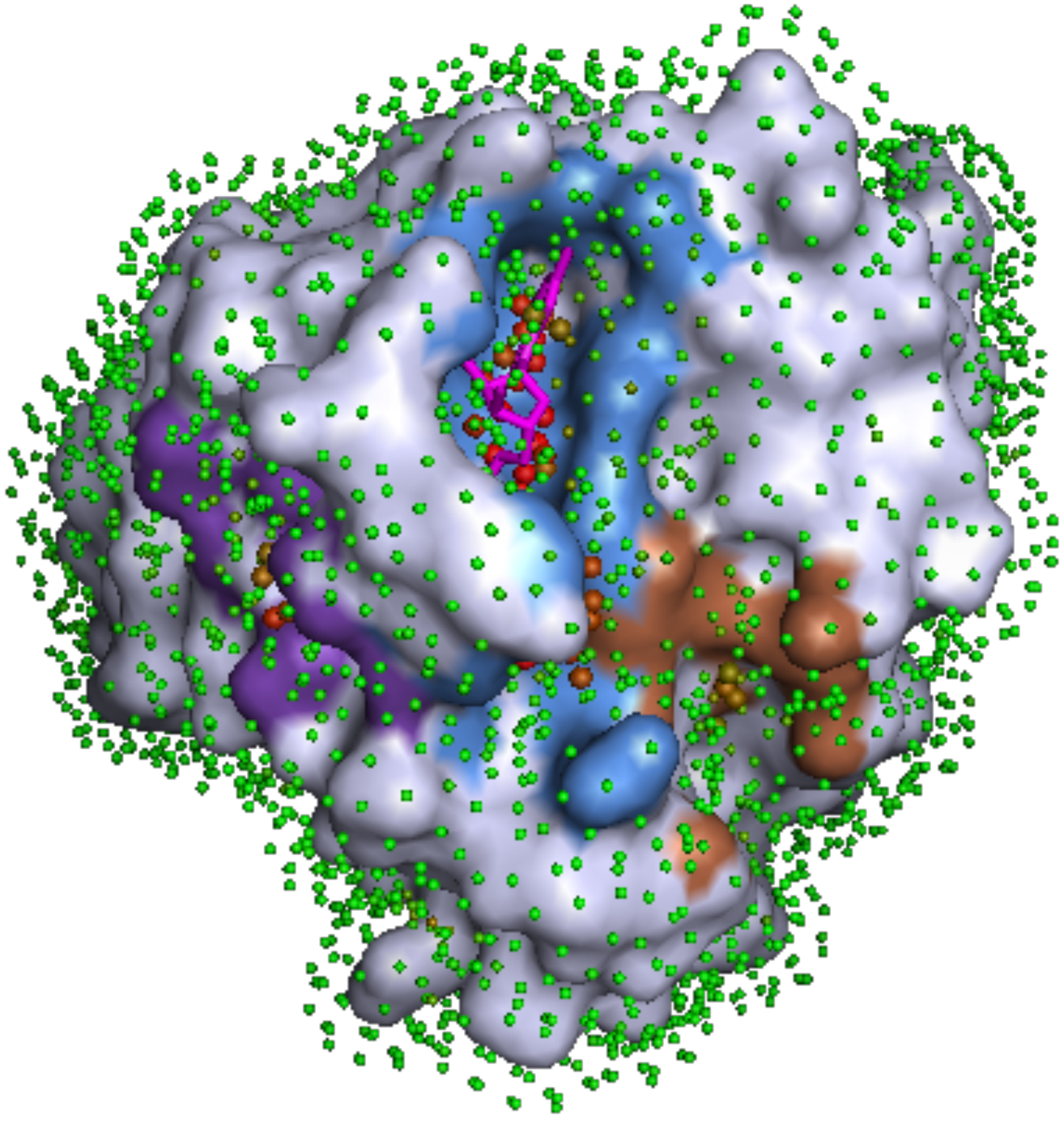




Detection of binding sites

1. Cover the surface with a **mesh of SAS points**
2. **Apply the model to every point of the mesh** → ligandability score
3. **Filter** out points with low ligandability score
4. **Cluster** the remaining **points** → **binding pocket**
5. **Score the pockets** – cumulative ligandability score → raw score → confidence score
6. **Map** pocket SAS points onto atoms

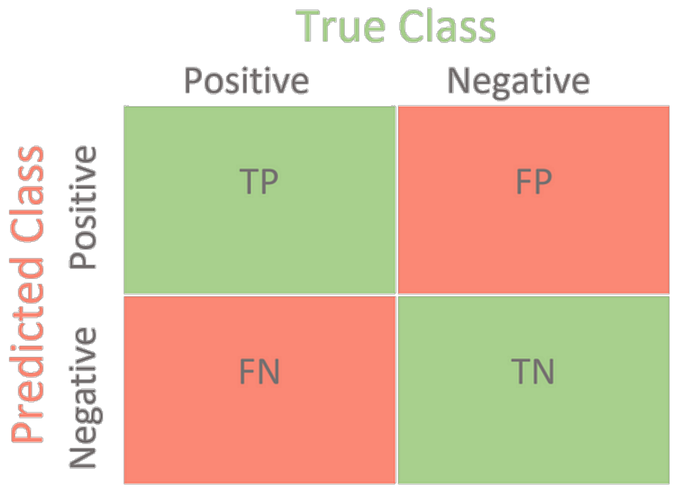




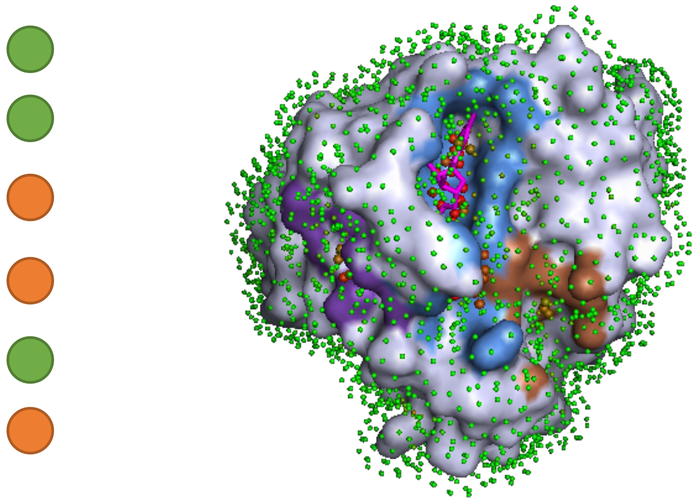
Binding sites prediction evaluation

Binding site evaluation metric

- Typical binary classification problem metrics not suitable
  - **No true negatives**



- **Success rate** with respect to **Top-$n + k$** pockets
  - n – number of true pockets in a protein
  - k - room for error



Pocket detection criteria

- Distance-based
  - **DCA** < threshold
    - the minimal distance between the center of the predicted pocket and any atom of the ligand
  - **DCC** < threshold
    - the distance between the centers of the predicted and true binding sites

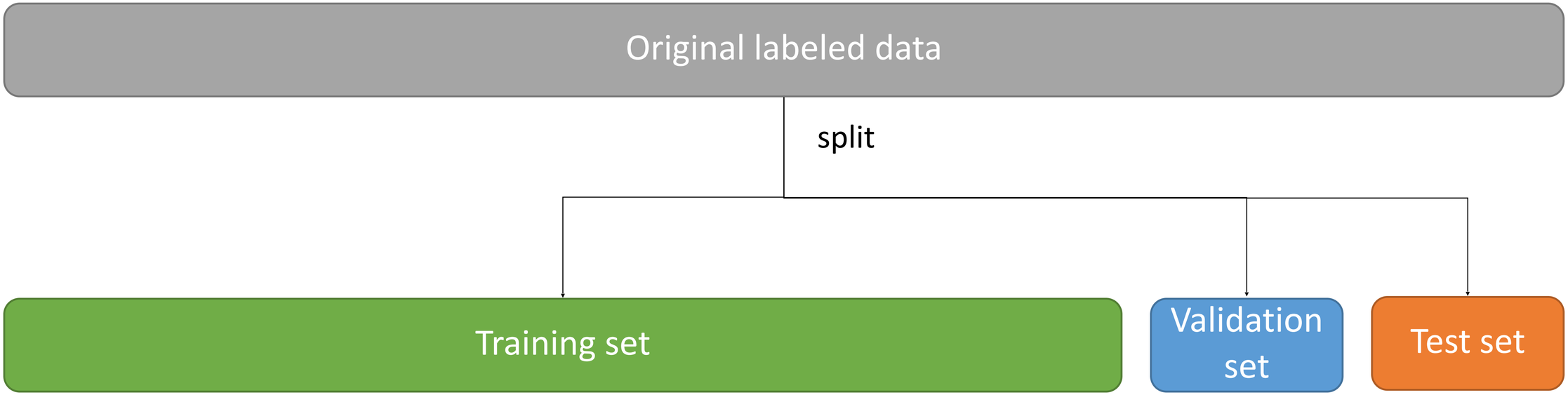
DCA(4) … distance from the **center** of the pocket to **any ligand atom** <= 4 Å

Top-(n+k) … On a given protein structure
n = number of true binding sites (bound ligands)

If there is only 1 ligand:

Top-n ~ Top-1
Top-(n+2) ~ Top-3

ML evaluation methodology



Evaluation datasets

- **Training** set – *CHEN11*
  - 251 proteins, 476 ligands
  - Non-redundant
  - Superimposed ligands from close homologs
- **Validation** set – *JOINED*
  - B48/U48 - 48 proteins in a bound and unbound state
  - B210 - 210 proteins in bound state
  - DT198 - 198 drug-target complexes
  - ASTEX - 85 proteins that was introduced as a benchmarking
  - Dataset for molecular docking methods
- **Testing sets**
  - *COACH420*
    - 420 single chain structures that contain a mix of drug targets and naturally occurring ligands
  - *HOLO4K*
    - >4000 structures
    - Larger multi-chain structures

CHEN11: Chen K, Mizianty M, Gao J, Kurgan L (2011) A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. Structure 19(5):613–621
B48/U48, B210: Huang B, Schroeder M (2006) Ligsitecsc: predicting ligand binding sites using the connolly surface and degree of conservation. BMC Struct Biol 6(1):19
DT198 : Zhang Z, Li Y, Lin B, Schroeder M, Huang B (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. Bioinformatics 27(15):2083–2088
ASTEX: Hartshorn M, Verdonk M, Chessari G, Brewerton S, Mooij W, Mortenson P, Murray C (2007) Diverse, high-quality test set for the validation of proteinligand docking performance. J Med Chem 50(4):726–741
COACH: Roy A, Yang J, Zhang Y (2012) Cofactor: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 40(W1):471–477
HOLO4K: Schmidtke P, Souaille C, Estienne F, Baurin N, Kroemer R (2010) Largescale comparison of four binding site detection algorithms. J Chem Inf Model 50(12):2191–200

Evaluation

	COACH420		HOLO4K	
	Top-n	**Top-(n+2)**	**Top-n**	**Top-(n+2)**
Fpocket	56.4	68.9	52.4	63.1
Fpocket+PRANK[a]	63.6	76.5	62.0	71.0
SiteHound[†]	53.0	69.3	50.1	62.1
MetaPocket 2.0[†]	63.4	74.6	57.9	68.6
DeepSite[†]	56.4	63.4	45.6	48.2
P2Rank[protrusion][b]	64.2	73.0	59.3	67.7
P2Rank	*72.0*	*78.3*	*68.6*	*74.0*

Krivák, R., & Hoksza, D. (2018). P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. Journal of cheminformatics, 10, 1-12.

feature	importance
protrusion	0.084528
bfactor	0.013888
apRawInvalids	0.011785
vsAromatic	0.010165
apRawValids	0.009403
atomO	0.009275
hydrophobic	0.008630
hydrophilic	0.007643
vsAcceptor	0.006244
vsHydrophobic	0.005273
atoms	0.005188
aromatic	0.004433
atomN	0.004236
hydrophatyIndex	0.004232
atomC	0.003687
vsDonor	0.003451
aliphatic	0.003350
atomicHydrophobicity	0.002663
hBondDonorAcceptor	0.002650
hDonorAtoms	0.002626
atomDensity	0.002549
polar	0.002402
ionizable	0.002142
hAcceptorAtoms	0.001904
hBondAcceptor	0.001705
sulfur	0.001621
negCharge	0.001538
acidic	0.001504
basic	0.001467
hydroxyl	0.001328
vsAnion	0.001072
hBondDonor	0.001059
posCharge	0.001021
vsCation	0.000832
amide	0.000831

Effect of conservation

	COACH420		HOLO4K	
	Top-n	Top-(n+2)	Top-n	Top-(n+2)
Fpocket 1.0	56.4	68.9	52.4	63.1
Fpocket 3.1	42.9	56.9	54.9	64.3
SiteHound[a]	53.0	69.3	50.1	62.1
MetaPocket 2.0[a]	63.4	74.6	57.9	68.6
DeepSite[a]	56.4	63.4	45.6	48.2
P2Rank	72.0	**78.3**	68.6	74.0
P2Rank+Cons.[b]	**73.2**	77.9	**72.1**	**76.7**

Table 2. Number of predicted binding sites and dataset statistics.

	COACH420	HOLO4K
Proteins	420	4009
Avg. protein atoms	2179	3908
Avg. ligands	1.2	2.4
Fpocket 1.0	14.6	27.0
Fpocket 3.1	13.9	16.0
SiteHound	66.2	99.5
MetaPocket 2.0	6.3	6.4
DeepSite	3.2	2.8
P2Rank	6.3	12.6
P2Rank+Conservation	3.4	7.7

Displayed is the average total number of binding sites predicted per protein by each method on a given dataset.

Jendele, L., Krivak, R., Skoda, P., Novotny, M., & Hoksza, D. (2019). PrankWeb: a web server for ligand binding site prediction and visualization. Nucleic acids research, 47(W1), W345-W349.

Runtime

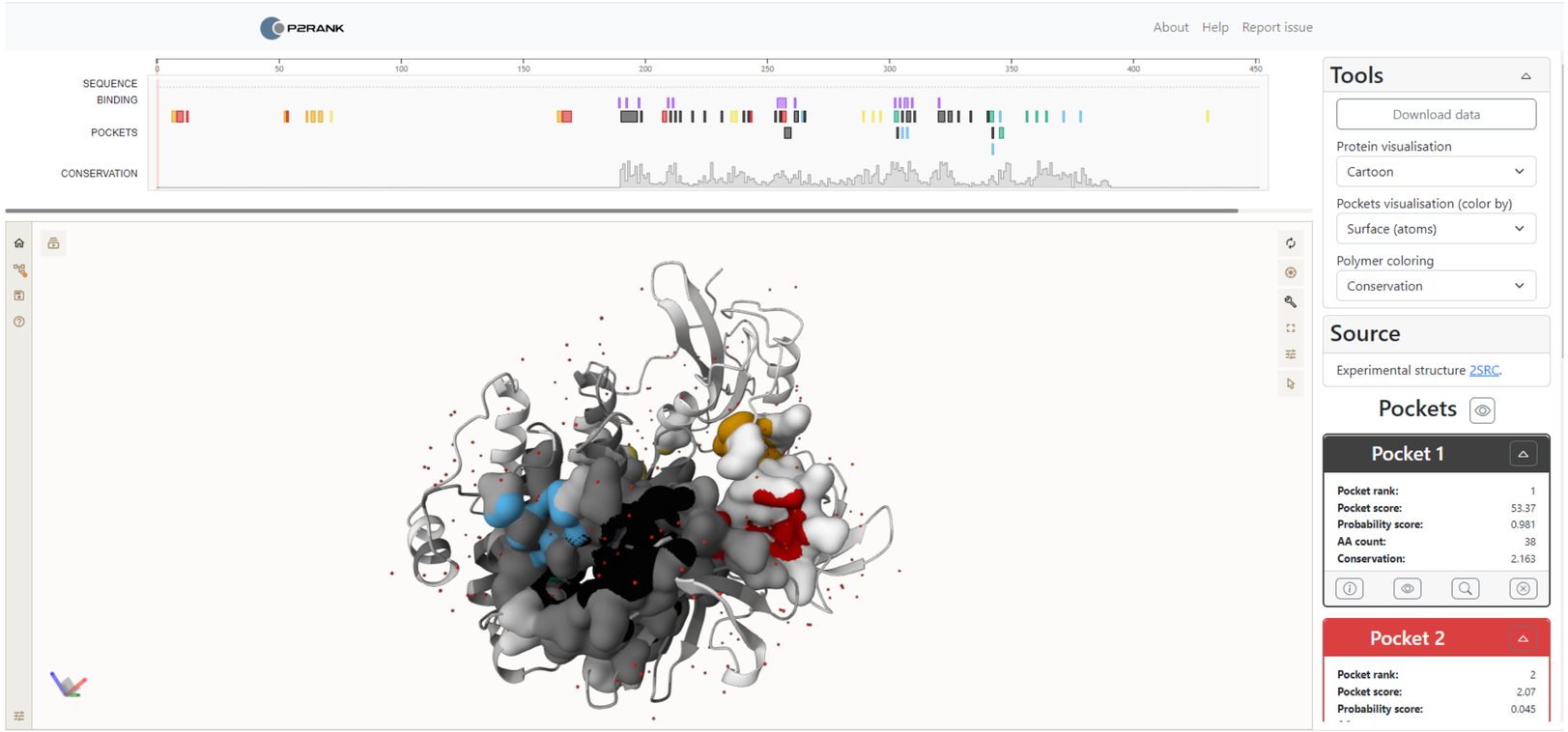
Method	Time[†]
COACH (web server)	15 h (self reported estimate)
eFindSite (web server)	6.9 ± 0 h
COACH (stand-alone)	6.4 ± 2 h
GalaxySite (web server)	2 h (self reported estimate)
3DLigandSite (web server)	1–3 h (self reported estimate)
ISMBLab-LIG (web server)	71 ± 2 min
FTSite (web server)	39 ± 3 min
LISE (web server)	39 ± 0.1 min
MetaPocket 2.0 (web server)	2.8 ± 0.4 min
DeepSite (web server)	38 ± 0.03 s
SiteHound (stand-alone)	12 ± 0.5 s
P2Rank (stand-alone)	6.8 ± 0.2 s (cold start*)
	0.9 s (in larger dataset*)
Fpocket (stand-alone)	0.2 ± 0.01 s

[†] Average time required for LBS prediction on a single protein. Displayed is self reported estimate or a result of our test on a small dataset of 5 proteins á ~ 2500 atoms. Stand-alone tools were tested on a single 3.7 GHz CPU core. For web servers the wall time from submitting a job to receiving the result was measured.

*Difference is due to JVM initialization and model loading cost

Availability

- **Command line app**
  - https://github.com/rdk/p2rank
  - Java, PyMOL
- **Webserver**
  - https://prankweb.cz
  - AlphaFold, conservation, API



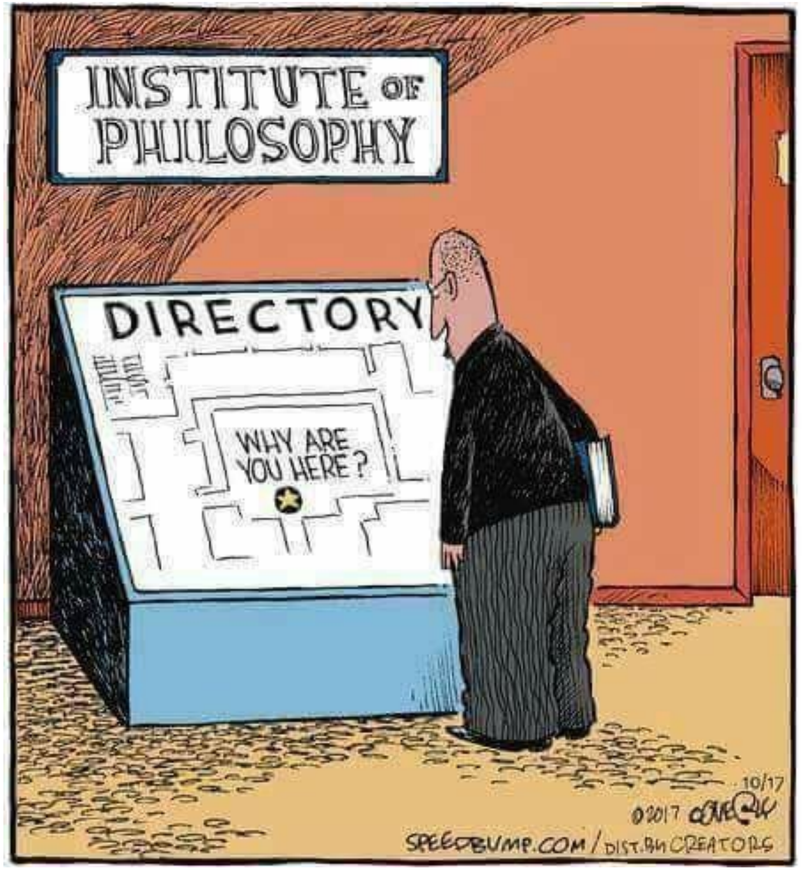
AHoJ-DB:
A PDB-wide assignment of apo & holo relationships based on individual protein-ligand interactions

Motivation for AHoJ – can we test better?

A. Suspiciously good results from P2Rank -> need for harder/more realistic targets (= apo structures)

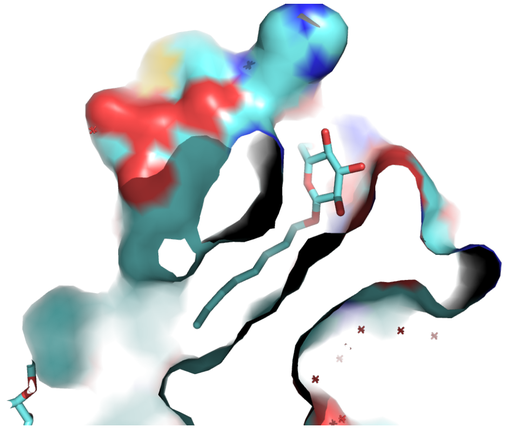
B. No apo-holo dataset in literature

- Specifications
  - ✓ Search by ligand binding site
  - ✓ Apply quality filters (resolution, experimental method)
  - ✓ Accept multiple search
  - ✓ Visualization

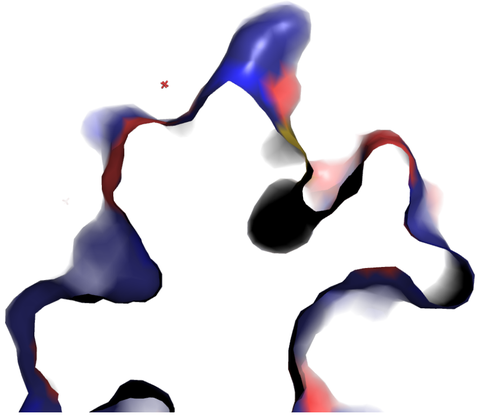


Cryptic binding sites

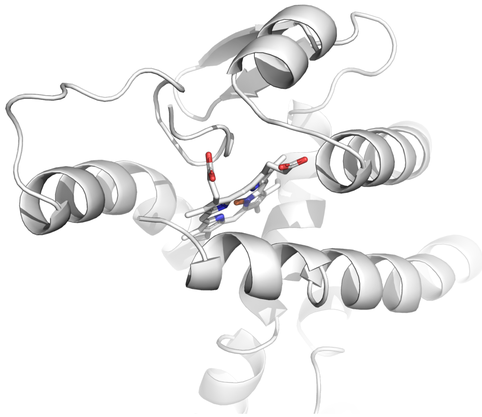
2npq A BOG

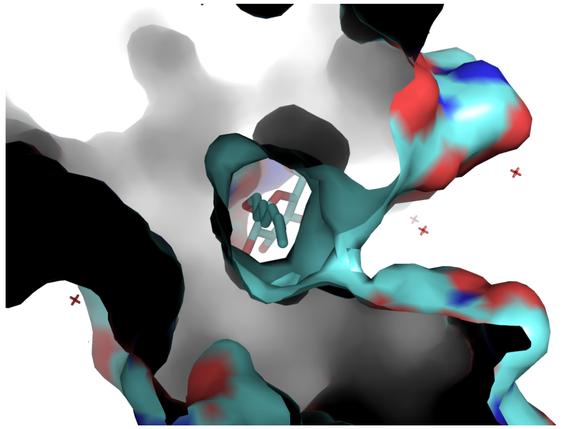


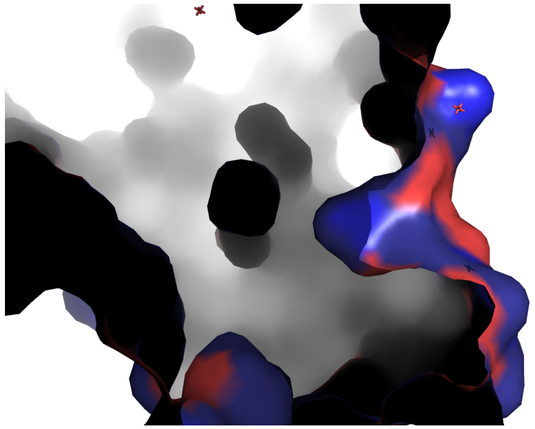
6sfi A



3cqv A HEM







2v0v A

AHoJ – Apo-Holo Juxtaposition BETA

AHoJ-DB Start My Jobs Help API About Us GitHub

Introduction

Search for Apo-Holo pairs

Query / Queries example1 example2 example3

2SRC A PTR

Job Name (optional)

Email for notification (optional)

Options

- [ ] X-ray structures only
- [ ] Exclude NMR structures
- [x] Ligand-free sites
- [ ] Consider water as ligand
- [x] Consider non-standard residues as ligands
- [ ] Consider D-amino acids as ligands
- [x] Save aligned Apo chains
- [x] Save aligned Holo Chains

Binding residues threshold: 1.0 %

Sequence overlap threshold: 0 %

Resolution threshold: 4.0 Å

Minimum TM-score: 0.5

Ligand scanning radius: 4.5 Å

Submit Job

apoholo.cz

Feidakis et al., 2022 - Bioinformatics

Job: 4QT6H
Query: 2SRC A PTR
Status: done

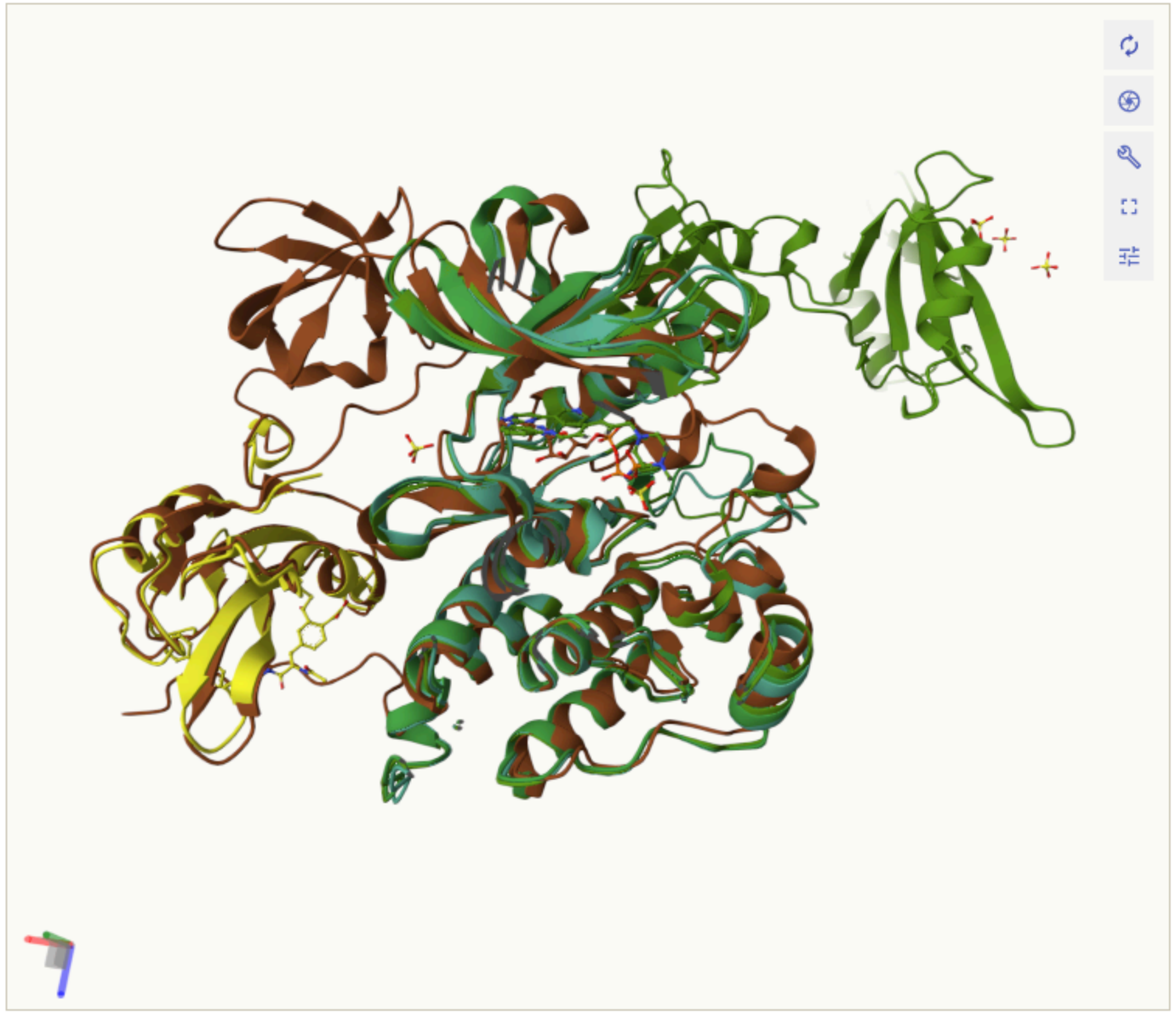
Download

Found APO Chains / Structures: 35 / 18

Chain	AT	Res	SO	MBR	RMSD	TM-sc	Lig
1y57 / A	A	1.91	100.0	100.0	2.49	0.541	
7yqe / A	A	3.5	35.6	67.0	1.55	0.368	
7yqe / B	A	3.5	35.6	67.0	1.44	0.362	
4f59 / A	A	1.71	23.8	67.0	0.89	0.238	
1yi6 / A	A	2.0	61.2	33.0	2.53	0.526	
1yi6 / B	A	2.0	61.2	33.0	2.39	0.534	
7ng7 / A	A	1.5	60.6	33.0	1.66	0.565	
4mxo / A	A	2.105	58.8	33.0	2.32	0.524	
6e6e / A	A	2.15	58.8	33.0	2.25	0.527	
6e6e / B	A	2.15	58.8	33.0	2.22	0.528	

Found HOLO Chains / Structures: 53 / 45

Chain	AT	Res	SO	MBR	RMSD	TM-sc	Lig
1ksw / A	A	2.8	100.0	100.0	0.91	0.986	PTR
4k11 / A	A	2.3	99.6	100.0	0.89	0.988	PTR
2h8h / A	A	2.2	98.9	100.0	0.84	0.982	PTR
1fmk / A	A	1.5	96.7	100.0	1.14	0.95	PTR
4f5b / A	A	1.57	23.8	67.0	0.84	0.238	PTR
4f5a / A	A	1.8	23.8	67.0	0.89	0.237	PO4
1o43 / A	A	1.5	23.4	67.0	0.9	0.233	821
1o4a / A	A	1.5	23.4	67.0	0.89	0.233	197
1o48 / A	A	1.55	23.4	67.0	0.91	0.233	853
1o4g / A	A	1.55	23.4	67.0	1.0	0.23	CSO I59
1o4k / A	A	1.57	23.4	67.0	0.91	0.233	PSN
1o4n / A	A	1.6	23.4	67.0	0.86	0.233	OXD



apoholo.cz

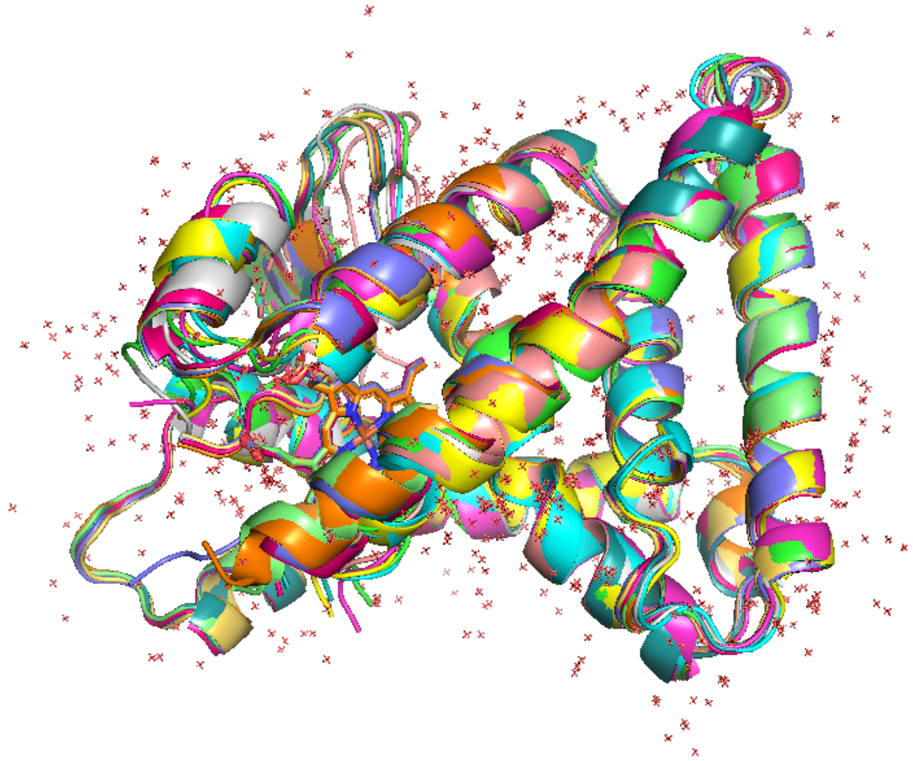
PyMOL

File Edit Build Movie Display Setting Scene Mouse Wizard Plugin Help

```
 Setting: sphere_mode set to 9.
 Setting: nb_spheres_quality set to 3.
PyMOL>rebuild
 Setting: bg_rgb set to white.
 Setting: bg_rgb set to grey80.
 Setting: bg_rgb set to grey50.
 Setting: opaque_background set to on.
 Setting: bg_rgb set to white.
 Setting: opaque_background set to off.
```

PyMOL>

Reset Zoom Orient Draw/Ray
Unpick Deselect Rock Get View
|< < Stop Play > >| MClear
Builder Properties Rebuild



all A S H L C
apo_2v0vA_aligne A S H L C
apo_2v0vB_aligne A S H L C
apo_2v0vC_aligne A S H L C
apo_2v0vD_aligne A S H L C
apo_2v7cA_aligne A S H L C
apo_2v7cB_aligne A S H L C
holo_4n73A_align A S H L C
(COH_A601_4n73A- A S H L C
holo_6wmqA_align A S H L C
(HEM_A601_6wmqA- A S H L C
holo_6wmqB_align A S H L C
(HEM_B601_6wmqB- A S H L C
holo_6wmsA_align A S H L C
(HEM_A601_6wmsA- A S H L C
holo_6wmsB_align A S H L C
(HEM_B601_6wmsB- A S H L C
query_3cqv 1/1 A S H L C
(HEM_A601_query) A S H L C

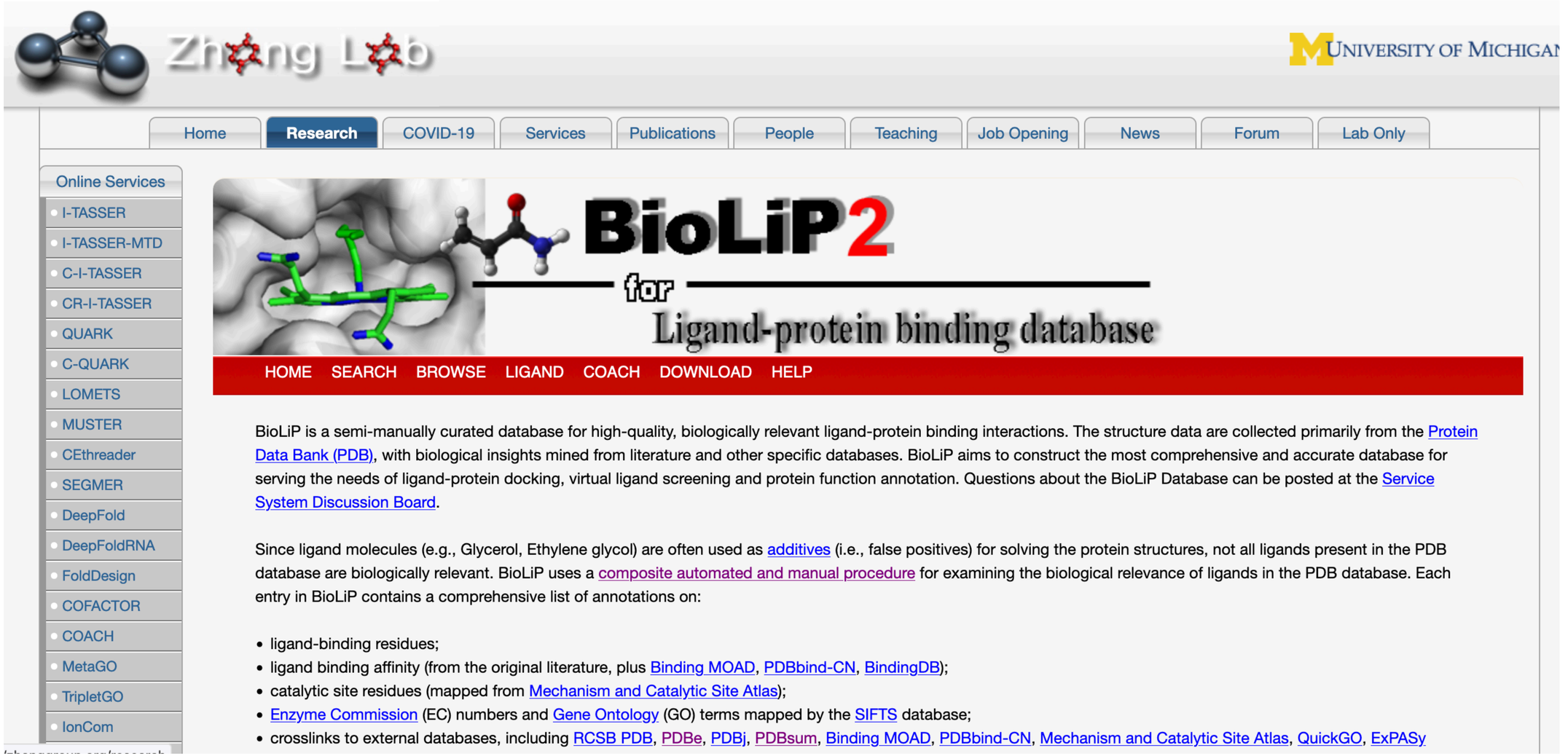
Mouse Mode 3-Button Viewing
Buttons L M R Wheel
& Keys Rota Move MovZ Slab
Shft +Box -Box Clip MovS
Ctrl Move PkAt Pk1 MvSZ
CtSh Sele Orig Clip MovZ
SnglClk +/- Cent Menu
DblClk Menu - PkAt
Selecting Residues
State 1/ 1

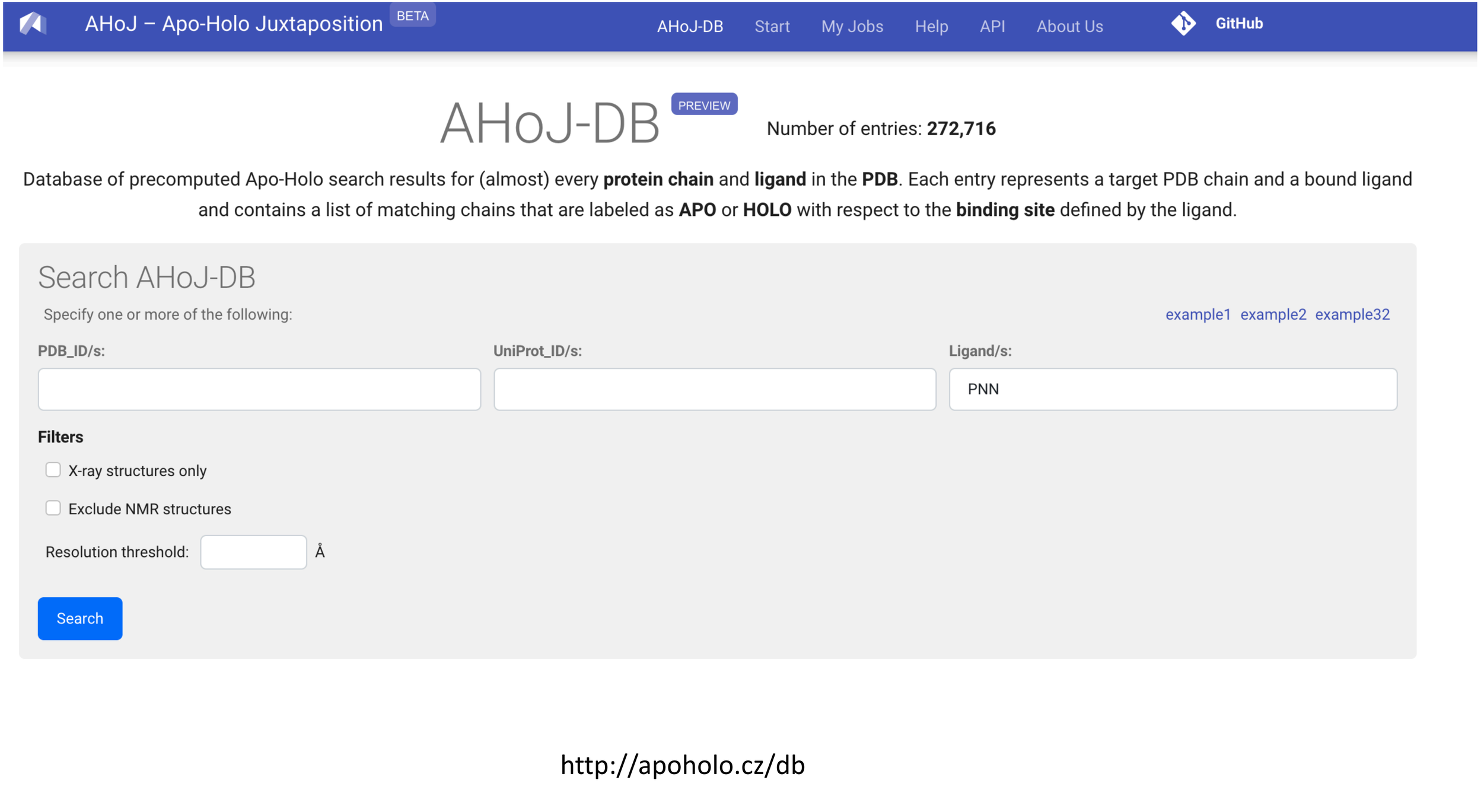
PyMOL>_

AHoJ-DB - PDB- wide identification of apo/holo structure pairs

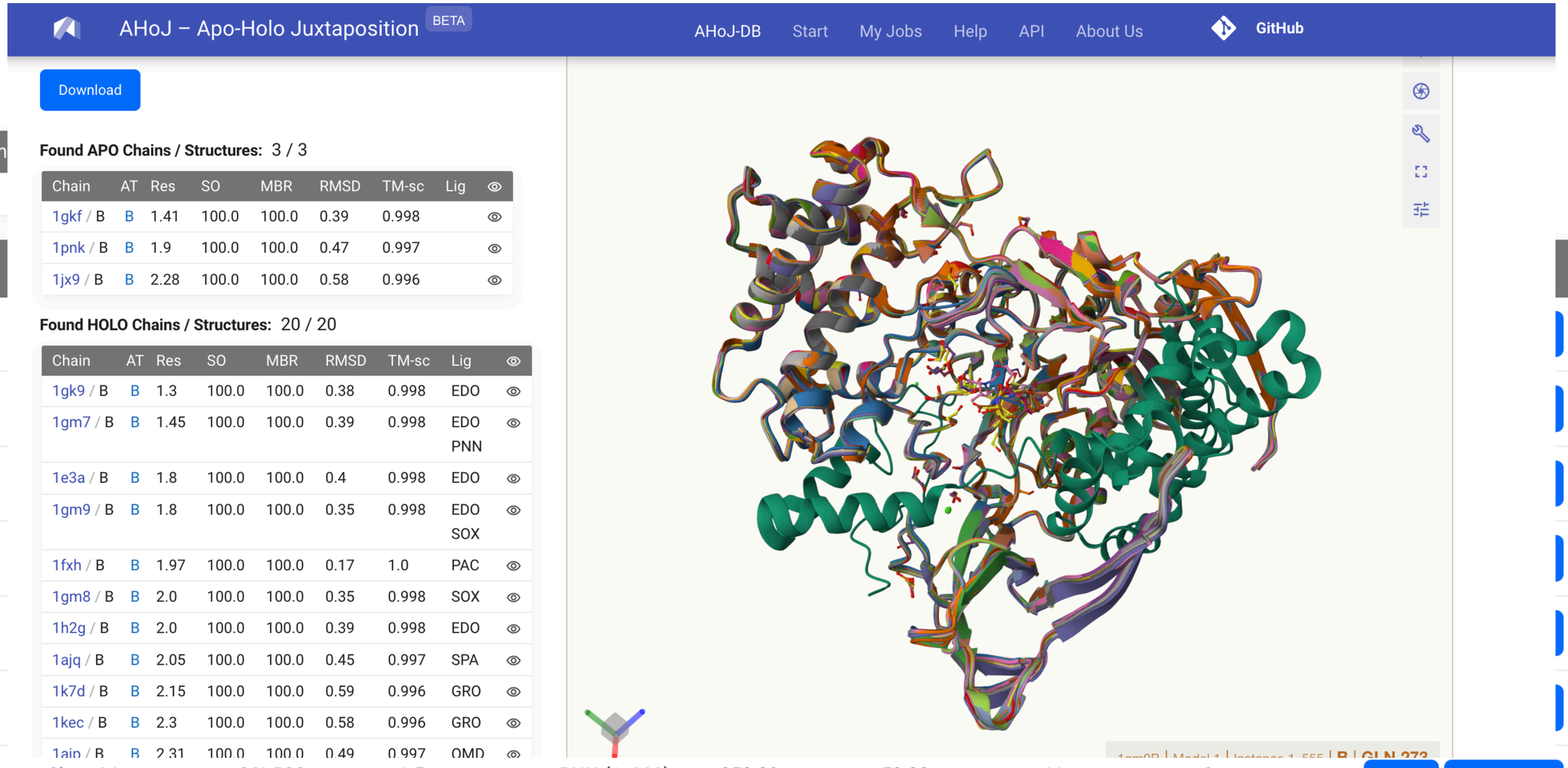
- database of precalculated apo/holo pairs for individual binding sites
- biologically relevant ligands
- search by binding site, uniprot id or ligand id

Biologically- relevant ligands





http://apoholo.cz/db



3huo-A-PNN-303	3huo / A	Q9L5C8	1.5	PNN (A_303)	353.29	58.88	61	9	View Download
3huo-A-PNN-304	3huo / A	Q9L5C8	1.5	PNN (A_304)	322.66	64.53	69	1	View Download
3huo-B-PNN-301	3huo / B	Q9L5C8	1.5	PNN (B_301)	567.94	31.55	10	60	View Download

Most common binding sites

	Holo (input)				Apo (output)				Coverage as % of input			
Ligand	#UniProt	#structures	#chains	#sites	#UniProt	#structures	#chains	#sites	%UniProt	%structures	%chains	%sites
ZN	1937	8259	16681	22635	745	3646	6713	8127	38	44	40	36
MG	2188	5943	13149	17515	1527	4170	9388	12794	70	70	71	73
CLA	200	147	1637	16114	40	49	386	4441	20	33	24	28
CA	1360	4378	8163	14269	778	2524	4415	6616	57	58	54	46
HEM	360	2147	4200	4741	36	196	432	451	10	9	10	10
MN	510	1675	3233	4679	292	854	1572	2455	57	51	49	52
SF4	275	872	1997	3221	26	105	195	227	9	12	10	7
ADP	584	1301	2797	3089	366	873	1932	2087	63	67	69	68
GLC	317	719	1164	2850	241	545	874	2038	76	76	75	72
CU	160	733	1505	2801	76	340	669	924	48	46	44	33
FE	254	835	1891	2522	78	268	721	1087	31	32	38	43
ATP	434	923	1891	2353	290	630	1206	1508	67	68	64	64
FAD	273	1006	1909	2163	23	46	105	135	8	5	6	6
BGC	276	581	921	2157	209	423	673	1458	76	73	73	68
NAD	272	686	1662	1939	128	328	814	886	47	48	49	46
MAN	238	556	1030	1910	157	349	636	942	66	63	62	49
BCL	33	69	585	1702	2	2	2	2	6	3	0	0
17				106660				46178				41

Conclusions

- Are the holo structures the right one for testing machine learning-based tools?

- Ahoj allows for specifc ligand binding site identification of apo/holo structure pairs

- Ahoj-DB – PDB-wide assignment of apo/holo pairs for BioLIP database

Acknowledgements

GAČR
CZECH SCIENCE FOUNDATION

elixir
CZECH REPUBLIC

David Hoksza
Faculty of Mathematics and Physics, Associate Professor

Radoslav Krivák
Faculty of Mathematics and Physics, PhD student

Lukas Jendele

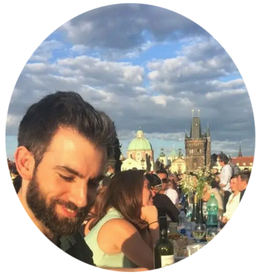
Faculty of Mathematics and Physics, Charles University

Department of Computer Science, ETH Zurich

Lukáš Polák

Faculty of Mathematics and Physics, Charles University

admin (at) lukaspolak.cz



Christos Feidakis
Faculty of Science, PhD student

Petr Škoda
Faculty of Mathematics and Physics, Assistant Professor

Dávid Jakubec
Faculty of Mathematics and Physics, Postdoc

Charles University Structural Bioinformatics Group

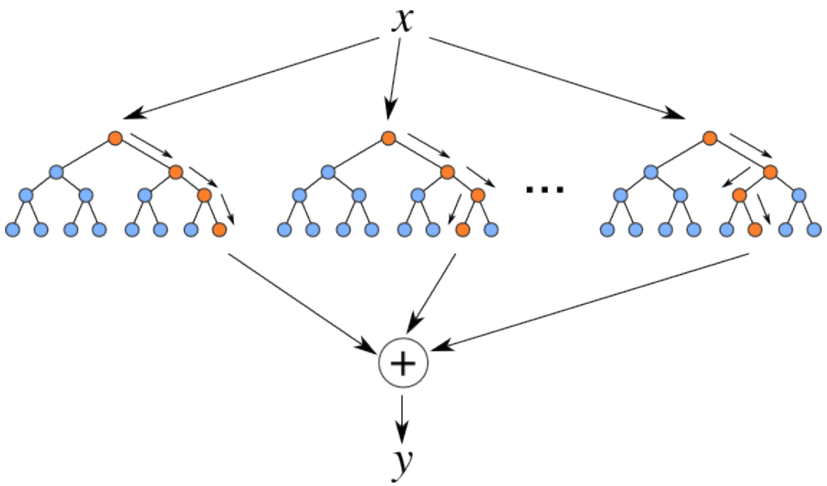
https://bioinformatika.mff.cuni.cz/cusbg/

•

Random Forests classification

- **Ensemble** of decision trees

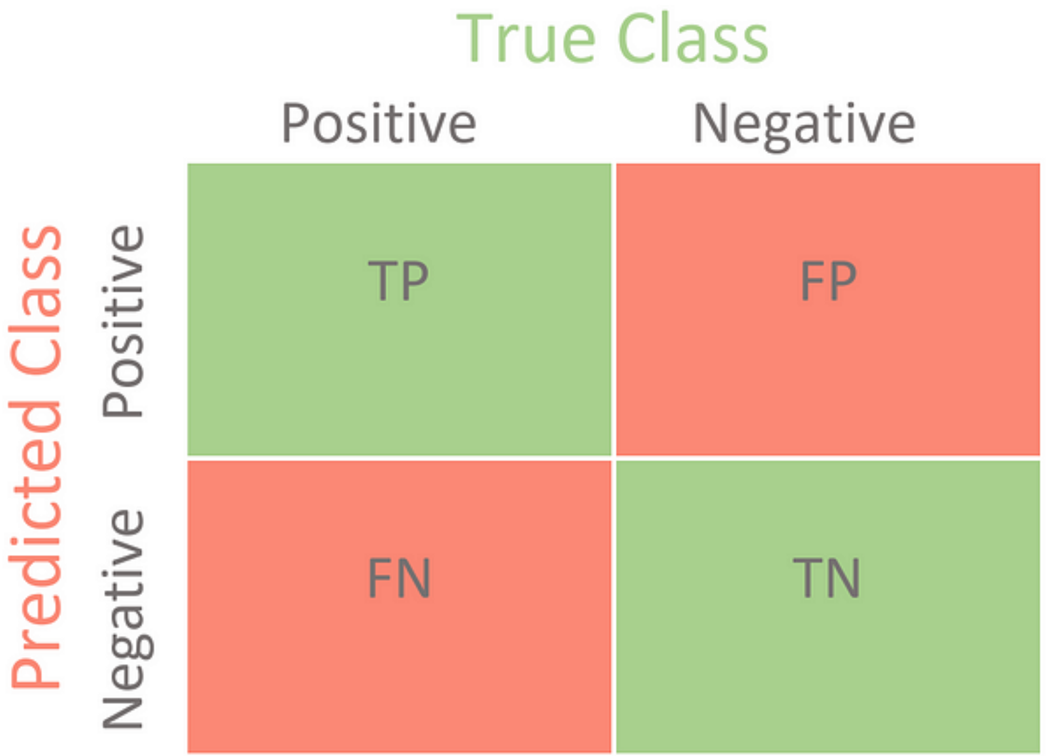
- **Single** decision **tree**
  - Unstable
  - Overfits



- **Solution**
  - Aggregate **multiple decision trees** on **bootstrapped data**
  - **Random choice of descriptors**

- Advantages
  - Suitable for **imbalanced data sets**
  - Estimates of what **variables** are **important** in the classification

Classification evaluation metrics



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 * \frac{Recall * Precision}{Recall + Precision}$$

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