

Binding site identification

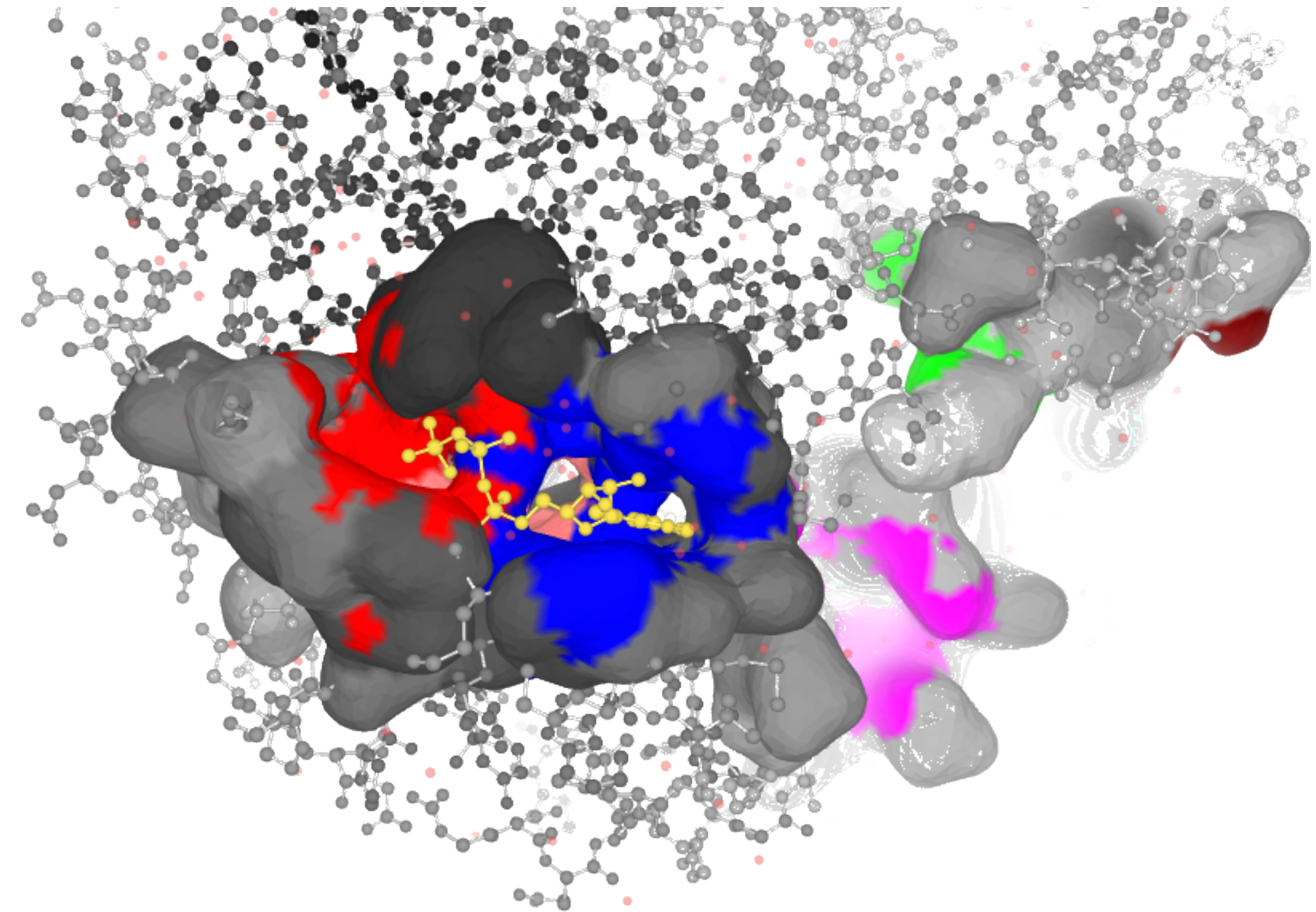
The Pocket Art

7ADD

Marian Novotný

Binding site

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



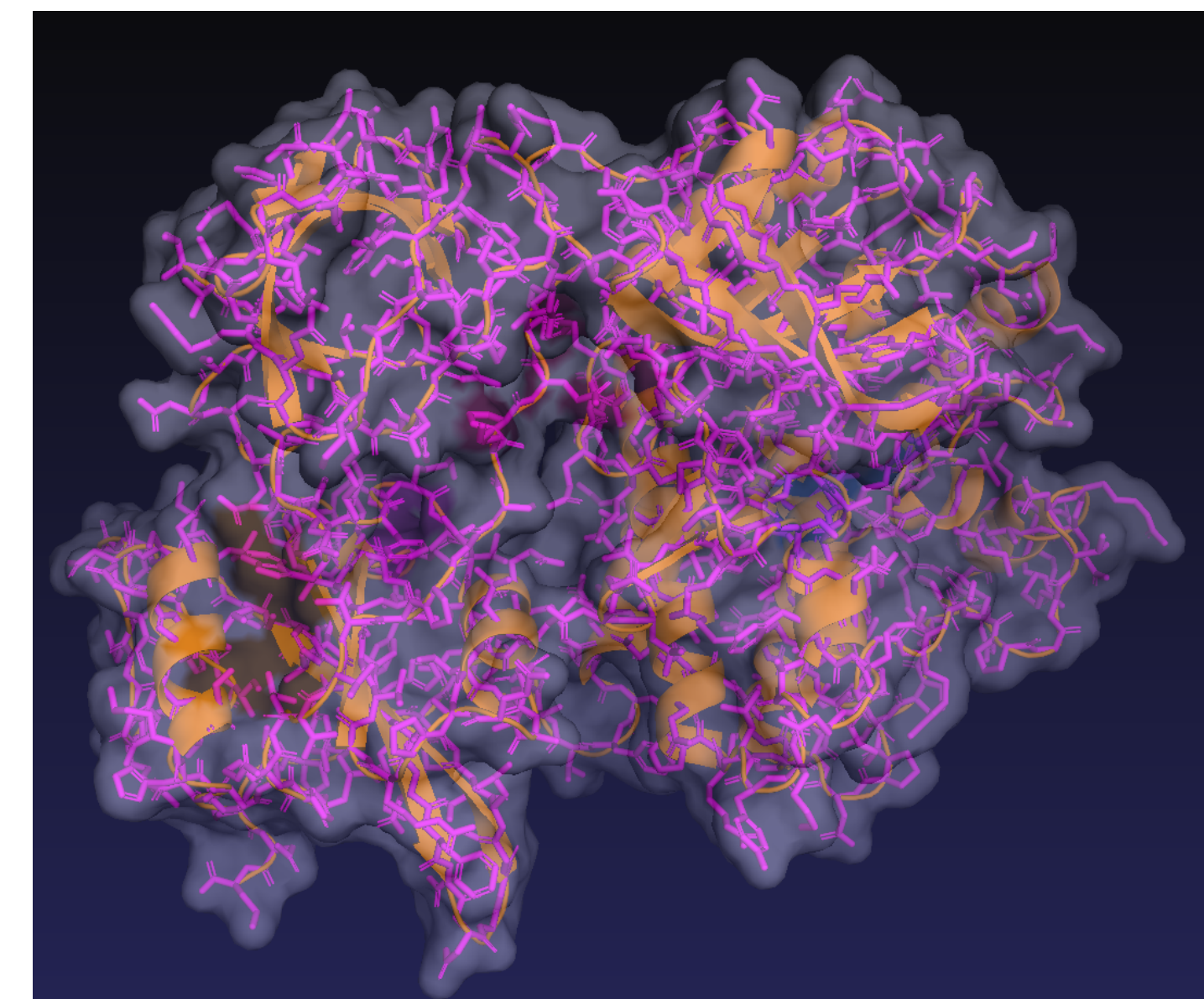
....FPWFGMDIGGTLVKLSYFEPIDITAEEEQEEVES....

Ligand binding sites prediction approaches

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

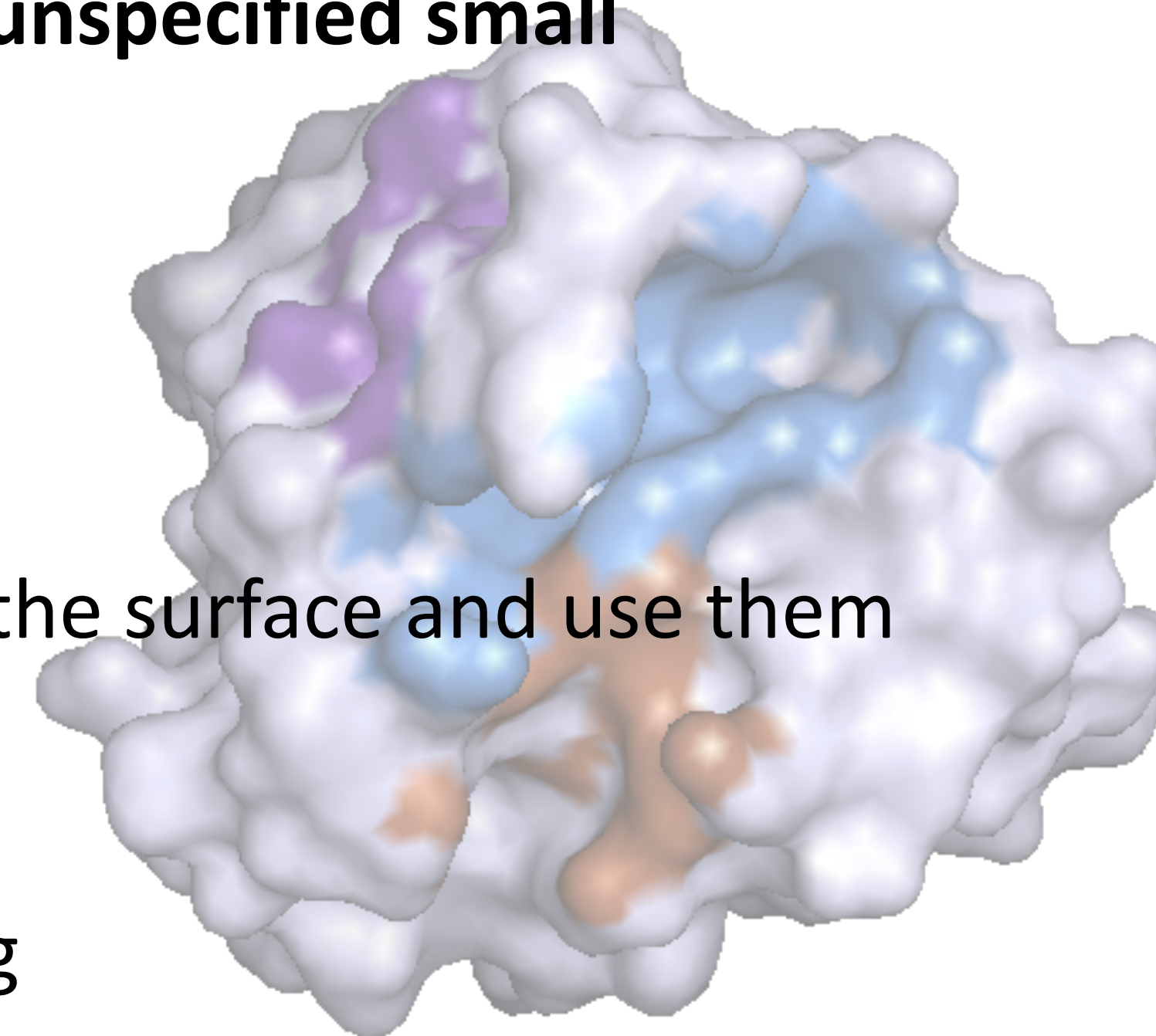
Over 50 methods developed, the comparison difficult!

Q5E940_BOVIN	-----MREDRATWKSNSYELKIIQLDDVYKCFIVGADNVGSKMOMQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--FALE	76
RLA0_HUMAN	-----MREDRATWKSNSYELKIIQLDDVYKCFIVGADNVGSKMOMQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--FALE	76
RLA0_MOUSE	-----MREDRATWKSNSYELKIIQLDDVYKCFIVGADNVGSKMOMQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--FALE	76
RLA0_RAT	-----MREDRATWKSNSYELKIIQLDDVYKCFIVGADNVGSKMOMQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--FALE	76
RLA0_CHICK	-----MREDRATWKSNSYEMKIIQLDDVYKCFIVGADNVGSKMOMQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--FALE	76
RLA0_RANSY	-----MREDRATWKSNSYELKIIQLDDVYKCFIVGADNVGSKMOMQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--FALE	76
Q7ZUG3_BRARE	-----MREDRATWKSNSYELKIIQLDDVYKCFIVGADNVGSKMOMQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--FALE	76
RLA0 ICTPU	-----MREDRATWKSNSYELKIIQLDDVYKCFIVGADNVGSKMOMQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--FALE	76
RLA0_DROME	-----MRENKAAWKAQYIKVYVLFDEPKCFIVGADNVGSKMOMQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--POLE	76
RLA0_DICDI	-----MSAE-SKRKLFTEKATLFTTKMIVAEADVFGSGLKIKRSIRGI-GAVLMGKNTIRVIRDLADSK--PELD	75
Q54LP0_DICDI	-----MSAE-SKRKNVTEKATLFTTKMIVAEADVFGSGLKIKRSIRGI-GAVLMGKNTIRVIRDLADSK--PELD	75
RLA0_PLAF8	-----MAKLSQOQKQMYTEKLSLLOQSKLIVHVDNVGSKMOMQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--FALE	76
RLA0_SULAC	-----MGLAVTTTKKAKKVDVVAELTEKLTHTLIIANIEGFPADKLHDIRKKLRGK-ADIKVTKNLFNIAKKNAG--VDK	79
RLA0_SULTO	-----MRIMAVITQERKIAKWKIEVKELEKLRHTLIIANIEGFPADKLHDIRKKLRGK-ADIKVTKNLFNIAKKNAG--LDVS	80
RLA0_SULSO	-----MKRLALALKQKRVASWKEVKELELTKNSNTLIGNLEGFPADKLHDIRKKLRGK-ADIKVTKNLFNIAKKNAG--LDIE	80
RLA0_AERPE	-----MSVVSIVQMYKREKPIPEKTLMLRELEELFSKRVVLFADLTGEPFVVRVKKLWKK-YPMVAKKRILLRAMEKAGLE--LDDN	86
RLA0_PYRAE	-----MMLAIGKRRYYVRQARKVKIVSEATLLOKPYVFLDLHLSRILHEFYRLLRY-GVILIKPTEKIAKTVYGG--IPAE	85
RLA0_METAC	-----MAEERHHEHHPQWKDEIEMIKELIQSHKVFQMVREGLLTKIKIRDLKQV-AVLKVSNTLLEKAINOLG--ETIP	78
RLA0_METMA	-----MAEERHHEHHPQWKDEIEMIKELIQSHKVFQMVREGLLTKIKIRDLKQV-AVLKVSNTLLEKAINOLG--ETIP	78
RLA0_ARCFU	-----MAAVRS--PEYKRAVEEIKRMISSKVVAIVSRNVVAGDMQKIRREFRGK-AEIKVVKLLEKADALG--EDVL	75
RLA0_METKA	-----MAVKAIGQPSQYEKVAEWRREVKELKEMDEVNGLVDLEGIPAPOLQETRAKLRERD-LIRMSRNTLLEKRAAELEKPELE	88
RLA0_METTH	-----MAHVAEWKKEVQELHDLIKGVVGVANLADIPAROLQKMRQTLRDS-ALIRMSKNTLLEKRAAELEKPELE--ENVD	74
RLA0_METTL	-----MITAESEHKIAPWKIEEWNKLELKNQIIVALVDMMEVPAVLOQETRDTR-DQMLKMSRNTLLEKRAAELEKPELE	82
RLA0_METVA	-----MIDAKSEHKIAPWKIEEWNKLELKNQIIVALVDMMEVPAVLOQETRDTR-DQMLKMSRNTLLEKRAAELEKPELE	82
RLA0_METJA	-----METKVAHVAPWKIEEWNKLELKNQIIVALVDMMEVPAVLOQETRDTR-DQMLKMSRNTLLEKRAAELEKPELE	81
RLA0_PYRAB	-----MAHVAEWKKEVEELANIKSVPVVALVDSMPAYLSQMRRLIRENGLLRVSRNTLLEKRAAELEKPELE	77
RLA0_PYRHO	-----MAHVAEWKKEVEELANIKSVPVVALVDSMPAYLSQMRRLIRENGLLRVSRNTLLEKRAAELEKPELE	77
RLA0_PYRFU	-----MAHVAEWKKEVEELANIKSVPVVALVDSMPAYLSQMRRLIRENGLLRVSRNTLLEKRAAELEKPELE	77
RLA0_PYRKO	-----MAHVAEWKKEVEELANIKSVPVVALVDSMPAYLSQMRRLIRENGLLRVSRNTLLEKRAAELEKPELE	76
RLA0_HALMA	-----MSAESEKRTETIPWKQEVDAIVEMIESVSGVVNLAGIPERLQDMRRDLHGT-AELRVSNTLLEKRAAELEKPELE--DGLE	79
RLA0_HALVO	-----MSAESEVRQTEVPEWKQEVDAIVEMIESVSGVVNLAGIPERLQDMRRDLHGT-AELRVSNTLLEKRAAELEKPELE--DGFE	79
RLA0_HALSA	-----MSAESEFRTTEVPEWKQEVDAIVEMIESVSGVVNLAGIPERLQDMRRDLHGT-AELRVSNTLLEKRAAELEKPELE--DGLD	79
RLA0_THEAC	-----MKEVSOQKELWNETTRIKASRSVAIVDAGIRIRIDTRGKNRGK-INLKVTKLLEKRAAELEKPELE--EKLS	72
RLA0_THEVO	-----MRKINPKKEIYSELQETKSKAVAVDLEKVRIRMDIRAKRQK-VKIVVKLLEKRAAELEKPELE--EKLT	72
RLA0_PICTO	-----MTEPQWKIDFVKNLENEINSKVAIVSFKLRNNEFKIENSIRDK-ARIVKLEKRAAELEKPELE--NNV	72
RLA0_ruler	1.....10.....20.....30.....40.....50.....60.....70.....80.....90	



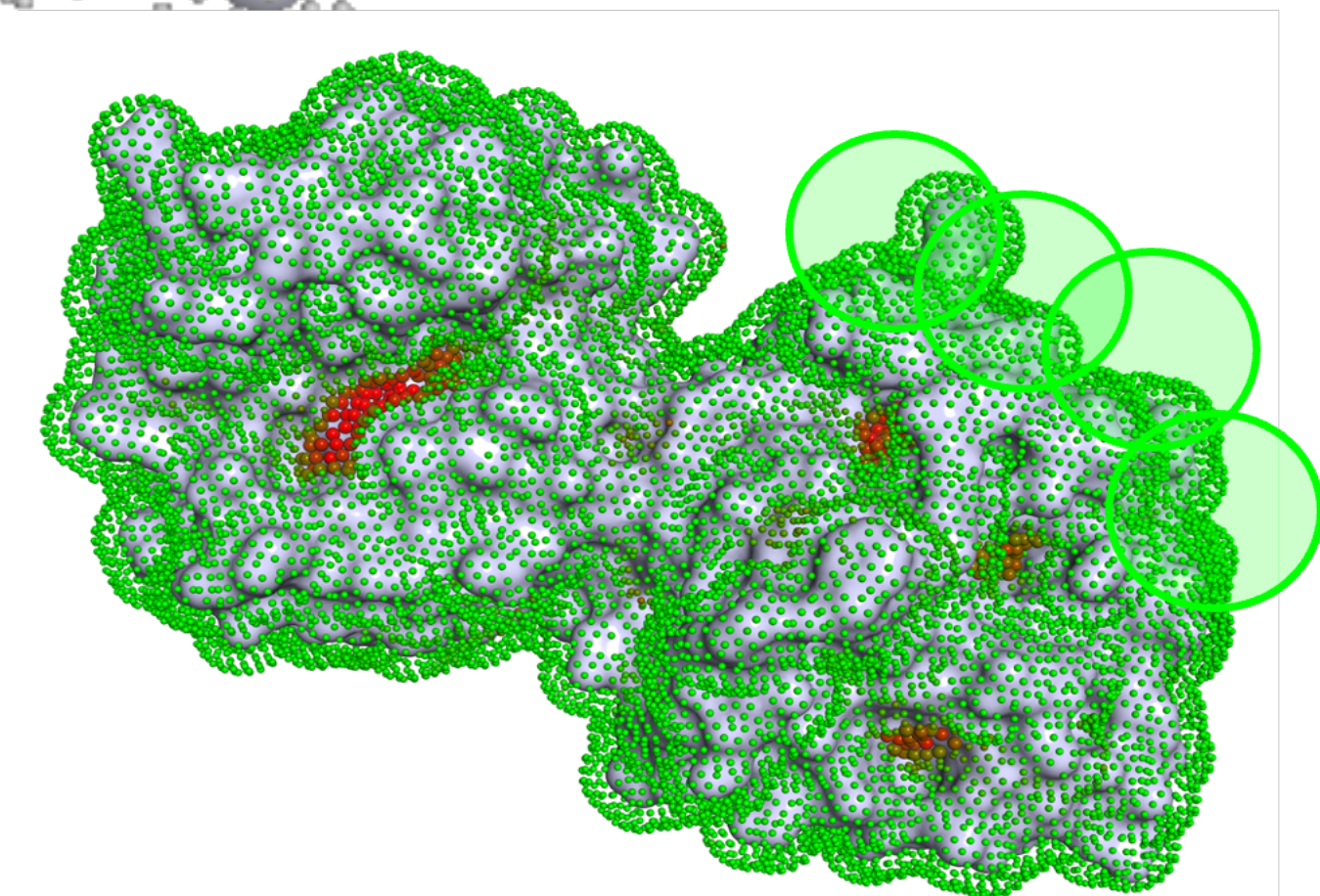
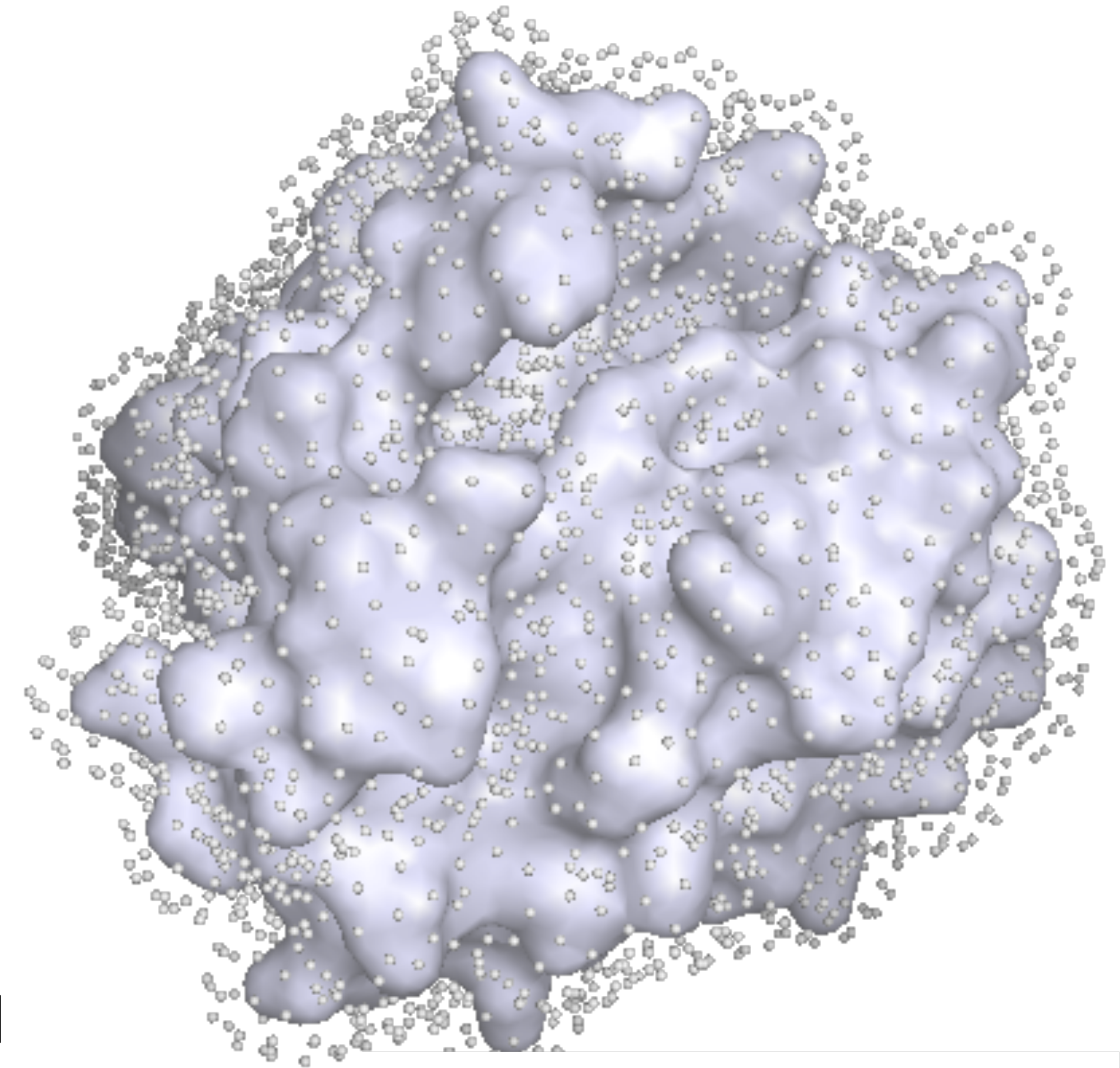
P2RANK

- 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



Model construction

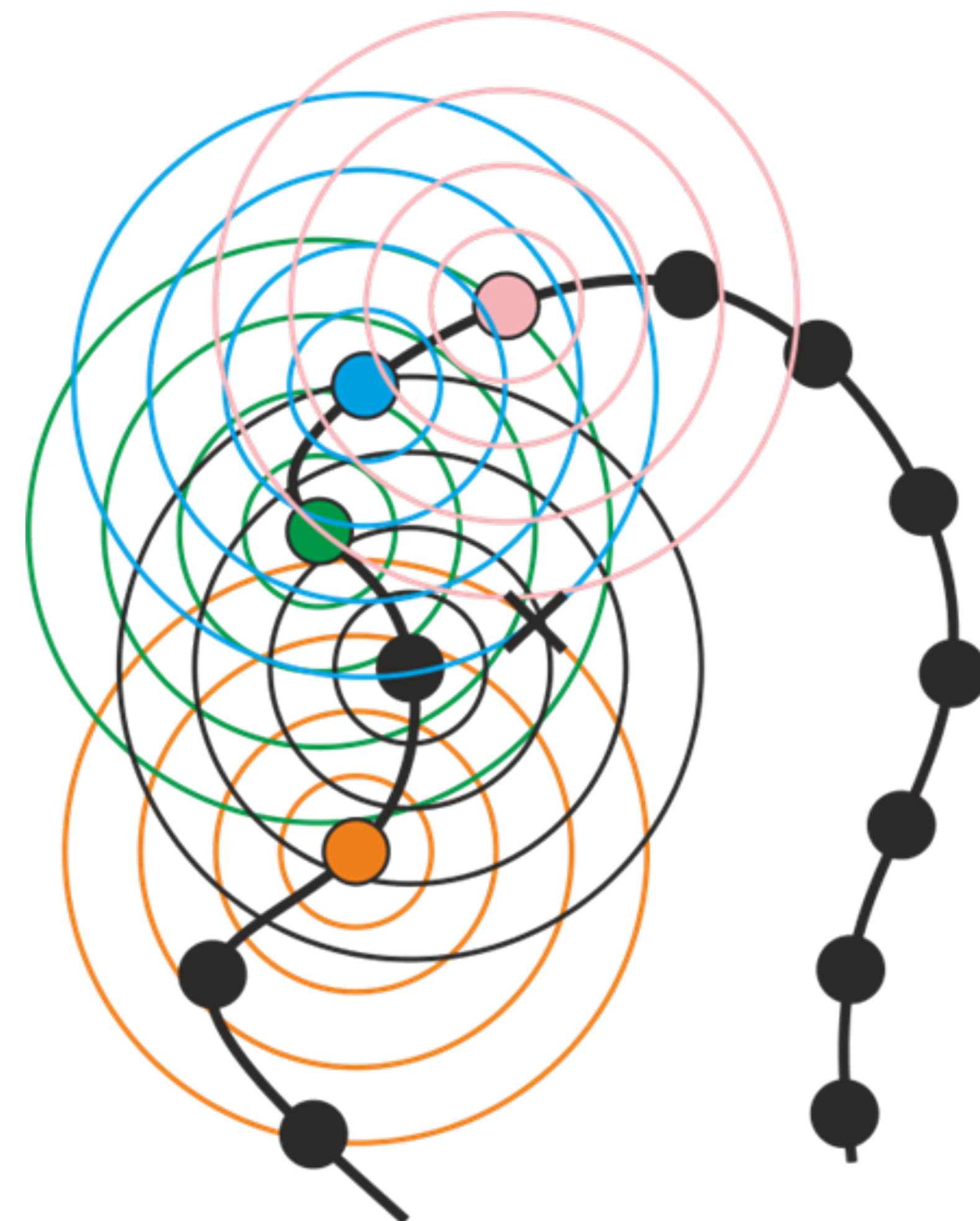
1. Obtaining known protein-ligand complexes
2. Cover the surface with a **mesh of points** (solvent accessible surface – SAS points)
3. **Label points** as binding/nonbinding
4. **Extract a vector of physico-chemical and structural features** for each SAS point of each protein
5. **Build a model**, which 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

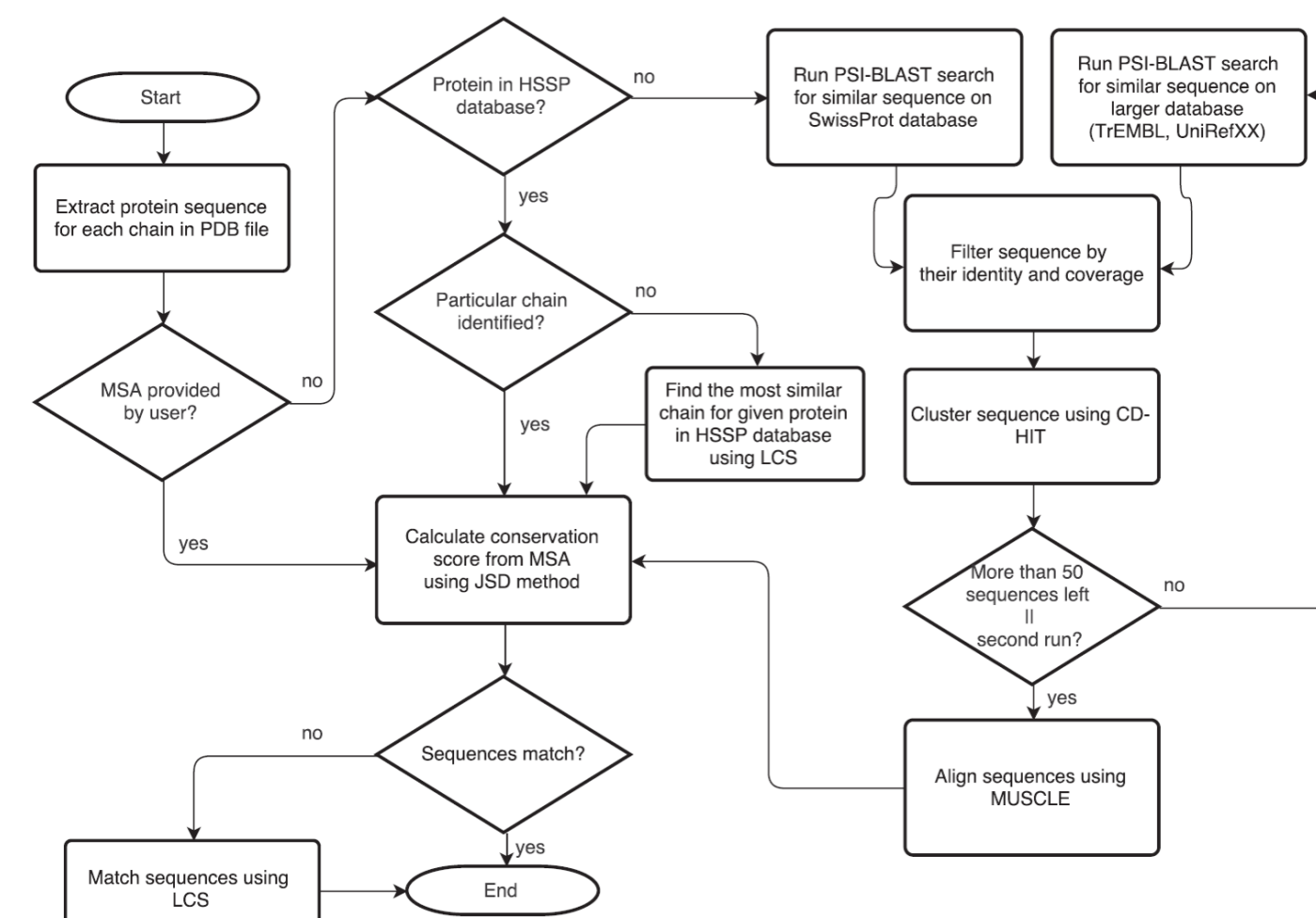
$$\text{IFV}(P) = \sum_{A_i \in A(P)} \text{AFV}(A_i) \cdot w(\text{dist}(P, A_i)) \quad || \quad \text{FV}(P)$$



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]
apRawValid	a	AT tab.	Ligand binding propensity for biologically valid ligands [4]
apRawInvalid	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

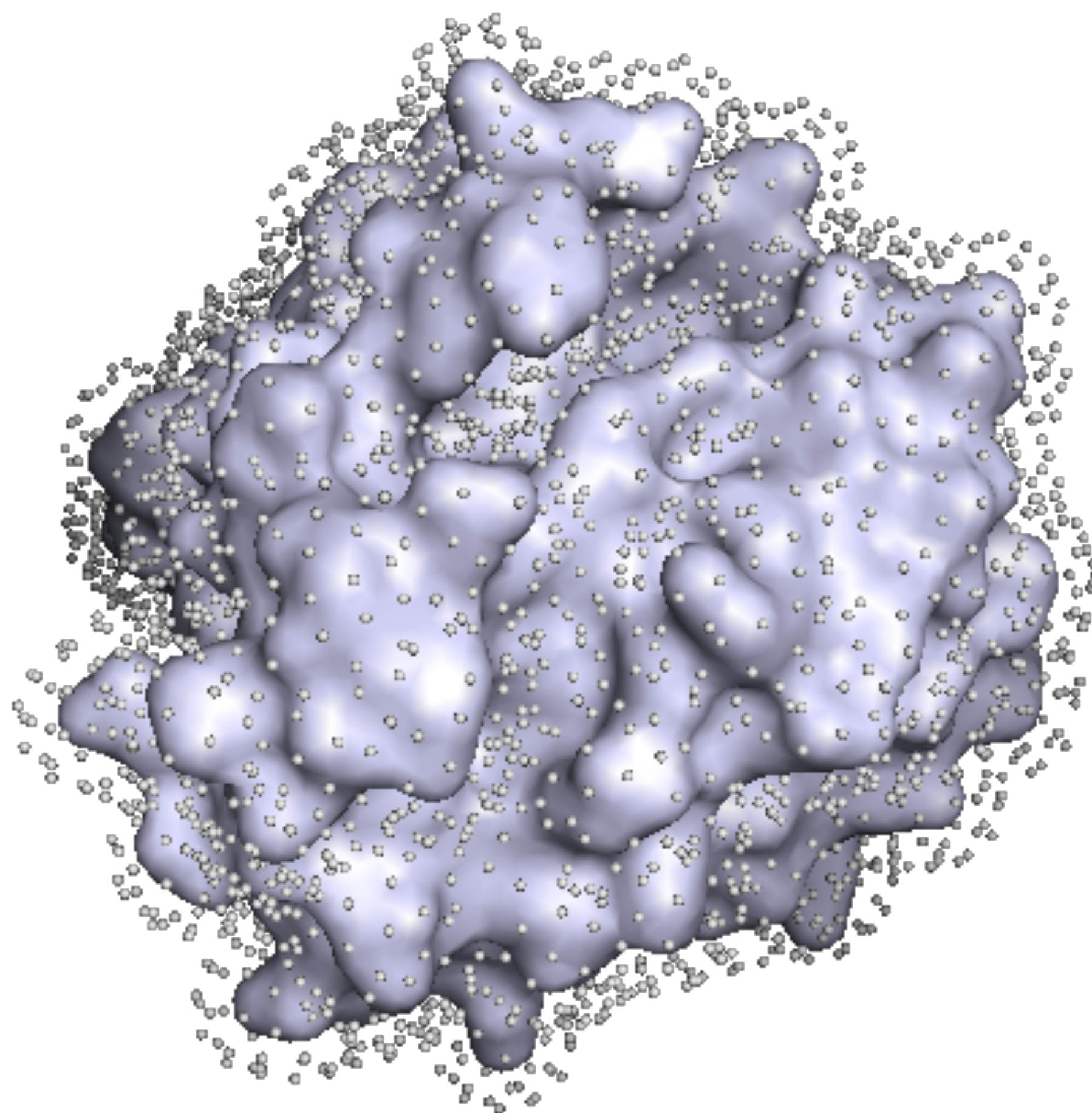
Features extraction - conservation

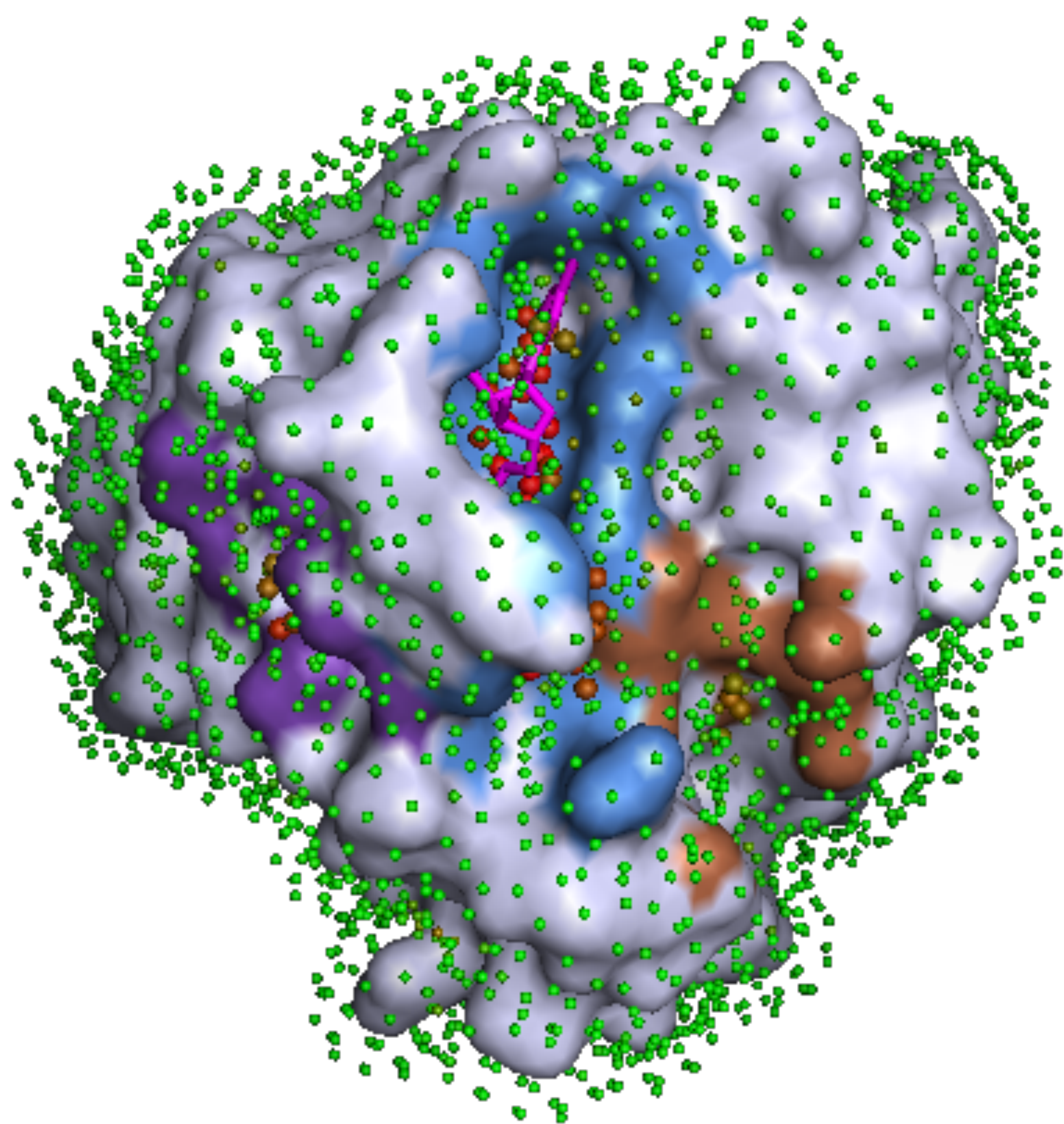
Q5E940	BOVIN	-----MPREDRATWKS	NYFLKIIQLDDY	PKCFIVGADNV	GSKMQQIRMS	LRGK-AVVL	MGKNTMMRKAIRGHLENN--PALE	76						
RLA0	HUMAN	-----MPREDRATWKS	NYFLKIIQLDDY	PKCFIVGADNV	GSKMQQIRMS	LRGK-AVVL	MGKNTMMRKAIRGHLENN--PALE	76						
RLA0	MOUSE	-----MPREDRATWKS	NYFLKIIQLDDY	PKCFIVGADNV	GSKMQQIRMS	LRGK-AVVL	MGKNTMMRKAIRGHLENN--PALE	76						
RLA0	RAT	-----MPREDRATWKS	NYFLKIIQLDDY	PKCFIVGADNV	GSKMQQIRMS	LRGK-AVVL	MGKNTMMRKAIRGHLENN--PALE	76						
RLA0	CHICK	-----MPREDRATWKS	NYFMKIIQLDDY	PKCFIVGADNV	GSKMQQIRMS	LRGK-AVVL	MGKNTMMRKAIRGHLENN--PALE	76						
RLA0	RANSY	-----MPREDRATWKS	NYFLKIIQLDDY	PKCFIVGADNV	GSKMQQIRMS	LRGK-AVVL	MGKNTMMRKAIRGHLENN--PALE	76						
Q7ZUG3	BRARE	-----MPREDRATWKS	NYFLKIIQLDDY	PKCFIVGADNV	GSKMQQIRMS	LRGK-AVVL	MGKNTMMRKAIRGHLENN--PALE	76						
RLA0	ICTPU	-----MPREDRATWKS	NYFLKIIQLDDY	PKCFIVGADNV	GSKMQQIRMS	LRGK-AVVL	MGKNTMMRKAIRGHLENN--PALE	76						
RLA0	DROME	-----MVRENKA	AWKAQYFIKVV	ELDFEPPKCFIV	GADNVGSKMOM	IRTSLRGL-AVVL	MGKNTMMRKAIRGHLENN--PQLE	76						
RLA0	DICDI	-----MSGAG-S	KRKKLFIEKAT	KLFTTYDKMIV	AEADFGSSQLOK	IRKSIRGI-GAVL	MGKNTMIRKVI	IRDLADSK--PELD	75					
Q54LP0	DICDI	-----MSGAG-S	KRKNVFEKAT	KLFTTYDKMIV	AEADFGSSQLOK	IRKSIRGI-GAVL	MGKNTMIRKVI	IRDLADSK--PELD	75					
RLA0	PLAF8	-----MAKLS	QKQKQMYEKL	SSLIQYKILIV	HVDNVGSKMAS	VRKSLRGK-ATIL	MGKNTIRIR	TALKKNLQAV--PQIE	76					
RLA0	SULAC	-----MIGLAV	TTTKKIAK	WVDEVAELTEK	LKTKHTIIIANIE	GFPADKLHE	IRKKLRGK-ADIK	VTKNNL	FNIALKNAG----YDVK	79				
RLA0	SULTO	-----MRIMAV	ITQERKIA	KWKEEVEKLEOK	REYHTIIIANIE	GFPADKLHD	IRKKMRGM-AEIK	VTKNTL	FGIAAKNAG----LDVS	80				
RLA0	SULSO	-----MKRLA	LALKQRK	VASWVKEEVEKLE	TELKNSNTILIGN	LEGFPADKLHE	IRKKLRGK-ATIK	VTKNTL	FKIAAKNAG----IDIE	80				
RLA0	AERPE	-----MSVVS	LVGQMYKRE	PIPEWKT	MLRELEELFSK	RVVLFADLTG	PTFVVQ	RVRK	LWKK-YPM	VAKKRIIL	RAMKAAGLE---LDDN	86		
RLA0	PYRAE	-----MMLA	IGKRRYV	TRQYARKVK	IVSEATELLQ	KYPYVFL	DLHGLSSRILHE	YRRLRRY-GVI	KIKP	PLFKIA	FTKVYGG---IPAE	85		
RLA0	METAC	-----MAEER	HTEHIPQ	WKDEIENIKEL	IQSHKVF	GMVIEGILAT	KMKIRRDLDV-AVL	KVSRNT	LTERAL	NQLG----ETIP	78			
RLA0	METMA	-----MAEER	HTEHIPQ	WKDEIENIKEL	IQSHKVF	GMVIEGILAT	KIKIRRDLDV-AVL	KVSRNT	LTERAL	NQLG----ESIP	78			
RLA0	ARCFU	-----MAAVR	GS---PPEY	KVRAVEEIKRM	ISSKPVVAIV	SFRNVPAGOM	KIRREFRGK-AEIK	VVKNTL	LERAL	DALG----GDYL	75			
RLA0	METKA	-----MAVKA	KQPPSGYE	PKVAE	WRREVEKLEELM	DEVENGLVDLE	GIPAPLOE	IRAKLRERD	IIRMS	RNTLMR	IALEEKLDER--PELE	88		
RLA0	METTH	-----MAHVA	EWKKEVE	QLHDLIKGYE	VVGIANLAD	IPARQLOK	MRQTLRDS-ALIR	MSKKT	LISL	ALEKAG	REL--ENVD	74		
RLA0	METTL	-----MITAE	SEHKIAP	WKIEEVN	KLKELLKNGQ	IVALVDMME	VPARQLOE	IRDKIR-CTM	LKMS	RNTLIE	RAI	KEVAEETGNPEFA	82	
RLA0	METVA	-----MIDAK	SEHKIAP	WKIEEVN	KLKELLKSN	VIALIDMME	VPAVQLOE	IRDKIR-DQ	MTLK	MSRNTL	IKR	AVEEVAEETGNPEFA	82	
RLA0	METJA	-----METK	VKAHVAP	WKIEEVKTL	KGLIKSKPV	VVAIVD	MMDVPAVQLOE	IRDKIR-DK	VKL	RMSRNTL	IRAL	KEAAEELNPKLA	81	
RLA0	PYRAB	-----MAHVA	EWKKEVEEELANL	IKSYPIAL	VDVSSMPAY	PLSQMRR	IRENG	LLRVS	RNTLIE	LAIKKAA	QELGKPELE	77		
RLA0	PYRHO	-----MAHVA	EWKKEVEEELAKL	IKSYPIAL	VDVSSMPAY	PLSQMRR	IRENG	LLRVS	RNTLIE	LAIKKAA	KELGKPELE	77		
RLA0	PYRFU	-----MAHVA	EWKKEVEEELANL	IKSYPIAL	VDVSSMPAY	PLSQMRR	IRENN	GLLRVS	RNTLIE	LAIKKVA	QELGKPELE	77		
RLA0	PYRKO	-----MAHVA	EWKKEVEEELANI	IKSYPIAL	VDVAVP	PAYPLSK	MRDKLE-GK	ALLRVS	RNTLIE	LAIKKAA	QELGQPELE	76		
RLA0	HALMA	-----MSAES	RKTETIP	EWKQEEV	DAIVEMIES	YESVGV	VNIAGIPSR	LODMRRD	LHGT-AEL	RVS	RNTLLE	RALDDVD---DGLE	79	
RLA0	HALVO	-----MSESE	VQTEVIP	QWKREEV	DELVDV	IESYESV	GVVGVAGIP	SRLOSMR	RELHGS-AAV	RMSRNT	LVNRAL	DEVN---DGFE	79	
RLA0	HALSA	-----MSAEE	QRTTEEV	PEWKRQ	EVAELVDL	LETYDS	VGVVNTGIP	SKLODMRR	GLHGQ-AAL	RMSRNT	LLVRALE	EAG---DGLD	79	
RLA0	THEAC	-----MKEV	SQKKELVNEIT	ORIKAS	RSVAIVDLAG	IRTRQIOD	IRGKNRGK-INL	KVIK	TL	LFKALE	NLGD---EKLS	72		
RLA0	THEVO	-----MRKIN	PKKKEIV	SELAOD	ITKSKAV	VDIKG	VTRMOD	IRAKNR	DK-VK	KVVK	TL	LFKALDS	IND---EKLT	72
RLA0	PICTO	-----MTEPA	QWKIDFV	KNLENE	INSR	KAIVS	IKLRNNE	FOKIRNS	IRDK-ARI	KVS	RARLL	RAL	ENLTKG---NNIV	72
ruler 1.....10.....20.....30.....40.....50.....60.....70.....80.....90														



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



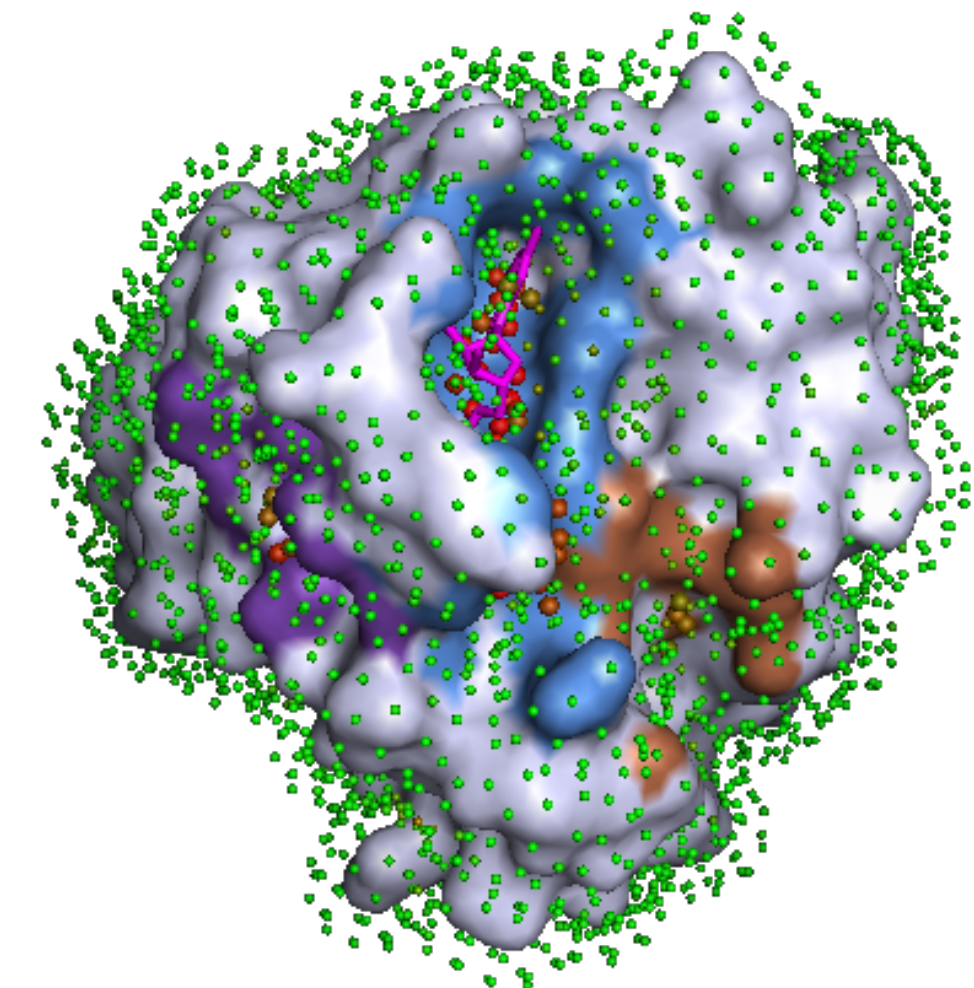


Binding sites prediction evaluation

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

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN



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

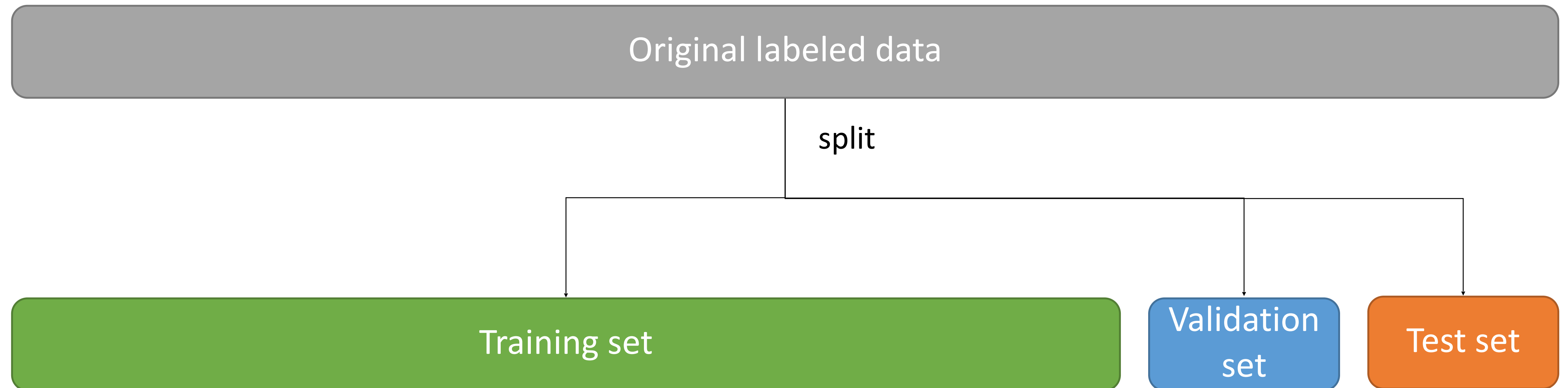
• **DCA(4)** ... distance from the **center** of the pocket to **any ligand atom** $\leq 4 \text{ \AA}$

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) Ligsitescs: 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	<u>72.0</u>	<u>78.3</u>	<u>68.6</u>	<u>74.0</u>

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.

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

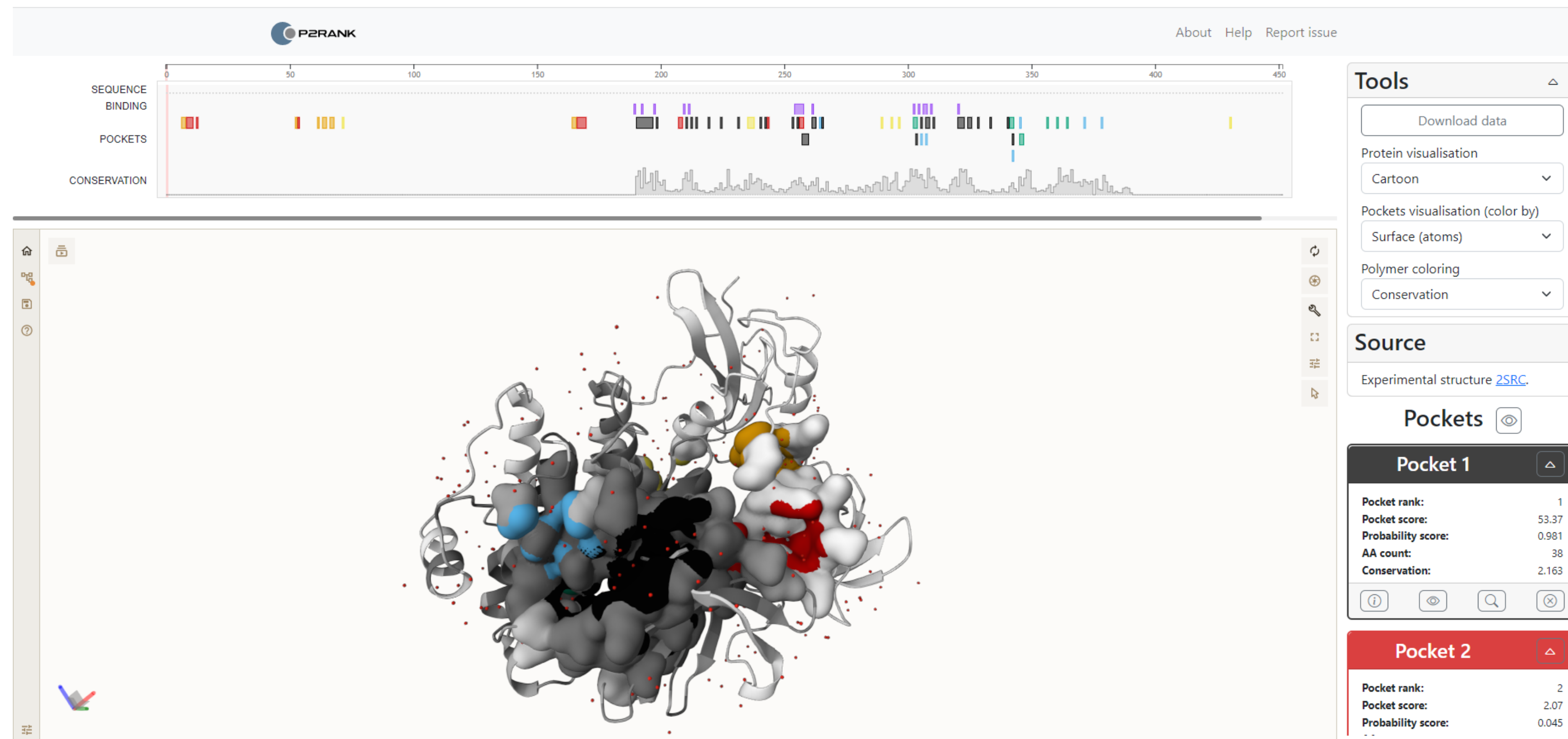
Availability

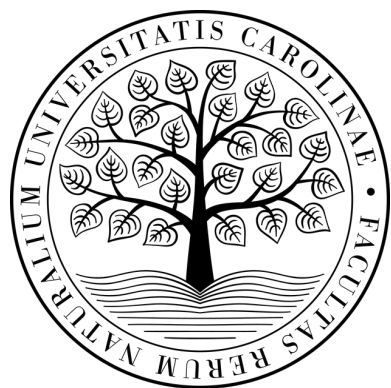
- **Command line app**

- <https://github.com/rdk/p2rank>
- Java, PyMOL

- **Webserver**

- <https://prankweb.cz>
- AlphaFold, conservation, API





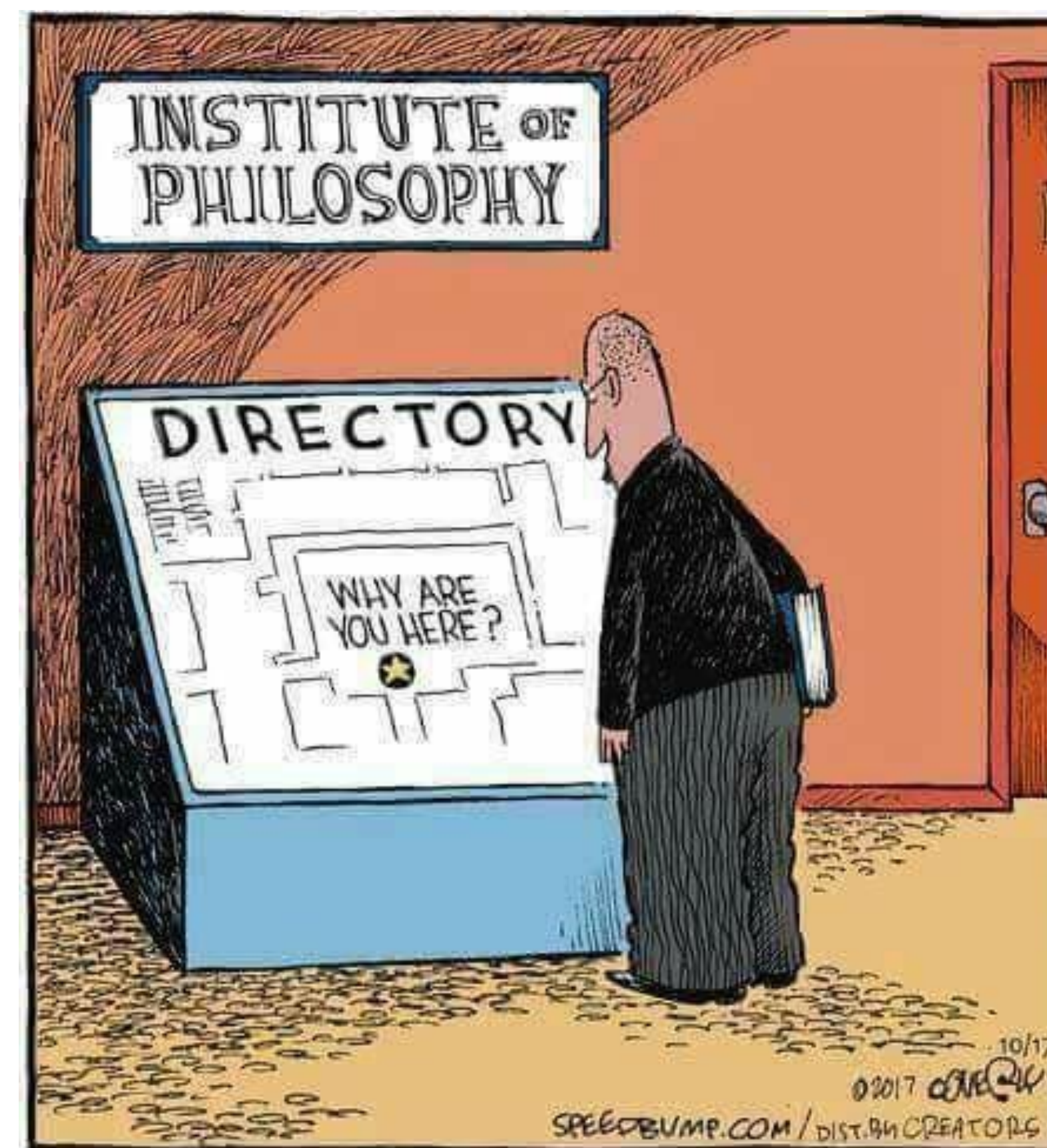
AHoJ-DB:
A PDB-wide
assignment of apo &
holo relationships
based on
individual protein-
ligand interactions



Motivation for AHoJ – can we test better?

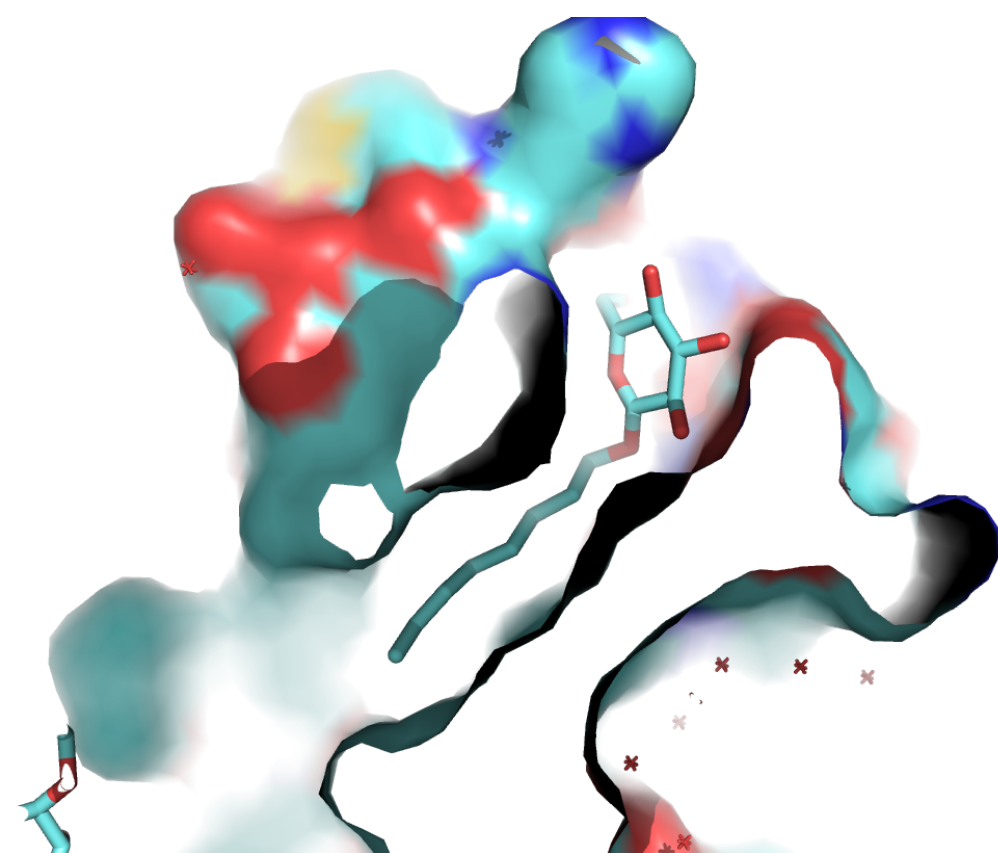
- A. Suspiciously good results from P2Rank -> need for harder/more realistic targets (= apo structures)
- B. No apo-holo dataset in literature

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

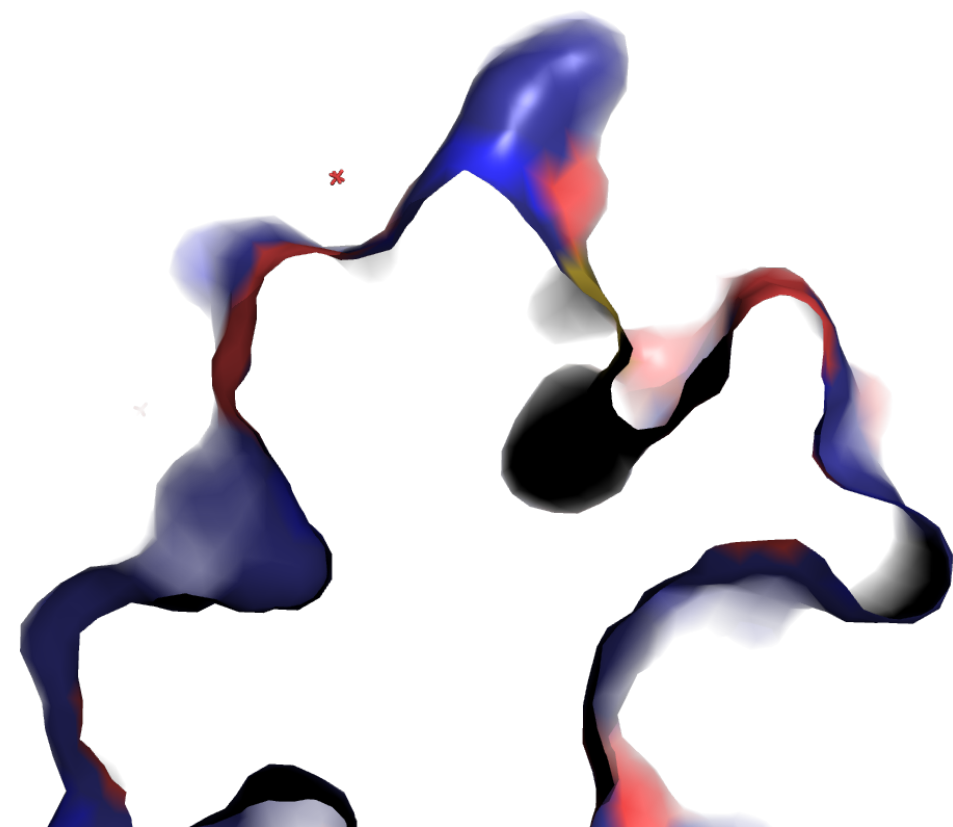


Cryptic binding sites

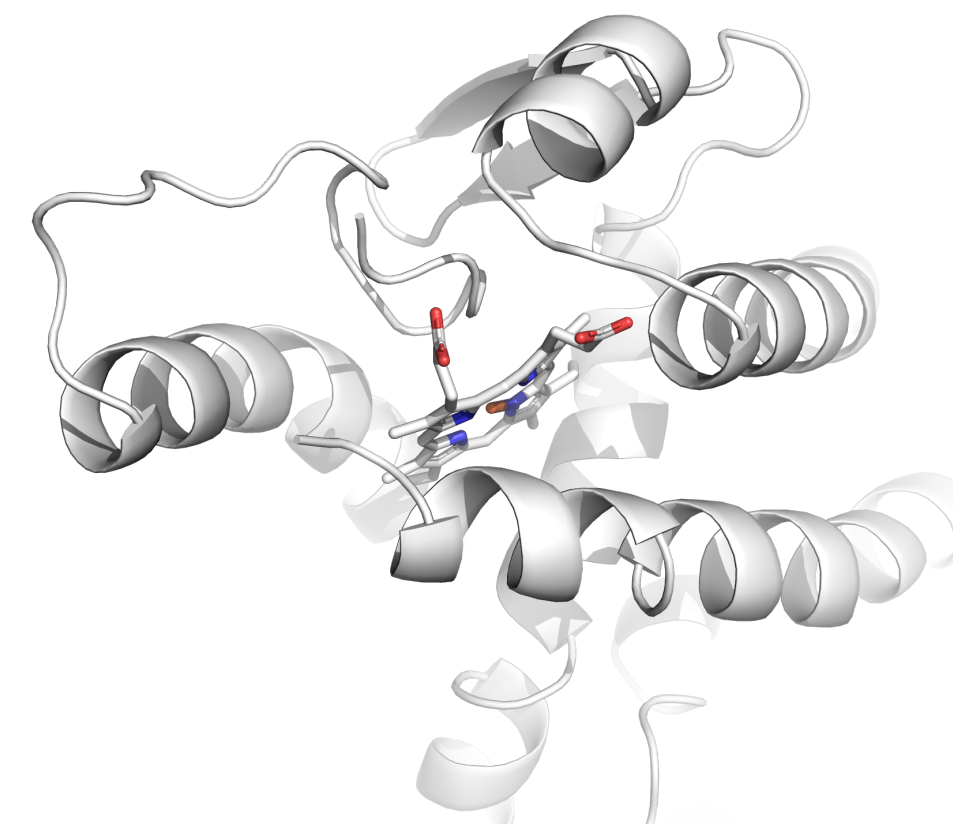
2npq A BOG



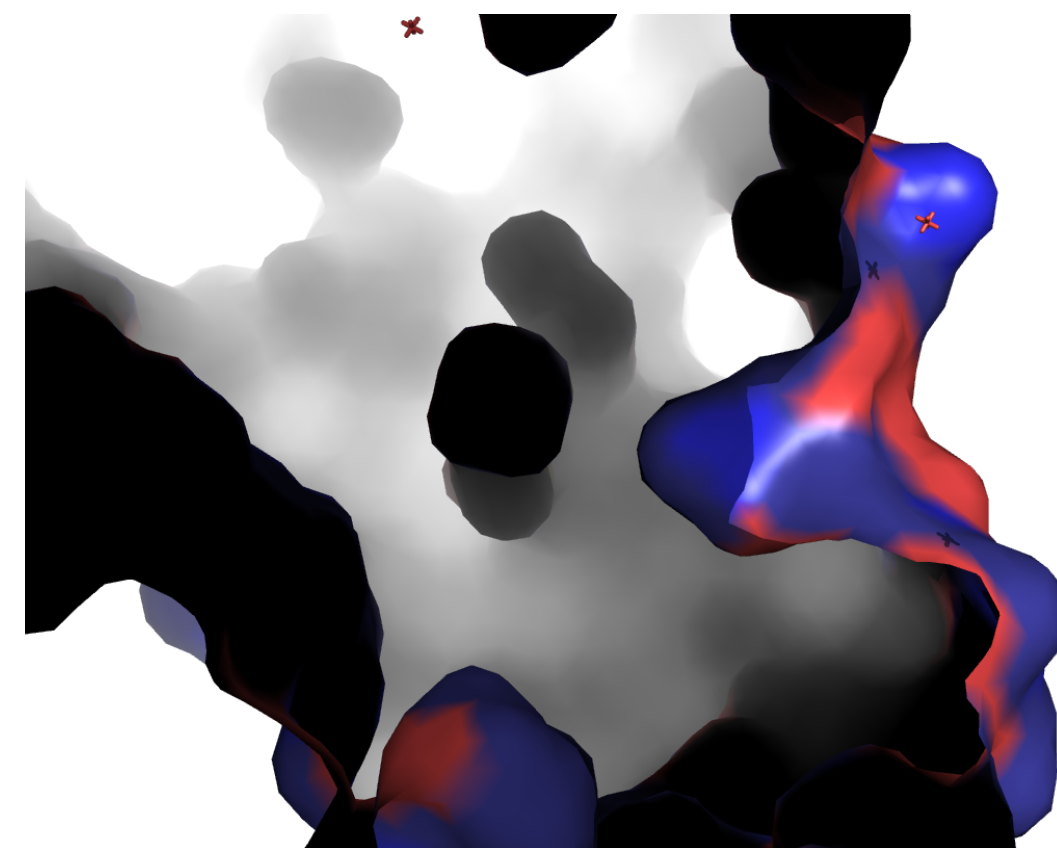
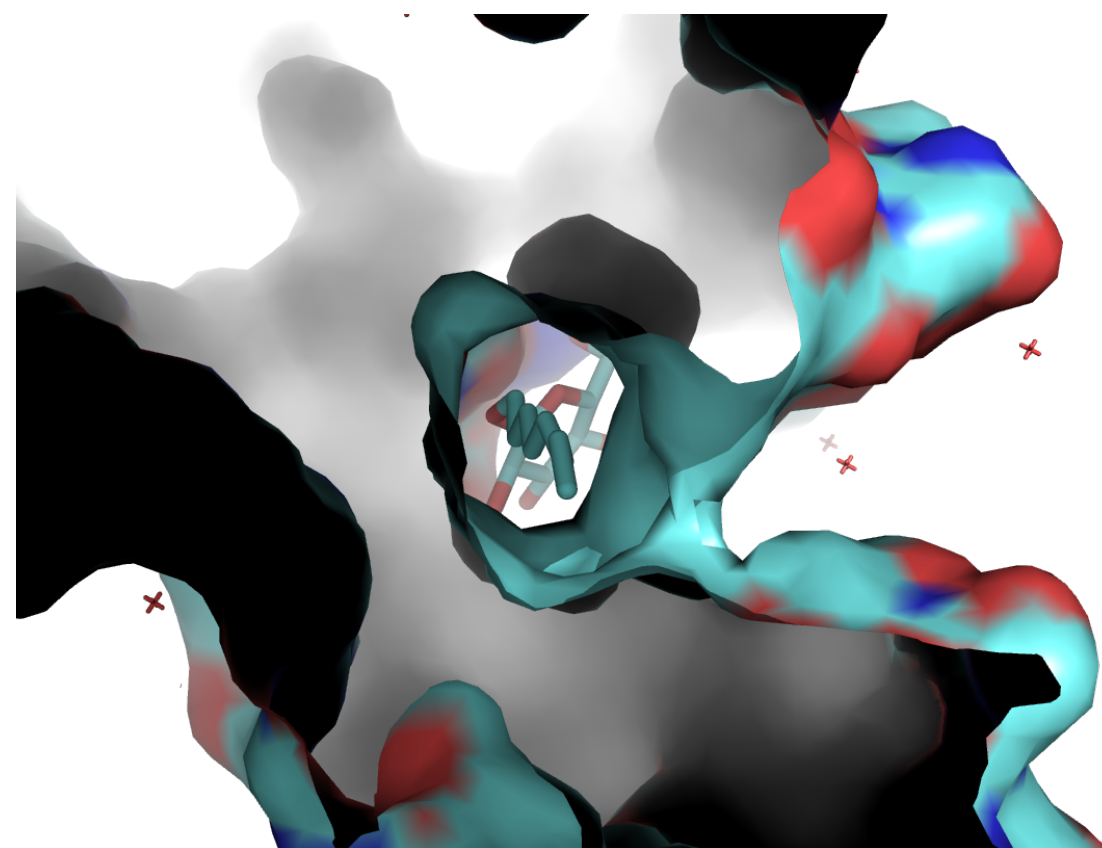
6sfi A



3cqy A HEM



2v0v A



[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
- Ligand-free sites
- Consider water as ligand
- Consider non-standard residues as ligands
- Consider D-amino acids as ligands
- Save aligned Apo chains
- Save aligned Holo Chains

Binding residues threshold: %Sequence overlap threshold: %Resolution threshold: ÅMinimum TM-score: Ligand scanning radius: Å[Submit Job](#)

Job: 4QT6H

Query: 2SRC A PTR

Status: done

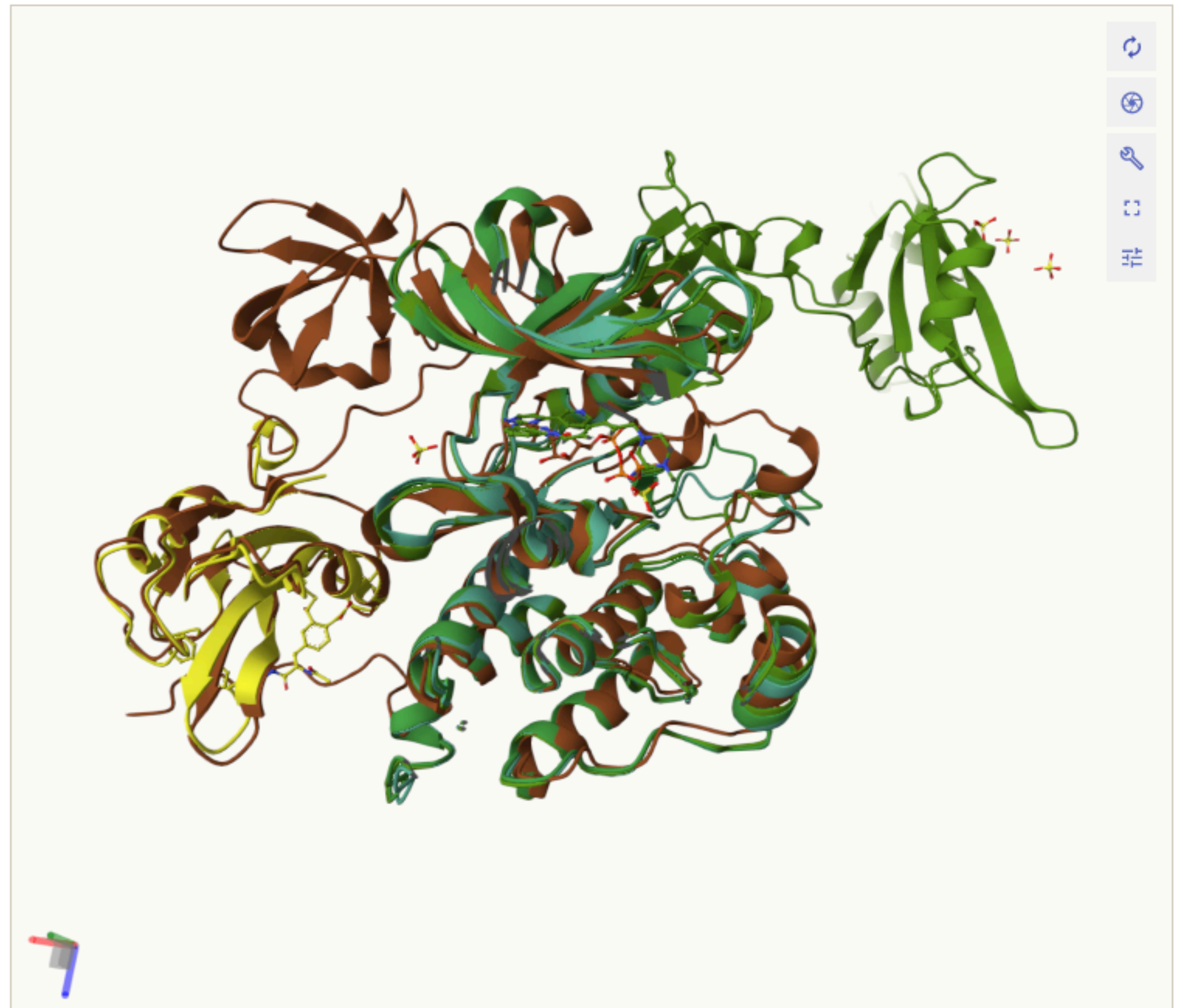
Download

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	👁



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> |



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

Zhang Lab UNIVERSITY OF MICHIGAN

Home **Research** COVID-19 Services Publications People Teaching Job Opening News Forum Lab Only

Online Services

- I-TASSER
- I-TASSER-MTD
- C-I-TASSER
- CR-I-TASSER
- QUARK
- C-QUARK
- LOMETS
- MUSTER
- CEthreader
- SEGMER
- DeepFold
- DeepFoldRNA
- FoldDesign
- COFACTOR
- COACH
- MetaGO
- TripletGO
- IonCom

BioLiP² for Ligand-protein binding database

HOME SEARCH BROWSE LIGAND COACH DOWNLOAD HELP

BioLiP is a semi-manually curated database for high-quality, biologically relevant ligand-protein binding interactions. The structure data are collected primarily from the [Protein Data Bank \(PDB\)](#), with biological insights mined from literature and other specific databases. BioLiP aims to construct the most comprehensive and accurate database for serving the needs of ligand-protein docking, virtual ligand screening and protein function annotation. Questions about the BioLiP Database can be posted at the [Service System Discussion Board](#).

Since ligand molecules (e.g., Glycerol, Ethylene glycol) are often used as [additives](#) (i.e., false positives) for solving the protein structures, not all ligands present in the PDB database are biologically relevant. BioLiP uses a [composite automated and manual procedure](#) for examining the biological relevance of ligands in the PDB database. Each entry in BioLiP contains a comprehensive list of annotations on:

- ligand-binding residues;
- ligand binding affinity (from the original literature, plus [Binding MOAD](#), [PDBbind-CN](#), [BindingDB](#));
- catalytic site residues (mapped from [Mechanism and Catalytic Site Atlas](#));
- [Enzyme Commission](#) (EC) numbers and [Gene Ontology](#) (GO) terms mapped by the [SIFTS](#) database;
- crosslinks to external databases, including [RCSB PDB](#), [PDBe](#), [PDBj](#), [PDBsum](#), [Binding MOAD](#), [PDBbind-CN](#), [Mechanism and Catalytic Site Atlas](#), [QuickGO](#), [ExpASY](#)



AHoJ-DB

PREVIEW

Number of entries: **272,716**

Database of precomputed Apo-Holo search results for (almost) every **protein chain** and **ligand** in the **PDB**. Each entry represents a target PDB chain and a bound ligand and contains a list of matching chains that are labeled as **APO** or **HOLO** with respect to the **binding site** defined by the ligand.

Search AHoJ-DB

Specify one or more of the following:

[example1](#) [example2](#) [example32](#)

PDB_ID/s:

UniProt_ID/s:

Ligand/s:

Filters

 X-ray structures only Exclude NMR structuresResolution threshold: Å



Download

Search Results Summary

Entries

Entry

1fxv-B-PNN-1001

1gm7-B-PNN-1577

1uob-A-PNN-1311

1uof-A-PNN-1312

3huo-A-PNN-300

3huo-A-PNN-302

3huo-A-PNN-303

3huo-A-PNN-304

3huo-B-PNN-301

Found APO Chains / Structures: 3 / 3

Chain	AT	Res	SO	MBR	RMSD	TM-sc	Lig	👁
1gkf / B	B	1.41	100.0	100.0	0.39	0.998		👁
1pnk / B	B	1.9	100.0	100.0	0.47	0.997		👁
1jx9 / B	B	2.28	100.0	100.0	0.58	0.996		👁

Found HOLO Chains / Structures: 20 / 20

Chain	AT	Res	SO	MBR	RMSD	TM-sc	Lig	👁
1gk9 / B	B	1.3	100.0	100.0	0.38	0.998	EDO	👁
1gm7 / B	B	1.45	100.0	100.0	0.39	0.998	EDO PNN	👁
1e3a / B	B	1.8	100.0	100.0	0.4	0.998	EDO	👁
1gm9 / B	B	1.8	100.0	100.0	0.35	0.998	EDO SOX	👁
1fxh / B	B	1.97	100.0	100.0	0.17	1.0	PAC	👁
1gm8 / B	B	2.0	100.0	100.0	0.35	0.998	SOX	👁
1h2g / B	B	2.0	100.0	100.0	0.39	0.998	EDO	👁
1ajq / B	B	2.05	100.0	100.0	0.45	0.997	SPA	👁
1k7d / B	B	2.15	100.0	100.0	0.59	0.996	GRO	👁
1kec / B	B	2.3	100.0	100.0	0.58	0.996	GRO	👁
1ain / B	B	2.31	100.0	100.0	0.49	0.997	OMD	👁

3huo / A

Q9L5C8

1.5

3huo / A

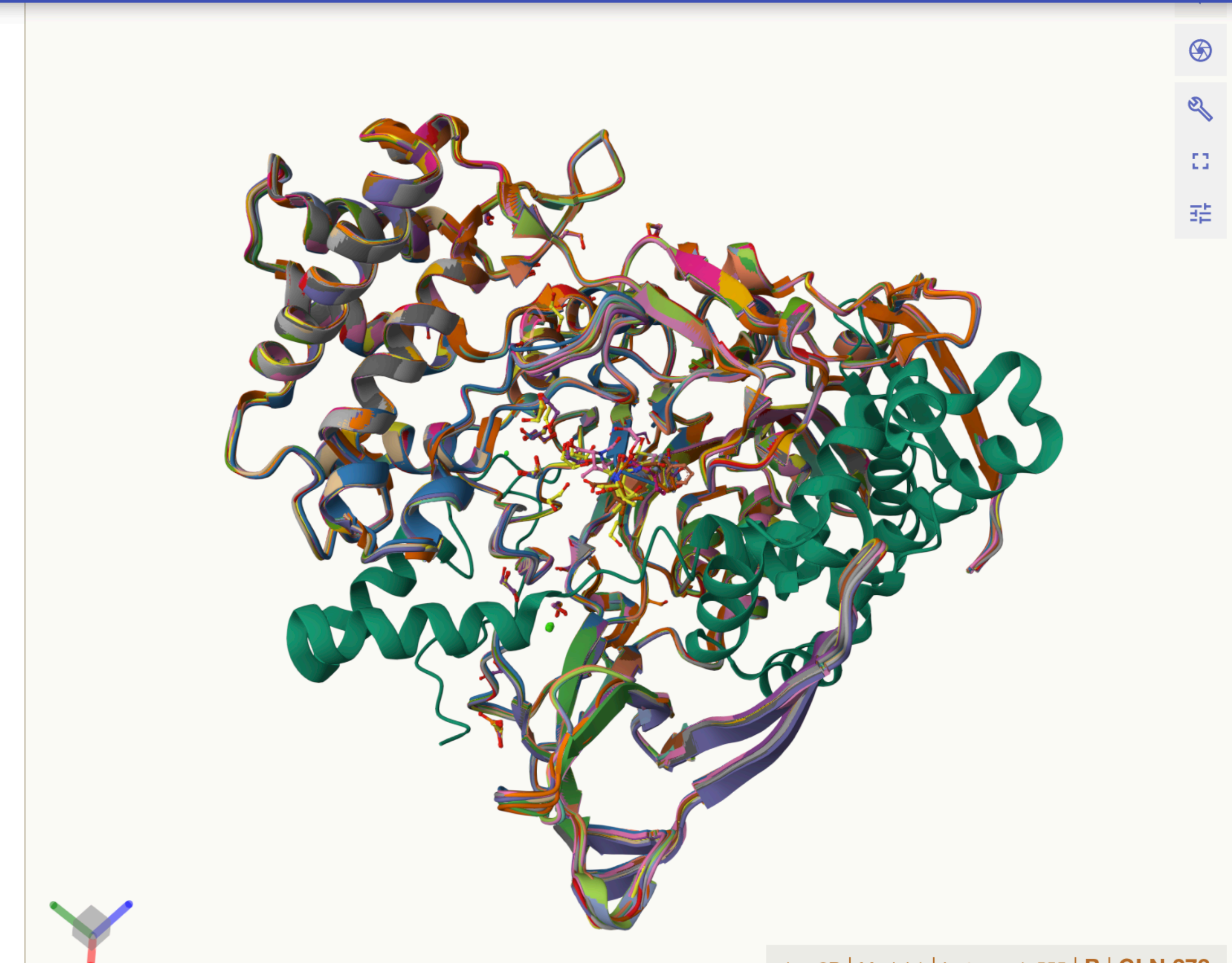
Q9L5C8

1.5

3huo / B

Q9L5C8

1.5



PNN (A_303)

353.29

58.88

61

9

View

Download

PNN (A_304)

322.66

64.53

69

1

View

Download

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 specific ligand binding site identification of apo/holo structure pairs
 - Ahoj-DB – PDB-wide assignment of apo/holo pairs for BioLIP database

Acknowledgements



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) lukapolak.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

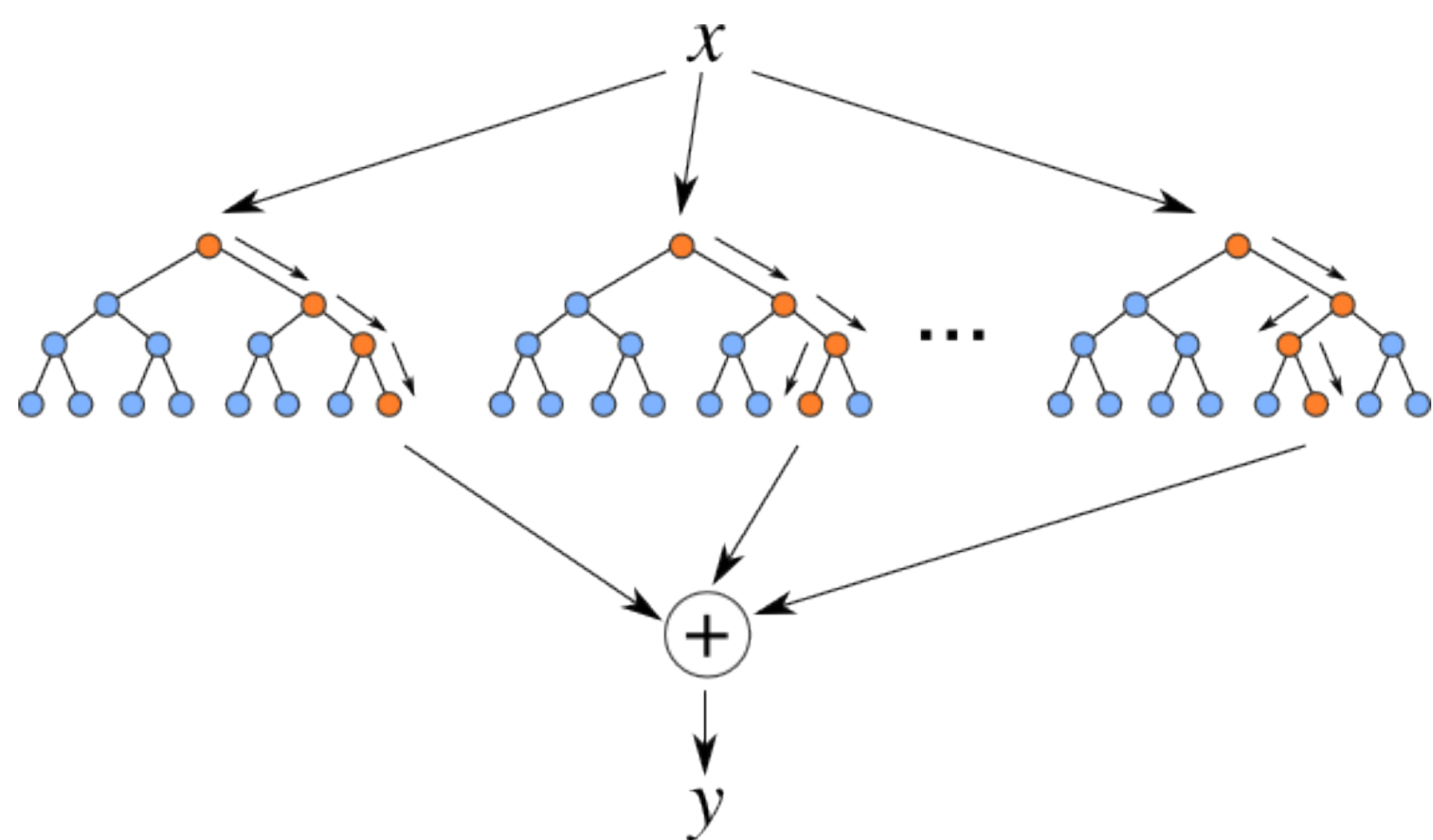
Charles University Structural Bioinformatics Group

<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

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

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

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

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

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

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