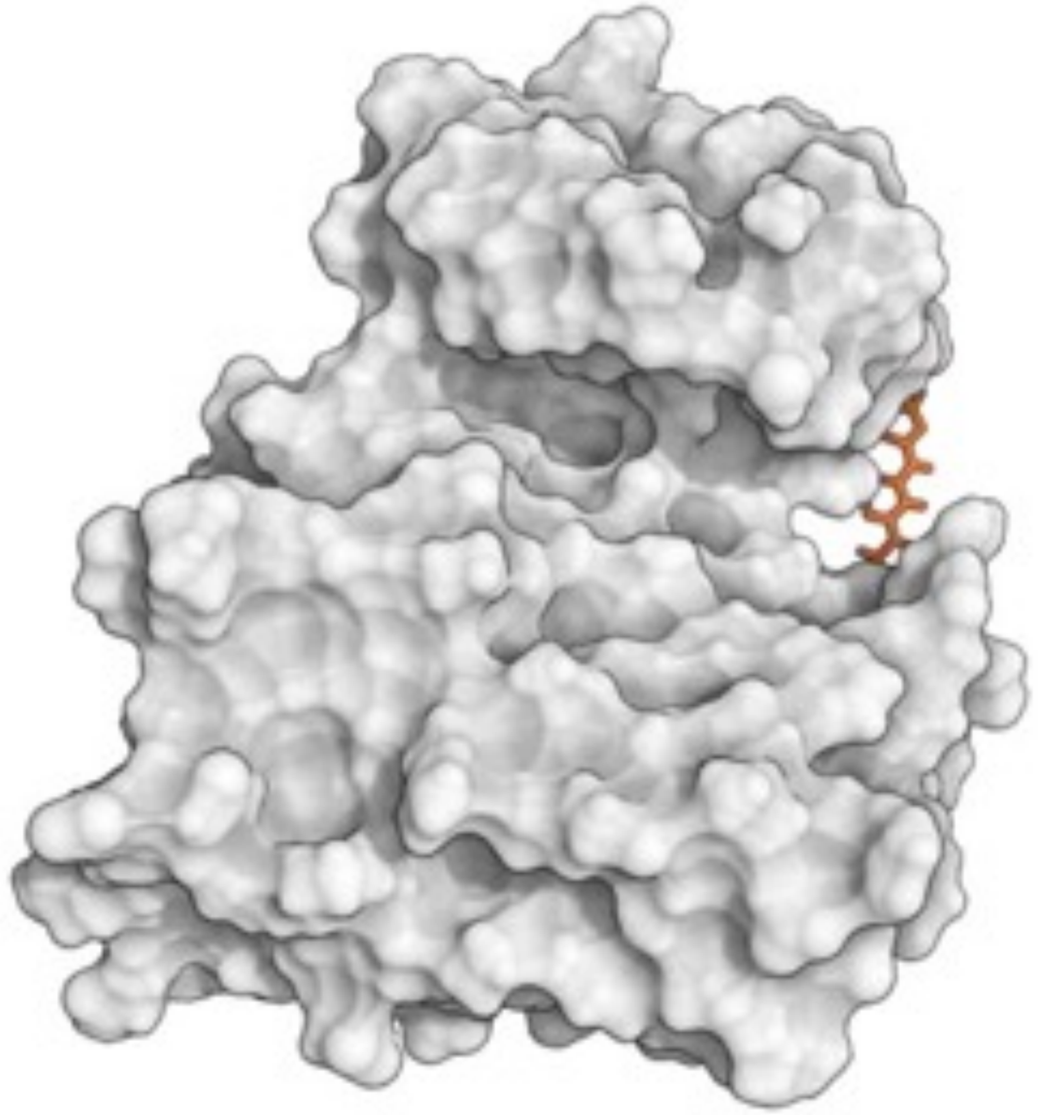
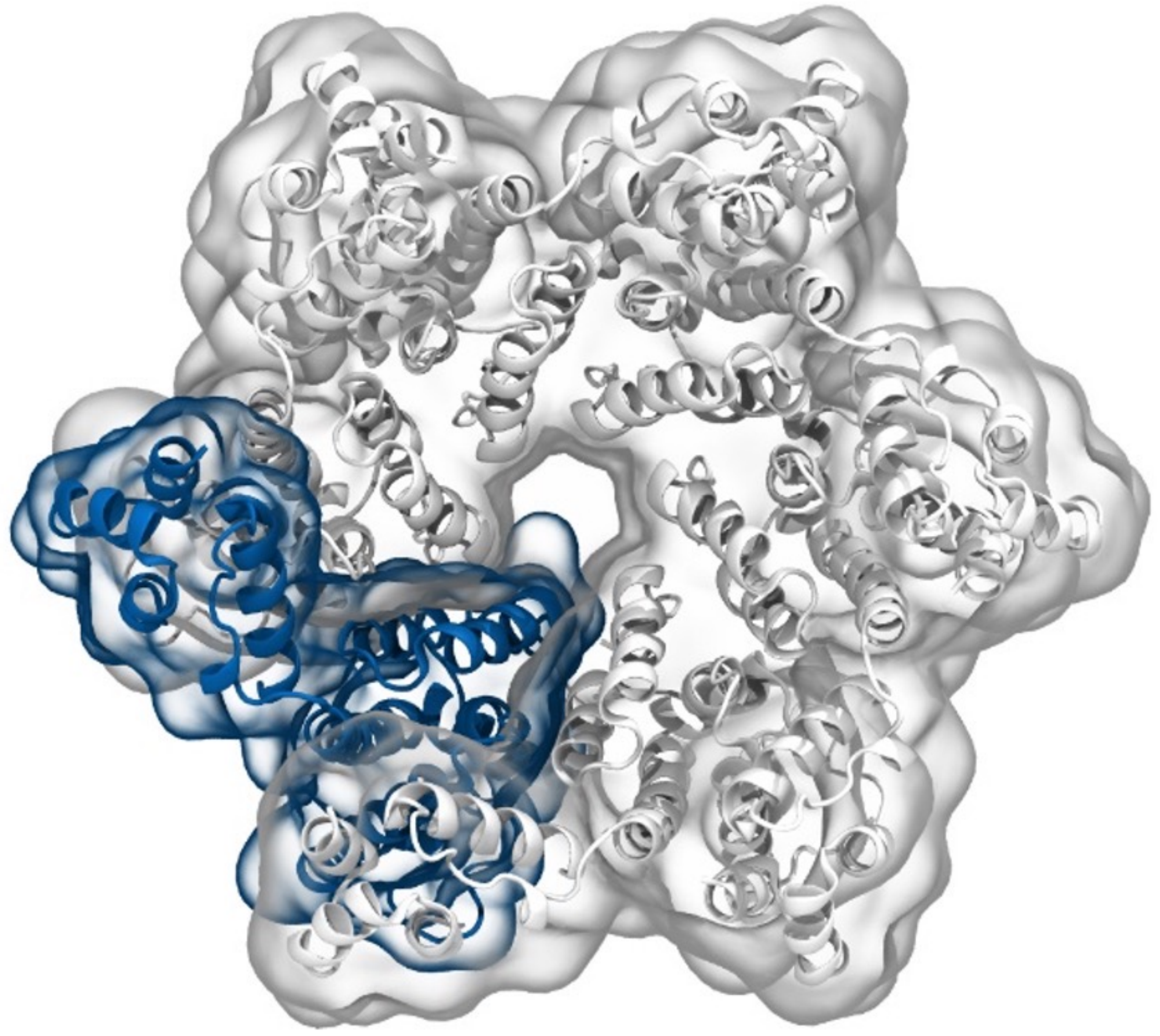
Simulation of Biomolecules



Lecture 2: Protein preparation



2024 CCP5 Summer School

Dr Matteo Degiacomi
Durham University
matteo.t.degiacomi@durham.ac.uk

Dr Antonia Mey
University of Edinburgh
antonia.mey@ed.ac.uk

A typical workflow for molecular dynamics

Get protein structure
↓
Prep for simulation
↓
Simulation setup
↓
Run Simulation Protocol
↓
Simulation Analysis

pdb api
pdb2gmx
SCHRÖDINGER Maestro
RDkit
OpenMM tools
SeeSAR The drug design dashboard
Amber tools
[...]

GUI, TCL, bash, python, Perl … get creative

Gromacs
OpenMM
Amber
NAMD
DLPoly
SOMD
Charmm
BioSimSpace
[...]

Python API

mostly C++
commandline

PyEMMA
MDAnalysis
md-analysis-tools
MDTraj
MDAnalysis
VMD
[...]

TCL, bash, python, Perl … get creative

Let’s get started with understanding protein structures

Get protein structure

↓

Prep for simulation

↓

Simulation setup

↓

Run Simulation Protocol

↓

Simulation Analysis

OpenEye SCIENTIFIC

pdb2gmx

pdb api

RDkit

SCHRÖDINGER Maestro

OpenMM tools

Amber tools

SeeSAR The drug design dashboard

[...]

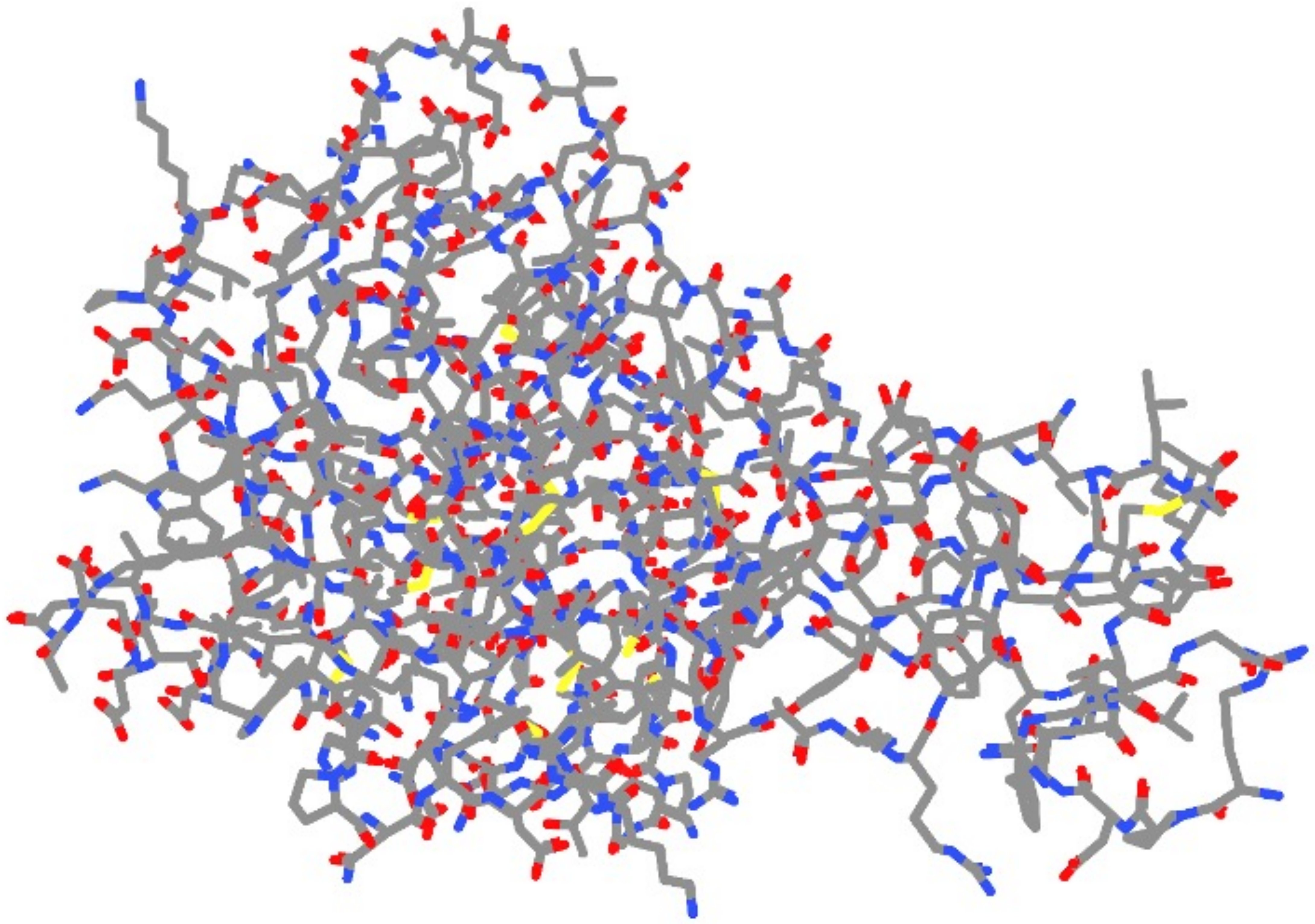
GUI, TCL, bash, python, Perl … get creative

Part 1: I have a protein I want to model, now what?

What things are important to look out for in protein prep?

https://www.menti.com/alc5aozap924   58950948

Crystal structures provide electron densities

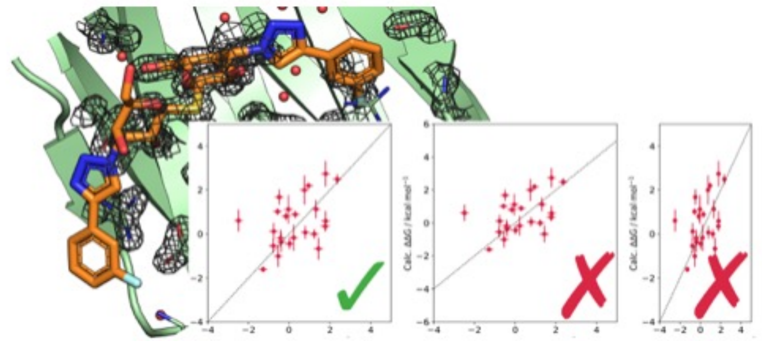


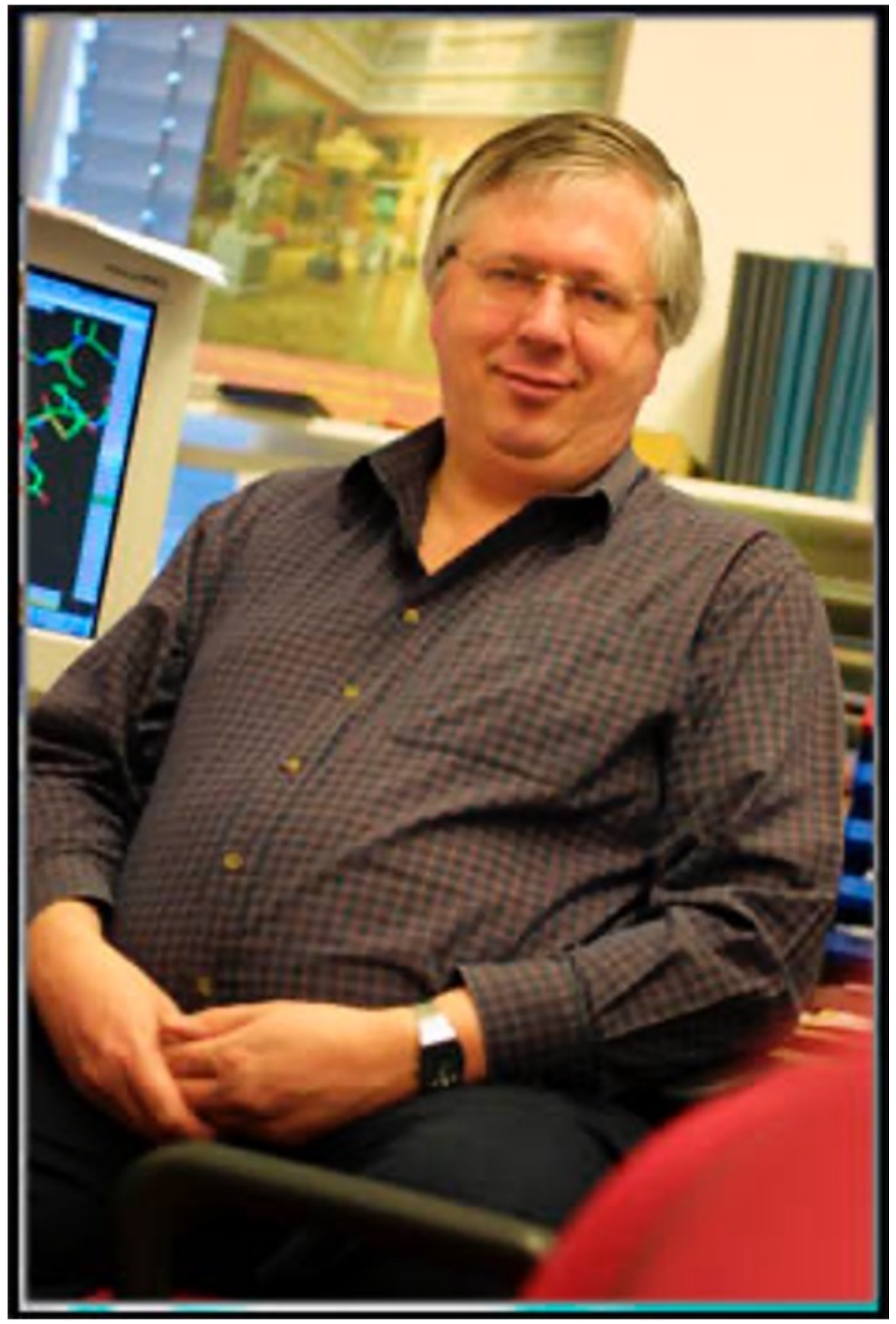
Crystal structures are models with potential errors

HOME / ARCHIVES / VOL. 4 NO. 1 (2022) / Articles

Best Practices for Constructing, Preparing, and Evaluating Protein-Ligand Binding Affinity Benchmarks [Article v1.0]

David F. Hahn
Computational Chemistry, Janssen Research & Development, Turnhoutseweg 30, Beerse B-2340, Belgium
https://orcid.org/0000-0003-2830-6880





Gert Vriend (author of WHAT_CHECK)

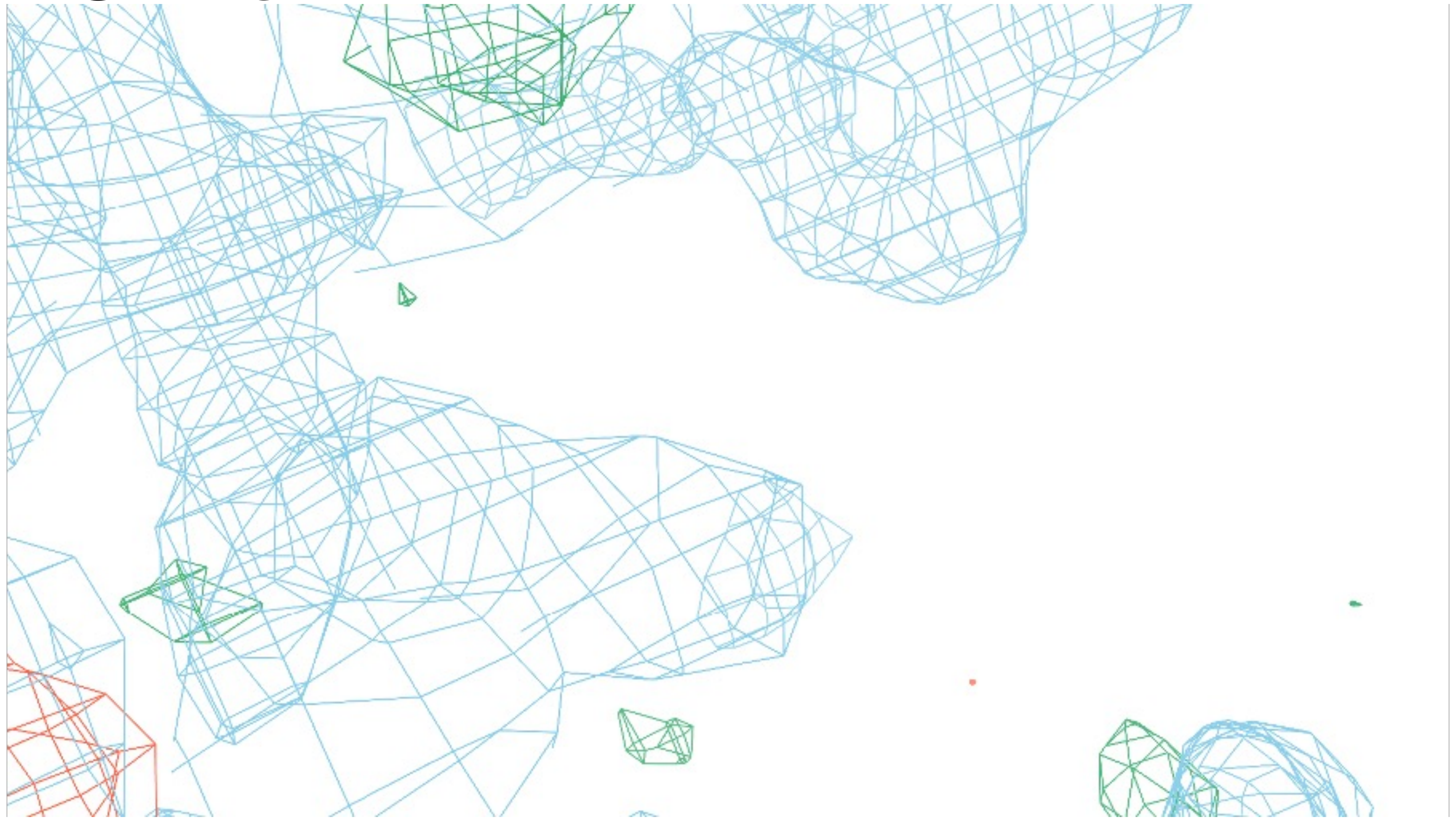
Examples of errors/oddities in PDB structures

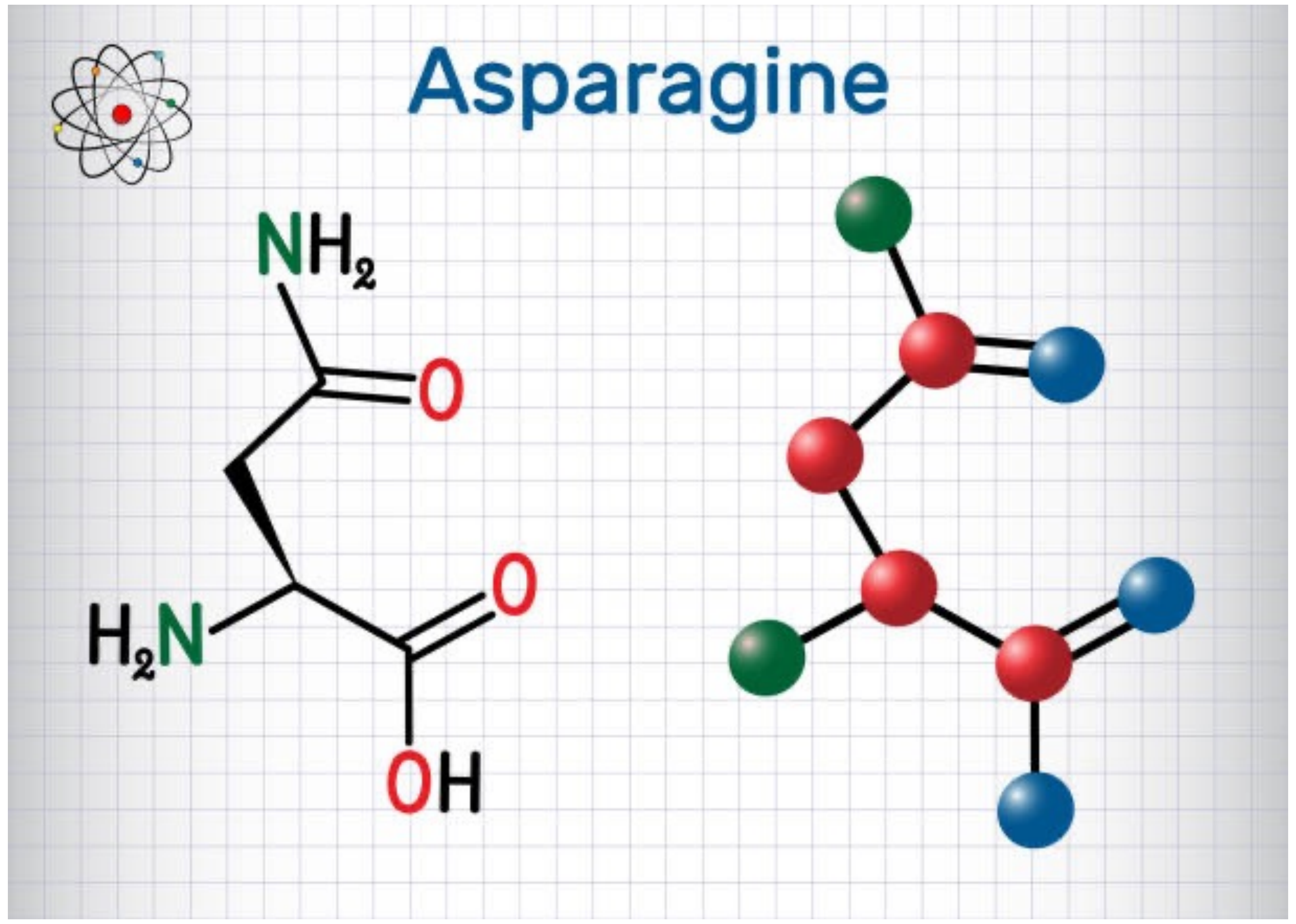
https://swift.cmbi.umcn.nl/teach/pdbad/

Crystal structures are models derived from electron densities

1T38: HUMAN O6-ALKYLGUANINE-DNA ALKYLTRANSFERASE

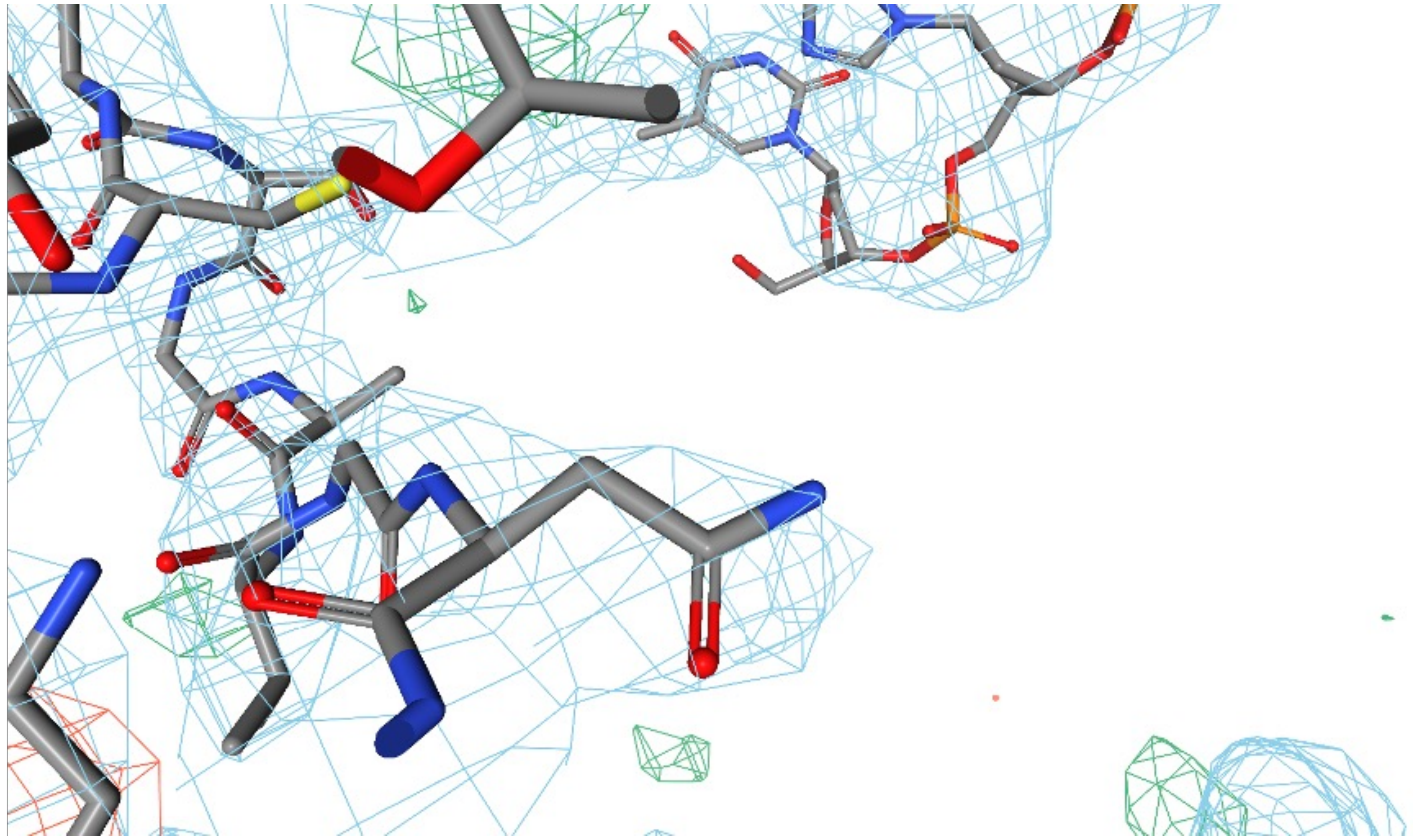
ASN157

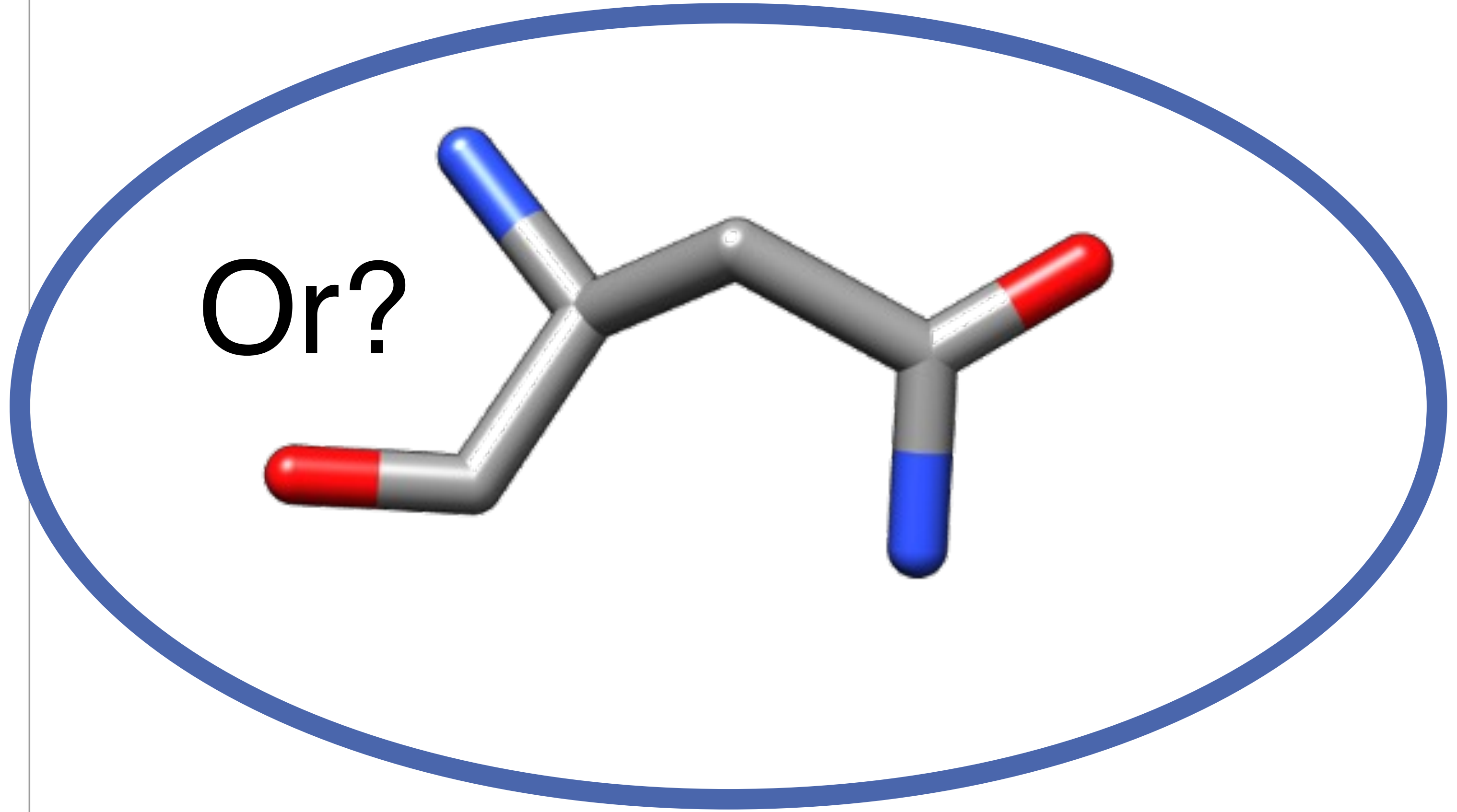




Crystal structures are models derived from electron densities

1T38: HUMAN O6-ALKYLGUANINE-DNA ALKYLTRANSFERASE

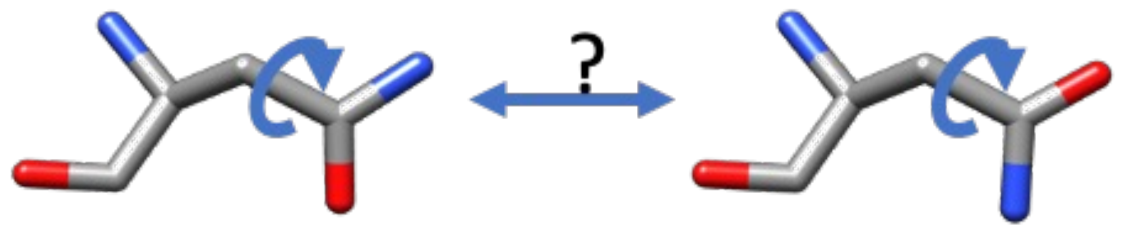




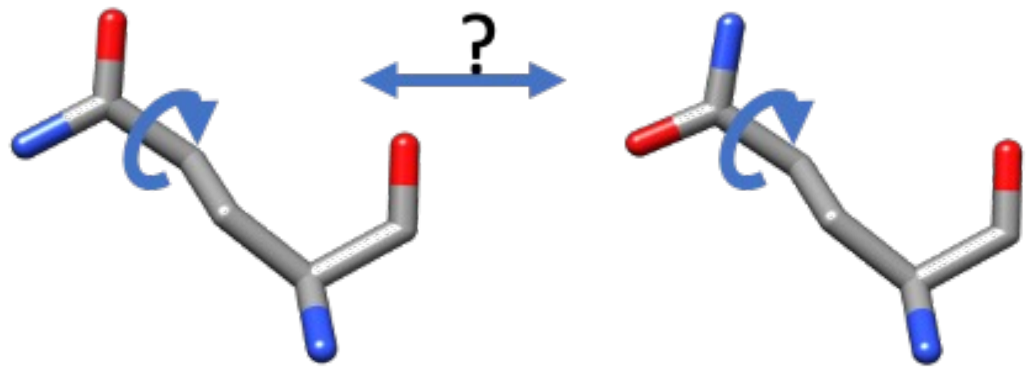
Alternative conformations of side chains - NGH flips

Typically the crystallographers will have assigned the orientation based on potential H-bonding interactions, etc.

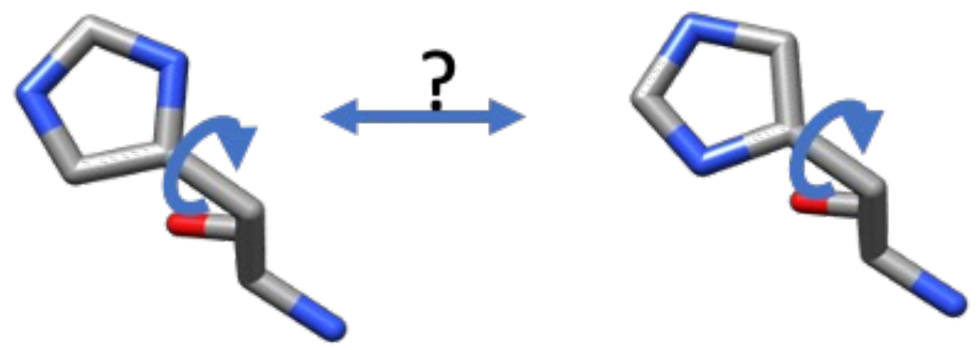
No standard protocol for this: always worth double-checking.



Asparagine (N)



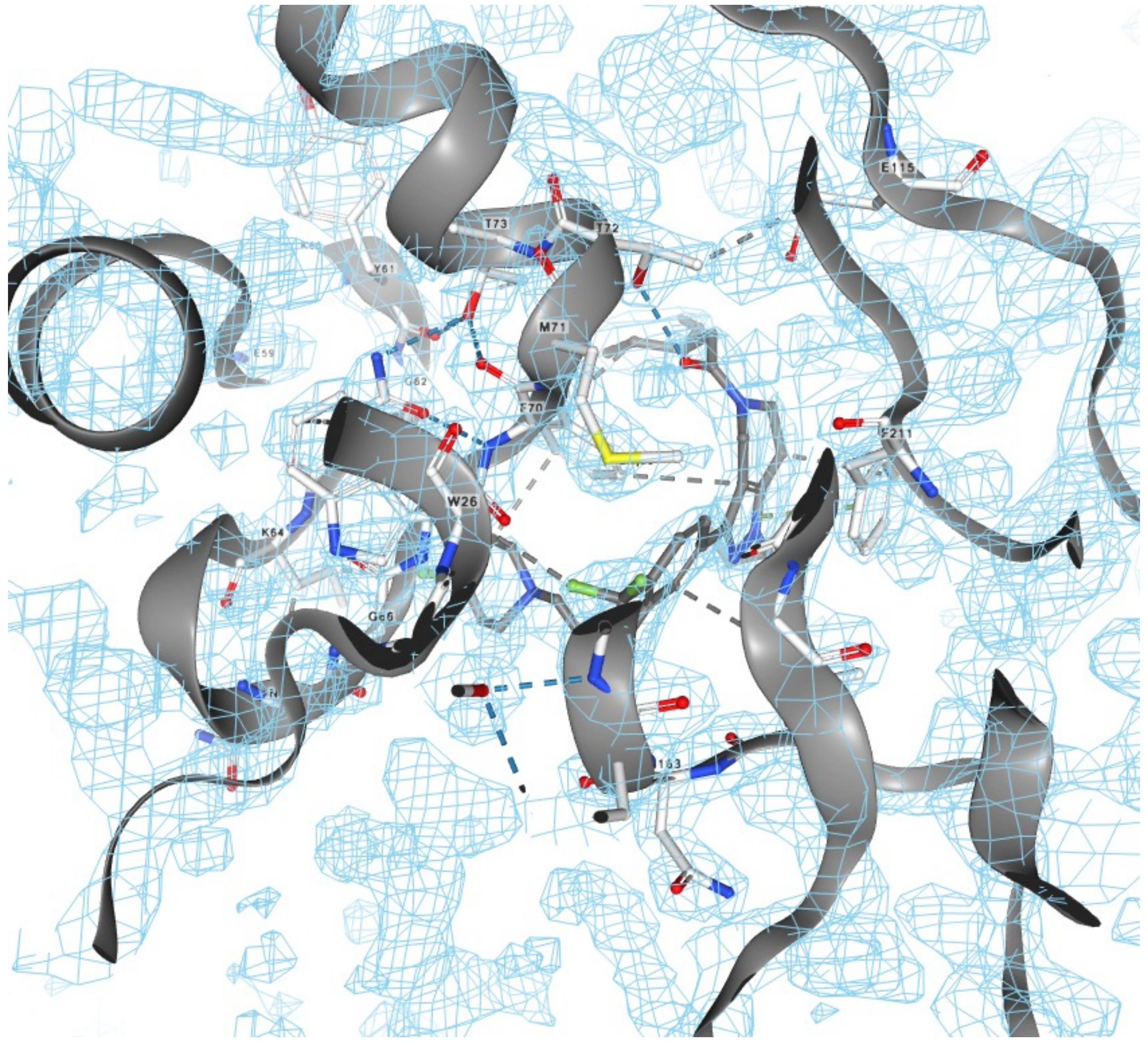
Glutamine (Q)



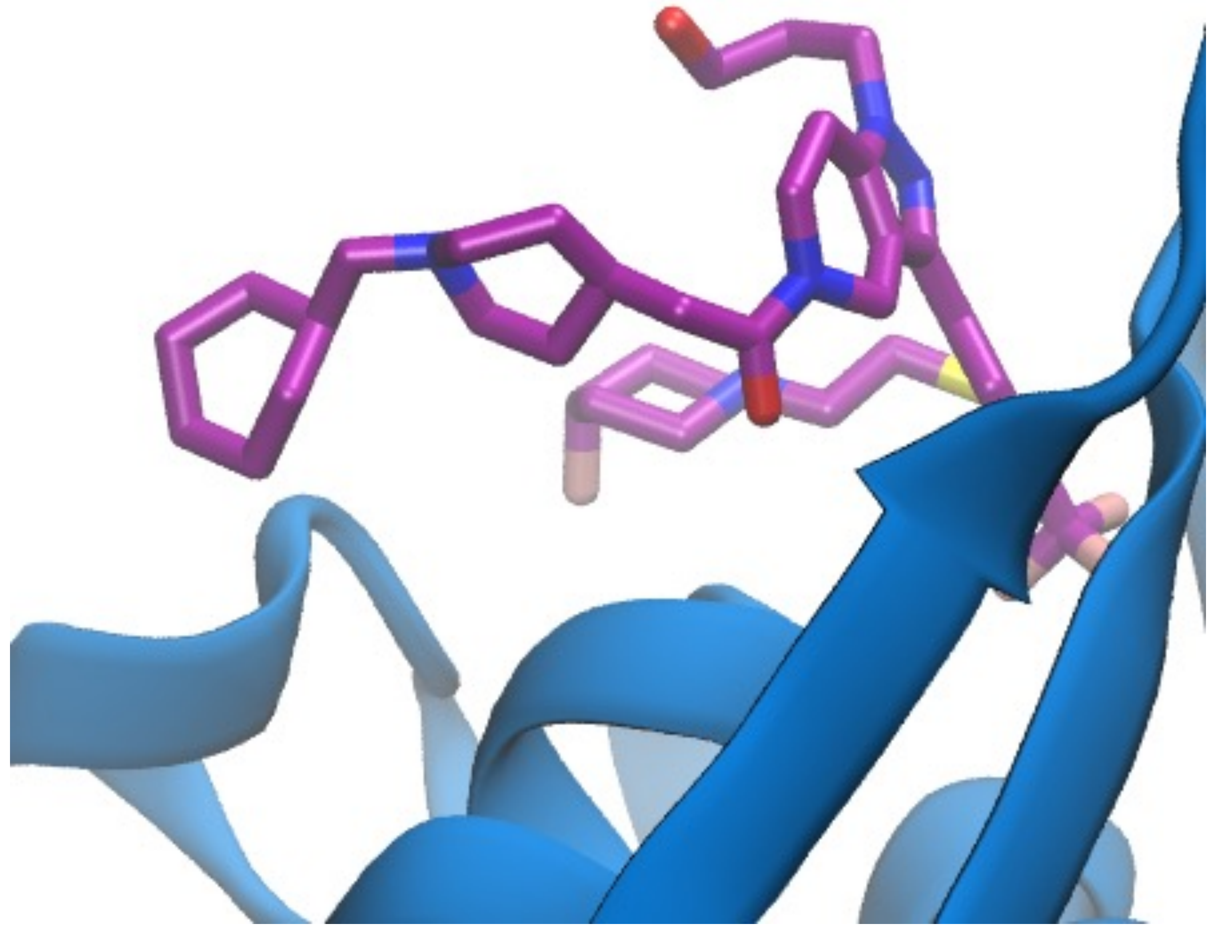
Histidine (H)

Ligand densities can also have creative input from the crystallographer

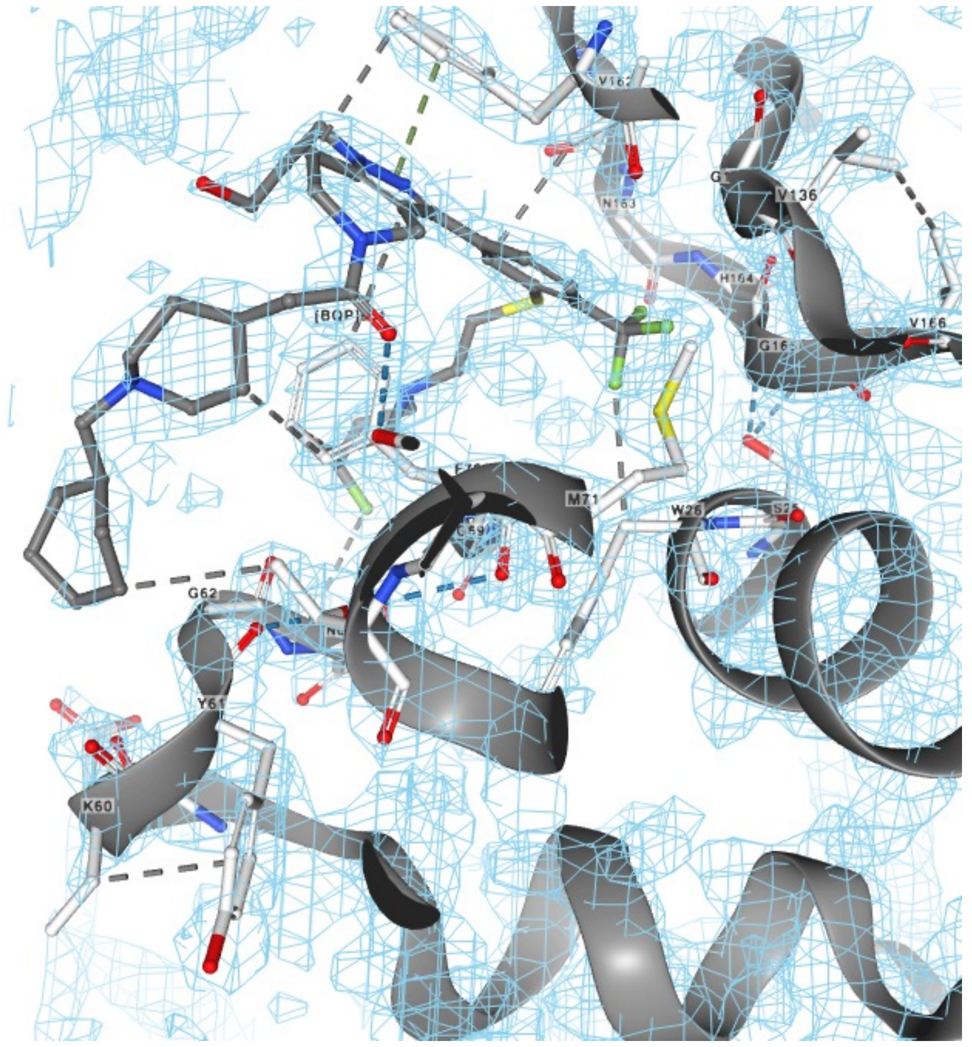
Cathepsin S



Cathepsin S
with ligand

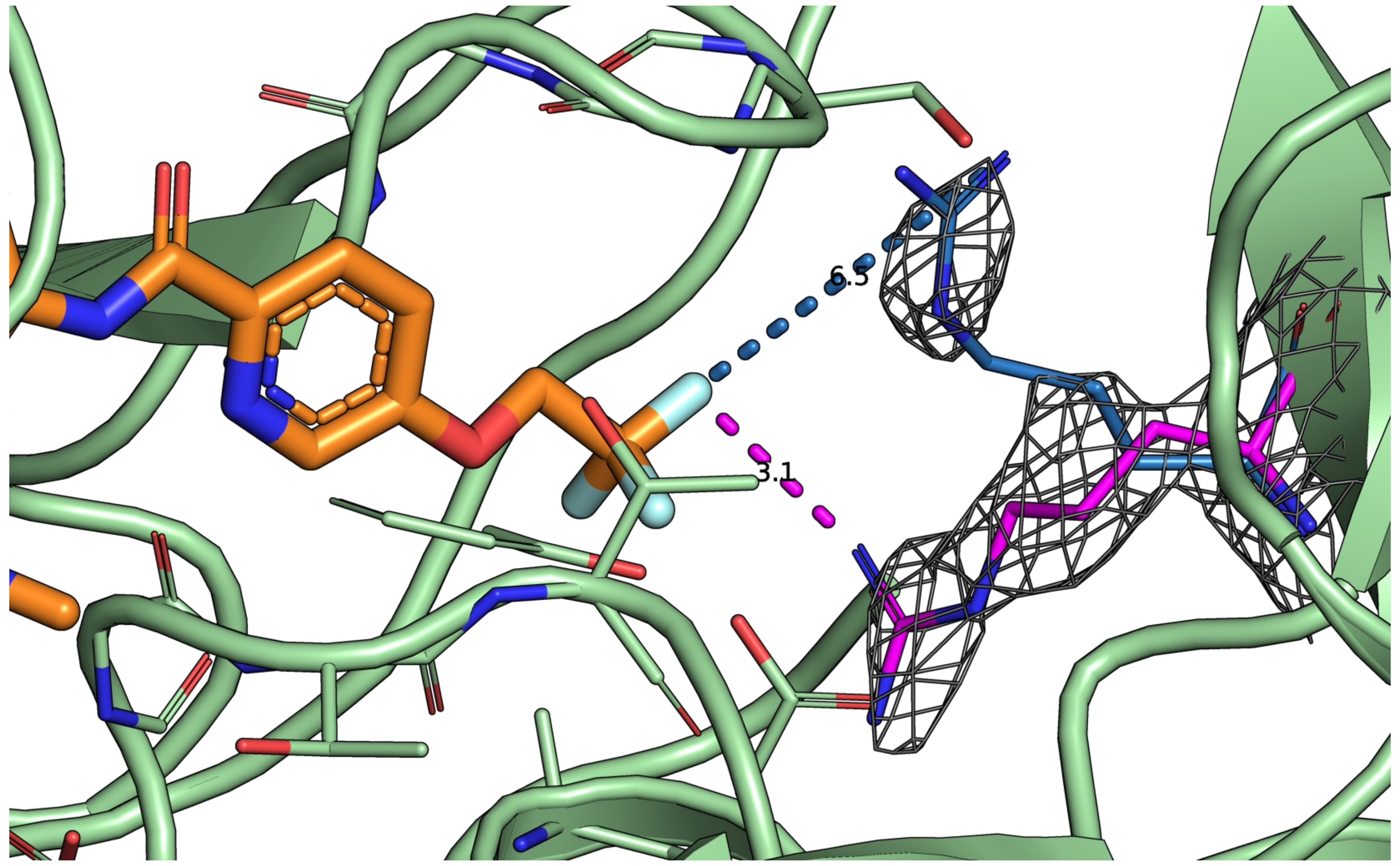


boat/chair?

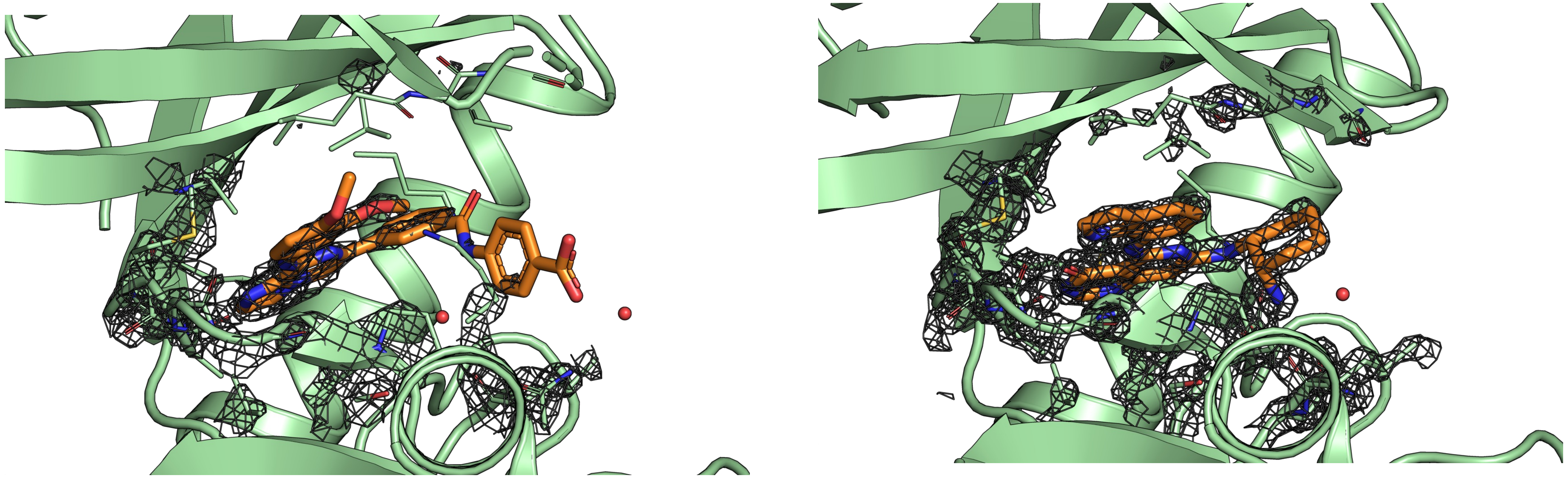


Picking the best crystal structure requires care

BACE (Hunt)



spleen tyrosine kinase



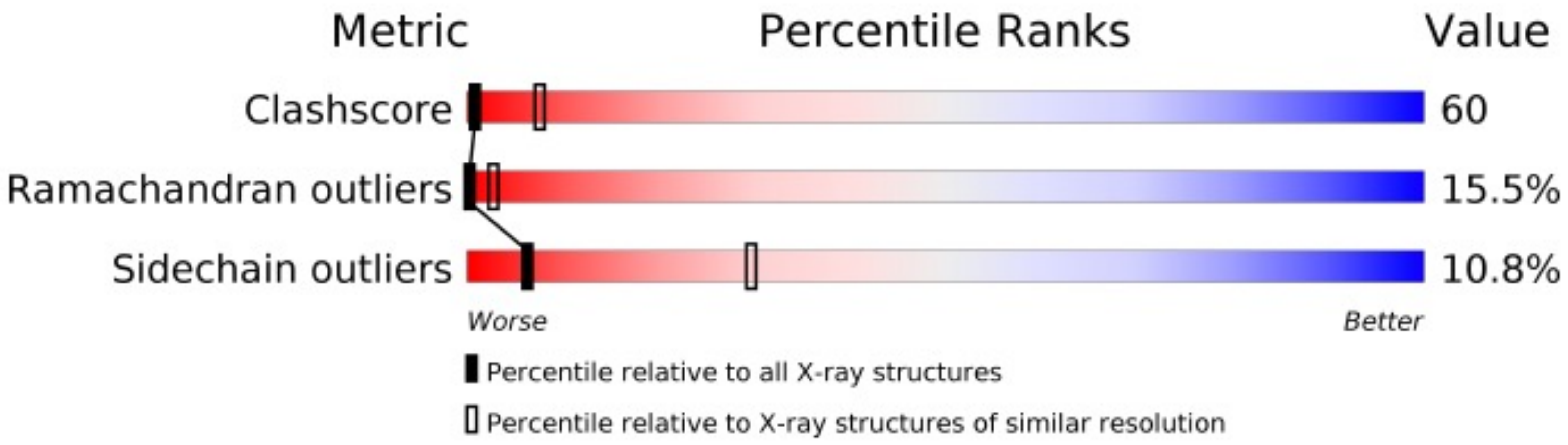
Which protein structure?

- Alternative side chain conformations need to be assessed carefully

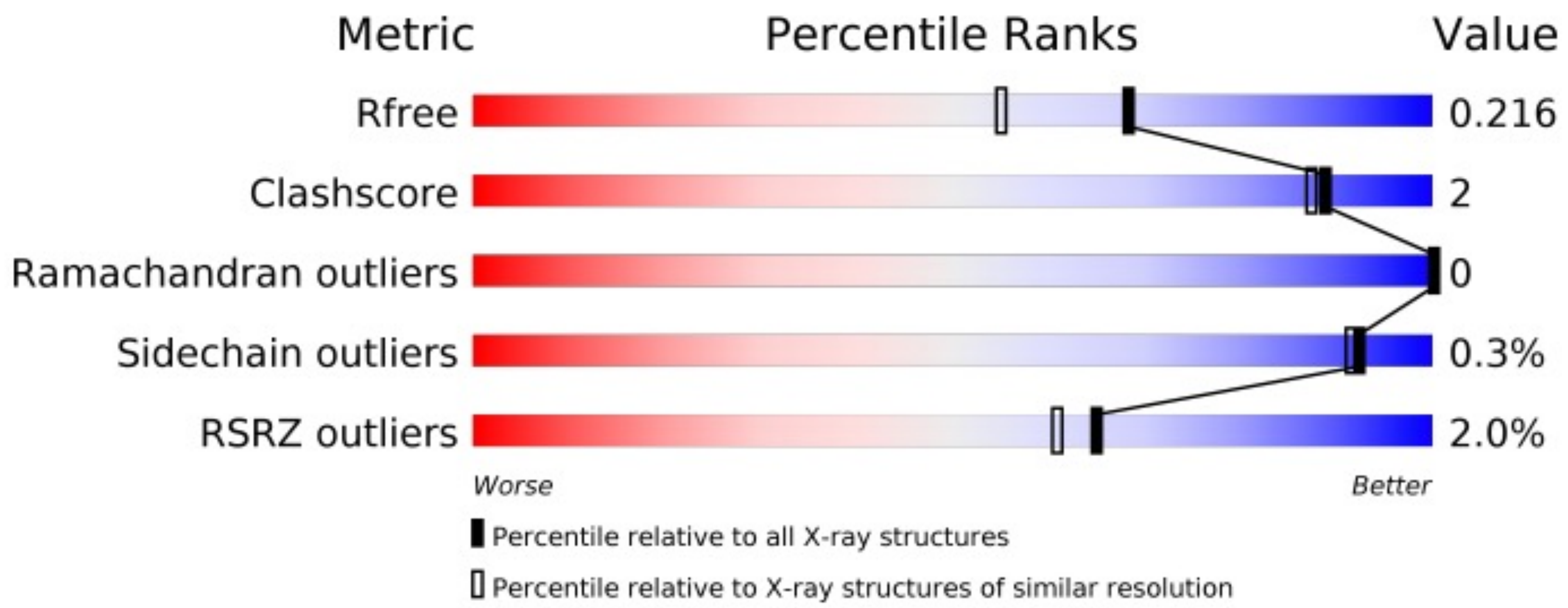
- Active site residue densities are important for choosing the right crystal structure

The RCSB PDB report can help with choosing structures

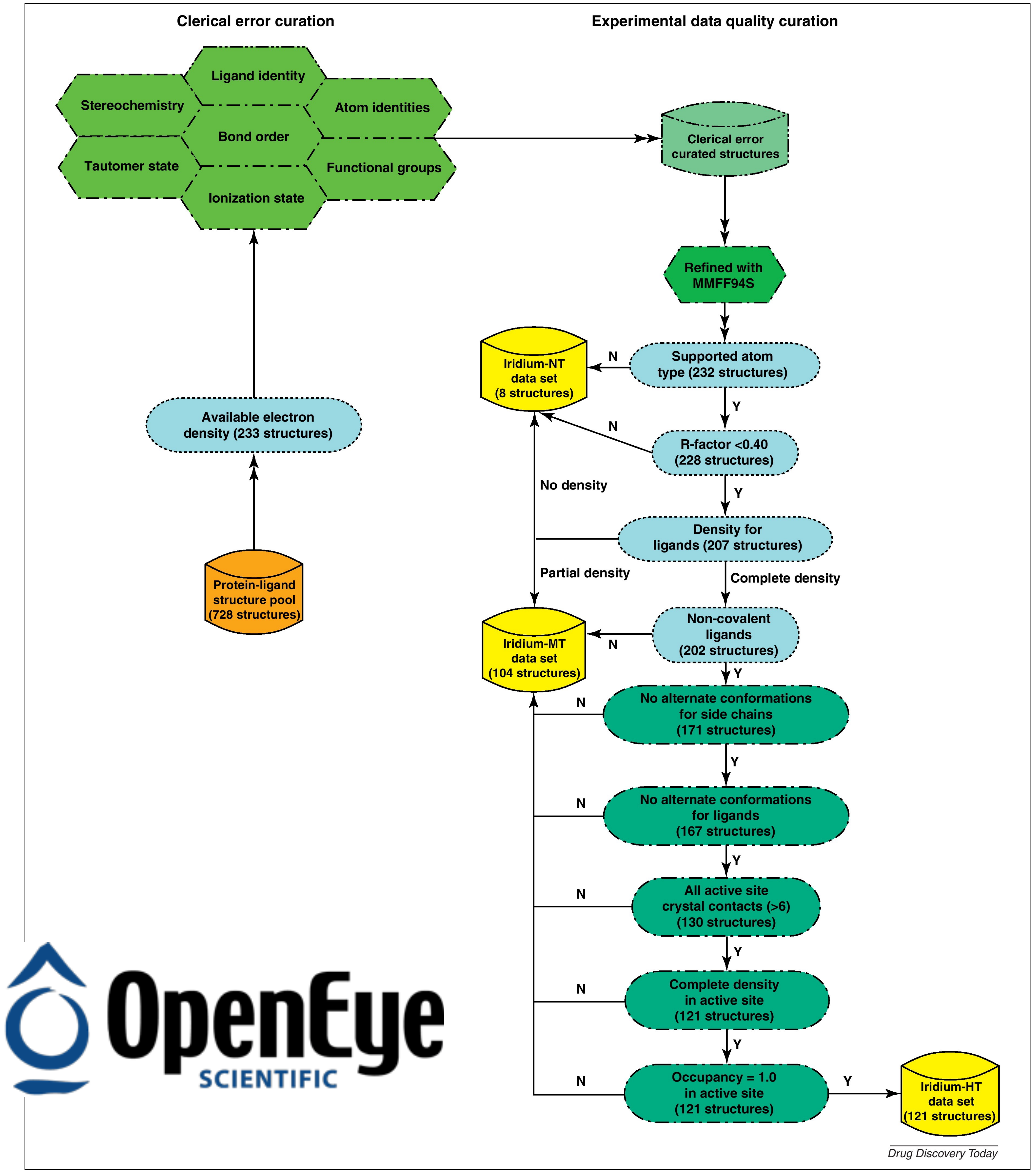
Jnk1 - 2GMX



Jnk1 - 3ELJ



The Iridium score can help assess the trustworthiness of an X-ray structure

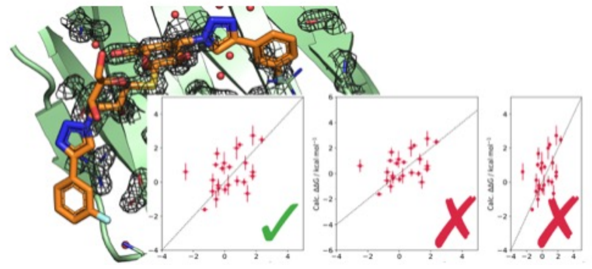


OpenEye SCIENTIFIC

HOME / ARCHIVES / VOL. 4 NO. 1 (2022) / Articles

Best Practices for Constructing, Preparing, and Evaluating Protein-Ligand Binding Affinity Benchmarks [Article v1.0]

David F. Hahn
Computational Chemistry, Janssen Research & Development, Turnhoutseweg 30, Beerse B-2340, Belgium
https://orcid.org/0000-0003-2830-6880



			6D09 (HT, 0.35)[b]			
Jnk1[74, 86]	2GMX (NT, -)[f]	0.77	3ELJ (MT, 0.31)[a]	3V3V (MT, 1.5)[b, h, e]	21	3.4
MCL1[74, 87]	4HW3 (HT, 0.41)	0.26	6O6F (HT, 0.30)	4WMU (HT, 0.41)[b]	42	4.2
			4ZBF (HT, 0.35)[b]	4ZBI (HT, 0.45)[b]		
			3WIX (HT, 0.37)[a, c]			

12 Features to inform the Iridium Score:

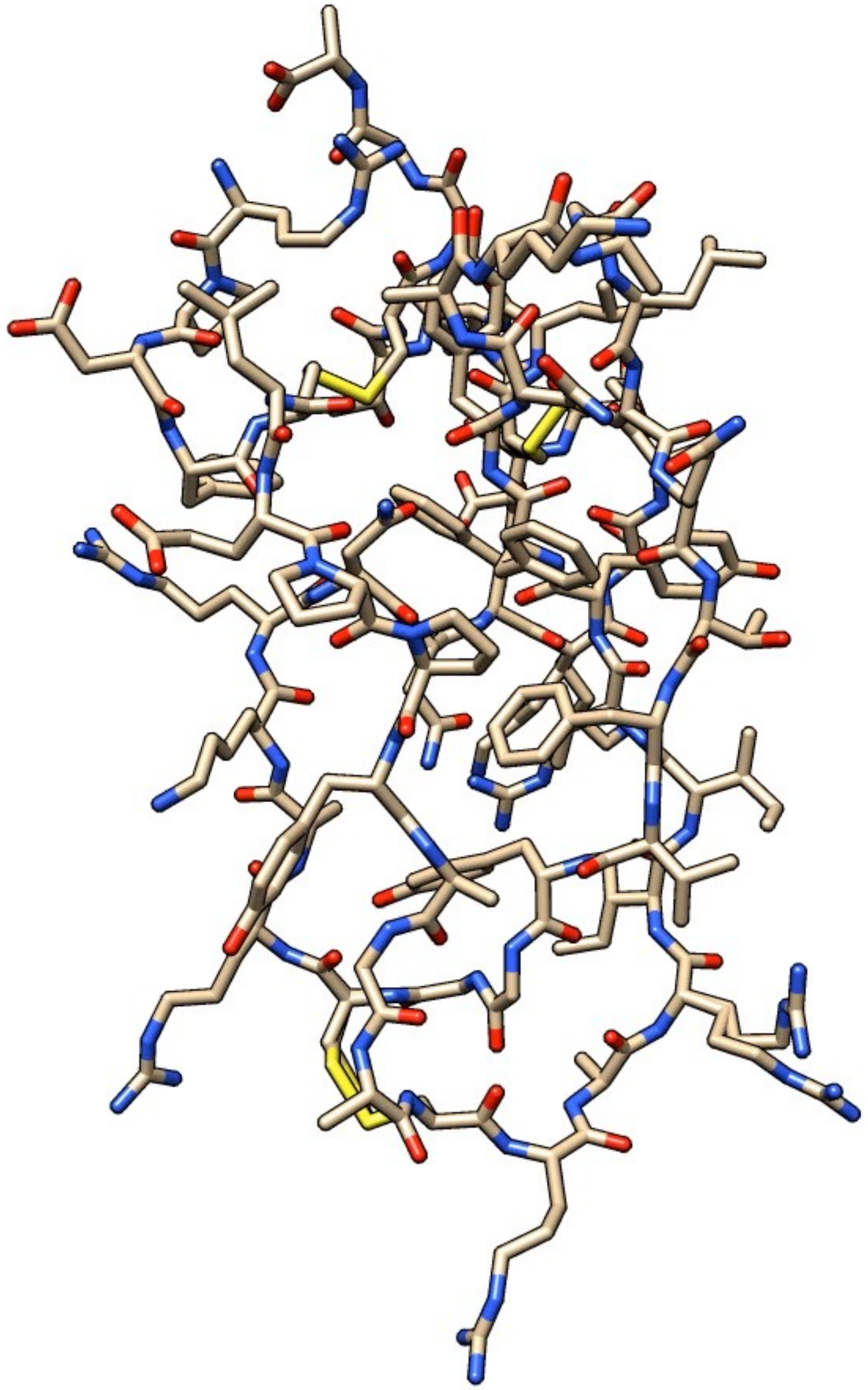
- R-free value
- Resolution
- Density coverage of ligand heavy atoms
- Active site density coverage
- Alternative locations active site/ligands
- […]

	HT	MT	NT
Ligand	> 0.9	< 0.9 and > 0.5	< 0.5
Active Site	> 0.95	< 0.95 and > 0.5	< 0.5

Warren et al. *Drug Discovery Today* (2012) **17** 270-1281

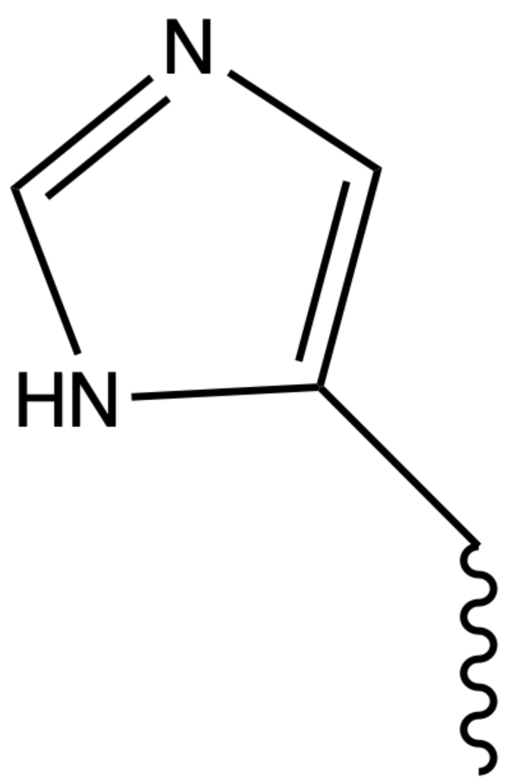
Missing information in structures: Hydrogen atoms

- Missing from most crystal structures
- Needed for molecular modelling
- Most MD packages (AMBER, CHARMM, GROMACS, etc.) include tools to “automatically” add H-atoms.
- Any issues?

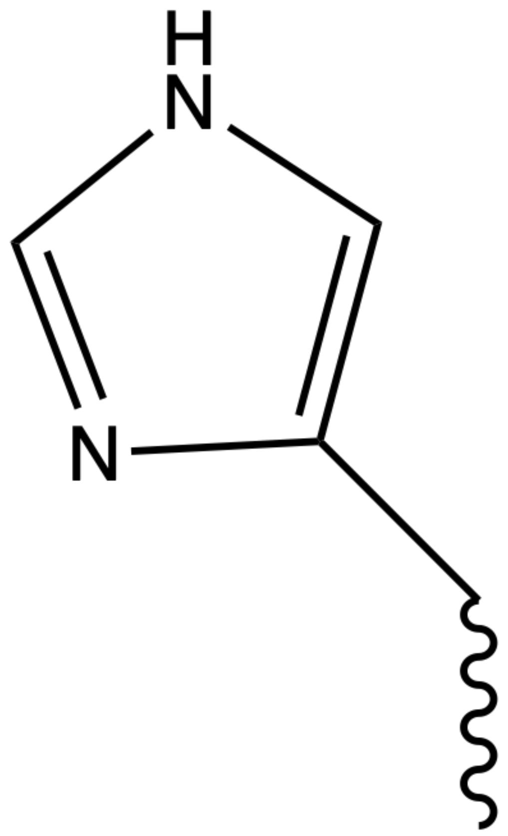


How to choose the right tautomer?

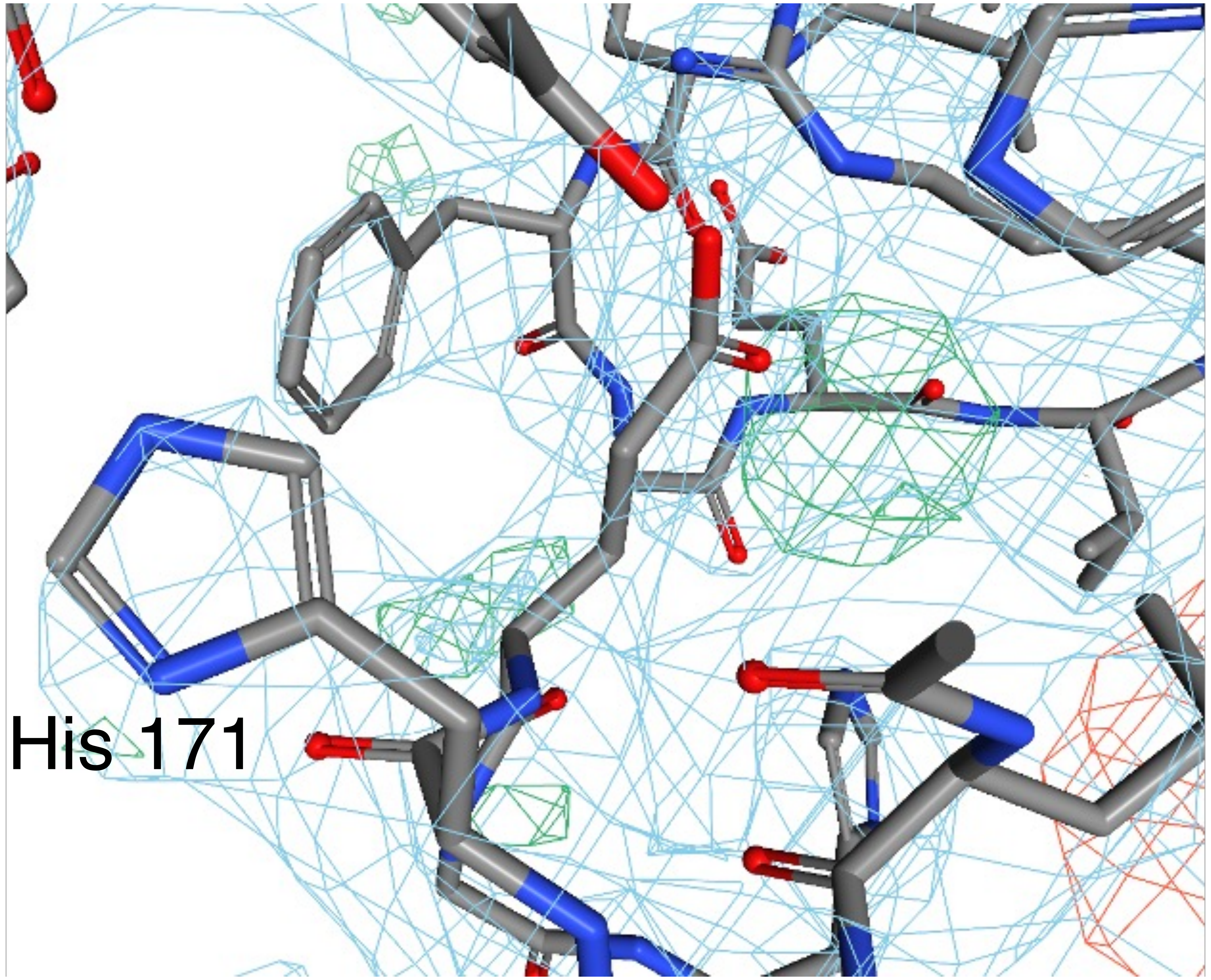
Which tautomer?



δ-tautomer

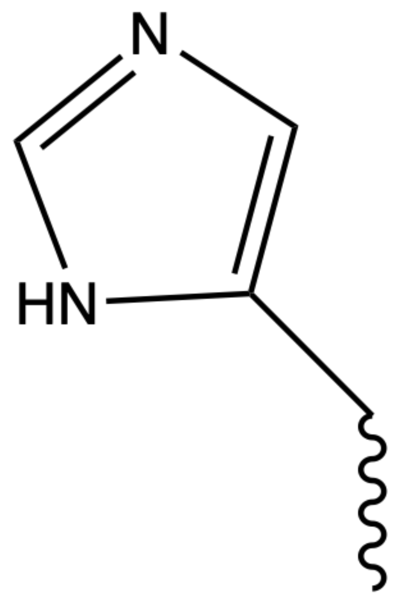


ε-tautomer

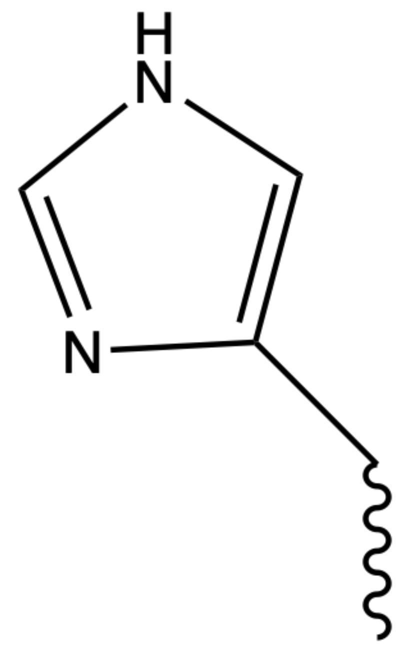


How to choose the right tautomer and protonation state?

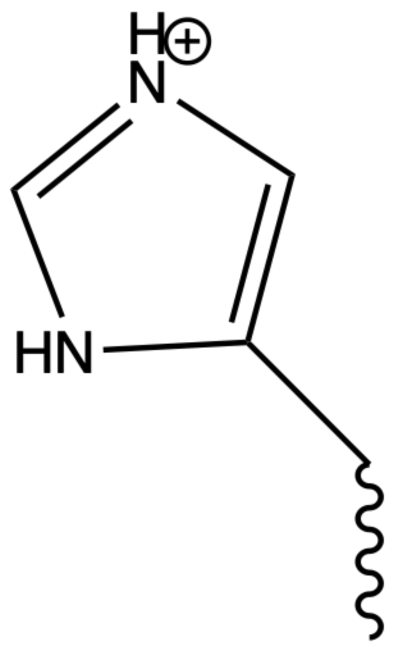
Which tautomer and protonation?



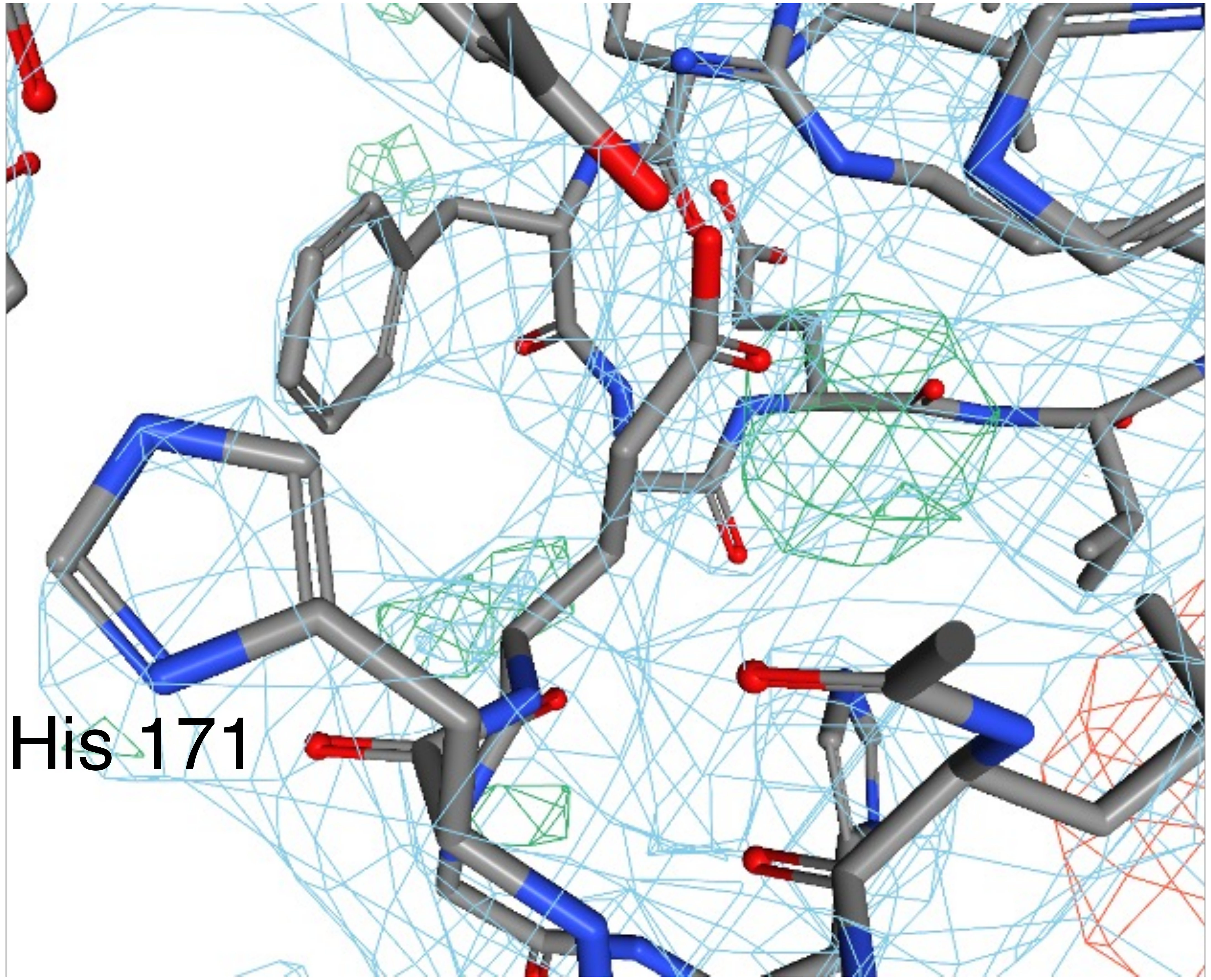
δ-tautomer



ε-tautomer

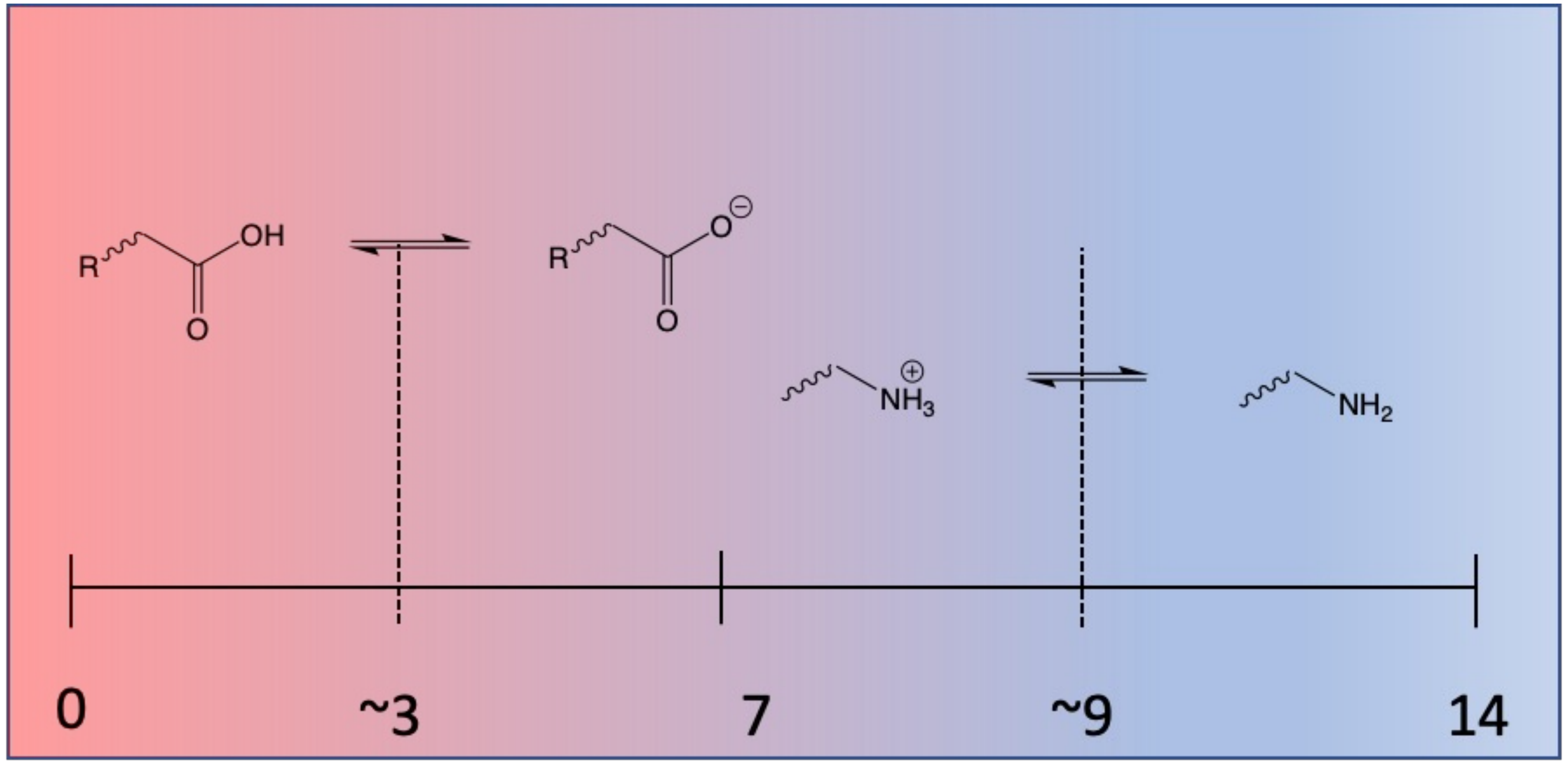


Προτονατεδ



We need to worry about pKa to decide protonation

- pKa: pH at which an acidic/basic group is 50% protonated/deprotonated.
- pKas are not fixed things!
- "Standard" values refer to the situation when the group is in dilute aqueous solution.
- Groups buried in the centre of hydrophobic proteins or close to other charged groups can show large pKa shifts.

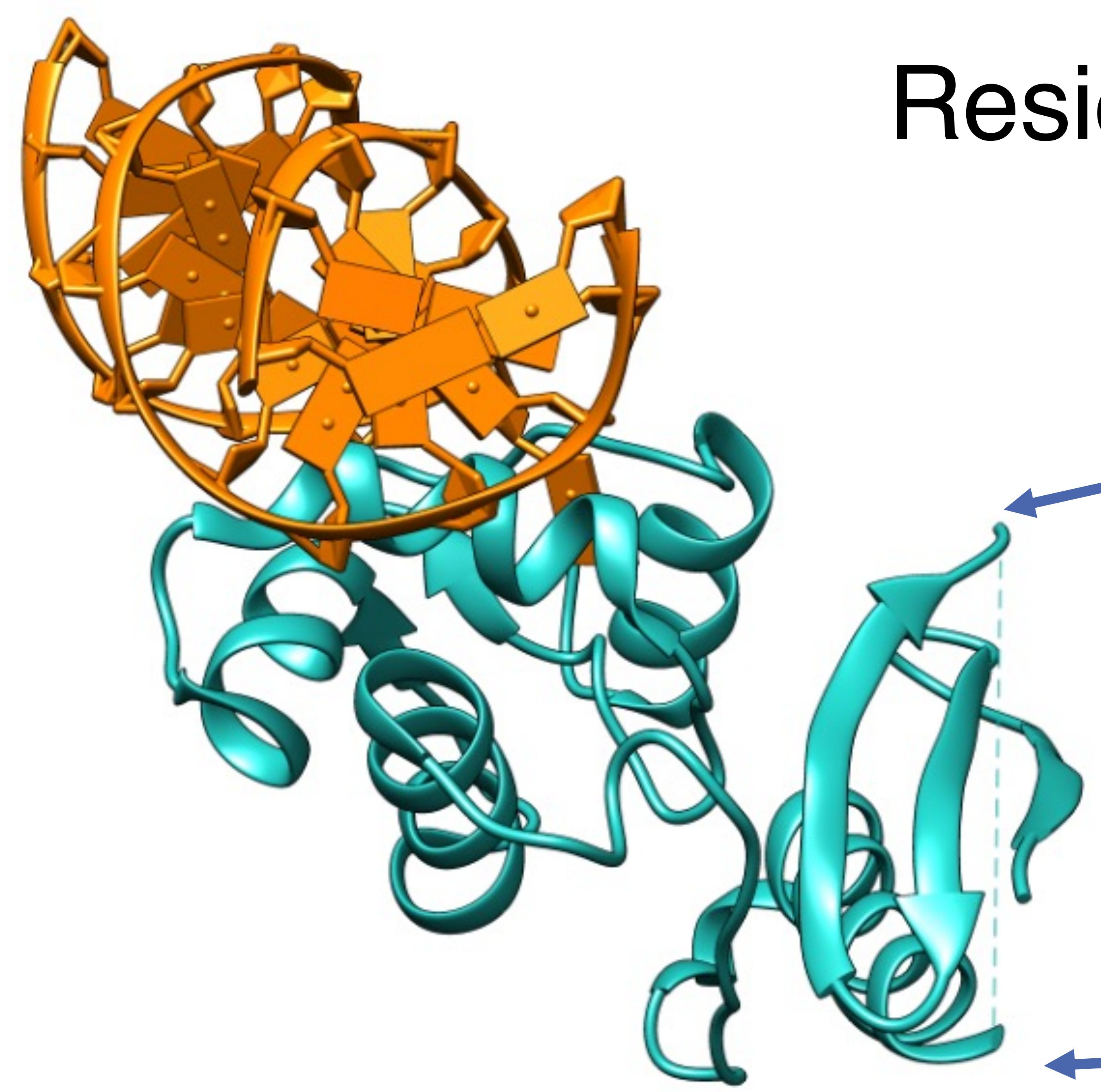


Amino acids that need protonation state consideration

Amino acid	pKa	options	Significance*
Aspartic acid	3.65	-COOH instead of COO^-?	possible
Glutamic acid	4.25	-COOH instead of COO^-?	possible
Histidine	6.00	protonated instead of neutral?	very possible
Cysteine	8.18	$-S^-$ instead of SH?	very possible
Tyrosine	10.07	$-O^-$ instead of OH?	possible
Lysine	10.53	$-NH_2$ instead of NH_3^+?	possible
Arginine	12.48	neutral instead of protonated?	unlikely

*for a simulation around physiological pH

What if the best structure has missing residues?



Residues 36 to 55 missing

```
REMARK 465 MISSING RESIDUES
REMARK 465 THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465 EXPERIMENT. (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 465 IDENTIFIER; SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465
REMARK 465   M RES C SSSEQI
```

```
REMARK 465     LYS A    36
REMARK 465     GLY A    37
REMARK 465     THR A    38
REMARK 465     SER A    39
REMARK 465     ALA A    40
REMARK 465     ALA A    41
REMARK 465     ASP A    42
REMARK 465     ALA A    43
REMARK 465     VAL A    44
REMARK 465     GLU A    45
REMARK 465     VAL A    46
REMARK 465     PRO A    47
REMARK 465     ALA A    48
REMARK 465     PRO A    49
REMARK 465     ALA A    50
REMARK 465     ALA A    51
REMARK 465     VAL A    52
REMARK 465     LEU A    53
REMARK 465     GLY A    54
REMARK 465     GLY A    55
```

Modeller

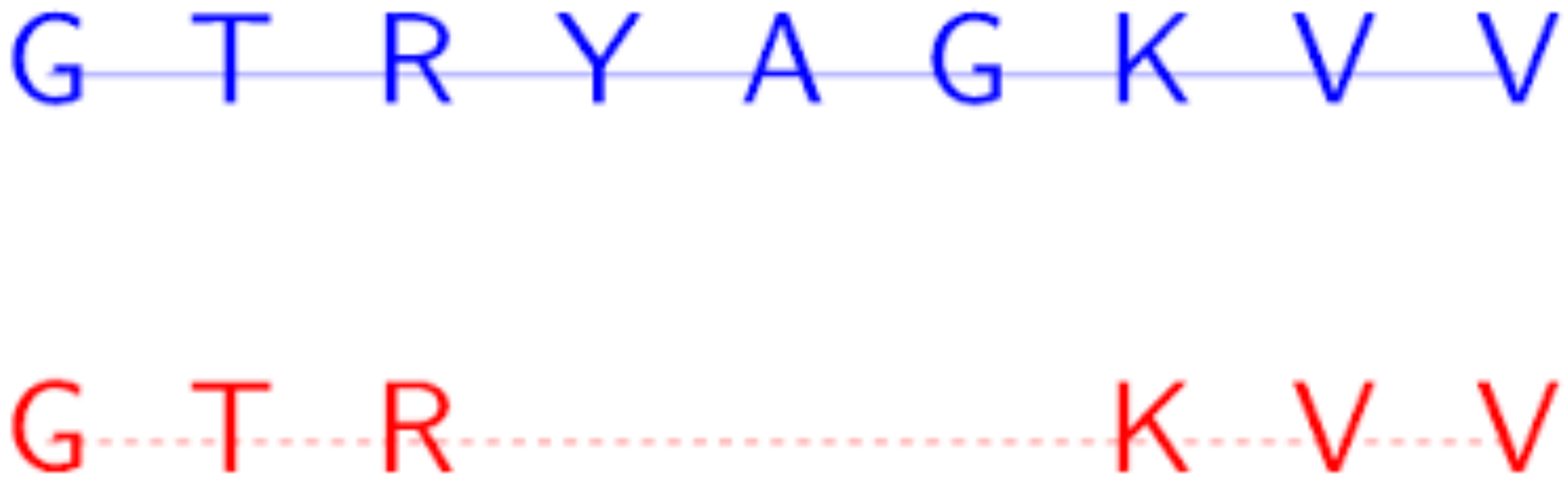
Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

PDB FIXER

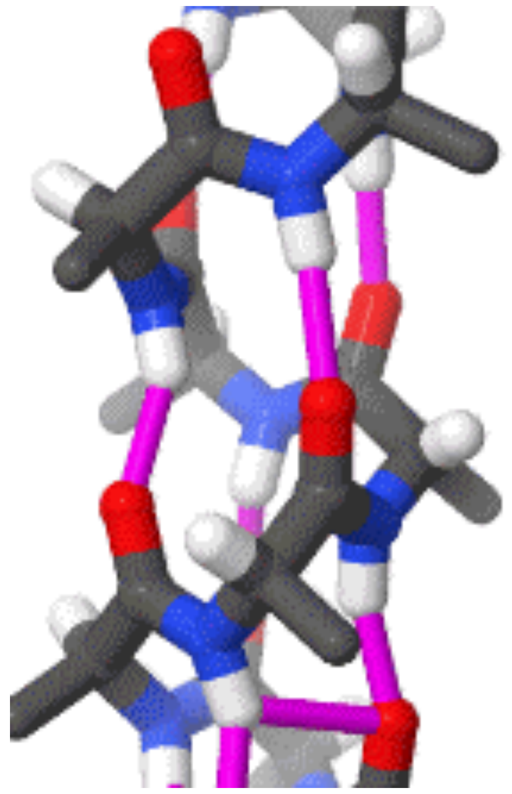
AlphaFold

AlphaFold2 structures have no missing residues or atoms

AlphaFold



No missing residues



Has hydrogens

AlphaFold2 structures can have unreliable areas

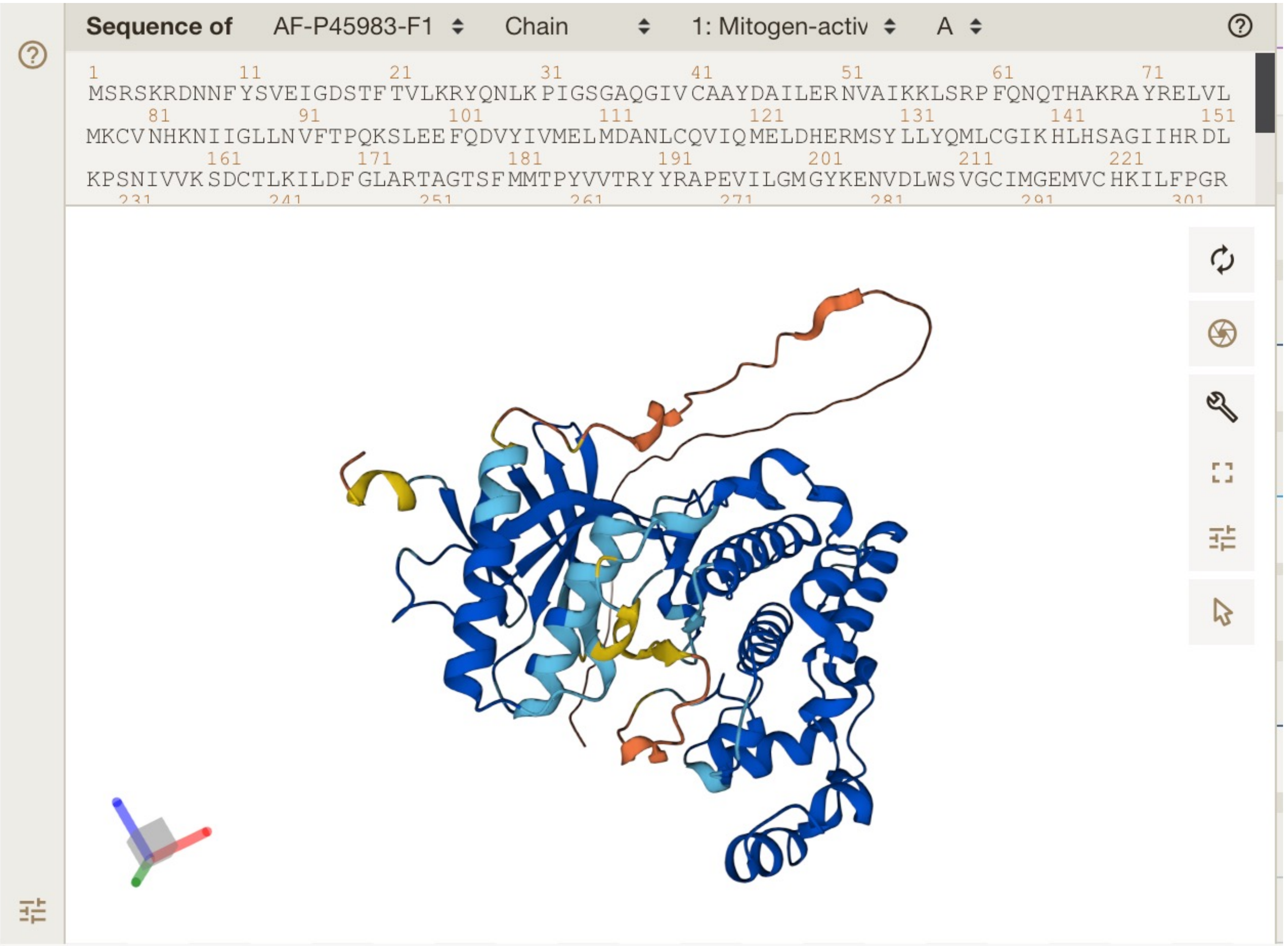
JNK1 — again

3D viewer

Model Confidence:

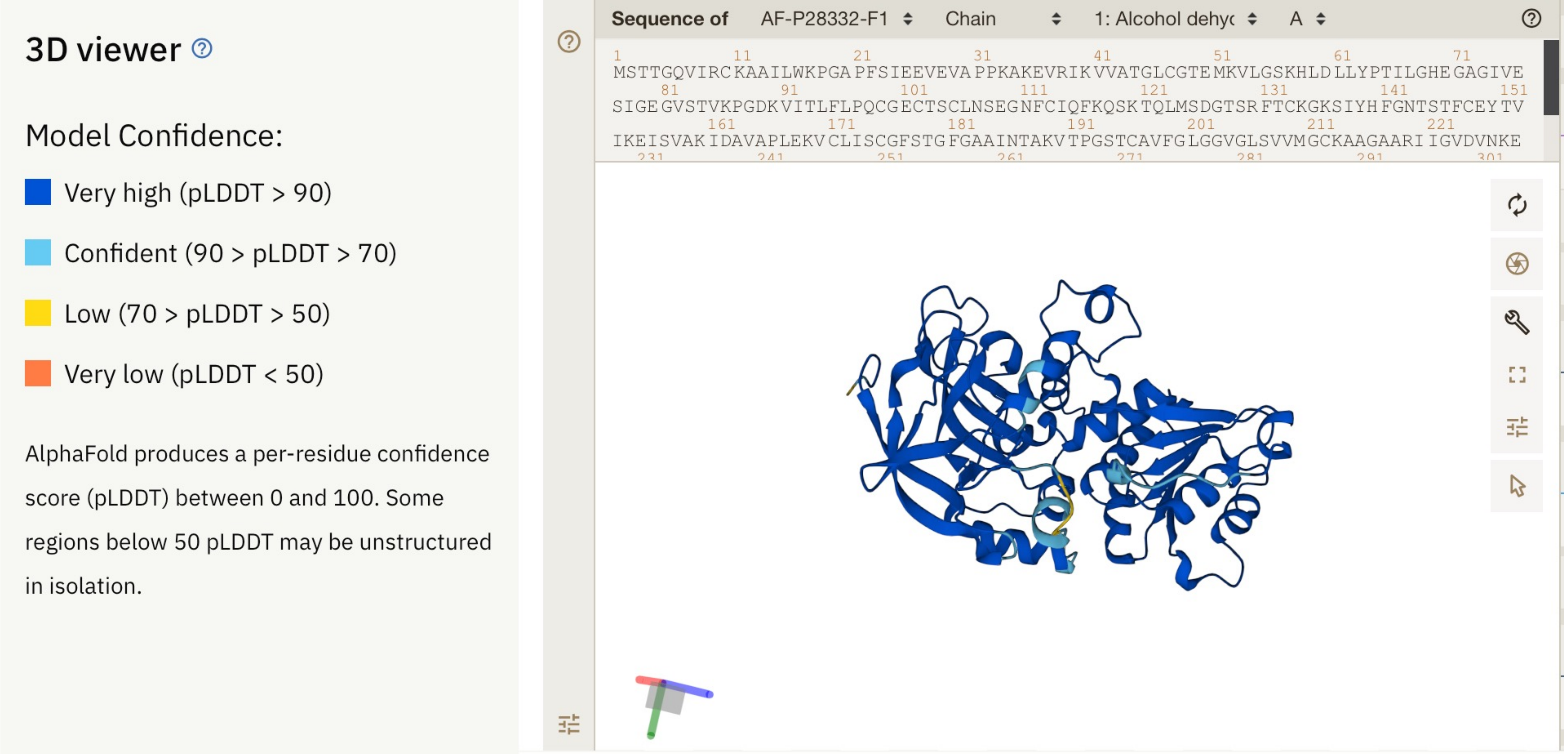
- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.



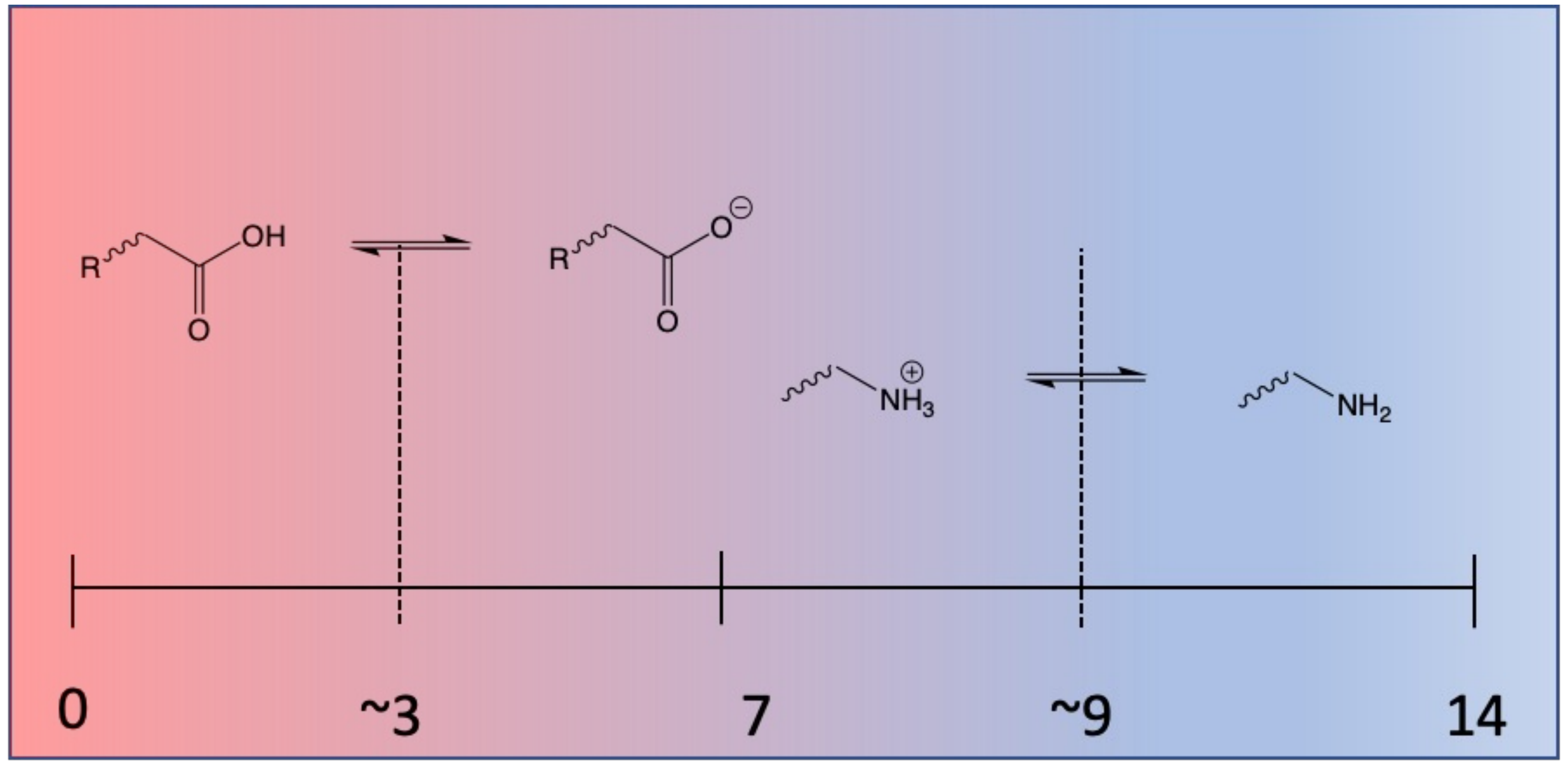
What is missing in this very good structure?

Alcohol dehydrogenase

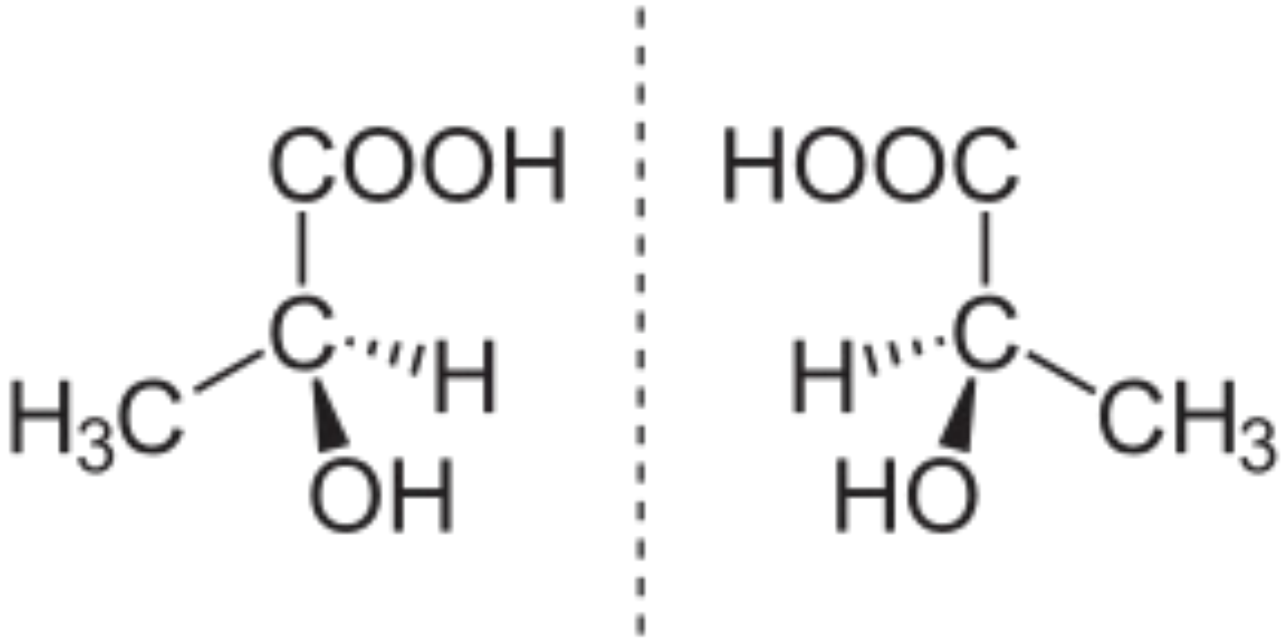


What about ligands and co-factors?

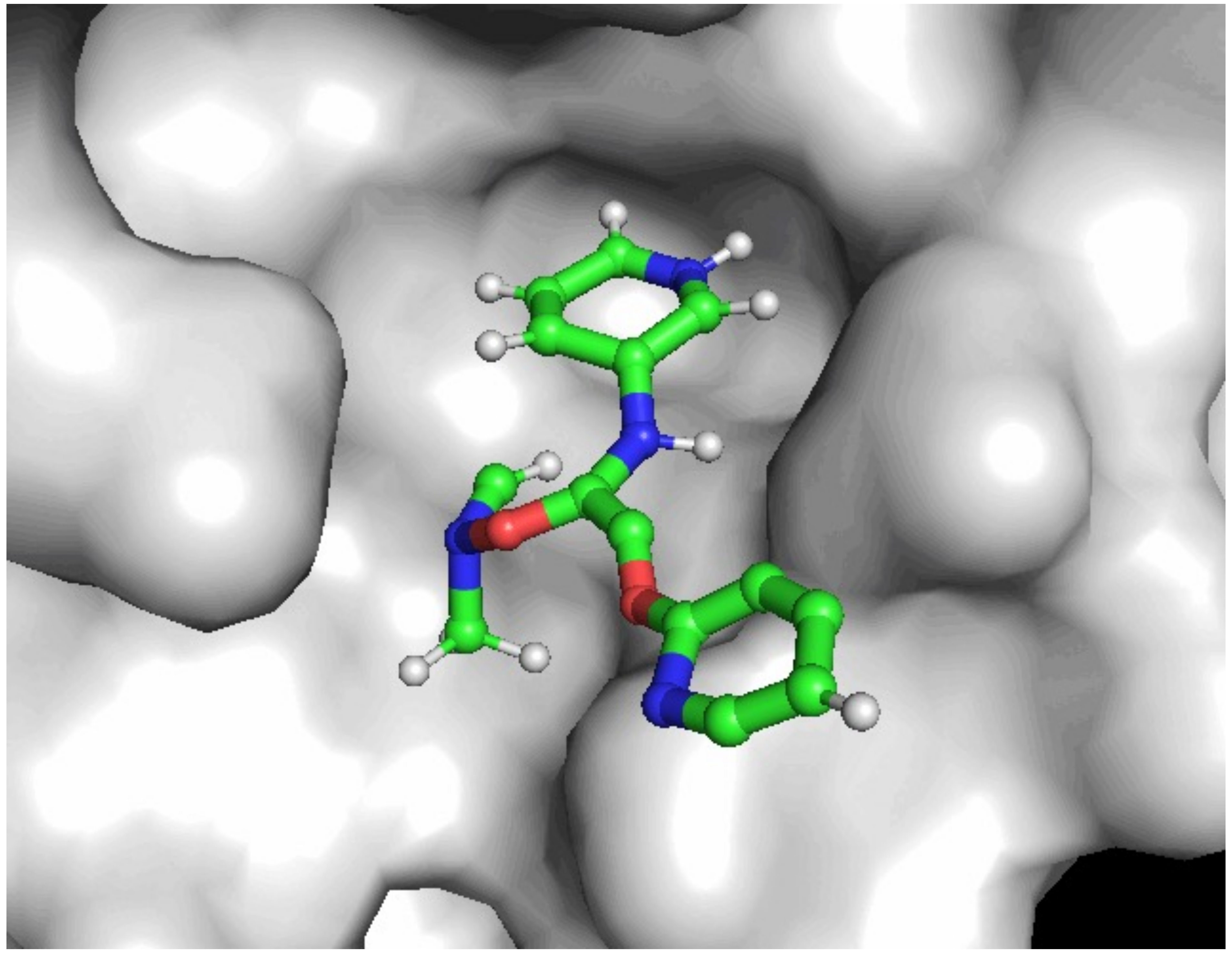
pKa for ligands is also important!



Choosing the right enantiomer



How do you get a ligand position when there is no crystal structure?



Trypsin: 1TRN

Task 1:
Your PhD supervisor has told you to go and simulate Trypsin and gave you 1TRN. Think about how to make choices in modelling

- Have a look at the protein 1TRN in the pdb and note down what you see and choices you make
- Stoichiometry (monomer or dimer protein?)
- PTMs (Does it have post translational modifications?)
- Non protein molecules (simulate or remove?)
- Disulphides (does it have disulphides?)
- Histidines (What will be the protonation state of histidines?)

https://www.rcsb.org/structure/1TRN

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | About | Documentation | Careers | COVID-19 | MyPDB | Contact us

Display Files | Download Files | Data API

Biological Assembly 1

Explore in 3D: Structure | Sequence Annotations | Electron Density | Validation Report | Ligand Interaction (ISP)

Global Symmetry: Asymmetric - C1

Global Stoichiometry: Homo 2-mer - A2

1TRN

CRYSTAL STRUCTURE OF HUMAN TRYPSIN 1: UNEXPECTED PHOSPHORYLATION OF TYROSINE 151

PDB DOI: https://doi.org/10.2210/pdb1TRN/pdb

Classification: HYDROLASE (SERINE PROTEINASE)
Organism(s): Homo sapiens
Mutation(s): No

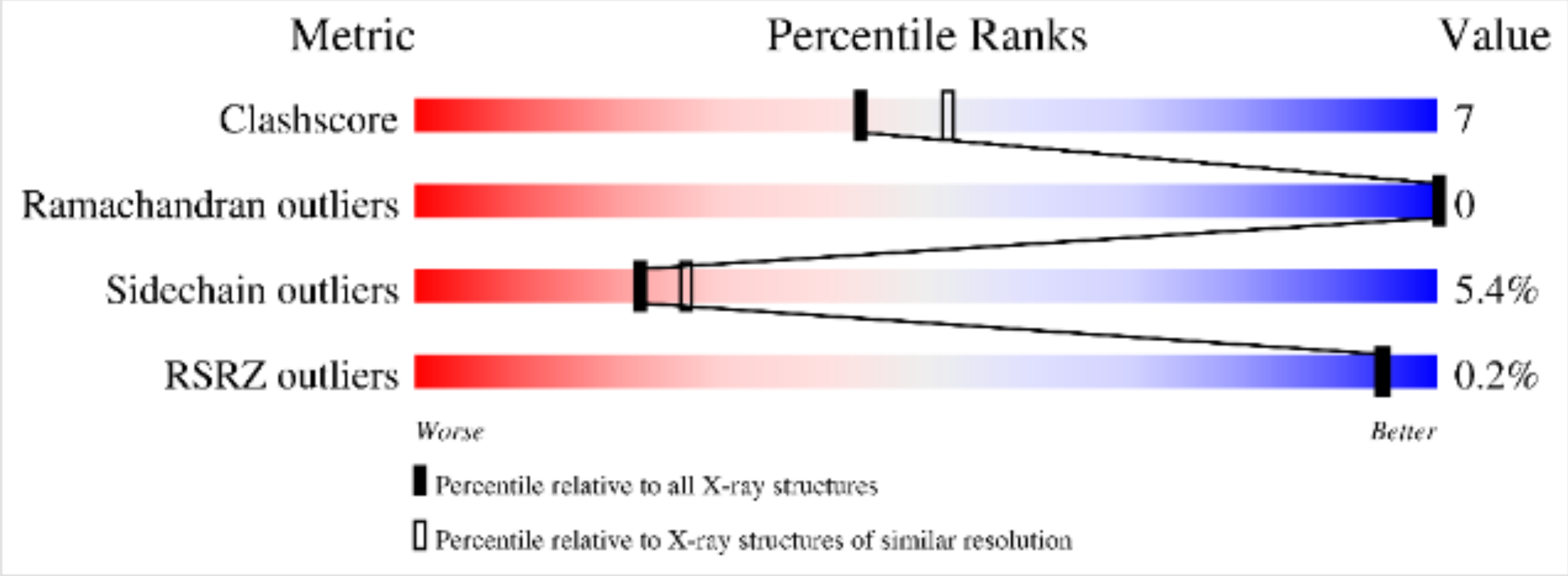
Deposited: 1995-03-16 **Released:** 1995-06-03
Deposition Author(s): Gaboriaud, C., Fontecilla-Camps, J.C.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.20 Å
R-Value Work: 0.177
R-Value Observed: 0.177

wwPDB Validation

3D Report | Full Report

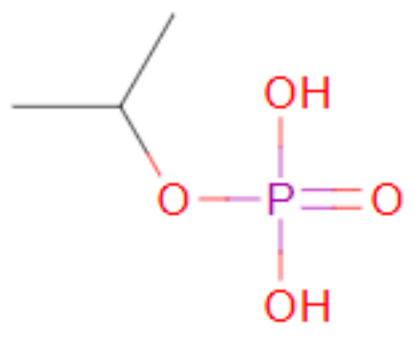


https://www.rcsb.org/structure/1TRN

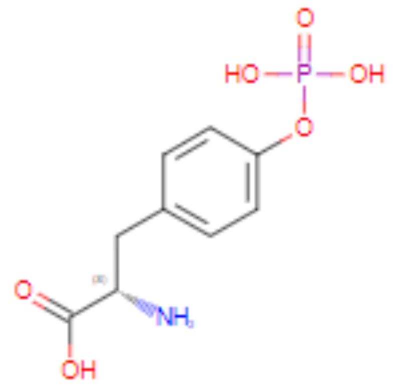
RCSB PDB Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

Small Molecules

Ligands 1 Unique

ID	Chains	Name / Formula / InChI Key	2D Diagram	3D Interactions
ISP **Query on ISP** Download Ideal Coordinates CCD File Download Instance Coordinates	C [auth A], D [auth B]	**PHOSPHORYLISOPROPANE** $C_3 H_9 O_4 P$ QPPQHRDVPBTVEV-UHFFFAOYSA-N		Interactions Interactions & Density

Modified Residues 1 Unique

ID	Chains	Type	Formula	2D Diagram	Parent
PTR **Query on PTR**	A, B	L-PEPTIDE LINKING	$C_9 H_{12} N O_6 P$		TYR


https://www.rcsb.org/structure/1TRN
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
About
Documentation
Careers
COVID-19
MyPDB
Contact us
Find proteins for P07477 (Homo sapiens)
Explore P07477
Go to UniProtKB: P07477
GTEx: ENSG00000204983
Entity Groups
Sequence Clusters
30% Identity
50% Identity
70% Identity
90% Identity
95% Identity
100% Identity
UniProt Group
P07477
Sequence Annotations
Expand
Reference Sequence
1TRN_1
0
20
40
60
80
100
120
140
160
180
200
220
1TRN_1
UNIPROT P07477
MODIFIED MONOMER
HYDROPATHY
DISORDER
DISORDERED BINDING
PFAM

https://www.uniprot.org/uniprotkb/P07477/entry

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search Help

Protein	Serine protease 1	Amino acids	247 (go to sequence)
Gene	PRSS1	Protein existence	Evidence at protein level
Status	UniProtKB reviewed (Swiss-Prot)	Annotation score	5/5
Organism	Homo sapiens (Human)		

Function
Names & Taxonomy
Subcellular Location
Disease & Variants
PTM/Processing
Expression
Interaction
Structure
Family & Domains
Sequence
Similar Proteins

Entry | Variant viewer 524 | Feature viewer | Genomic coordinates | Publications | External links | Hi

BLAST | Download | Add | Add a publication | Entry feedback

Feedback

Help

Function

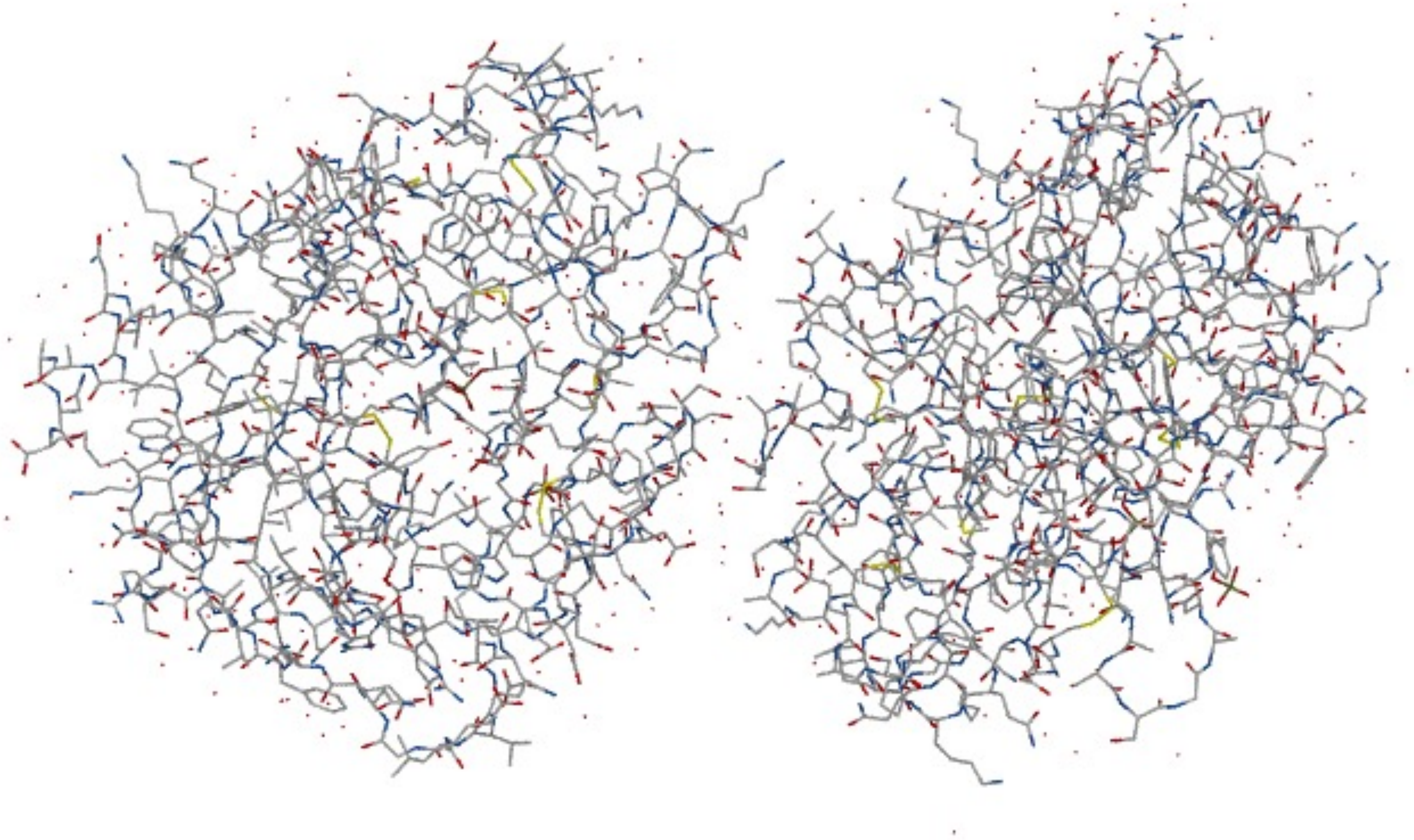
Has activity against the synthetic substrates Boc-Phe-Ser-Arg-Mec, Boc-Leu-Thr-Arg-Mec, Boc-Gln-Ala-Arg-Mec and Boc-Val-Pro-Arg-Mec. The single-chain form is more active than the two-chain form against all of these substrates.

1 Publication

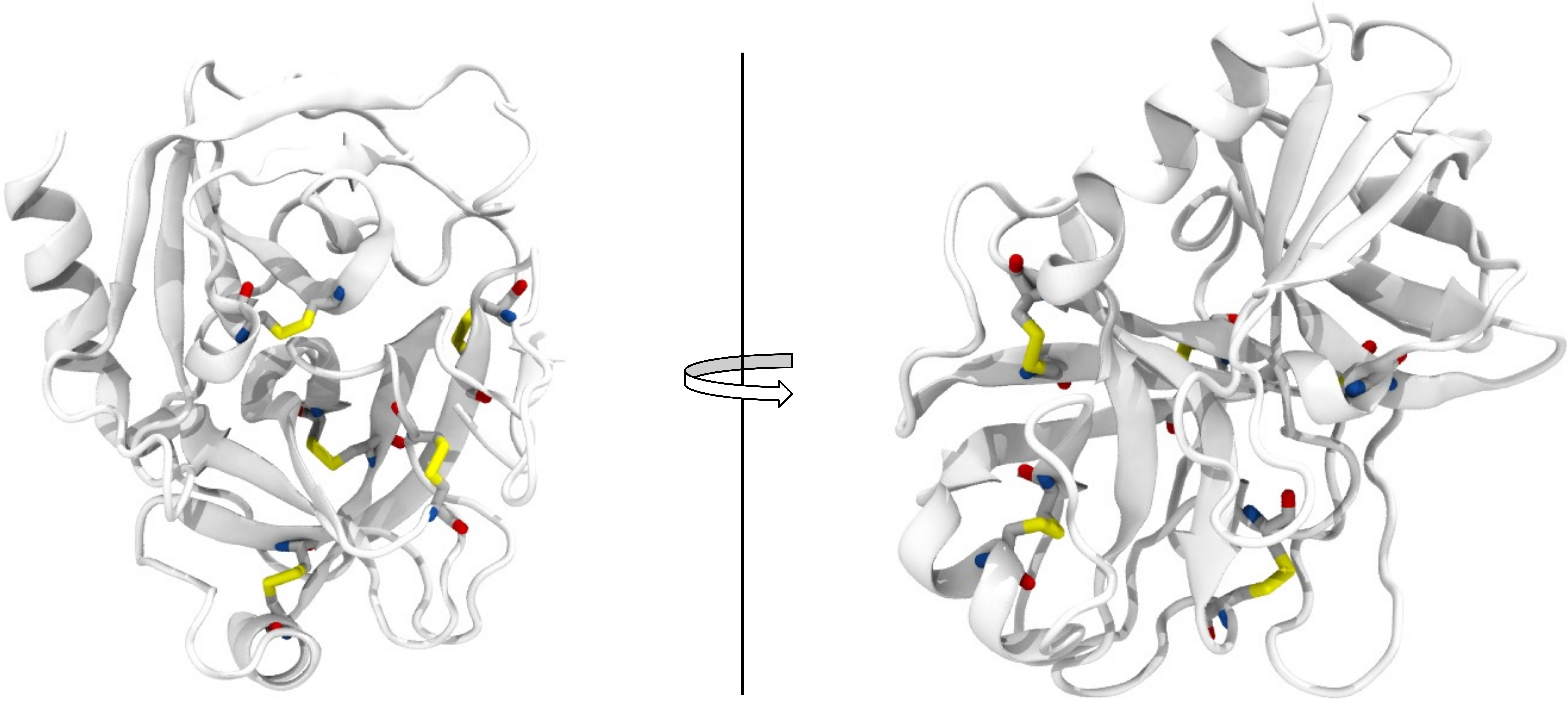
Caution

Tyr-154 was proposed to be phosphorylated (PubMed:8683601) but it has been shown (PubMed:17087724) to be sulfated instead. Phosphate and sulfate groups are similar in mass and size, and this can lead to erroneous interpretation of the results. 2 Publications

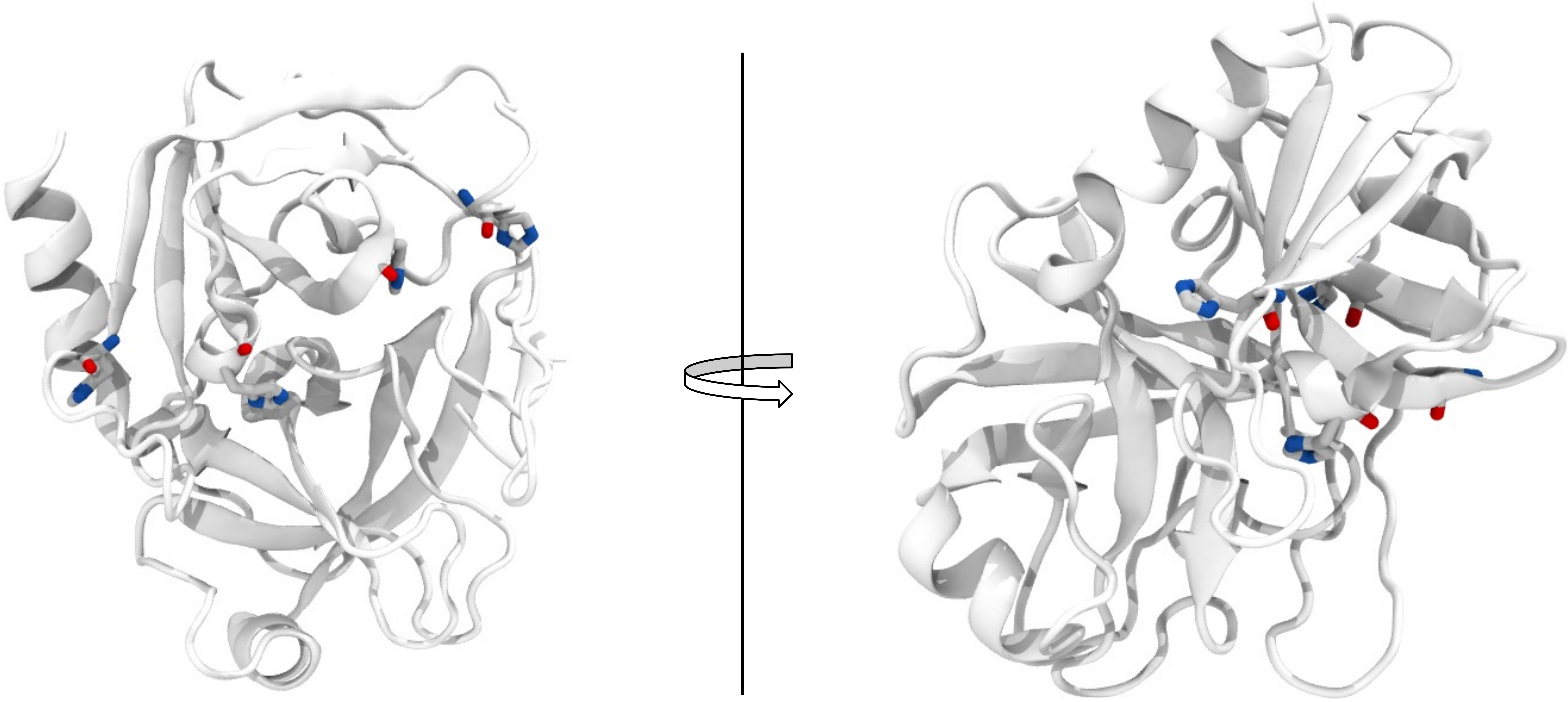
Non-protein molecules



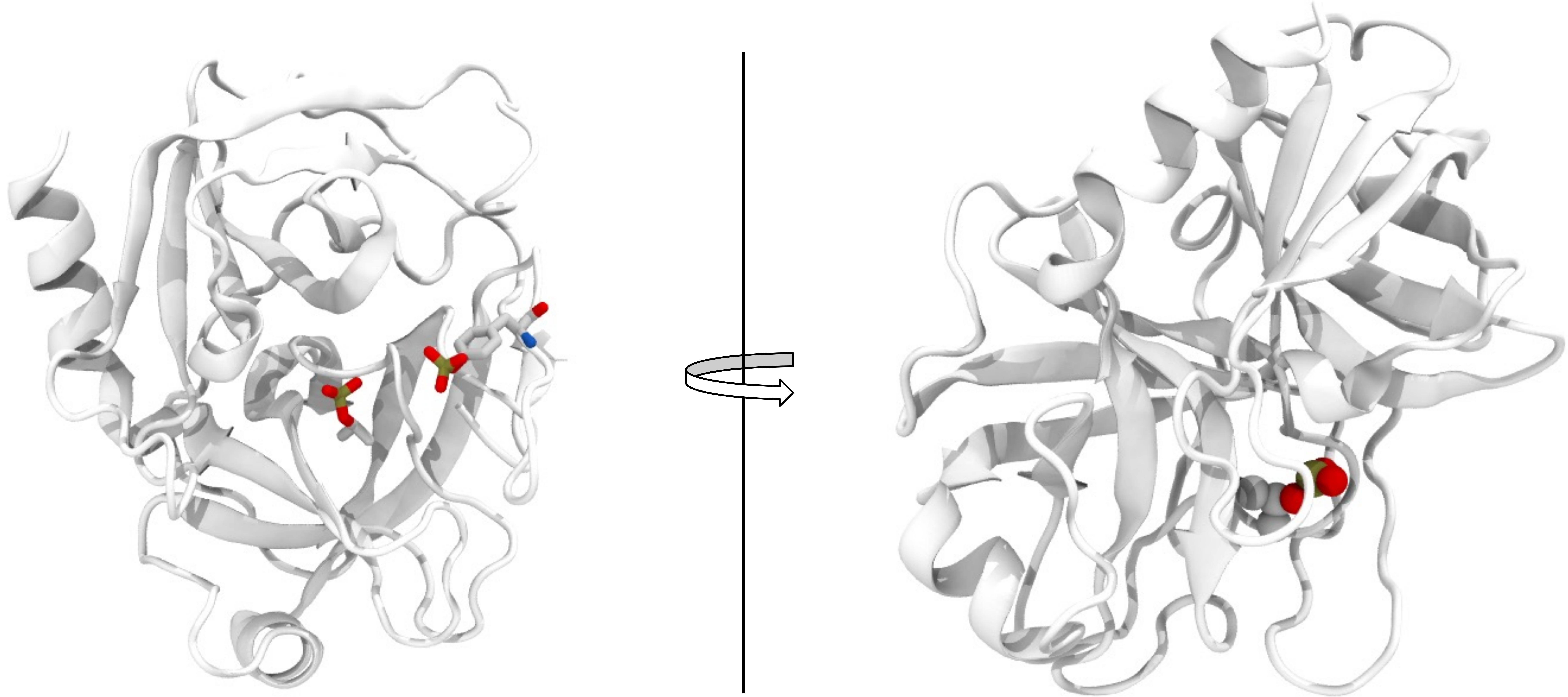
Disulphide bridges



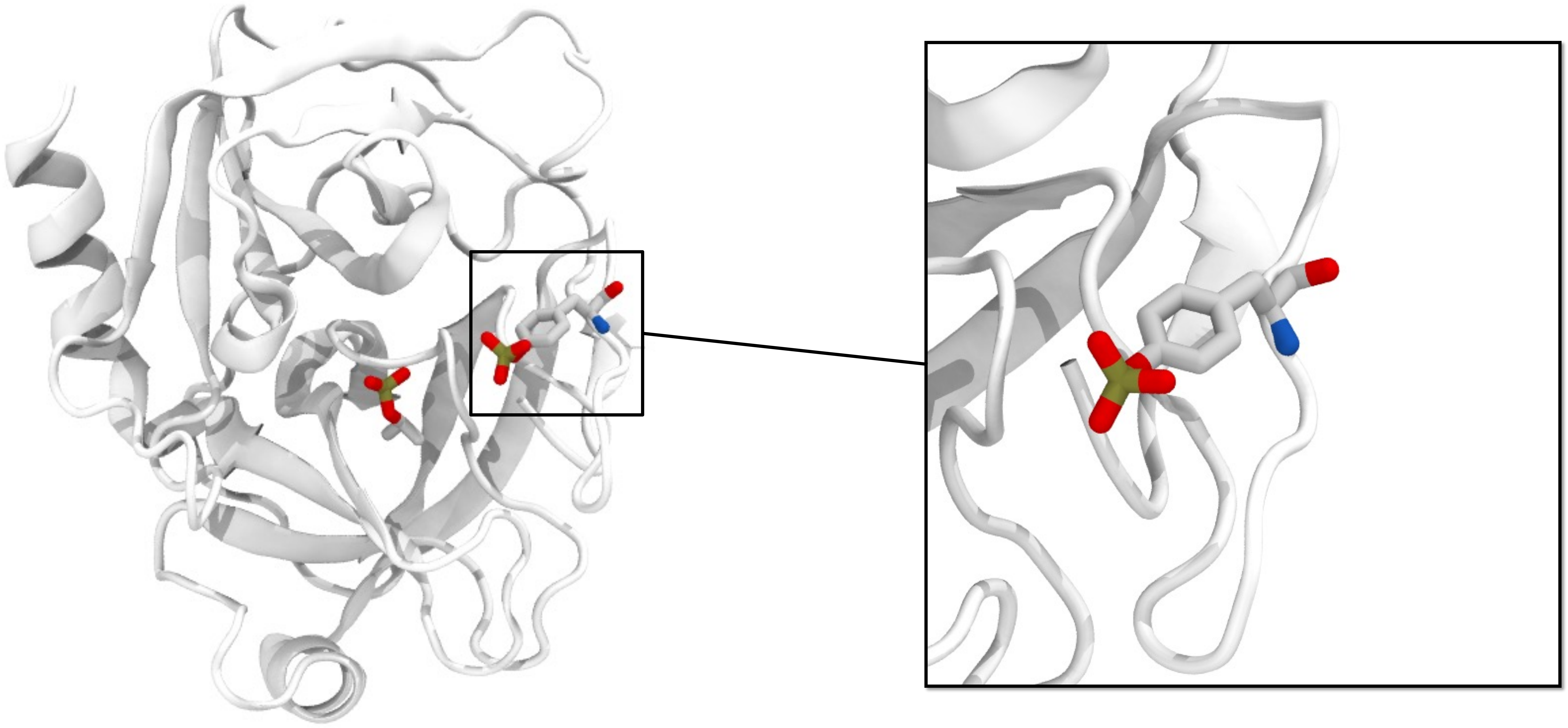
Histidines



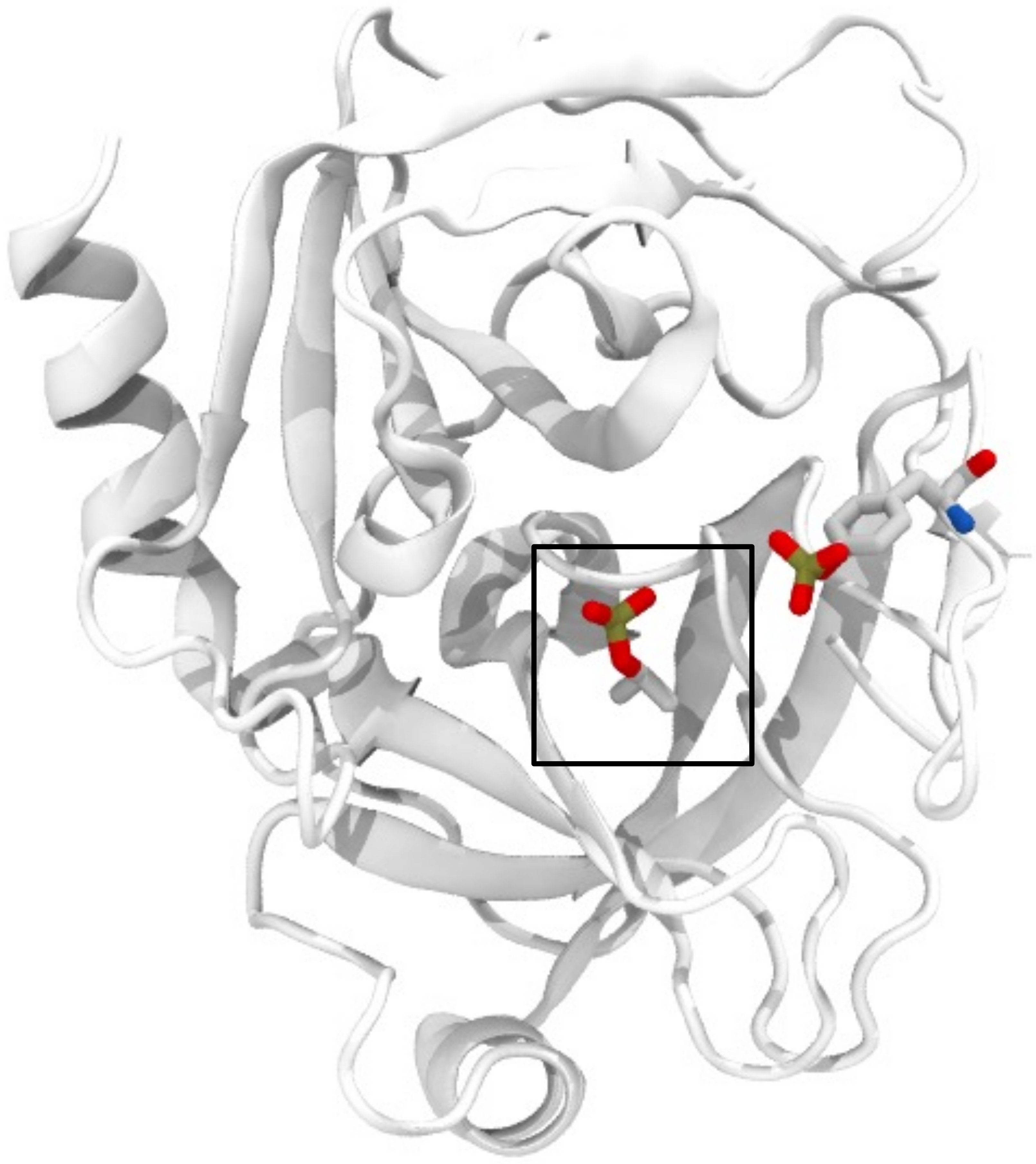
Amino acid modifications: TYR151

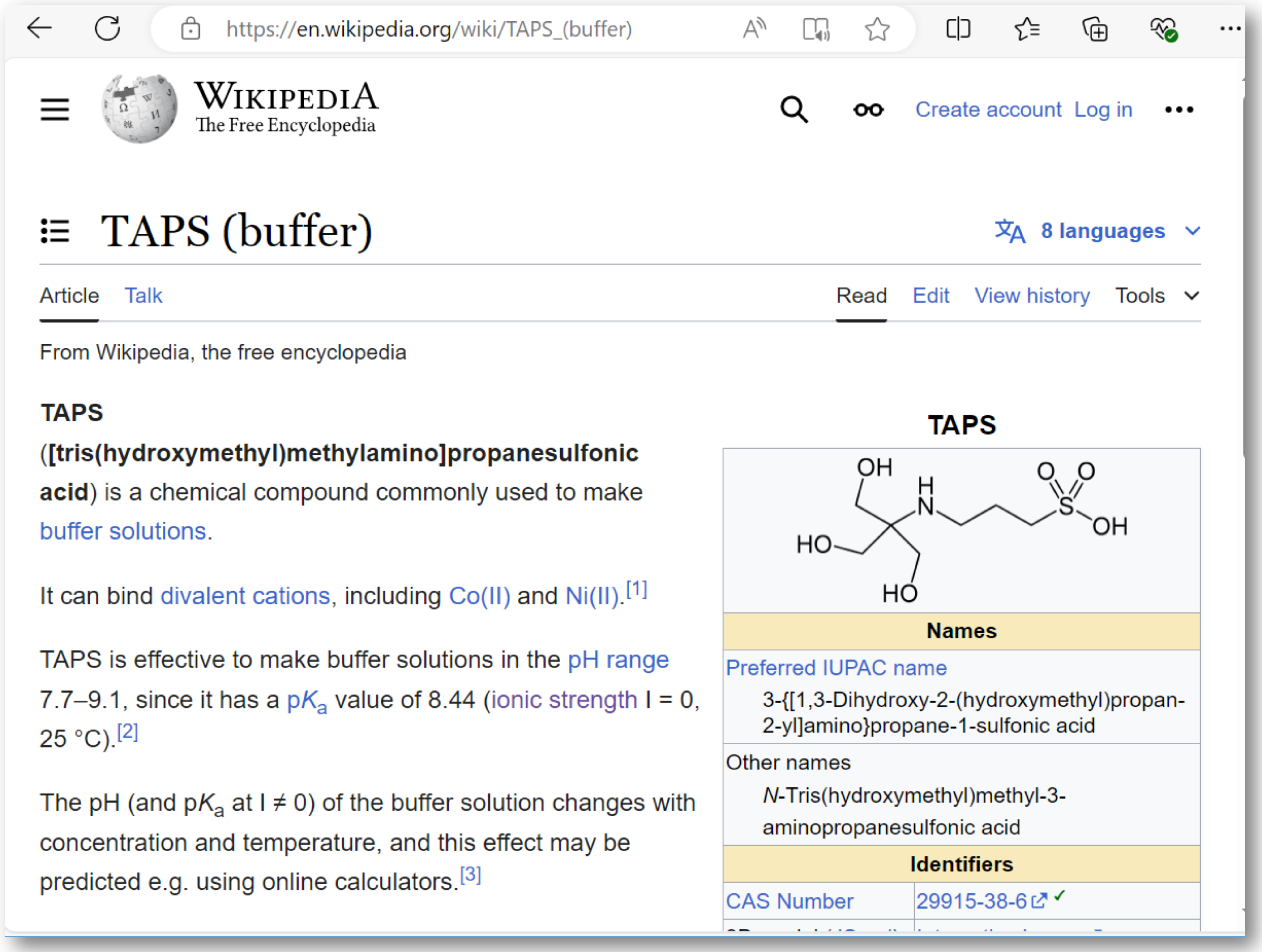


Amino acid modifications: TYR151



Ligand?




https://en.wikipedia.org/wiki/TAPS_(buffer)
Wikipedia
The Free Encyclopedia
Create account
Log in
TAPS (buffer)
8 languages
Article
Talk
Read
Edit
View history
Tools
From Wikipedia, the free encyclopedia
TAPS ([tris(hydroxymethyl)methylamino]propanesulfonic acid) is a chemical compound commonly used to make buffer solutions.
It can bind divalent cations, including Co(II) and Ni(II).[1]
TAPS is effective to make buffer solutions in the pH range 7.7–9.1, since it has a pKa value of 8.44 (ionic strength I = 0, 25 °C).[2]
The pH (and pKa at I ≠ 0) of the buffer solution changes with concentration and temperature, and this effect may be predicted e.g. using online calculators.[3]
TAPS
OH
H
N
O
O
S
OH
HO
HO
Names
Preferred IUPAC name
3-{[1,3-Dihydroxy-2-(hydroxymethyl)propan-2-yl]amino}propane-1-sulfonic acid
Other names
N-Tris(hydroxymethyl)methyl-3-aminopropanesulfonic acid
Identifiers
CAS Number
29915-38-6

Tools that will help with all of this

Most simulation packages contain tools to help add missing H-atoms. Not all work in the same way.

Where there is more than one possible answer, not all packages will **make the same decision**.

There are two independent, well-established tools designed specifically to look at these issues, and a webserver that puts them together:

Reduce

Adds hydrogens, and tests for 'NQH flips'

Download:
http://kinemage.biochem.duke.edu/software/reduce.php

PropKa

Predicts amino acid protonation states

Download:
https://github.com/jensengroup/propka-3.1

PDB2PQR server

Interface to both Reduce and proPka

Website:
http://server.poissonboltzmann.org/pdb2pqr

...

Let’s try some protein prep out!

uniprot.org

rcsb.org

https://server.poissonboltzmann.org

Task 2:
Identify a good crystal structure for **Carboxypeptidase B2** and use the Poisson Boltzmann server to produce output for amber.
Starting structure PDB: 3D67 use uniprot to find out more!

Task 3:
Take a look at the AlphaFold 2 structure, what do you need to watch out for here?

Next: Docking