



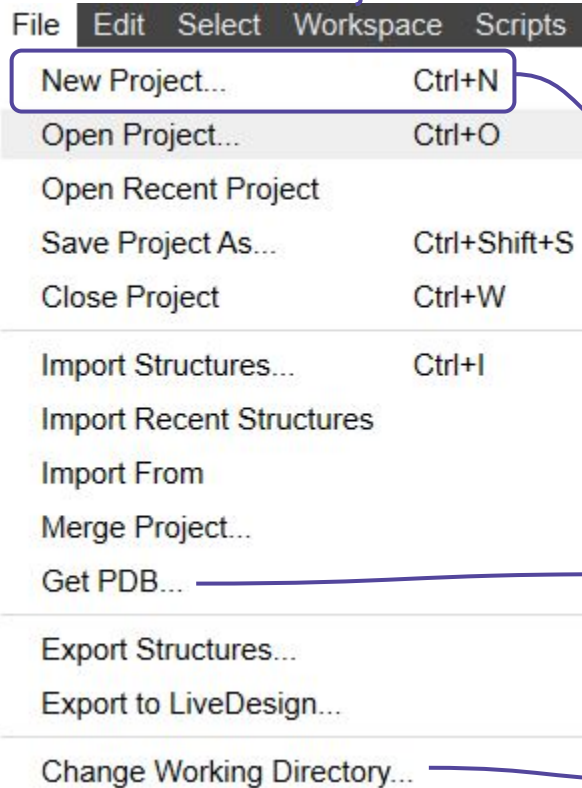
**Schrödinger**

# **A deep dive into structure preparation and analysis**

Mila Krämer, Rita Podžuna  
LRZ, 2022



# Project Setup for Day 2

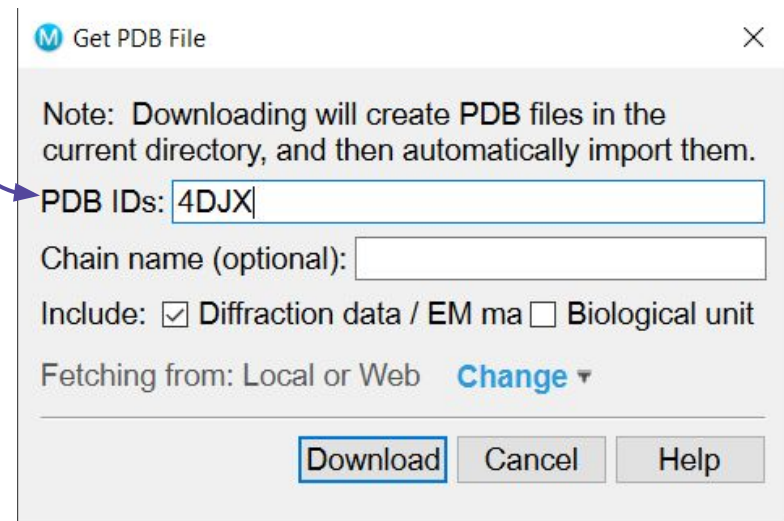
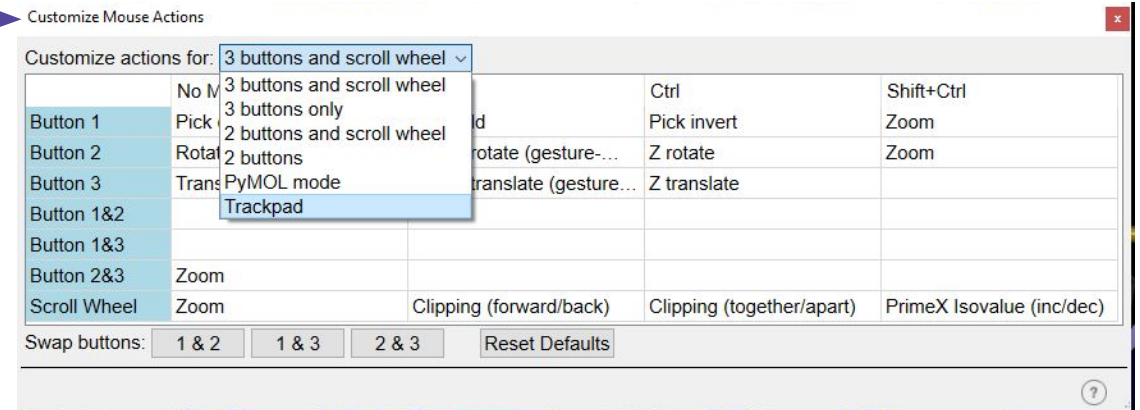


1. Choose where project data should be saved

3. Fetch BACE-1 structure from the PDB

2. Set to where Maestro should put results of calculations and other output  
My recommendation: inside project folder

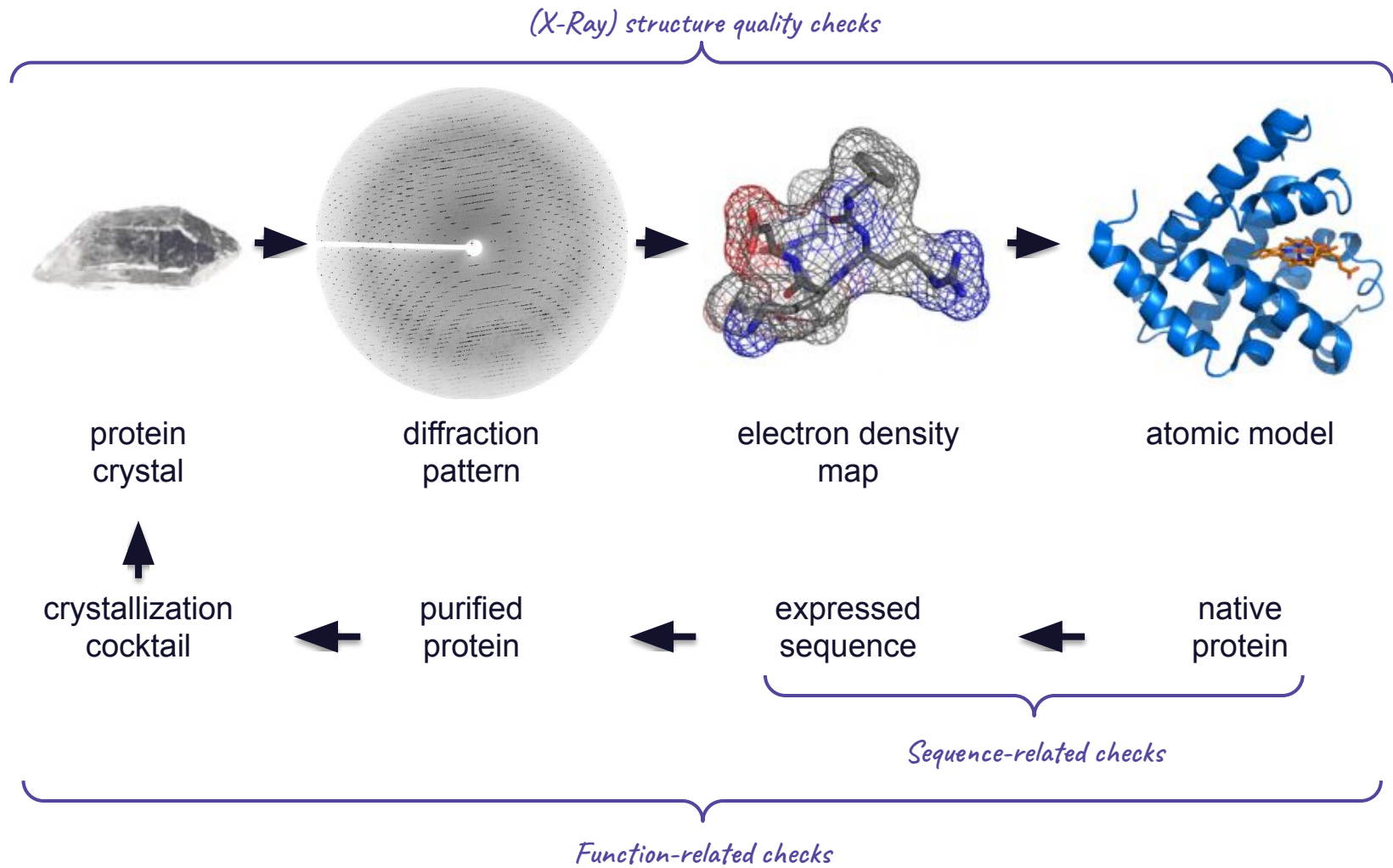
0. In case you're using a trackpad or are used to PyMOL:



# What do I need to check before using a protein structure?

What question am I trying to answer?

What calculations will I be running?





# Function-related checks

## ❑ Is the protein a monomer or a multimer?

You can find this information in UniProt in the **Interaction** section.

### Interaction<sup>i</sup>

#### Subunit structure<sup>i</sup>

Homotrimer.

Interacts with SPPL2B.

1 Publication

TNF $\alpha$

UniProt ID: P01375

#### Binary interactions<sup>i</sup>

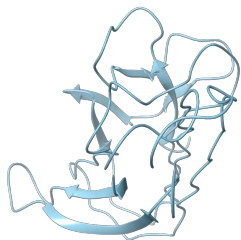
Subcellular location

Diseases

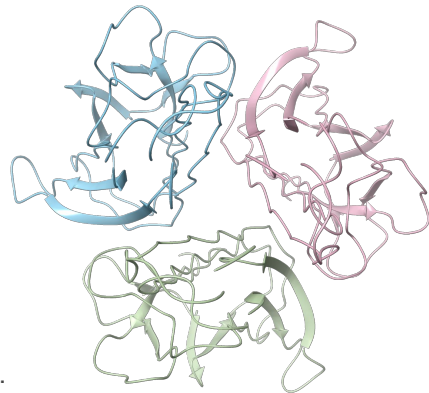
Reset filters

P01375 has binary interactions with 43 proteins

The multimeric structure is often only available through crystallographic symmetry. You can access it by downloading the **biological unit** from the PDB.



Regular entry vs the biological unit of TNF $\alpha$  trimer (PDB ID: 4TSV).



## ❑ If the protein is a multimer, is it a homomer or a heteromer?

Again, UniProt's **Interaction** section is the place to check.

**Note:** If the subunits of the multimer are encoded by different genes, then each subunit will have its own UniProt entry.

### Interaction<sup>i</sup>

#### Subunit structure<sup>i</sup>

Heterotetramer of two alpha chains and two beta chains in adult hemoglobin A (HbA); two alpha chains and two delta chains in adult hemoglobin A2 (HbA2); two alpha chains and two epsilon chains in early embryonic hemoglobin Gower-2; two alpha chains and two gamma chains in fetal hemoglobin F (HbF).

(Microbial infection) Interacts with Staphylococcus aureus protein isdB.

1 Publication

#### Binary interactions<sup>i</sup>

Subcellular location

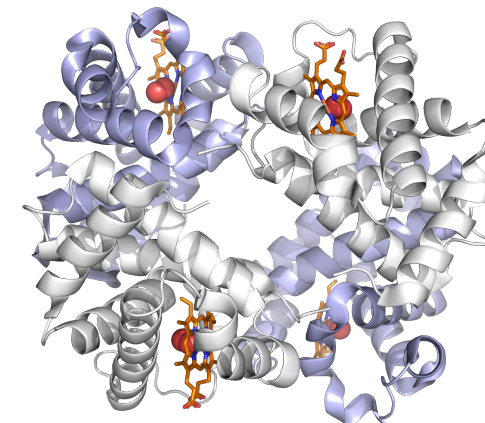
Diseases

Reset filters

P69905 has binary interactions with 13 proteins

Hemoglobin subunit  $\alpha$

UniProt ID: P69905



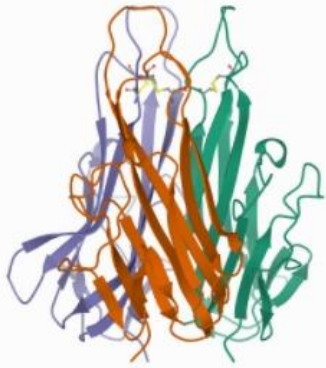
Adult hemoglobin A with bound oxygen (PDB ID: 1GZX).



# M How to get to the biological unit?

Structure Summary 3D View Annotations Experiment Sequence Genome

Biological Assembly 1 ?



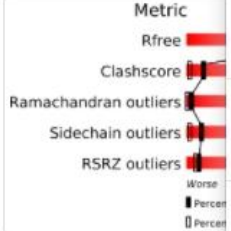
**4TSV**  
HIGH RESOLUTION CRYSTAL STRUCTURE OF A HUMAN  
DOI: 10.2210/pdb4TSV/pdb  
Classification: **LYMPHOKINE**  
Organism(s): Homo sapiens  
Expression System: Escherichia coli BL21(DE3)  
Mutation(s): Yes ⓘ

Deposited: 1997-10-29 Released: 1998-12-30  
Deposition Author(s): Cha, S.-S., Kim, J.-S., Cho, H.-S., Oh, B.-H.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION  
Resolution: 1.80 Å  
R-Value Free: 0.262  
R-Value Work: 0.200  
R-Value Observed: 0.200

wwPDB Validation ⓘ



Display Files Download Files

- FASTA Sequence
- PDB Format
- PDB Format (gz)
- PDBx/mmCIF Format
- PDBx/mmCIF Format (gz)
- PDBML/XML Format (gz)
- Biological Assembly 1**
- Structure Factors (CIF)
- Structure Factors (CIF - gz)
- Validation Full PDF
- Validation XML
- fo-fc Map (DSN6)
- 2fo-fc Map (DSN6)
- Map Coefficients (MTZ format)

3D View: Structure | Electron Density

Global Symmetry: Cyclic - C3 ⓘ (3D View)  
Global Stoichiometry: Homo 3-mer - A3 ⓘ

Find Similar Assemblies

Biological assembly 1 assigned by authors and generated by PISA,PQS (software)

This is version 1.3 of the entry. See complete [history](#).

# M How to get to the biological unit?

There are two options to download and view biological units within Maestro from the PDB:

- File -> Get PDB
- Tasks -> Browse -> Protein Preparation and Refinement -> Protein Preparation Workflow (should also be in your Favourites toolbar under Protein Preparation)

M Get PDB File

Note: Downloading will create PDB files in the current directory, and then automatically import them.

PDB IDs: 4tsv

Chain name (optional):

Include:  Diffraction data  Biological unit

Fetching from: Local or Web [Change](#)

Download Cancel Help

# M How to get to the biological unit?

If you're working with an internal structure, you can still generate the biological unit using our **command line tools**.

However, your PDB file has to have the **REMARK350** fields containing the **BIOMT** symmetry operators.

The script\* below takes a PDB file as input and creates a .mae file of the input PDB with the complete biological unit.

```
REMARK 350 BIOMOLECULE: 1
REMARK 350 AUTHOR DETERMINED BIOLOGICAL UNIT: TRIMERIC
REMARK 350 SOFTWARE DETERMINED QUATERNARY STRUCTURE: TRIMERIC
REMARK 350 SOFTWARE USED: PISA,PQS
REMARK 350 TOTAL BURIED SURFACE AREA: 5860 ANGSTROM**2
REMARK 350 SURFACE AREA OF THE COMPLEX: 18080 ANGSTROM**2
REMARK 350 CHANGE IN SOLVENT FREE ENERGY: -36.0 KCAL/MOL
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A
REMARK 350 BIOMT1 1 1.000000 0.000000 0.000000 0.000000
REMARK 350 BIOMT2 1 0.000000 1.000000 0.000000 0.000000
REMARK 350 BIOMT3 1 0.000000 0.000000 1.000000 0.000000
REMARK 350 BIOMT1 2 -0.500000 -0.866025 0.000000 233.52000
REMARK 350 BIOMT2 2 0.866025 -0.500000 0.000000 -57.78121
REMARK 350 BIOMT3 2 0.000000 0.000000 1.000000 0.000000
REMARK 350 BIOMT1 3 -0.500000 0.866025 0.000000 166.80000
REMARK 350 BIOMT2 3 -0.866025 -0.500000 0.000000 173.34364
REMARK 350 BIOMT3 3 0.000000 0.000000 1.000000 0.000000
```

```
> $SCHRODINGER/run generate_biounit.py 4tsv.pdb 4tsv_biounits.mae
```





# Function-related checks

## ❑ Is the protein known for multiple conformational states?

Some enzymes can adopt **multiple conformational states** that can significantly differ from one to another in terms of RMSD.

Working with the wrong conformation of your target can be a **huge waste** of time and/or computational resources.

Unfortunately, UniProt has no sections that can help in this case.

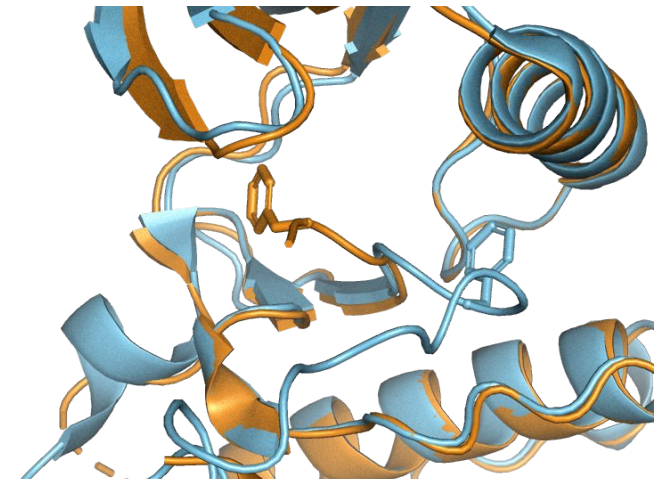
Your **best options** are to:

- Check the literature to see if any such states have been identified - structural papers are especially valuable.
- Compare all available PDB structures and make sure that the site of interest is more or less the same (i.e., w/o any large conformational changes).

Kinases are rather notorious for the amount of conformational states they explore.

For example, **DFG-in** (PDB ID: 3S3I) and **DFG-out** (PDB ID: 1KV1) are conformations in which the phenylalanine of the DFG motif in the activation loop undergoes a large shift within the ATP binding site.

Kinase inhibitors are developed to specifically target one of these two states.



p38 $\alpha$  kinase in DFG-in and DFG-out conformations.

# How to align protein structures?

Protein structures can be difficult to align due to:

- low structural similarity
- low (or no) sequence identity
- a different number of residues
- different sequence numbering

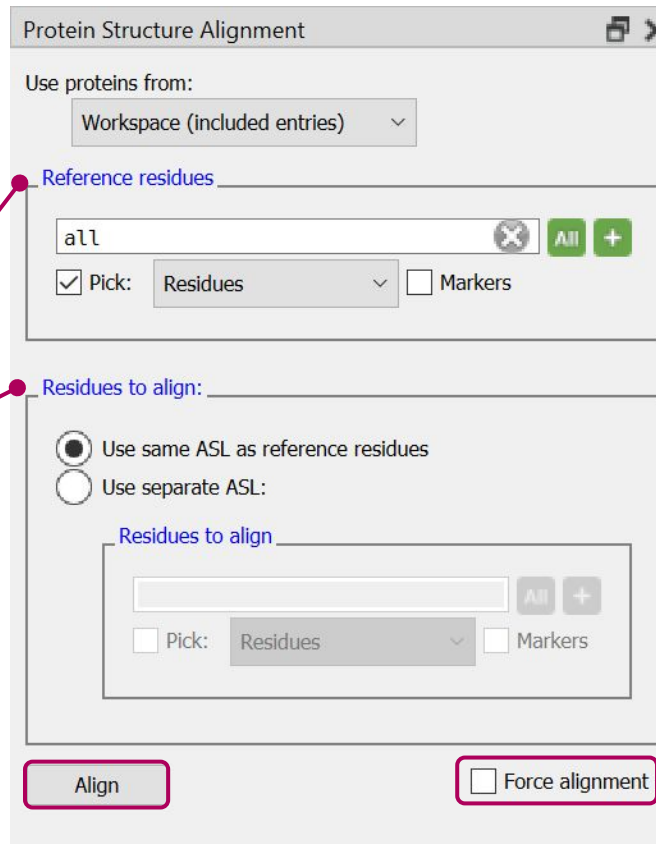
There are two options to align protein structures within Maestro:

- Tasks -> Browse -> Protein Preparation and Refinement -> Protein Structure Alignment
- Tasks -> Browse -> Protein Preparation and Refinement -> Binding Site Alignment

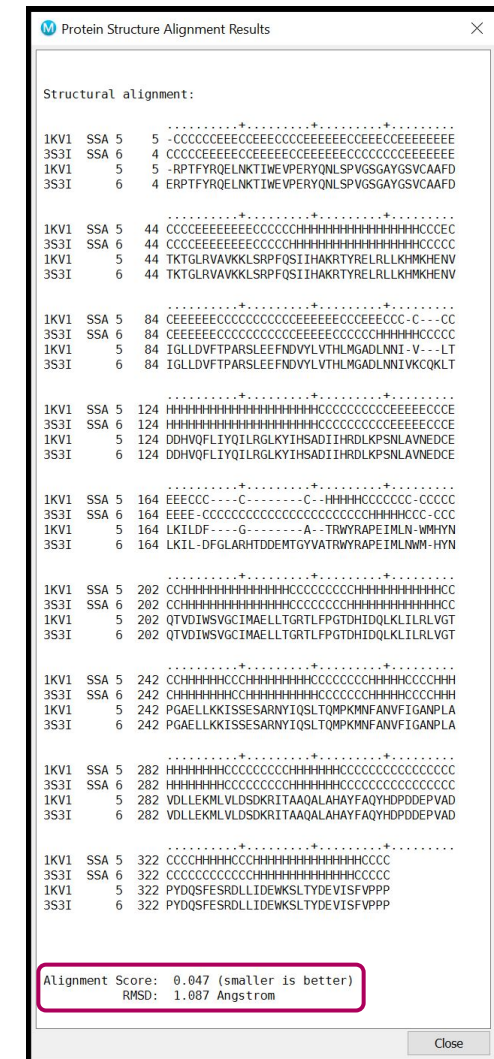
# M How to align protein structures?

Protein Structure Alignment uses **secondary structure elements** for the alignment to the **reference structure** (the one with the lowest entry number).

You can use ASL to define which residues are used for the alignment. Make sure they contain **at least one secondary structure element** to obtain a meaningful alignment.



*aligns selected residues even w/o sufficient similarity*



*alignment scores above 0.7-0.8 indicate insufficient similarity*



# M How to align protein structures?

The binding site alignment algorithm first runs a **global structural alignment** and then automatically generates the list of  $C\alpha$  atoms to use in a **pairwise alignment** from the selected residues.

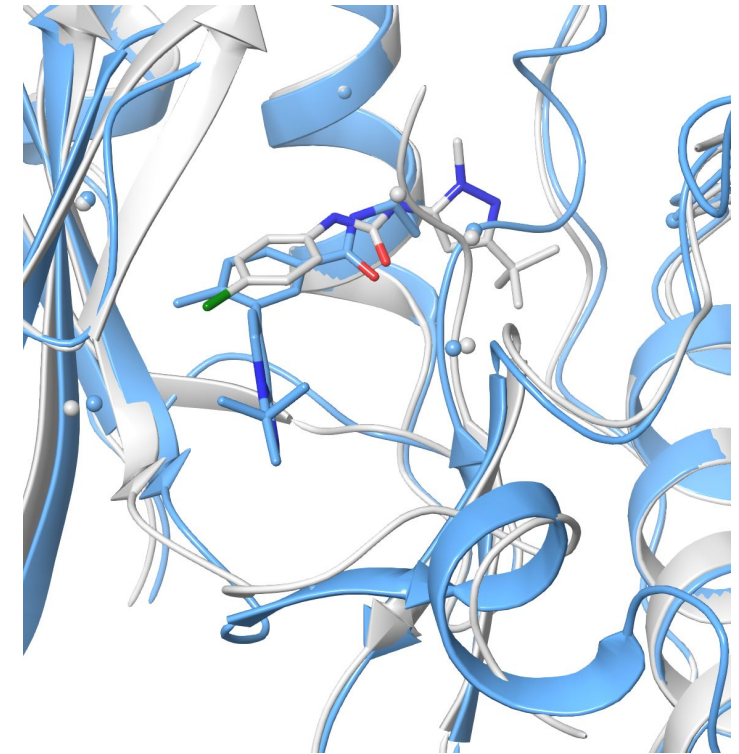
*either detect the residues automatically based on the distance from the ligand or pick them yourself*

*determines which atoms are used for the pairwise alignment*

*skips the global alignment*

*skips the global alignment and calculates  $C\alpha$  RMSD*

The screenshot shows the 'Align Binding Sites' dialog box. At the top, it says 'Use proteins from: Workspace (2 included entries)'. Under 'Residues for alignment', there are two main options: 'Automatically detect binding site residues' (selected) and 'Manually select residues'. The 'Automatically detect' option has a sub-option 'Detect ligand automatically' (selected) and 'Use molecule number:'. There is also a 'Pick ligand' checkbox. The 'Manually select residues' option has an 'Align residues:' field and a 'Select residues...' button, along with a 'Pick a residue to include/exclude' checkbox. Below this, there is an 'Ignore atom pairs greater than 5.0 Å apart' checkbox. At the bottom, there are checkboxes for 'Structures are pre-aligned' and 'In-place (calculate RMSD only)'. The 'Job name:' field contains 'align\_binding\_sites\_1' and a 'Run' button is highlighted with a red box. At the very bottom, it says 'Host=localhost:1, Incorporate=Append new entries as a new grc'.



Aligned binding sites of two PDB structures of p38 $\alpha$  kinase (white - PDB ID: 1KV1; blue - PDB ID: 3S3I).  $C\alpha$  atoms used for the pairwise alignment are shown as white and blue spheres.

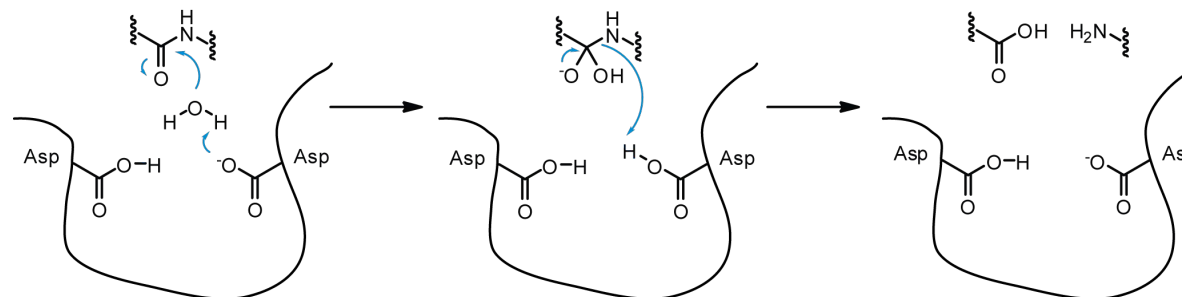
# Function-related checks

## ❑ What about atypical chemical forms?

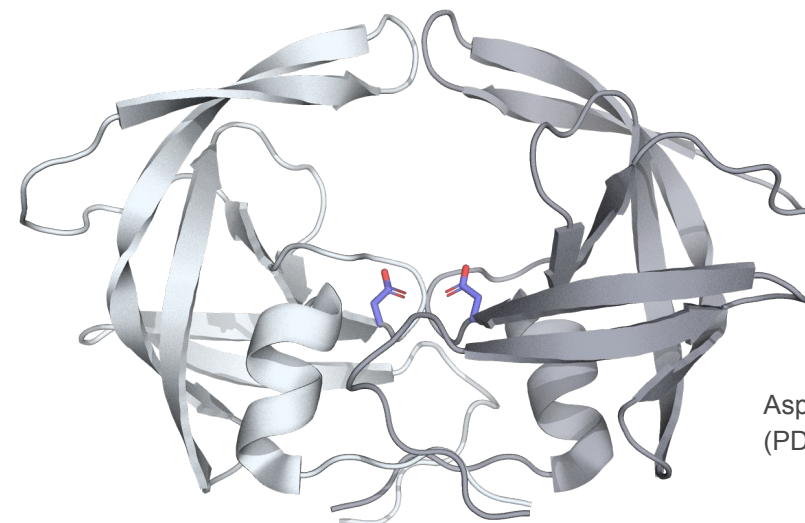
When preparing enzymes, you should also consider whether any of the residues assume an **uncommon protonation/tautomerization** state as a part of the mechanism of action.

It's always best to **check relevant literature** to make sure that you are working with the correct form of the protein for the scientific problem you're trying to solve. (No helpful UniProt section).

While **empirical methods** are certainly getting better at predicting pKa values, they can still make mistakes, so always double check whether the desired protonation state has been assigned to relevant residues.



The protonation state of **Asp-dyads in aspartic proteases** (where one Asp is protonated and the other deprotonated) is crucial for their mechanism of action and has been driving the inhibitor design.



Asp-dyad in HIV-1 protease (PDB ID: 3HVP).

# Function-related checks

## ❏ Maybe there are some PTMs?

Post-translational modifications are **covalent modifications** of proteins that involve either a proteolytic cleavage or the addition of a modifying group to an amino acid.

**More than 200 PTMs** have been characterized up to date.

PTMs can modulate protein's activity state, localization, turnover, and interactions with other proteins. It's therefore imperative to know if PTMs are involved in your scientific question.

### PTM / Processing<sup>i</sup>

Molecule processing					
Feature key	Position(s)	Description	Actions	Graphical view	Length
Chain <sup>1</sup> (PRO_0000185728)	1 – 636	Tumor protein p73	<a href="#">Add</a> <a href="#">BLAST</a>		636
Amino acid modifications					
Feature key	Position(s)	Description	Actions	Graphical view	Length
Modified residue <sup>i</sup>	27	Phosphothreonine; by PLK1 <a href="#">2 Publications</a>			1
Modified residue <sup>i</sup>	28	Phosphotyrosine; by SRC and HCK <a href="#">1 Publication</a>			1
Modified residue <sup>i</sup>	99	Phosphotyrosine; by ABL1 <a href="#">1 Publication</a>			1
Cross-link <sup>1</sup>	627	Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO); in isoform Alpha <a href="#">1 Publication</a>			
Cross-link <sup>1</sup>	627	Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2) <a href="#">Combined sources</a>			

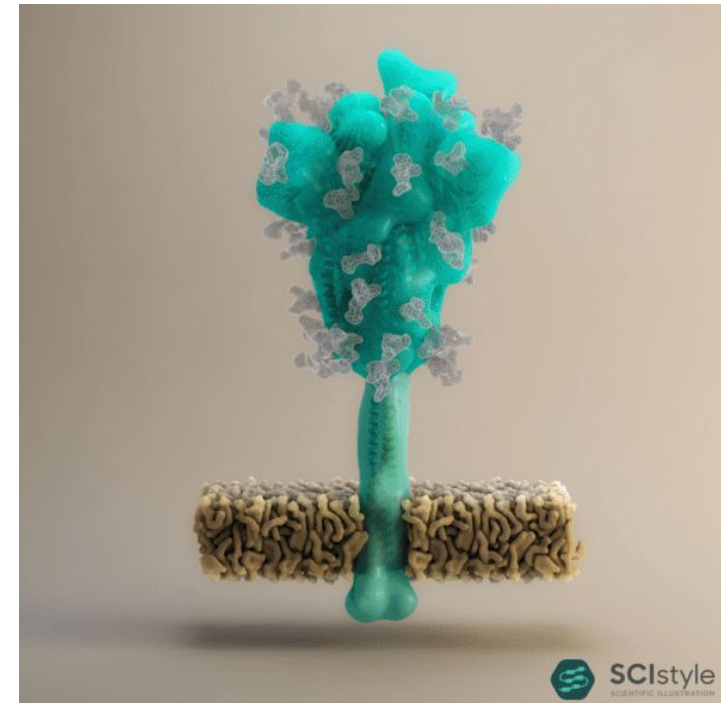
### Post-translational modification<sup>i</sup>

Isoform alpha (but not isoform beta) is sumoylated on Lys-627, which potentiates proteasomal degradation but does not affect transcriptional activity. Phosphorylation by PLK1 and PLK3 inhibits the transcription regulator activity and pro-apoptotic function. [1 Publication](#)

Higher levels of phosphorylation seen in the brain from patients with Huntington disease.

Polyubiquitinated by RCHY1/PIRH2; leading to its degradation by the proteasome. [2 Publications](#)

While **phosphorylation** is without a doubt the most common PTM, the recent COVID pandemic has shone a light on **glycosylation** as N-glycans linked to the spike protein help the SARS-CoV2 virus stay hidden from the host immune system.

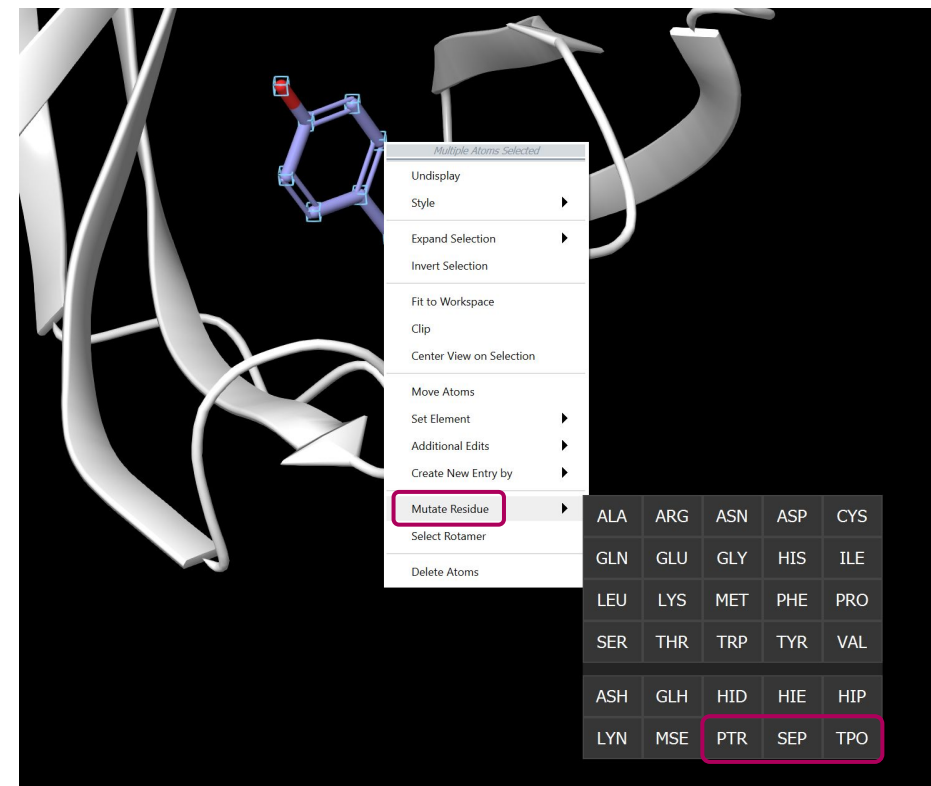
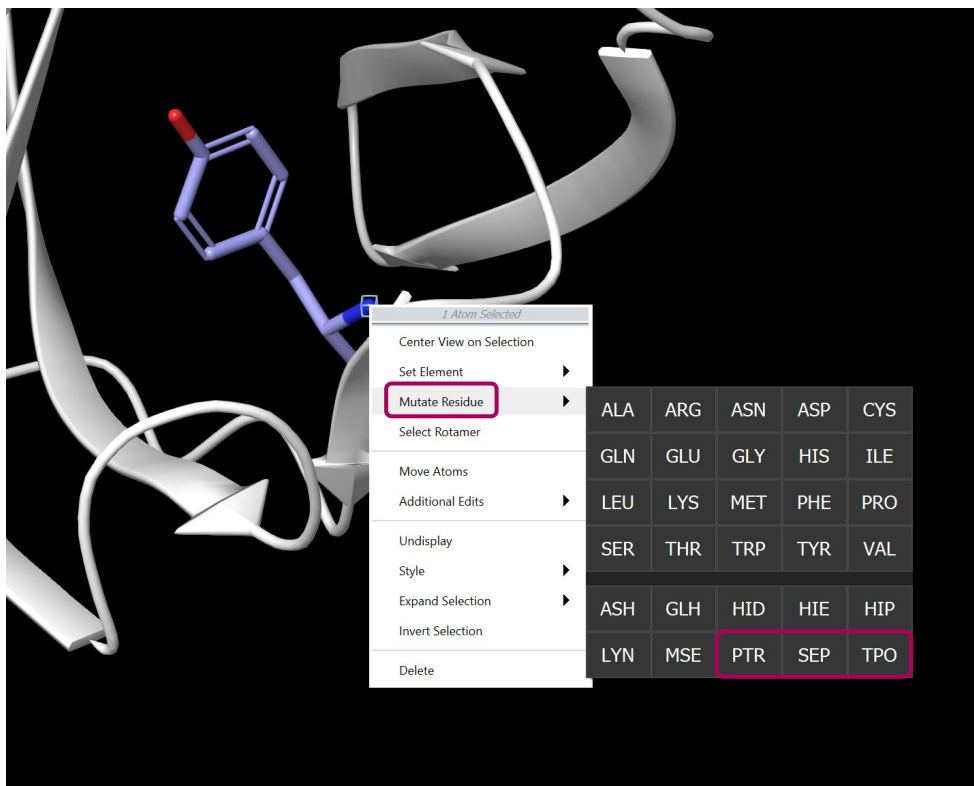


Tumor protein p73  
UniProt ID: O15350

Thomas Splettstößer @ scistyle.com

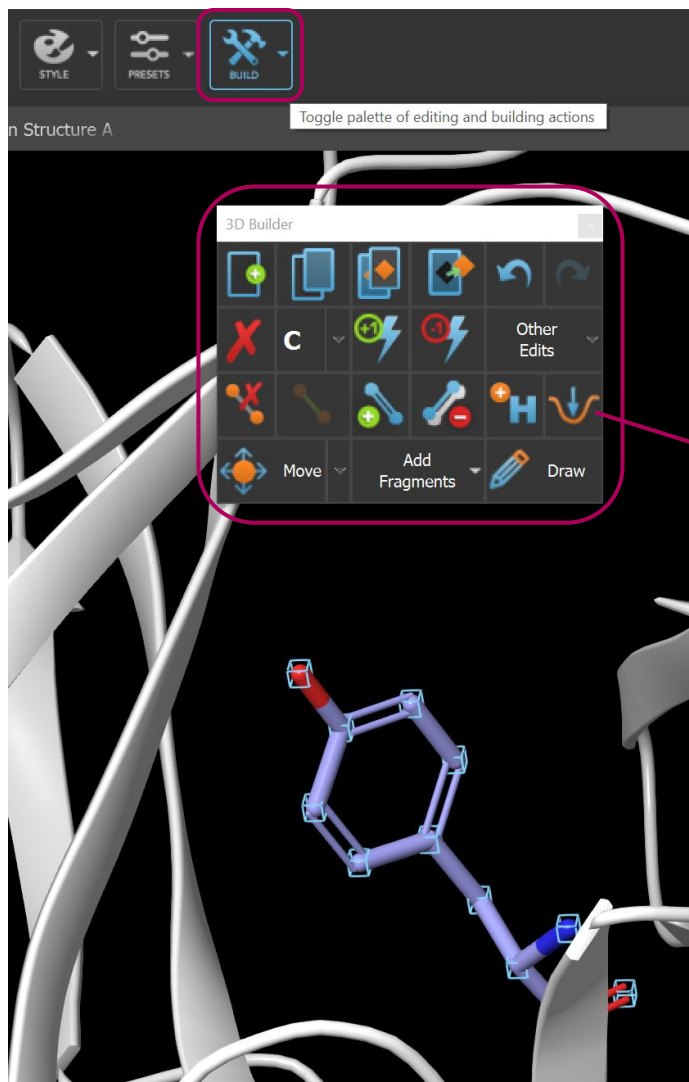


# M How to add PTMs?



Right-clicking on either a single-atom or a multi-atom selection will open a menu with 'Mutate Residue' option which allows you to **phosphorylate** Ser, Thr, or Tyr.

# M How to add PTMs?



- **3D builder** allows you can create any kind of custom residue.
- It's advisable to minimize the custom residue and its neighbors afterward.

- Another option is to use the **Nonstandard Residues** panel from **Biologics** which allows you to build your own database of nonstandard residues and use it in **Residue Scanning Calculations** to introduce mutations.

Name	Code	Structure	Locked	Mutat
ALA	A	<chem>CC(N)C(=O)O</chem>	Standard Built-in	<input checked="" type="checkbox"/>
ARG	R	<chem>CCC(NC(=[NH2+])N)C(=O)O</chem>	Standard Built-in	<input checked="" type="checkbox"/>
ARN	R	<chem>CCC(NC(=O)N)C(=O)O</chem>	Standard Built-in	<input checked="" type="checkbox"/>
ASH	D	<chem>CC(O)C(=O)O</chem>	Standard Built-in	<input checked="" type="checkbox"/>
ASN	N	<chem>CC(N)C(=O)O</chem>	Standard Built-in	<input checked="" type="checkbox"/>

# Function-related checks

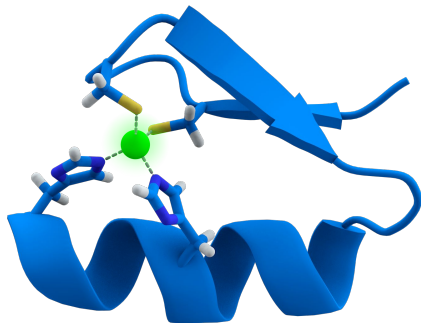
## ❑ Are any metals involved?

**Metalloproteins** require metal ions for their function.

In order to properly model such proteins, it is important to consider the **oxidation state** of the metal ion.

Low-resolution X-ray or cryo-EM structures can contain metal ions whose **coordination** by surrounding residues hasn't been modelled properly.

Such issues might require some manual building during protein preparation.



Zinc finger motif where Zn ion is coordinated by 2 Cys and 2 His residues (PDB ID: 1A1L).

Thomas Splettstößer @ scistyle.com

## ❑ Does the protein bind any other cofactors?

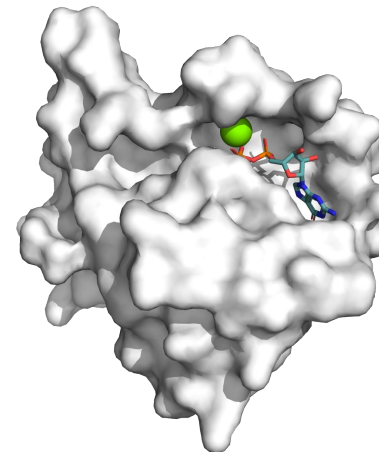
Cofactors are any **non-protein compounds** required for protein's activity.

Cofactor<sup>i</sup>

Zn<sup>2+</sup> 22 Publications, Co<sup>2+</sup> 1 Publication

Note: Zinc. Can also use cobalt(II) with lower efficiency, but not copper(II), nickel(II) and manganese(II). 1 Publication

UniProt doesn't list cofactors that are part of the **catalytic reaction** (e.g., NAD, FAD, ATP), so it's important to consult the literature to ensure you're modelling correct protein cofactors.



K-Ras has a pM affinity towards GTP/GDP (PDB ID: 4EPV) and there are almost no *apo* PDB structures of K-Ras.

# The final checklist

## Function-related checks

- What's the subcellular location of the protein?
- Is the protein a monomer or a multimer?  
If a multimer, is it a homomer or a heteromer?
- Is the protein known for multiple conformational states?
- What about atypical chemical forms?
- Maybe there are some PTMs?
- Are any metals involved?
- Does the protein bind any other cofactors?

# Sequence-related checks





# Sequence-related checks

## ❑ Is the whole protein there? Any missing (sub)domains?

Proteins very often get chopped up into smaller functionally relevant segments that are **easier to crystallise** than the full protein.

It's always a good idea to check UniProt and relevant literature to make sure you have the correct parts of the protein in your structure.

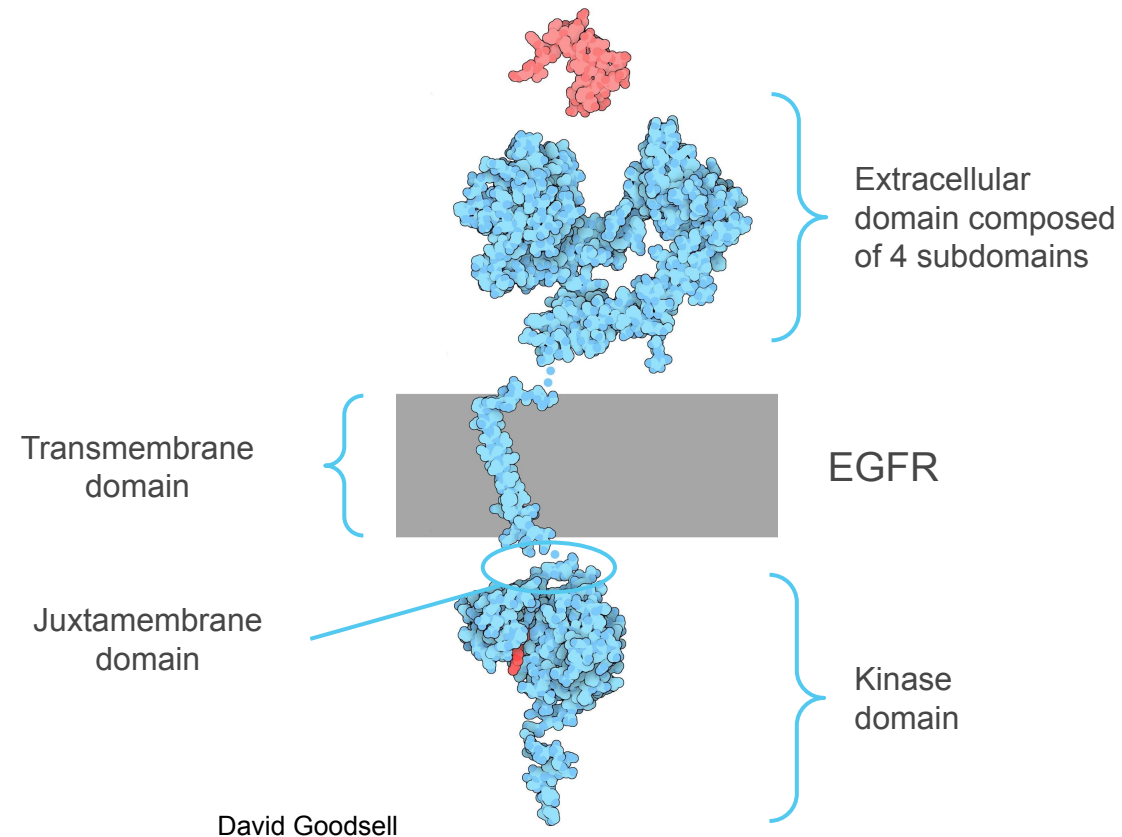
EGFR (UniProt ID: P00533)

### Topology

Feature key	Position(s)	Description	Actions	Graphical view	Length
Topological domain <sup>i</sup>	25 – 645	Extracellular <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		621
Transmembrane <sup>i</sup>	646 – 668	Helical <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		23
Topological domain <sup>i</sup>	669 – 1210	Cytoplasmic <a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		542

Whether you need the complete protein will certainly depend on the problem you want to study:

- If you are only interested in running a structure-based virtual screen to find kinase domain inhibitors, then that domain is sufficient.
- If you want to study the effects of somatic mutations in the extracellular domain on the kinase activity, then you will need the full protein.



# Sequence-related checks

## ❑ Are you working with the correct sequence?

Your protein structure could contain mutations that have been introduced:

- For functional reasons, e.g. to solve the structure of an (in)activating mutation.
- To facilitate protein crystallisation.

It's always recommended to compare the sequence of your structure with the **canonical** one (obtained from UniProt's **Sequence** section) using **sequence alignment**.

**Isoform 1** (identifier: **P00533-1**) [UniParc] [FASTA](#) [Add to basket](#)

Also known as: p170

*This isoform has been chosen as the canonical<sup>1</sup> sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.*

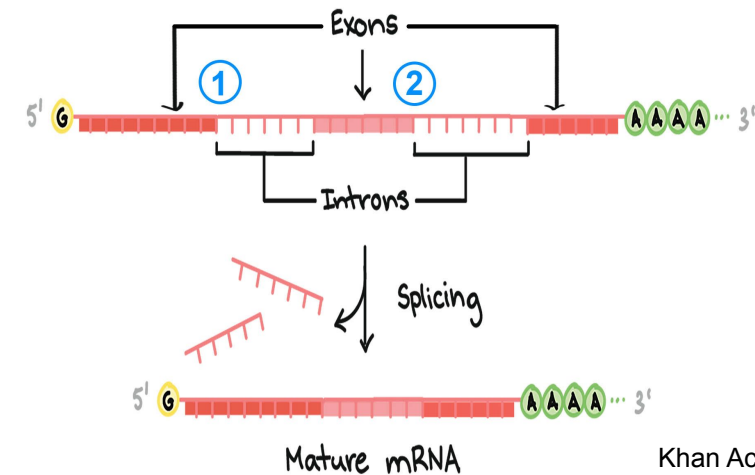
[« Hide](#)

10	20	30	40	50
MRPSGTAGAA	LLALLAALCP	ASRALEEKKV	CQGTSNKLTK	LGTFEDHFLS
60	70	80	90	100
LQRMFNCEV	VLGNLEITYV	QRNYDLSFLK	TIQEVAGYVL	IALNTVERIP
110	120	130	140	150
LENLQIIRGN	MYYENSYALA	VLSNYDANKT	GLKELPMRNL	QEILHGAVRF
160	170	180	190	200
SNNPALCNVE	SIQWRDIVSS	DFLSNMSMDF	QNHGSCQKC	DPSCPNGSCW

Length: 1,210  
Mass (Da): 134,277  
Last modified: November 1, 1997 - v2  
Checksum: <sup>i</sup>D8A2A50B4EFB6ED2

Proteins can have multiple **splice variants** which could be of therapeutic interest. For example, JNK kinases comprise 3 isoforms encoded by 3 distinct genes which can be spliced into 10 variants.

Sequences of splice variants are also available in UniProt's **Sequence** section.



# Sequence-related checks

- ❑ Are there any “extras”, e.g. signalling peptides or expression tags?

**Sequence alignments** to the canonical sequence will also reveal whether additional amino acids are present.

Proteins can contain a **signalling peptide** that determines their subcellular location and/or a **propeptide** part that is cleaved in the mature form of the protein.

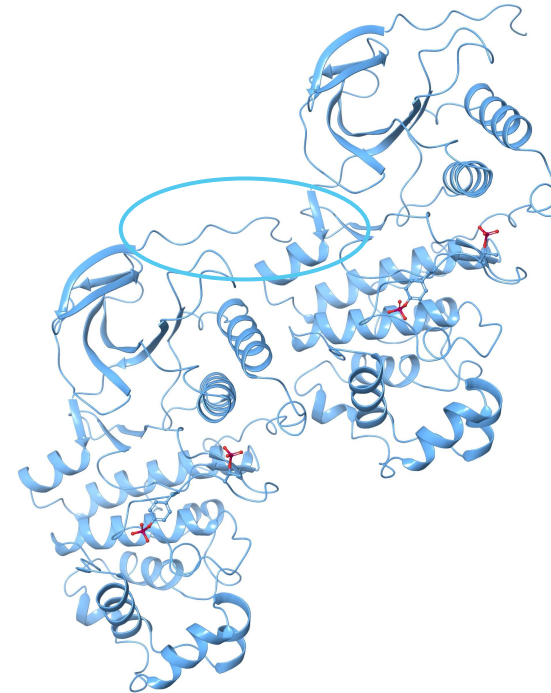
NGF (UniProt ID: P01138)

## PTM / Processing<sup>i</sup>

### Molecule processing

Feature key	Position(s)	Description	Actions	Graphical view	Length
Signal peptide <sup>i</sup>	1 – 18	<a href="#">Sequence analysis</a>	<a href="#">Add</a> <a href="#">BLAST</a>		18
Propeptide <sup>i</sup> (PRO_0000019599)	19 – 121		<a href="#">Add</a> <a href="#">BLAST</a>		103
Chain <sup>i</sup> (PRO_0000019600)	122 – 241	Beta-nerve growth factor	<a href="#">Add</a> <a href="#">BLAST</a>		120

Most often, short expression tags are added to the N- or C-terminus for protein purification. While in most cases **expression tags** are considered to be of no consequence for the protein structure, they can at times cause artifacts.



In p38 $\alpha$ , a 20-aa-long His-tag found its way to the kinase interaction motif docking site and caused a large conformational change (PDB ID: 3PY3).

- ❑ Are there any homologues?

**Rat and mouse proteins** tend to have a very high degree of similarity to human proteins.

It makes sense to calculate the sequence alignment across a few species as this could easily expand the starting pool of structures.

# M How to compare sequences?

Step 1. Get the sequence of your structure from the PDB and the canonical one from UniProt.

**Isoform 1** (identifier: **P00533-1**) [UniParc] [FASTA](#) [Add to basket](#)

Also known as: p170  
This isoform has been chosen as the canonical<sup>1</sup> sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

« Hide

```

10      20      30      40      50
MRPSGTAGAA LLALLAALCP ASRALEEKKV CQGSNKLQT LGTFEDHFLS
60      70      80      90     100
LQRMFNCEV VLGNIETIV QRNYDLSFLK TIQEVAGYVL IALNTVERIP
110     120     130     140     150
LENLQIIRGN MYYENSVALA VLSNYDANK GLKELPMRNL QEILHGAVRF
160     170     180     190     200
SNNPALCNVE SIQRDVISS DFLSNMMSDF QNHLGSCQKC DPSCPNGSCW
    
```

**Length:** 1,210  
**Mass (Da):** 134,277  
**Last modified:** November 1, 1997 - v2  
**Checksum:** <sup>1</sup>D8A2A50B4EFB6ED2  
BLAST

Step 2. Use a multiple sequence alignment tool like Clustal Omega.

Step 3. Result!

```

sp|P00533|EGFR_HUMAN  MRPSGTAGAAALLALLAALCPASRALEEKKVCGQTSNKLQTFEDHFLSLQRMFNCEV 60
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  VLGNIETIVQRNYDLSFLKTIQEVAGYVLIALNTVERIPLNQLTIQRMYYENSVALA 120
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  VLSNYDANKTGLKELPMRNLQEILHGAVRFSSNNPALCNVESTQRDVISSDFLSNMSMDF 180
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  QNHLGSCQKCDPSCPNGSCWGAGEENCKLTKITCAQCSCGRKSGSPDCCHNCAAGC 240
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  TGPRESDCLVCRFDEATKDCDTPPLNLYPTTIQMDVFNPEGYSFGATCVKCKPRIVY 300
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  VTDHGSVCRACGADSYEMEDGVKCKCEGPKRCKVNGIGTFEKDLSINATNIKHKF 360
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  NCTSISGDHLPLVAFRGSFTHPLDPLQELDLTKVKEITGFLLIQWPNRDLHAF 420
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  ENLEIIRGTRKQHQQFSLAVVSLNITSLGRLSKLEISDGDVVISGKMLCYANTINKKL 480
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  FGTSGQTKIISNRGENSKATGQVCHALCSPEGCGPEPRDCSRVNSGRGCDKCN 540
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  LLEGEPRFVSEICIQHPCELPQWNIITCTGRGPNICIQCAHYIDGPHCVKTPAGVM 600
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  GENITLWKYADAGHYCHLPHICTYGTGPGLEGCTHGPKIPSTATGMVALLLLW 660
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  ALGTGLFRRRHIVKRTLRLLQERLVEPLTPSEAPNQLRLILKETFKIKIKVLGS 720
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  GAFGTYYKGLMPEGEKVKIPVAIKELREATSPKANKEILDEAYVNASVDNPHVCRLLGI 780
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  CLTSLVLIQLMPEGCLLDVYREHKDNIQSYLLNLCVQIAKGMVLEDRRLVHRDLAA 840
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  RINLVKTPQHWKITDFGLAKLLGAEKEYHAEGKVPKWMALSTLHRIYTHQSDWHSY 900
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  GVTVMELHTFGSKPYDGIPIASEISLLEKGERLPPPTCTIDVYHMKCMIDADSRRK 960
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  FRELLIEFSKWARDPQRYLVIQGDERHPLSPDTSNFRALHDEEDIDVDADEVLIPO 1020
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  QGFSSPSTRTPLLSSLATSNHSTVACIDRNLQSCPIKESDFLQRYSSOPTGALTED 1080
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  SIDDTFLPVPEYINQSVKRPAGSVQNPVYHNPQVLPNAPSRRDPHYQPHSTAVGNPEYL 1140
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  TVQPTCVNSTFSDPSAHAQKGSQISLDNPDYQDFFPKAEKPNIGFGKSTAEAEYLRV 1200
5Y9T_1|Chain  -----
sp|P00533|EGFR_HUMAN  APOSSEFTGA 1210
5Y9T_1|Chain  -----
    
```



# M How to compare sequences?

Tasks -> Browse -> Protein Preparation and Refinement -> Multiple Sequence Viewer/Editor

Multiple Sequence Viewer/Editor

File Edit Select View Align

Workspace View 1 \* +

Always linked to Workspace contents Find / Fetch: enter sequence substring or PDB IDs Homologs... Align Other Tasks

TITLE	CHN	10	20	30	40	50	60	ID %
5Y9T	A	G S H M A S G E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	100					
5Y9T	A	S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I M Q L M P F G C L L D Y V R E H K D N I G S Q Y	100					
5Y9T	A	L L N W C V Q I A K G M N Y L E D R R L V H R D L A A R N V L V K T P Q H V K I T D F G L A K L L G A E E K E Y H A E G G K	100					
5Y9T	A	V P I K W M A L E S I L H R I Y T H Q S D V W S Y G V T V W E L M T F G S K P Y D G I P A S E I S I L E K G E R L P Q P P	100					
5Y9T	A	I C T I D V Y M I M V K C W M I D A D S R P K F R E L I I E F S K M A R D P Q R Y L V I Q G D E R M H L P S P T D S N F Y R	100					
5Y9T	A	A L M D E E D M D D V V D A D E Y L I P Q Q G	100					

SEQUENCES 1 selected 1 total REFERENCE 5Y9T (A)

STRUCTURES 1 in Workspace (1 total) OTHER TABS 0 sequences (1 tab)

recolouring button

- MSV will automatically load the sequences of included entries (in this case, the EGFR kinase domain).
- The residues are coloured based on the side-chain properties, with the **missing residues shown in darker shades**.



# M How to compare sequences?

Tasks -> Browse -> Protein Preparation and Refinement -> Multiple Sequence Viewer/Editor

Multiple Sequence Viewer/Editor

File Edit Select View Align

Workspace View 1

Find / Fetch: enter sequence substring or sequence code / name

Homologs... Align Other Tasks

TITLE	CHN	Sequence	ID %
5Y9T	A	G S H M A S G E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I	100
5Y9T	A	P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I M Q L M P F	100
5Y9T	A	N W C V Q I A K G M N Y L E D R R L V H R D L A A R N V L V K T P Q H V K I T D F G L A K L L G A E E K E Y H A E G G K V P I	100
5Y9T	A	K W M A L E S I L H R I Y T H Q S D V W S Y G V T V W E L M T F G S K P Y D G I P A S E I S S I L E K G E R L P Q P P I C T I	100
5Y9T	A	D V Y M I M V K C W M I D A D S R P K F R E L I I E F S K M A R D P Q R Y L V I Q G D E R M H L P S P T D S N F Y R A L M D E	100
5Y9T	A	E D M D D V V D A D E Y L I P Q Q G	100

SEQUENCES 0 selected 1 total  
STRUCTURES 1 in Workspace (1 total)

REFERENCE 5Y9T (A)  
OTHER TABS 0 sequences (1 tab)

fetch the sequence using UniProt or PDB ID

choose whether you want to download the sequence from the UniProt (always canonical!) or sequence/structure from the PDB

- You can fetch more sequences using UniProt/PDB ID.

# M How to compare sequences?

Multiple Sequence Viewer/Editor

File Edit Select View Align

Workspace View 1 Workspace Copy

Load from: Workspace Find / Fetch: P00533 Homologs... Align Other Tasks

TITLE	CHN	Sequence	ID %
5Y9T	A	G S H M A S G E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T S	100
P00533 EGFR_HU...		M R P S G T A G A A L L A L L A L C P A S R A L E E K K V C Q G T S N K L T Q L G T F E D H F L S L Q R M F N N C E V V L G	6
5Y9T	A	P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I M Q L M P F G C L L D Y V R E H K D N I G S Q Y L L	100
P00533 EGFR_HU...		N L E I T Y V Q R N Y D L S F L K T I Q E V A G Y V L I A L N T V E R I P L E N L Q I I R G N M Y Y E N S Y A L A V L S N Y D	6
5Y9T	A	N W C V Q I A K G M N Y L E D R R L V H R D L A A R N V L V K T P Q H V K I T D F G L A K L L G A E E K E Y H A E G G K V P I	100
P00533 EGFR_HU...		A N K T G L K E L P M R N L Q E I L H G A V R F S N N P A L C N V E S I Q W R D I V S S D F L S N M S M D F Q N H L G S C Q K	6
5Y9T	A	K W M A L E S I L H R I Y T H Q S D V W S Y G V T V W E L M T F G S K P Y D G I P A S E I S S I L E K G E R L P Q P P I C T I	100
P00533 EGFR_HU...		C D P S C P N G S C W G A G E E N C Q K L T K I I C A Q Q C S G R C R G K S P S D C C H N Q C A A G C T G P R E S D C L V C R	6
5Y9T	A	D V Y M I M V K C W M I D A D S R P K F R E L I I E F S K M A R D P Q R Y L V I Q G D E R M H L P S P T D S N F Y R A L M D E	100
P00533 EGFR_HU...		K F R D E A T C K D T C P P L M L Y N P T T Y Q M D V N P E G K Y S F G A T C V K K C P R N Y V V T D H G S C V R A C G A D S	6
5Y9T	A	E D M D D V V D A D E Y L I P Q Q G	100
P00533 EGFR_HU...		Y E M E E D G V R K C K C E G P C R K V C N G I G I G E F K D S L S I N A T N I K H F K N C T S I S G D L H I L P V A F R G	6
5Y9T	A		100

SEQUENCES 1 selected 2 total  
STRUCTURES 1 in Workspace (1 total)  
REFERENCE 5Y9T (A)  
OTHER TABS 1 sequence (2 tabs)

- Downloaded sequences need to be aligned.

# M How to compare sequences?

The screenshot displays the Multiple Sequence Viewer/Editor interface. The main window shows a multiple sequence alignment of protein sequences. The sequences are color-coded by amino acid type. A menu is open over the alignment, showing options for alignment. The 'Align' button is highlighted with a red box. The 'Align' menu is open, showing options for alignment. The 'Align' menu is open, showing options for alignment. The 'Align' menu is open, showing options for alignment.

Multiple Sequence Viewer/Editor

File Edit Select View Align

Workspace View 1 Workspace Copy

Load from: Workspace Find / Fetch: P00533 Homologs... Align Other Tasks

Align: Sequences Structures

Using: Multiple sequence alignment

Find globally conserved residues (Pfam)

Superimpose structures following alignment

Selected only

Align

SEQUENCES 1 selected 2 total REFERENCE 5Y9T (A)

STRUCTURES 1 in Workspace (1 total) OTHER TABS 1 sequence (2 tabs)

- Downloaded sequences need to be aligned.



# M How to compare sequences?

Multiple Sequence Viewer/Editor

File Edit Select View Align

Workspace View 1 Workspace Copy +

Load from: Workspace Find / Fetch: P00533 Homologs... Align Other Tasks

640 650 660 670 680 690

5Y9T A P G L E G C P T N G P K I P S I A T G M V G A L L L L V V A L G I G L F M R R R H I V R K R T L R R L L Q E R E L V E P L T 100  
P00533|EGFR\_HU... P G L E G C P T N G P K I P S I A T G M V G A L L L L V V A L G I G L F M R R R H I V R K R T L R R L L Q E R E L V E P L T 98

700 710 720 730 740 750

5Y9T A A S G E A P N Q A L L R I L K E T E F F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T S P K A N 100  
P00533|EGFR\_HU... P S G E A P N Q A L L R I L K E T E F F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T S P K A N 98

760 770 780 790 800 810

5Y9T A K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I M Q L M P F G C L L D Y V R E H K D N I G S Q Y L L N W C V 100  
P00533|EGFR\_HU... K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I T Q L M P F G C L L D Y V R E H K D N I G S Q Y L L N W C V 98

820 830 840 850 860 870 880

5Y9T A Q I A K G M N Y L E D R R L V H R D L A A R N V L V K T P Q H V K I T D F G L A K L L G A E E K E Y H A E G G K V P I K W M A 100  
P00533|EGFR\_HU... Q I A K G M N Y L E D R R L V H R D L A A R N V L V K T P Q H V K I T D F G L A K L L G A E E K E Y H A E G G K V P I K W M A 98

890 900 910 920 930 940

5Y9T A L E S I L H R I Y T H Q S D V W S Y G V T V W E L M T F G S K P Y D G I P A S E I S S I L E K G E R L P Q P P I C T I D V Y M 100  
P00533|EGFR\_HU... L E S I L H R I Y T H Q S D V W S Y G V T V W E L M T F G S K P Y D G I P A S E I S S I L E K G E R L P Q P P I C T I D V Y M 98

950 960 970 980 990 1000

5Y9T A I M V K C W M I D A D S R P K F R E L I I E F S K M A R D P Q R Y L V I Q G D E R M H L P S P T D S N F Y R A L M D E E D M D 100  
P00533|EGFR\_HU... I M V K C W M I D A D S R P K F R E L I I E F S K M A R D P Q R Y L V I Q G D E R M H L P S P T D S N F Y R A L M D E E D M D 98

1010 1020 1030 1040 1050 1060 1070

5Y9T A D V V D A D E Y L I P Q Q G F F S S P S T S R T P L L S S L S A T S N N S T V A C I D R N G L Q S C P I K E D S F L Q R Y S S 100  
P00533|EGFR\_HU... D V V D A D E Y L I P Q Q G F F S S P S T S R T P L L S S L S A T S N N S T V A C I D R N G L Q S C P I K E D S F L Q R Y S S 98

SEQUENCES 2 selected 2 total REFERENCE 5Y9T (A)  
STRUCTURES 1 in Workspace (1 total) OTHER TABS 1 sequence (2 tabs)

- It's still hard to see where the differences are.

# M How to compare sequences?

The screenshot displays the Multiple Sequence Viewer/Editor interface. The main window shows a sequence alignment of P00533 (EGFR\_HU...) against the reference sequence 5Y9T (A). The alignment is color-coded by residue type. A 'COLOR SEQUENCES' dialog box is open, showing the 'COLOR SCHEMES & SETTINGS' section. The 'Apply to:' dropdown is set to 'All residues on tab'. The 'Color by:' dropdown is set to 'Side Chain Chemistry', and the 'Residue Similarity' option is selected. The 'HIGHLIGHT SELECTED RESIDUES' section is empty. The 'COLOR LINKED STRUCTURES' section has 'Apply colors to structures:' checked, and 'Whenever they change', 'Color entire residues', and 'Color Carbons' are all checked.

SEQUENCES 2 selected 2 total  
STRUCTURES 1 in Workspace (1 total)

REFERENCE 5Y9T (A)  
OTHER TABS 1 sequence (2 tabs)

*colour by residue similarity*

*the same colour scheme can also be applied to linked structures*



# M How to compare sequences?

Multiple Sequence Viewer/Editor

File Edit Select View Align

Workspace View 1 Workspace Copy

Load from: Workspace Find / Fetch: P00533 Homologs... Align Other Tasks

5Y9T A  
P00533|EGFR\_HU... H P N C T Y G C T G P G L E G C P T N G P K I P S I A T G M V G A L L L L V V A L G I G L F M R R R H I V R K R T L R R L

5Y9T A  
P00533|EGFR\_HU... L Q E R E L V E P L T P S G E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I

5Y9T A  
P00533|EGFR\_HU... K E L R E A T S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L M Q L M P F G C L L D Y V R E H K

5Y9T A  
P00533|EGFR\_HU... D N I G S Q Y L L N W C V Q I A K G M N Y L E D R R L V H R D L A A R N V L V K T P Q H V K I T D F G L A K L L G A E E K E

5Y9T A  
P00533|EGFR\_HU... Y H A E G G K V P I K W M A L E S I L H R I Y T H Q S D V W S Y G V T V W E L M T F G S K P Y D G I P A S E I S S I L E K G

5Y9T A  
P00533|EGFR\_HU... E R L P Q P P I C T I D V Y M I M V K C W M I D A D S R P K F R E L I I E F S K M A R D P Q R Y L V I Q G D E R M H L P S P

5Y9T A  
P00533|EGFR\_HU... T D S N F Y R A L M D E E D M D D V V D A D E Y L I P Q Q G F F S S P S T S R T P L L S S L S A T S N N S T V A C I D R N G

SEQUENCES 2 selected 2 total REFERENCE 5Y9T (A)  
STRUCTURES 1 in Workspace (1 total) OTHER TABS 1 sequence (2 tabs)

- Now it's obvious the sequence differences come from a point mutation and the addition of an expression tag at the N-terminus.
- Note that the missing residues are shown by darker colour shades.

# M How to find and compare homologues?

The screenshot displays the Multiple Sequence Viewer/Editor interface. The top menu includes File, Edit, Select, View, and Align. The workspace shows 'View 1' and 'Workspace Copy'. The 'Load from:' dropdown is set to 'Workspace'. The 'Find / Fetch:' field contains 'P00533'. The 'Align' button is highlighted with a red box. The main area shows a sequence alignment of P00533 and 5Y9T. The reference sequence is 5Y9T (A). The alignment shows the following sequences:

```
5Y9T A P00533|EGFR_HU... HPNCTYGCTGPGLEGCP TNGPKIPSIATGMV GALLLLLVVALGIGLFMRRRHIVRKRRTLRL L
5Y9T A P00533|EGFR_HU... GSHMASG EAPNQALLRILKETEFKKIKVLGSGAFGTVYKGLWIP EGEKVKIPVAI
LQERELVEPLTPSGEAPNQALLRILKETEFKKIKVLGSGAFGTVYKGLWIP EGEKVKIPVAI
5Y9T A P00533|EGFR_HU... KELREA TSPKANKEILDEAYVMASVDNPHVCRLLGICLTSTVQLIMQLMPFGCLLDYVREHK
KELREATSPKANKEILDEAYVMASVDNPHVCRLLGICLTSTVQLI TQLMPFGCLLDYVREHK
5Y9T A P00533|EGFR_HU... DNIGSQYLLNWCVQIAKGMNYLEDRRLLVHRDLAARNV LVKTPQHVKITDFGLAKLLGAE EKE
DNIGSQYLLNWCVQIAKGMNYLEDRRLLVHRDLAARNV LVKTPQHVKITDFGLAKLLGAE EKE
5Y9T A P00533|EGFR_HU... YHAEGGKVP I KWMAL E SILHRIYTHQS DVWSYGVTVWELMTFGSKPYDGI PASEI SSILEKG
YHAEGGKVP I KWMAL E SILHRIYTHQS DVWSYGVTVWELMTFGSKPYDGI PASEI SSILEKG
5Y9T A P00533|EGFR_HU... ERLPQPP ICTIDVYMIMVKCWMIDADSRPKFRELI IEF SKMARDPQRYLV IQGDERMHLPS P
ERLPQPP ICTIDVYMIMVKCWMIDADSRPKFRELI IEF SKMARDPQRYLV IQGDERMHLPS P
5Y9T A P00533|EGFR_HU... TDSNFYRALMDEED MDDVVDADEYLIPQQG
TDSNFYRALMDEED MDDVVDADEYLIPQQG FFS SPSTSRTPLLSSLSATSNNSTVACIDRNG
```

At the bottom, the status bar shows: SEQUENCES 1 selected 2 total, STRUCTURES 1 in Workspace (1 total), REFERENCE 5Y9T (A), and OTHER TABS 1 sequence (2 tabs). The bottom right corner contains icons for editing, analysis, and adding new elements.

# M How to find and compare homologues?

The screenshot shows the Multiple Sequence Viewer/Editor interface. The main window displays a sequence alignment of 5Y9T (P00533) with a reference sequence highlighted in red. A dialog box titled "Search for Homo..." is open, showing the reference sequence (5Y9T:A), the algorithm (BLAST), and the search database (PDB). The dialog box also includes a "Run Search" button and a "Close" button. The interface includes a menu bar (File, Edit, Select, View, Align), a toolbar with various icons, and a status bar at the bottom showing "SEQUENCES 1 selected 2 total" and "STRUCTURES 1 in Workspace (1 total)".

- We will use the sequence from the PDB structure as the reference because it contains just the kinase domain, while the UniProt sequence contains the whole protein.
- If the PDB structure contains multiple copies of the same molecule, you can also fetch their structures.





# M How to find and compare homologues?

Multiple Sequence Viewer/Editor

File Edit Select View Align

Workspace Workspace Copy

Load from: Workspace Find / Fetch: P00533 Homologs... Align Other Tasks

Homolog Search Results - 5Y9T:A

Results of search run from: Workspace Copy

Reference Sequence: 5Y9T:A Algorithm: BLAST

Select one or more homologs to import into viewer: Choose top: 10 Select

Name	E-value	Score	Identity %	Positive %	Gaps %	Description
5Y9T_A	0	1669	100	100	0	Crystal Structure of EGFR T790M mutant in complex with naquotinib [Homo sapiens]
1M14_A	0	1664	99.6997	99.6997	0	Tyrosine Kinase Domain from Epidermal Growth Factor Receptor [Homo sapiens]
6TFU_A	0	1663	99.6997	99.6997	0	Crystal Structure of EGFR T790M/V948R in Complex with Covalent Pyrrolopyrimidine 14d [Homo sapiens]
6S89_A	0	1647	98.7988	98.7988	0	Crystal Structure of EGFR-T790M/C797S in Complex with Covalent Pyrrolopyrimidine 19g [Homo sapiens]
4I24_A	0	1644	99.696	99.696	0	Structure of T790M EGFR kinase domain co-crystallized with dacomitinib [Homo sapiens]
2J1U_A	0	1644	100	100	0	Crystal structure of EGFR kinase domain T790M mutation in complex with AEE788 [Homo sapiens]
3IKA_A	0	1642	100	100	0	Crystal Structure of EGFR 696-1022 T790M Mutant Covalently Binding to WZ4002 [Homo sapiens]
4TKS_A	0	1642	99.3939	99.3939	0	Native-SAD phasing for human EGFR kinase domain. [Homo sapiens]
4G5P_A	0	1640	100	100	0	Crystal structure of EGFR kinase T790M in complex with BIBW2992 [Homo sapiens]
5CAV_A	0	1638	99.3921	99.3921	0	EGFR kinase domain with compound 41a [Homo sapiens]
4ZJV_A	0	1638	99.3921	99.3921	0	crystal structure of EGFR kinase domain in complex with Mitogen-inducibile gene 6 protein [Homo sapiens]
4I23_A	0	1638	99.3921	99.3921	0	Crystal structure of the wild-type EGFR kinase domain in complex with dacomitinib (soaked) [Homo sapiens]

Include structures when importing PDB sequences  Import into active tab (default is original search tab)

Export 10 sequences selected Import Cancel

SEQUENCES 0 selected 2 total REFERENCE 5Y9T (A) OTHER TABS 1 sequence (1 tab)

STRUCTURES 1 in Workspace (1 total)

- We can now select the entries we're interested in based on sequence identity and the provided description.
- We can also at the same time download their structures.
- You can always access the BLAST search results through:  
Other Tasks -> Homologs Search Results...





# How to find and compare homologues?

Multiple Sequence Viewer/Editor

File Edit Select View Align

Workspace Workspace Copy

Load from: Workspace Find / Fetch: P00533 Homologs... Align Other Tasks

Homolog Search Results - 5Y9T:A

Results of search run from: Workspace Copy

Reference Sequence: 5Y9T:A Algorithm: BLAST

Select one or more homologs to import into viewer: Choose top: 10 Select

Name	E-value	Score	Identity %	Positive %	Gaps %	Description
2RFD_A	0	1597	99.3769	99.6885	0	Crystal structure of the complex between the EGFR kinase domain and a Mig6 peptide [Homo sapiens]
2RGP_A	0	1570	99.6825	99.6825	0	Structure of EGFR in complex with hydrazone, a potent dual inhibitor [Homo sapiens]
5GNK_A	0	1467	99.6599	99.6599	0	Crystal structure of EGFR 696-988 T790M in complex with LXX-6-34 [Homo sapiens]
3LZB_A	0	1427	99.3056	99.3056	0	EGFR kinase domain complexed with an imidazo[2,1-b]thiazole inhibitor [Homo sapiens]
3PP0_A	0	1317	78.6585	89.3293	0.304878	Crystal Structure of the Kinase domain of Human HER2 (erbB2). [Homo sapiens]
3BBT_B	0	1314	77.5385	88.6154	0	crystal structure of the ErbB4 kinase in complex with lapatinib [Homo sapiens]
2R4B_A	0	1225	79.2642	88.2943	0	ErbB4 kinase domain complexed with a thienopyrimidine inhibitor [Homo sapiens]
60P9_A	3.20961e-122	872	59.0909	76.2238	0	HER3 pseudokinase domain bound to bosutinib [Homo sapiens]
3LMG_A	5.54179e-122	872	59.0909	76.2238	0	Crystal structure of the ERBB3 kinase domain in complex with AMP-PNP [Homo sapiens]
4RIW_A	5.58861e-121	864	59.0106	76.3251	0	Crystal structure of an EGFR/HER3 kinase domain heterodimer [Homo sapiens]
3KEX_A	9.35255e-121	862	59.2199	76.5957	0	Crystal structure of the catalytically inactive kinase domain of the human epidermal growth factor receptor
4RIX_A	2.97922e-120	859	58.6572	76.3251	0	Crystal structure of an EGFR/HER3 kinase domain heterodimer containing the cancer-associated HER3-Q790R mutation

Include structures when importing PDB sequences  Import into active tab (default is original search tab)

Export 1 sequence selected Import Cancel

SEQUENCES 1 selected 2 total REFERENCE 5Y9T (A)  
STRUCTURES 1 in Workspace (1 total) OTHER TABS 1 sequence (1 tab)

human homologues of EGFR

# M How to find and compare homologues?

The screenshot displays the Multiple Sequence Viewer/Editor interface. The main window shows a list of sequences on the left and a sequence alignment grid in the center. An 'Align' dialog box is open, showing options for alignment. The 'Align' button is highlighted in the top right of the main window. The 'Align' dialog box has the following options:

- Using: Multiple sequence alignment
- Find globally conserved residues (Pfam)
- Superimpose structures following alignment
- Selected only

The 'Align' button is highlighted in the dialog box. The main window also shows a 'Homologs...' search bar and an 'Other Tasks' dropdown menu. The status bar at the bottom indicates 'SEQUENCES 11 selected 11 total' and 'STRUCTURES 11 in Workspace (11 total)'.

- We once again have to align the downloaded sequences.
- We can also align the associated structures after the alignment based on the sequence alignment.



# M How to find and compare homologues?

- The sequence gaps in other structures indicate missing residues.

Multiple Sequence Viewer/Editor

File Edit Select View Align

Workspace Workspace Copy

Load from: Workspace Find / Fetch: P00533 Homologs... Align Other Tasks

TITLE	CHN	10	20	30	40	50	60	ID %
5Y9T	A	G S H M A S G	E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	100				
1M14	A		G E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	99				
6TFU	A			N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	100			
6S89	A		E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	99				
4I24	A			N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	100			
2JIU	A		G E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	98				
3IKA	A		G E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	100				
4TKS	A		E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	100				
4G5P	A		G E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	97				
5CAV	A		E A P N Q A L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	99				
4ZIV	A			L L R I L K E T E F K K I K V L G S G A F G T V Y K G L W I P E G E K V K I P V A I K E L R E A T	100			

TITLE	CHN	70	80	90	100	110	120	ID %
5Y9T	A	S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I M Q L M P F G C L L D Y V R E H K D N I G S Q Y	100					
1M14	A	S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I T Q L M P F G C L L D Y V R E H K D N I G S Q Y	99					
6TFU	A	S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I M Q L M P F G C L L D Y V R E H K D N I G S Q Y	100					
6S89	A	S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I M Q L M P F G S L L D Y V R E H K D N I G S Q Y	99					
4I24	A		A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I M Q L M P F G C L L D Y V R E H K D N I G S Q Y	100				
2JIU	A	S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I M Q L M P F G C L L D Y V R E H K D N I G S Q Y	98					
3IKA	A	S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I M Q L M P F G C L L D Y V R E H K D N I G S Q Y	100					
4TKS	A	S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I T Q L M P F G C L L D Y V R E H K D N I G S Q Y	100					
4G5P	A	S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I M Q L M P F G C L L D Y V R E H K D N I G S Q Y	97					
5CAV	A	S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I T Q L M P F G C L L D Y V R E H K D N I G S Q Y	99					
4ZIV	A	S P K A N K E I L D E A Y V M A S V D N P H V C R L L G I C L T S T V Q L I T Q L M P F G C L L D Y V R E H K D N I G S Q Y	100					

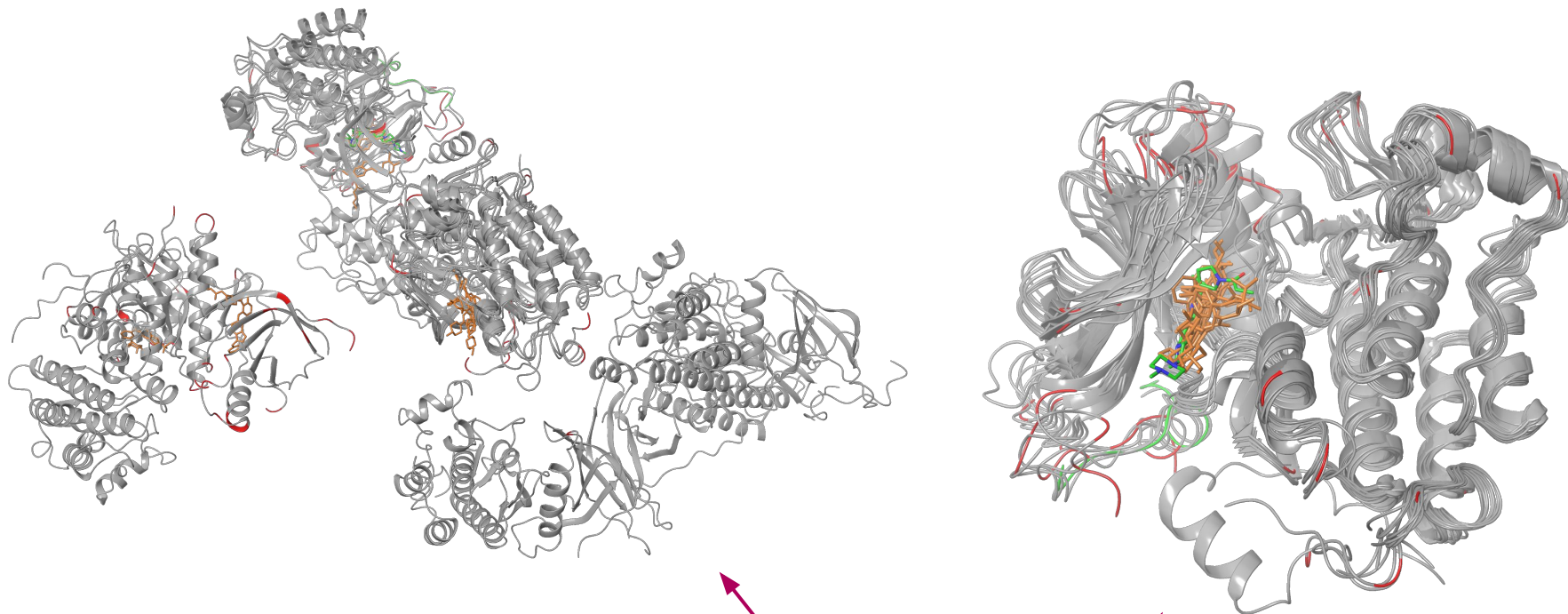
  

TITLE	CHN	130	140	150	160	170	180	ID %
5Y9T	A	L L N W C V Q I A K G M N Y L E D R R L V H R D L A A R N V L V K T P Q H V K I T D F G L A K L L G A E E K E Y H A E G G K	100					
1M14	A	L L N W C V Q I A K G M N Y L E D R R L V H R D L A A R N V L V K T P Q H V K I T D F G L A K L L G A E E K E Y H A E G G K	99					
6TFU	A	L L N W C V Q I A K G M N Y L E D R R L V H R D L A A R N V L V K T P Q H V K I T D F G L A K L L G A E E K E Y H A E G G K	100					

SEQUENCES 11 selected 11 total REFERENCE 5Y9T (A)

STRUCTURES 11 in Workspace (11 total) OTHER TABS 11 sequences (1 tab)

# M How to find and compare homologues?



**Note:** These entries are individual chains taken from the original PDB models, i.e. out of the deposited crystal context, so certain X-ray quality checks cannot be performed on them.

Workspace before and after the sequence-based structural alignment.







# The final checklist

## Function-related checks

- What's the subcellular location of the protein?
- Is the protein a monomer or a multimer?  
If a multimer, is it a homomer or a heteromer?
- Is the protein known for multiple conformational states?
- What about atypical chemical forms?
- Maybe there are some PTMs?
- Are any metals involved?
- Does the protein bind any other cofactors?

## Sequence-related checks

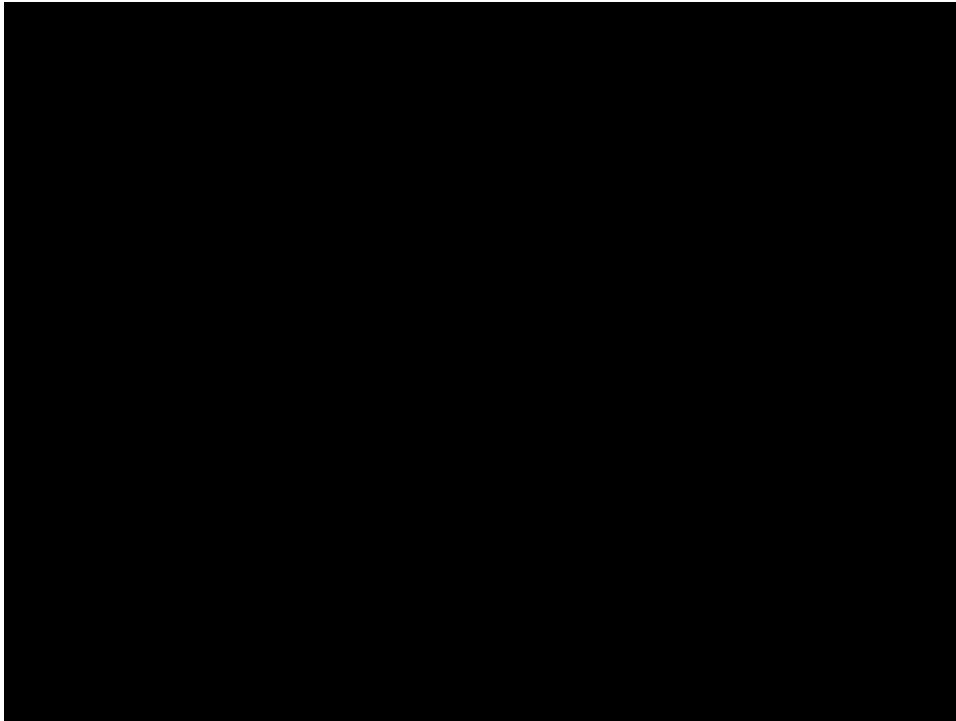
- Is the whole protein there? Any missing (sub)domains?
- Are you working with the correct sequence?
- Are there any “extras”, e.g. signalling peptides or expression tags?
- Are there any homologues?

# **X-ray quality checks**



# X-ray quality checks

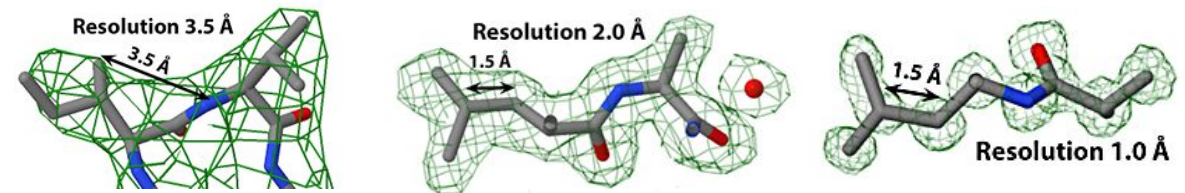
- ❑ Is the resolution high enough?



Video created by James Holton @ Berkley Lab showing how the electron density changes in the 0.5-5 Å resolution range. It's **not entirely realistic** as it's created w/o hydrogens and directly from the atomic model (so no errors, noise, etc.), but it's still highly illustrative.

Resolution - the distance at which you can tell objects apart.

Lower the number, higher the quality.



proteopedia.org

Quick resolution guide from Proteopedia:

- 1.2 Å Excellent -- backbone and most side chains very clear. Some hydrogens may be resolved.
- 2.5 Å Good -- backbone and many side chains clear.
- 3.5 Å OK -- backbone and bulky side chains mostly clear.
- 5.0 Å Poor -- backbone mostly clear; side chains not clear.

cut off

desperation



# X-ray quality checks

## ❑ Are the R and $R_{\text{free}}$ factors reasonably low?

**R factor** is the **measure of error** between the observed intensities used in the refinement process and the ones calculated from the structural model.

**$R_{\text{free}}$  factor** is calculated in the same manner, but on a subset of intensities that haven't been used in the refinement (5-10% of the data). Thus, it is used to estimate **model bias** on the refinement process.

Lower the numbers, smaller the error and better the fit to experimental data.

Rules of thumb:

- R factors < 0.2 are considered reliable.
- Random models give R factors in the 0.4-0.6 range.
- Lower the resolution, higher the R factors (i.e. higher the model errors).
- R factors shouldn't be > resolution/10.
- $R_{\text{free}} - R < 0.07$

# X-ray quality checks

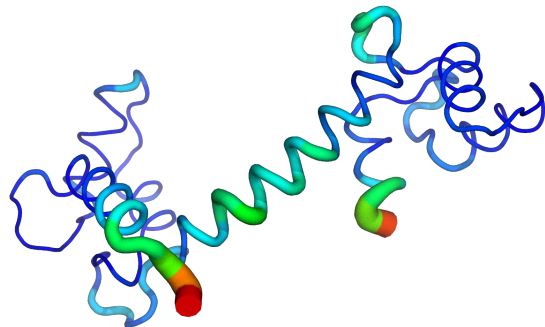
## ❑ What are the B-factors like?

Uncertainty of the modelled atomic coordinates increases with the disorder present in the crystal.

There are two types of disorder - **static** (parts of the protein are stable, but present in different conformations) and **dynamic** (some parts of every protein copy are subject to **thermal motion**).

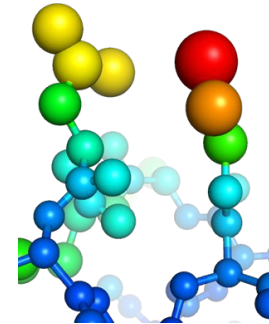
**B-factors** (temperature factors) reflect that disorder and are proportional to the mean square displacement of the atom.

Higher the number, higher the mobility/disorder of the atom.

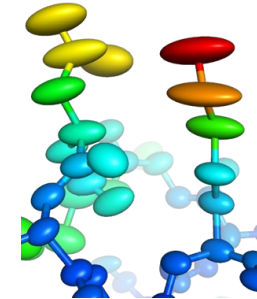


Representation of calmodulin in which the higher B-factor values are shown using warm colours and thicker tube (PDB ID: 1EXR).

B-factors can be modelled as **isotropic** or **anisotropic**, based on whether the displacement is considered to be identical in all directions or not.



isotropic B-factors



anisotropic B-factors

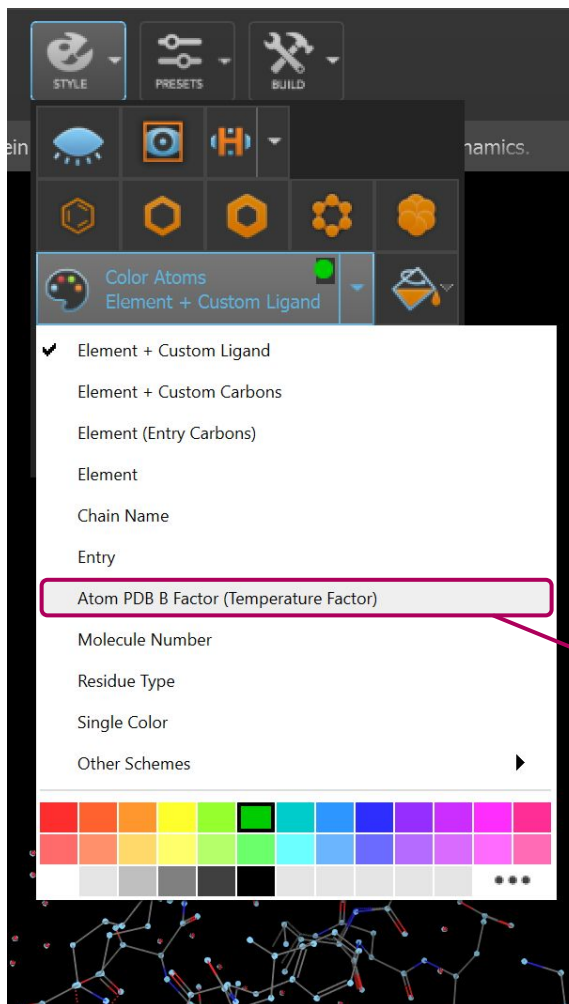
phenix-online.org

Rules of thumb and caveats:

- B-factors  $< 30 \text{ \AA}^2$  indicate reliable positions.
- B-factors  $> 60 \text{ \AA}^2$  signify disorder.
- Crystal contacts can lower B-factors of otherwise mobile regions.
- High B-factors can also arise from model errors.
- Comparison of B-factors across different PDB structures is meaningless (unless they were obtained under identical experimental conditions and refinement process - highly unlikely!).

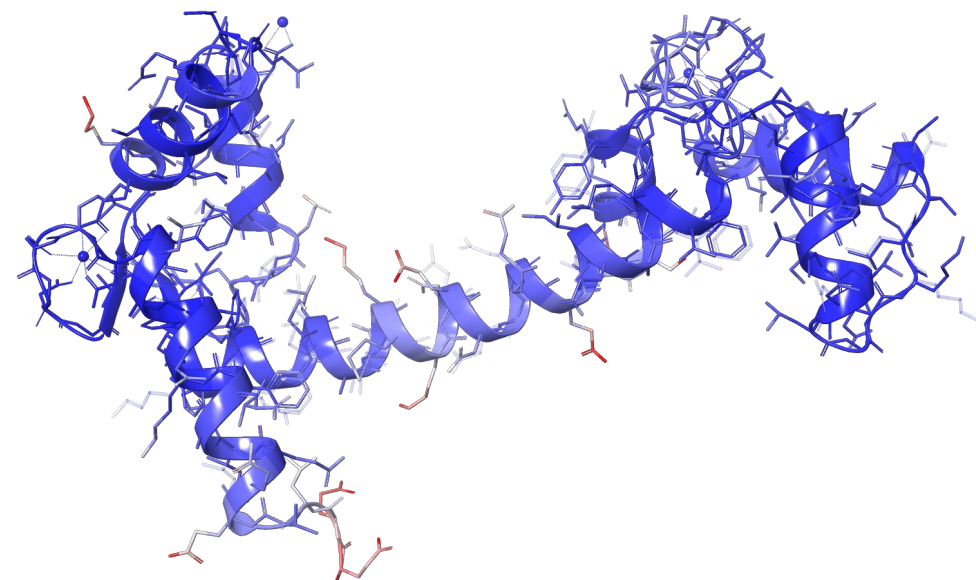
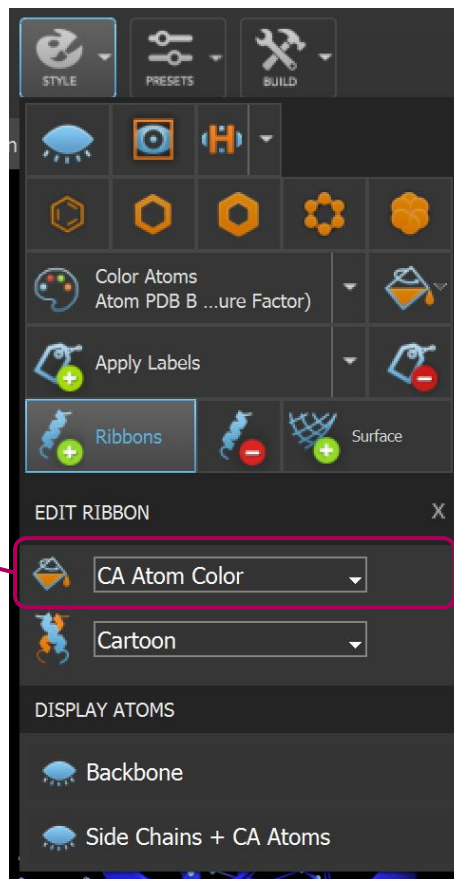


# M How to view B-factors?



*add ribbons coloured by the CA atoms*

*colour the selected atoms based on their B-factors*



Representation of calmodulin in which the higher B-factor values are shown using warmer colours (PDB ID: 1EXR).

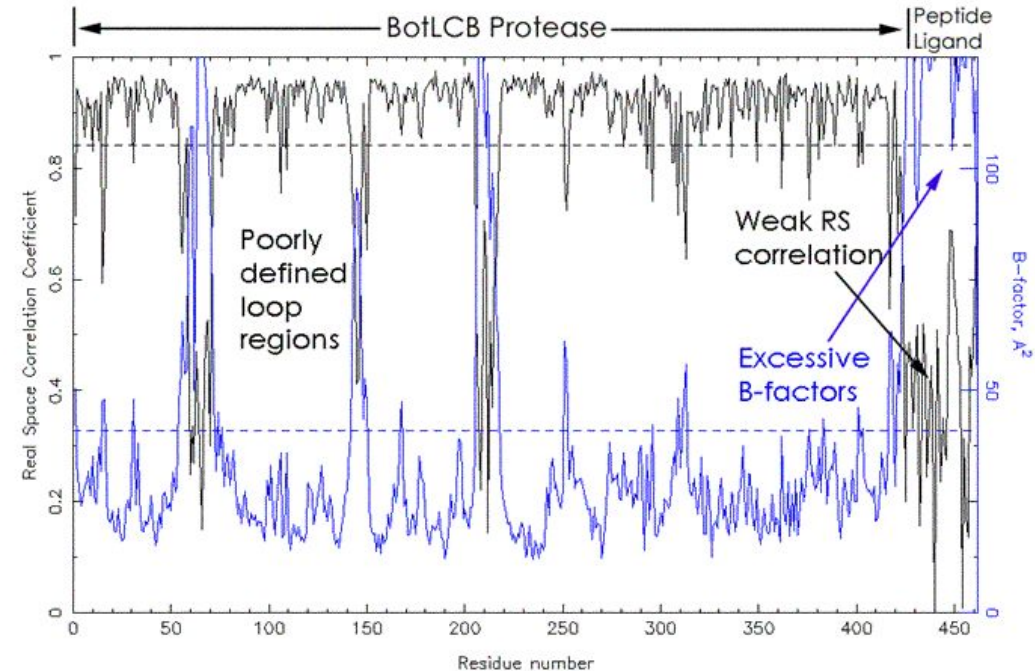
# X-ray quality checks

## ❑ What about the RSCC values?

**Real space correlation coefficient (RSCC)** is a measure of similarity between an electron density map calculated from the experimental data and the one calculated directly from a structural model.

RSCC corresponds to the **sample Pearson correlation coefficient**, so the values range from -1 (perfect anti-correlation) to 1 (perfect correlation), where 0 indicates no correlation. In practice, the expected values range from 0 to 1 and typically everything below 0.8 is considered to indicate a **poor density fit**.

RSCC is calculated per residue and it's often plotted together with the B-factors to identify problematic regions.



The RSCC (in black) and B-factor (in blue) plot for the BotLCB protease in complex with synaptobrevin-II (PDB ID: 1F83). The left part of the plot corresponding to the protease indicates a good density fit, with the exception of three loops. However, the synaptobrevin-II peptide on the right has low RSCC values and excessive B-factors indicating a very problematic region that should be carefully examined. Example taken from <https://www.ruppweb.org/Xray/tutorial/rscs.htm>.

# X-ray quality checks

LOCAL METRIC

## □ Are there any geometric outliers or clashes?

During the **refinement process**, the structural model is adjusted to produce a better fit to the experimental data, while typically keeping the geometry of the molecules as **close to ideal values** as possible.

However, **outliers still occur** and should be checked to make sure they're not affecting relevant protein sites.

Two types of **validation checks** are done during the deposition of structures to the PDB:

- Comparison of selected metrics to all the other PDB structures, both in the same resolution range and across all resolutions.
- Comparison of protein and ligand geometry to the ideal values.

4OW0

X-Ray Structural and Biological Evaluation of a Series of Potent and Highly Selective Inhibitors of Human Coronavirus Papain-Like Proteases

DOI: 10.2210/pdb4OW0/pdb

Classification: **HYDROLASE/HYDROLASE INHIBITOR**

Organism(s): SARS coronavirus Urbani

Expression System: Escherichia coli BL21(DE3)

Mutation(s): No

Deposited: 2014-01-28 Released: 2014-04-23

Deposition Author(s): Baez-Santos, Y.M., Mesecar, A.

Experimental Data Snapshot

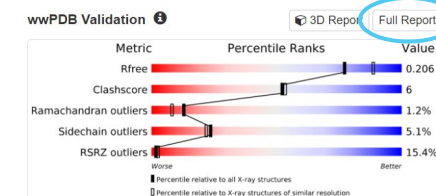
Method: X-RAY DIFFRACTION

Resolution: 2.10 Å

R-Value Free: 0.204

R-Value Work: 0.176

R-Value Observed: 0.177



Full validation report

Comparison to other PDB structures

## Protein geometry checks

- bond lengths
- bond angles
- chirality
- planarity (side chains, peptide bond, main chain)
- close contacts / clashes
- torsion angles
  - backbone (Ramachandran plot)
  - side chains

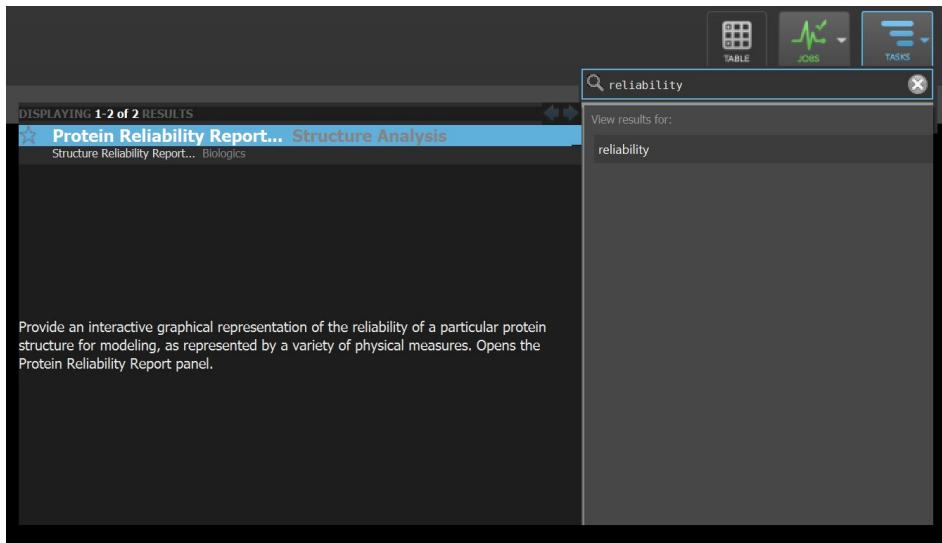
## Ligand geometry checks

- bond lengths
- bond angles
- chirality
- rings
- close contacts / clashes
- torsion angles

Outliers indicate possible model errors!



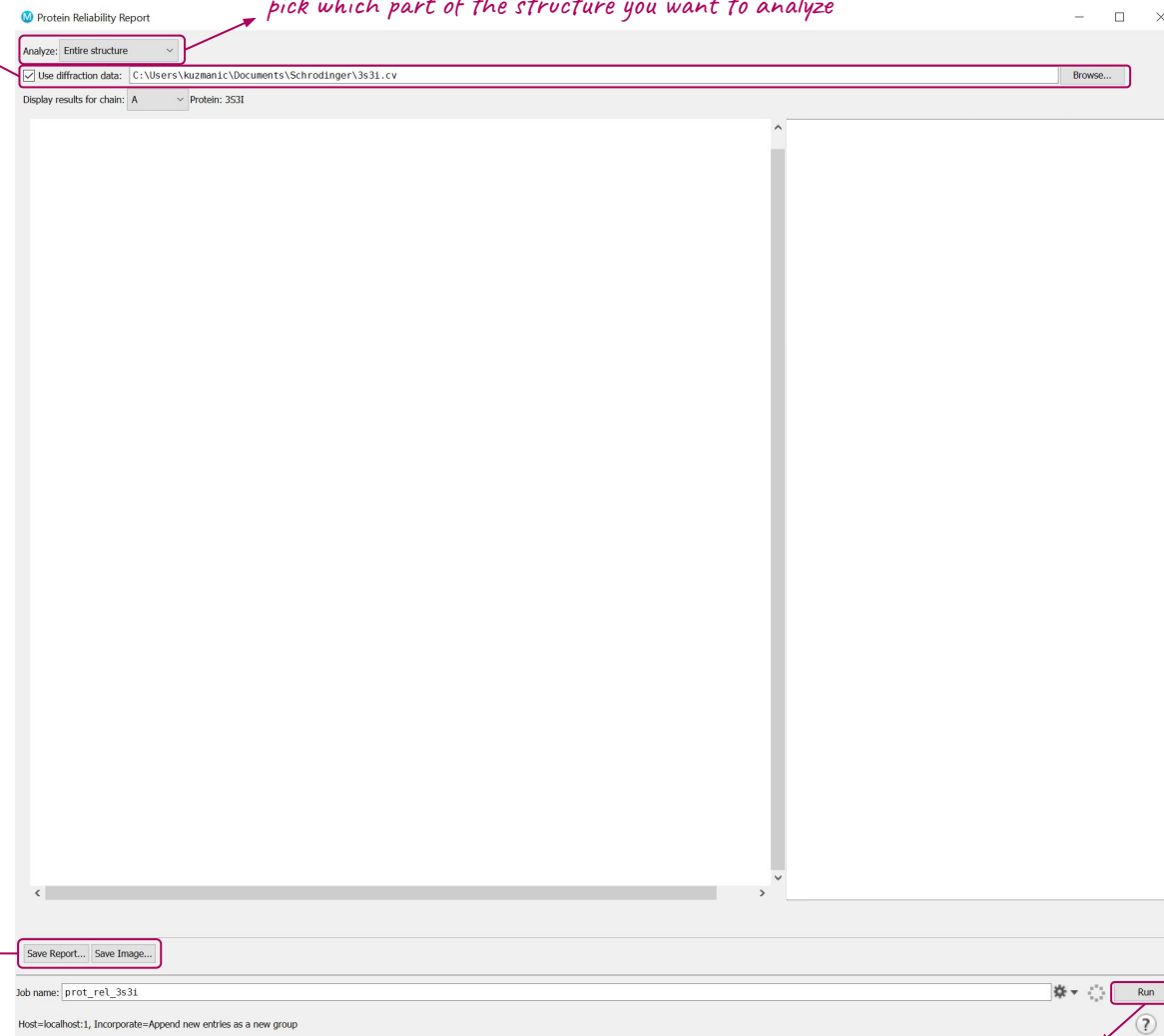
# How to check the protein structural quality?



- You can access **Protein Reliability Report** via:  
Tasks -> Browse -> Structure Analysis -> Protein Reliability Report
- Once you run the job, you can evaluate the quality of your structure across different categories based on the color and size of individual bubbles:  
Larger and redder the bubble, more issues are associated with that property.

*use diffraction data if available (needs to be pre-downloaded)*

*pick which part of the structure you want to analyze*



*save the report as a .txt file or an image*

*you need to run the job to get the report*



# M How to check the protein structural quality?

## 1 – Structure Quality in Binding Site

**Ligand and Binding Site RSCC > 0.9**

**Packing** - checks how different the environment of each fragment is compared to average environment in a curated dataset (calculated per residue as Z-score).

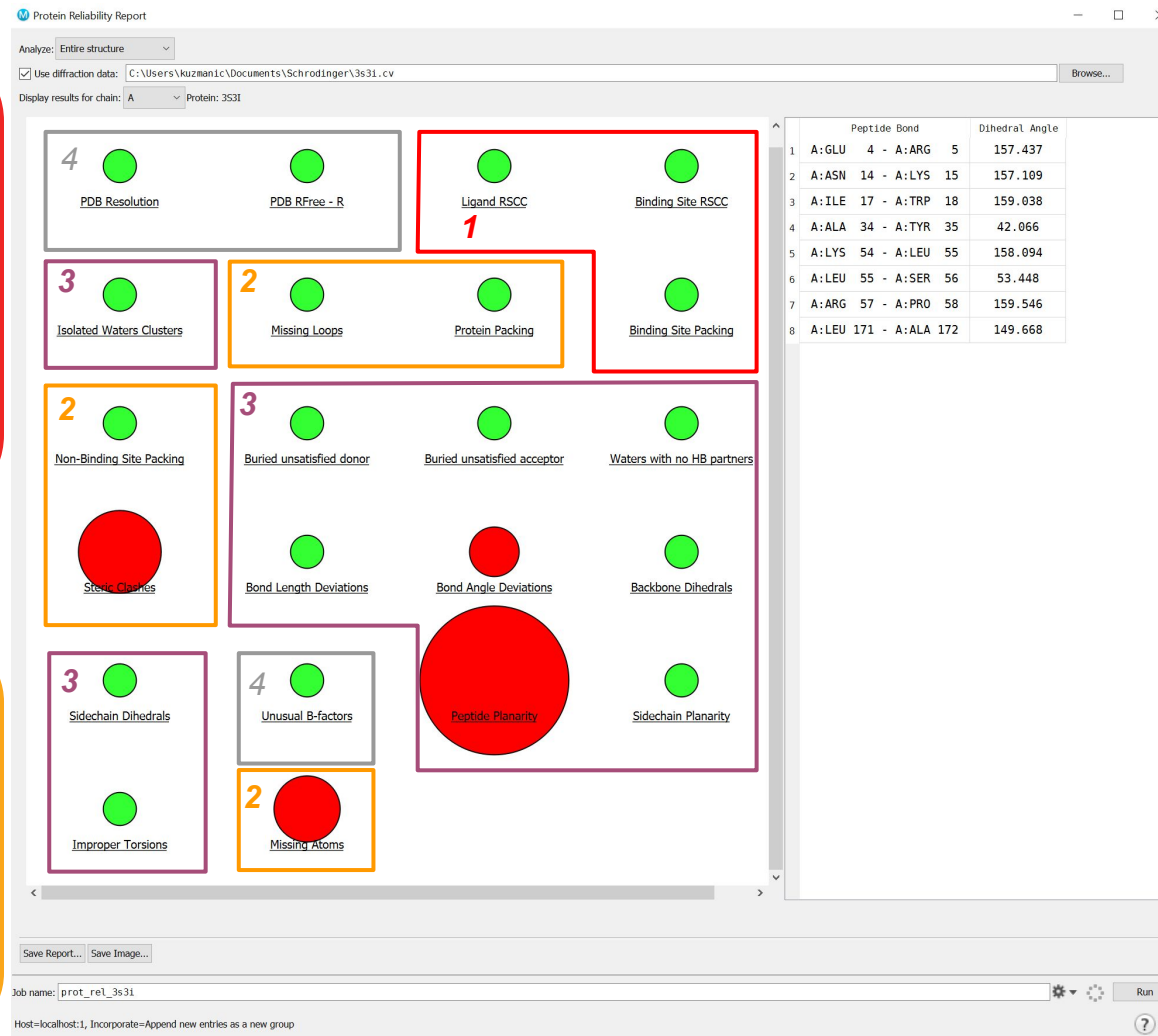
**Binding and Non-Binding Site Packing** report the number of consecutive residues with Z-score < -4 (possibly incorrectly built regions).

## 2 – Overall Structure Quality

**Steric Clashes**

**Missing Loops/Atoms** - lists loops missing from the structure and residues with missing heavy atoms in the analysed region.

**Protein Packing** reports whether the average protein Z-score falls below -1.2 which indicates a poorly built structure.



## 3 – Minor Structural Issues

**Isolated Water Clusters** - number of water clusters that are not hydrogen-bonded to anything else. Electron density possibly misidentified as water.

Number of **Buried unsatisfied HB donors and acceptors** or **Waters with no HB partners** can be indicative of modelling errors.

The same applies to a host of geometry checks: **Bond Length and Angle Deviations, Backbone and Sidechain Dihedrals, Peptide and Sidechain Planarity, and Improper Torsions.**

## 4 – Structure statistics

**PDB Resolution < 2.5 Å**

**PDB Rfree - R < 0.06**

**Unusual B-factors** - number of residues whose average B-factor is > 100 Å<sup>2</sup>.

# X-ray quality checks

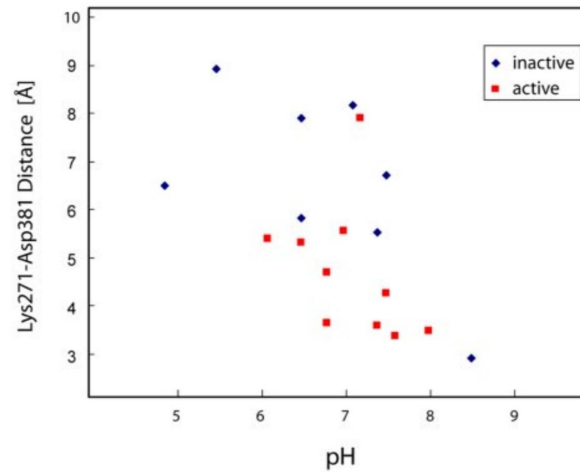
## ❑ What were the experimental conditions like?

Experimental conditions used to grow protein crystals can affect the protein conformation and even cause artifacts.

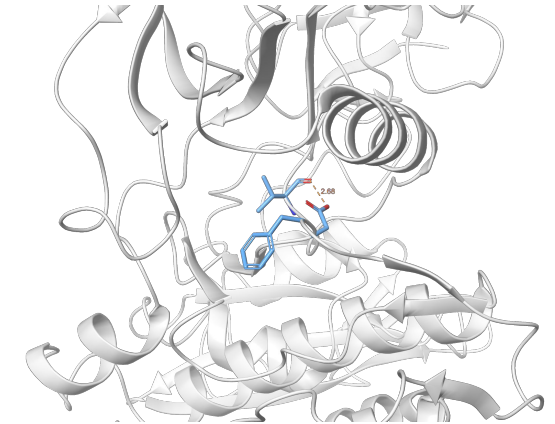
Pay close attention to:

- pH
- Salts and compounds used to facilitate crystallisation
- The method used for protein-ligand crystals (co-crystallisation vs soaking)

**pH** of the crystallisation buffer can cause residues to adopt **atypical protonation states** which could be functionally important.



Dependence of distance between DFG-Asp and ATP-coordinating Lys on the pH of the crystallisation buffer in *apo* Abl kinase structures. From Shan *et al.* PNAS 106 (1) 139-144 (2009).



Short distance between Asp and the backbone carbonyl oxygen indicates that Asp is protonated in the DFG-out conformation of Abl kinase (PDB ID: 1OPK).

# X-ray quality checks

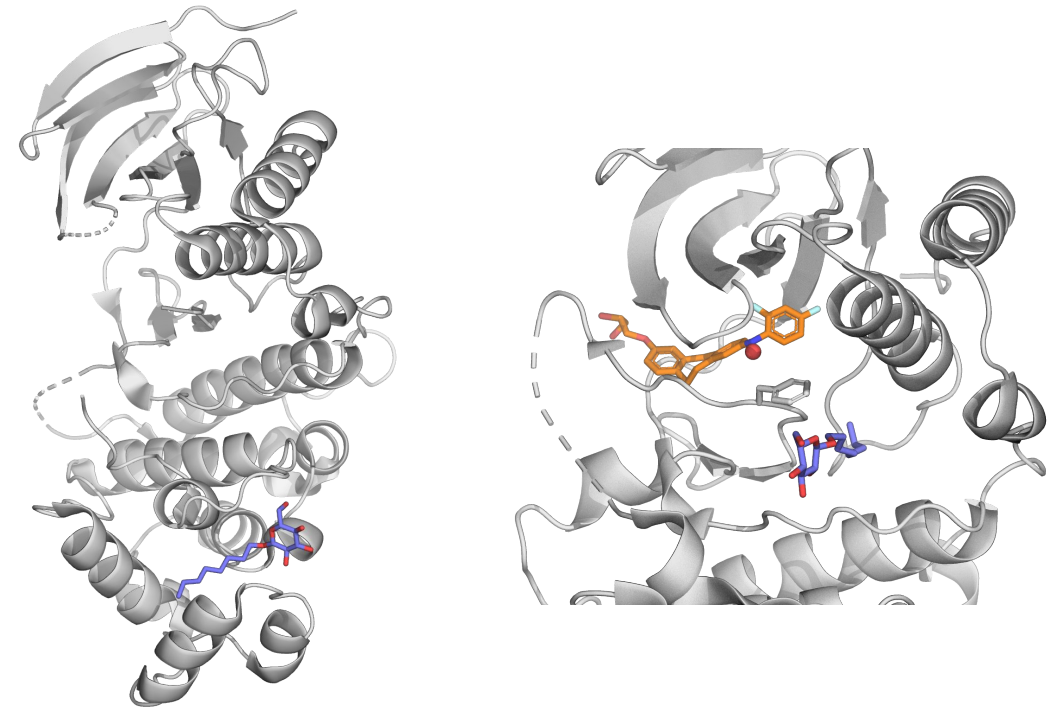
## ❑ What were the experimental conditions like?

Experimental conditions used to grow protein crystals can affect the protein conformation and even cause artifacts.

Pay close attention to:

- pH
- Salts and compounds used to facilitate crystallisation
- The method used for protein-ligand crystals (co-crystallisation vs soaking)

**Crystallising agents** can at times bind to proteins and alter their conformations. Such binding events can also reveal additional protein functionality.



N-octyl-beta-glucopyranoside is a detergent that has accidentally contributed to the discovery of **lipid binding capability** of p38 $\alpha$  kinase, as well as its alternative mechanism of activation, when it was found bound to the MAPK insert (left, PDB ID: 2NPQ). However, in some structures, the tail of the same detergent can **displace DFG-Phe**, irrespective of the presence of an inhibitor, and lead to a conformation between DFG-in and DFG-out states (right, PDB ID: 3QUE).

# X-ray quality checks

## ❑ What were the experimental conditions like?

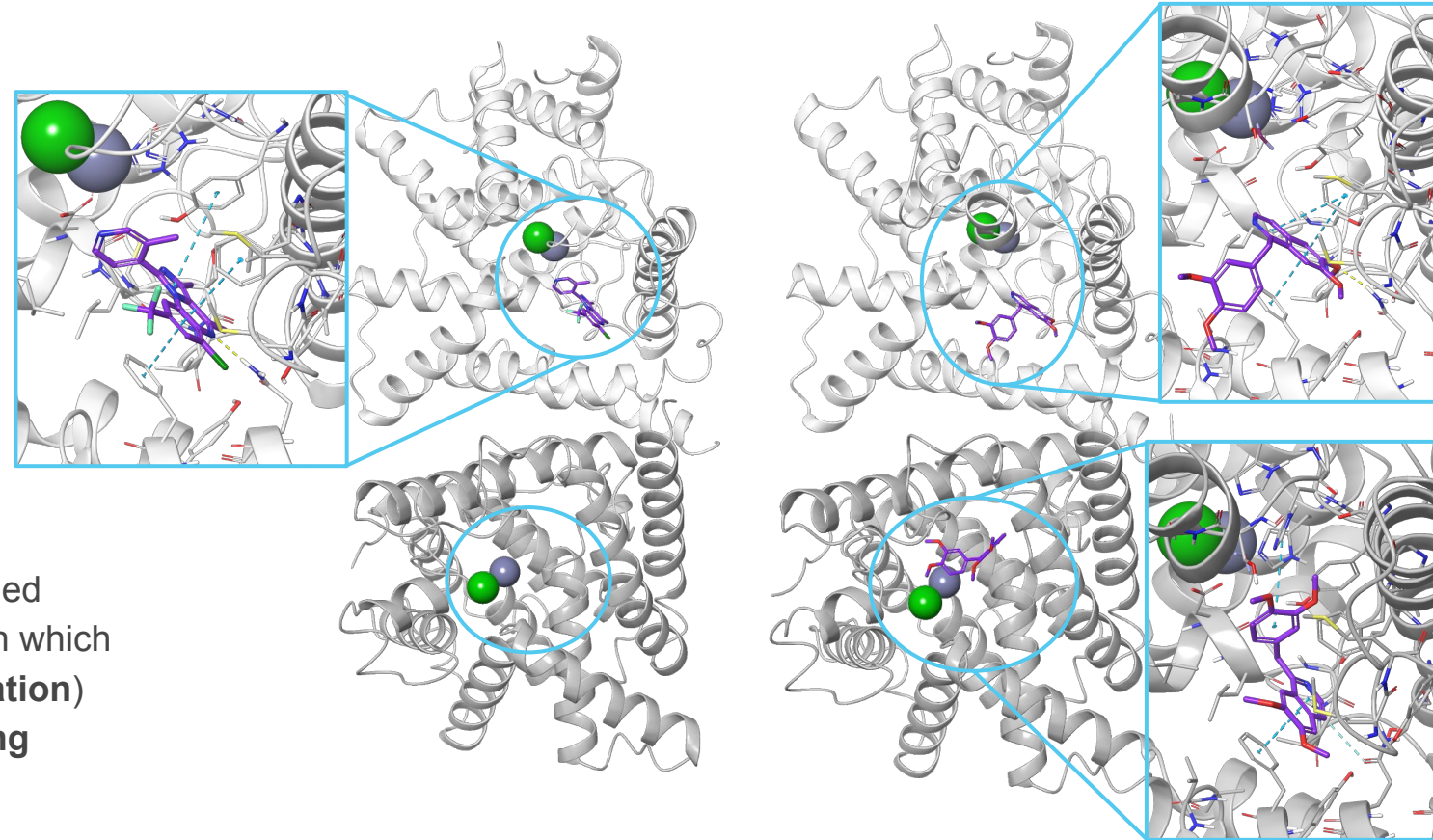
Experimental conditions used to grow protein crystals can affect the protein conformation and even cause artifacts.

Pay close attention to:

- pH
- Salts and compounds used to facilitate crystallisation
- The method used for protein-ligand crystals (co-crystallisation vs soaking)

Crystals of protein-ligand complexes can be obtained either by directly growing crystals from a solution in which both protein and ligand are present (**co-crystallisation**) or by first creating protein crystals and then **soaking** them in a ligand solution.

While soaking requires less time and resources, it can lead to **misleading binding poses** as the ligand is added after the crystal lattice has been formed and the binding site might not be fully available.



PDE10A protein typically has two chains in the asymmetric unit. When such crystals are soaked, the ligand can usually bind only in one of the chains due to crystal contacts that block the other site (left, PDB ID: 3SNL). Occasionally, a ligand can access the other site too, but it assumes a different binding mode (right, PDB ID: 2WEY). The enlarged ligand binding sites are aligned for easier comparison.



# X-ray quality checks

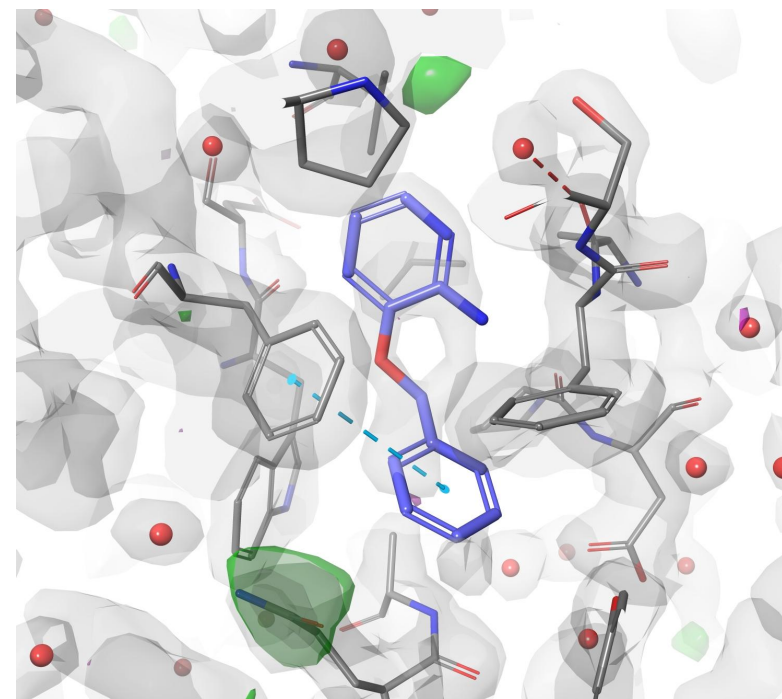
- ❑ Are there any parts of the structure that don't fit the electron density well?

Even in high-resolution structures, there could still be parts of the model that don't necessarily describe the experimental data well and their fit to the electron density is **poor**.

The density fit is **important to visualise** for sites of interest, **especially if they contain a ligand**, as any errors in such sites can lead the project in the wrong direction.

Most of the entries in the PDB contain the **diffraction data** that can be used together with the deposited model to create and view the electron densities.

If you observe any incorrect rotamers or poor density fits, you can either **consult a crystallographer**, try to **re-refine** the electron density map yourself (only if really desperate and/or confident in your skills), or check whether it has already been re-refined by the **PDB-REDO** project.



Electron density maps of the leukotriene A4 hydrolase binding site with a bound inhibitor (PDB ID: 3FTY). 2Fo-Fc map is shown in grey (contoured at 1  $\sigma$ ), while the Fo-Fc difference map (contoured at 3  $\sigma$ ) is shown in magenta (for density added by the model, but unsupported by diffraction data) and green (for density unaccounted for by the structural model).

# M How to view the electron density?

**3FTY**  
Leukotriene A4 hydrolase in complex with fragment 3-(benzyloxy)pyridin-2-amine

DOI: 10.2210/pdb3FTY/pdb

Classification: **HYDROLASE**  
Organism(s): Homo sapiens  
Expression System: Escherichia coli  
Mutation(s): No

Deposited: 2009-01-13 Released: 2009-07-28  
Deposition Author(s): Davies, D.R.

**Experimental Data Snapshot**

Method: X-RAY DIFFRACTION  
Resolution: 2.15 Å  
R-Value Free: 0.257  
R-Value Work: 0.206  
R-Value Observed: 0.208

**wwPDB Validation**

Metric	Percentile Ranks	Value
Rfree		0.251
Clashscore		5
Ramachandran outliers		0
Sidechain outliers		2.2%
RSRZ outliers		1.3%

This is version 1.2 of the entry. See complete history.

## 3FTY

Leukotriene A4 hydrolase in complex with fragment 3-(benzyloxy)pyridin-2-amine

Sequence of 3FTY | Leukotri... 1: Leukotriene ... A

**Structure**

3FTY | Leukotriene A4 hydrolase in ...

Type Assembly

Asm Id 1: Author And Softwar...

3IP 710 | D [auth A]

**Measurements**

**Components** 3FTY

Component	Type	Visibility	Actions
Polymer	Cartoon	<input type="checkbox"/>	...
Ligand	Ball & Stick	<input type="checkbox"/>	...
Water	Ball & Stick	<input type="checkbox"/>	...
Ion	Ball & Stick	<input type="checkbox"/>	...
[Focus] Target	Ball & Stick	<input type="checkbox"/>	...
[Focus] Surroundings (5 Å)	<input type="checkbox"/>	...	

**Density** 3FTY

Map	σ Level	Visibility
2Fo-Fc σ	1.5	<input type="checkbox"/>
Fo-Fc(+ve) σ	3	<input type="checkbox"/>
Fo-Fc(-ve) σ	-3	<input type="checkbox"/>

Entry 3fty  
View Around Focus ...

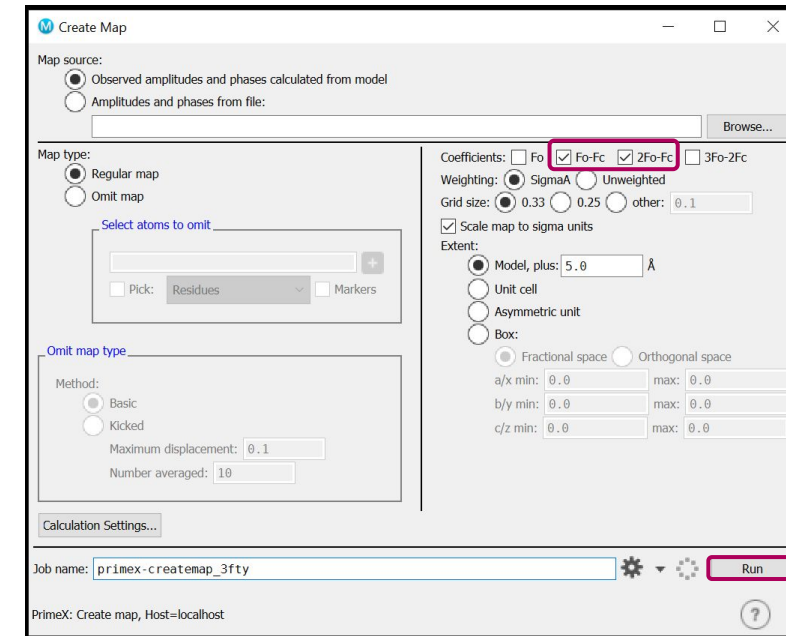
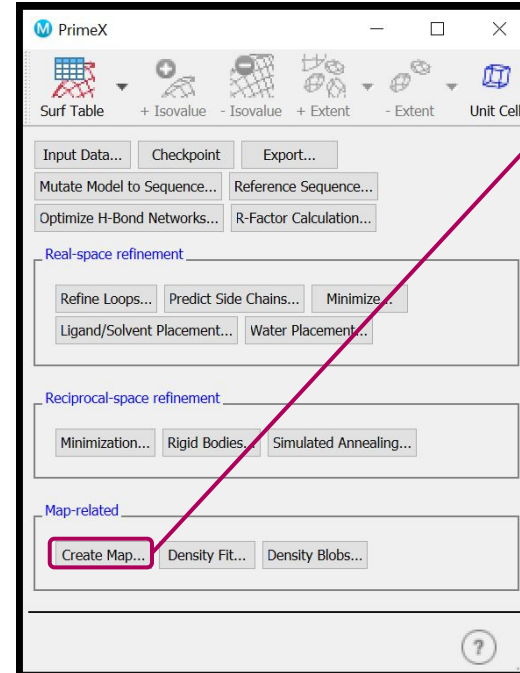
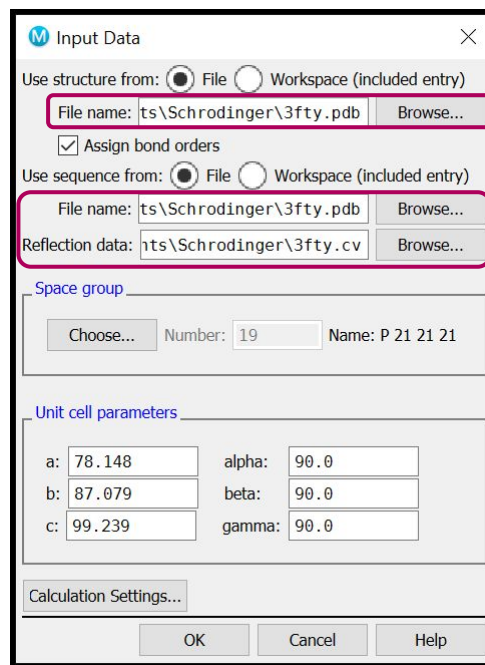
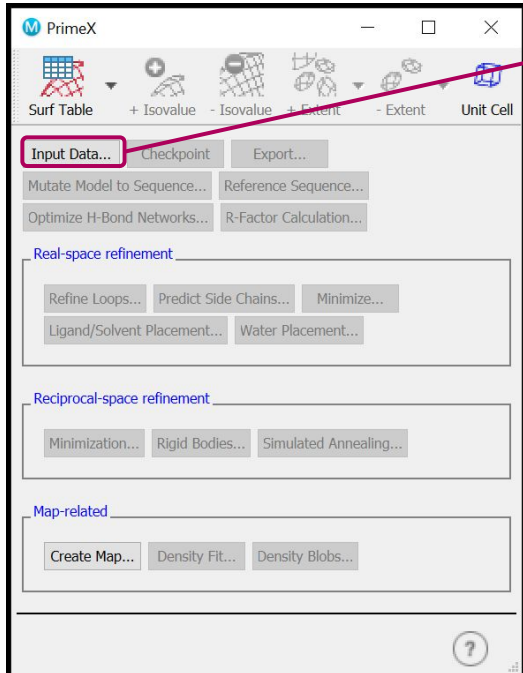
Nothing to Update

Controls Help

Once you select a residue, the view shows its electron density, as well as the density of residues within a 5-Å range. You can also control the  $\sigma$  level for each of the maps.

# M How to view the electron density?

Tasks -> Browse -> Other Applications -> PrimeX



You need to point PrimeX to your structural model and its sequence, as well as the reflection data (if using a PDB model, the reflections can be pre-downloaded through Get PDB File or Protein Preparation Wizard).

Make sure both the 2Fo-Fc and the Fo-Fc coefficients are selected to create the corresponding maps.

# M How to view the electron density?

Tasks -> Browse -> Other Applications -> PrimeX

*If map results do not automatically appear in the workspace, you can import them manually*

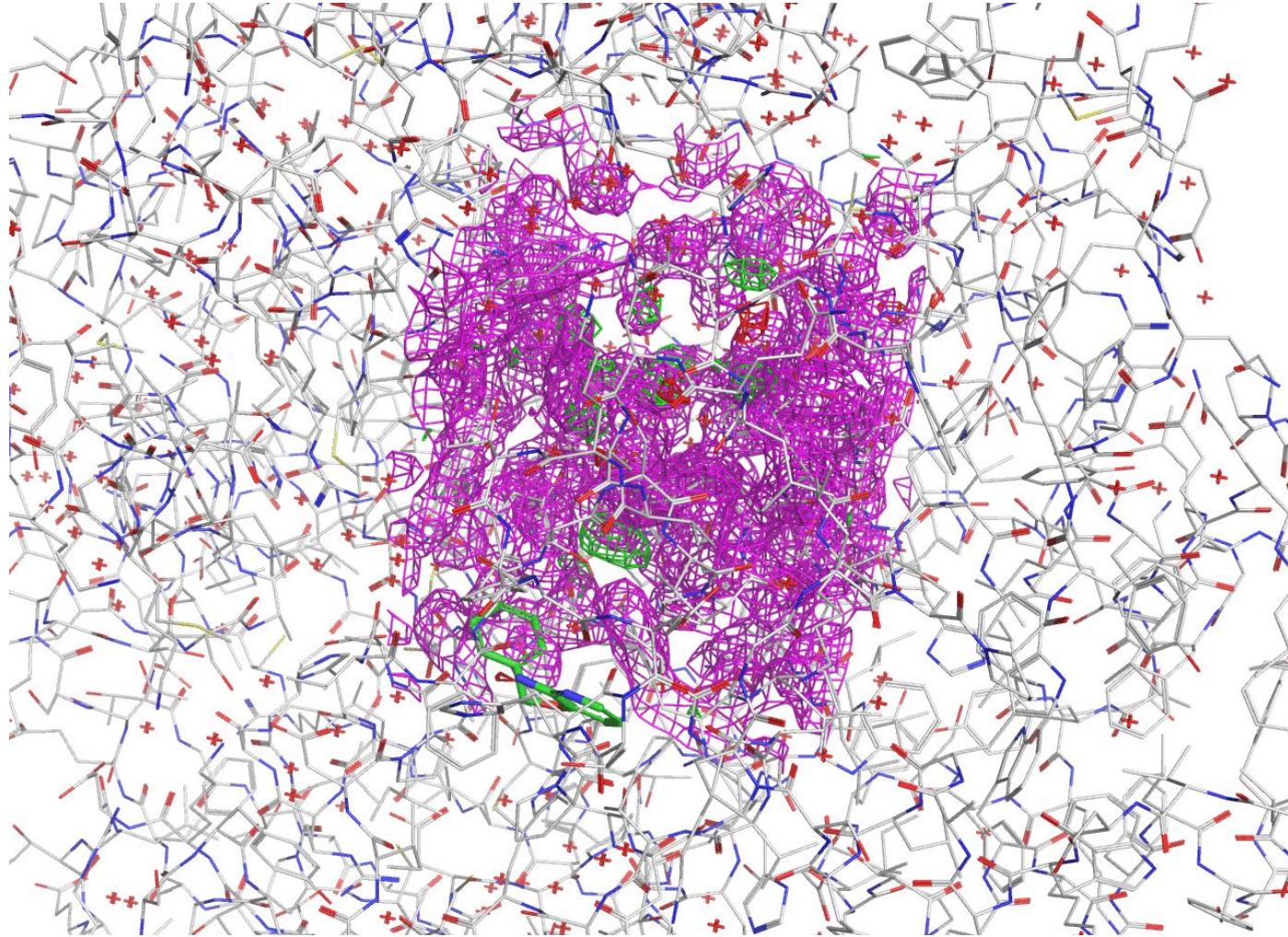
The image shows the PrimeX software interface. The 'Manage Surfaces' dialog is open, displaying a table of surfaces. The 'Import...' button is highlighted with a red box. Below it, the 'Import Surface / Volume File' dialog is open, showing a file explorer view of a directory. Two files, 'primex-createmap-4dix-out-0.cns' and 'primex-createmap-4dix-out-1.cns', are highlighted with a red box. The 'File name' field at the bottom of the file explorer contains the names of these two files.

In	Limit	Entry	Volume Name	Vol	Surface
<input type="radio"/>		16: 4DJX	primex-...	<input type="radio"/>	Fo-Fc
<input type="radio"/>		16: 4DJX	primex-...	<input type="radio"/>	2Fo-Fc

File name: "primex-createmap-4dix-out-1.cns" "primex-createmap-4dix-out-0.cns"



# M How to view the electron density?



# M How to view the electron density?

In	Limit	Entry	Volume Name	Vol	Surface Name	Comments	Surface Type	Isovalue	Area	Sigma
		1: 3FTY	primex-...		Fo-Fc	Regular	Fo-Fc	2.99844	3819.199	0.999
		1: 3FTY	primex-...		2Fo-Fc	Regular	2Fo-Fc	0.999304	1233.334	0.999

change how much density is shown on the screen (less is more or your computer might slow down dramatically)

change the contouring level (higher the value, stronger the density still shown)

change the colour, style, and transparency of the map

Name: Fo-Fc

Style:  Solid  Mesh  Dot

Transparency:

Front surface: 0

Back surface: 0

Color scheme: Constant

Color: ■ Negative: ■

Limit:

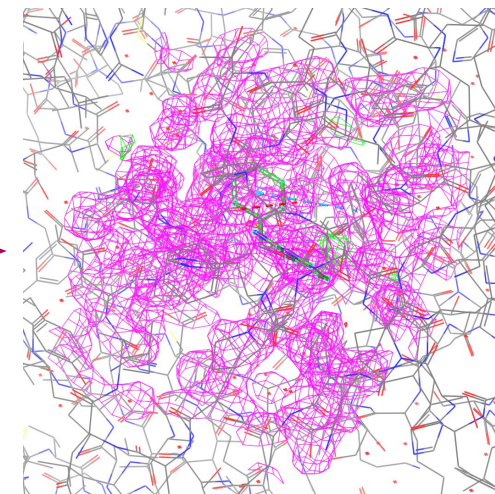
Display at most: 16.00 Å<sup>3</sup>

Isovalue: 3.00 Sigma

2.99844



Shortcut: click your mouse wheel on the area of interest to center

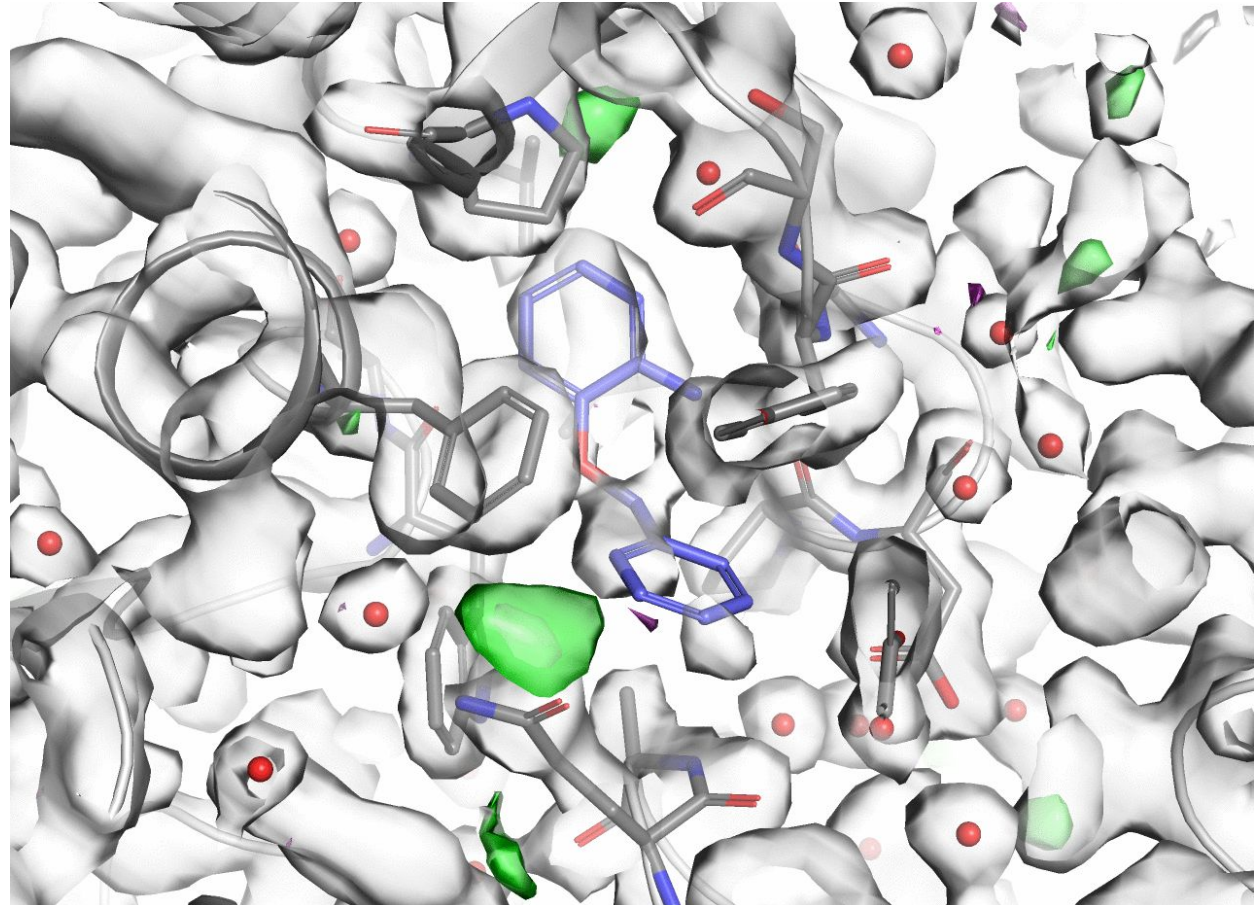
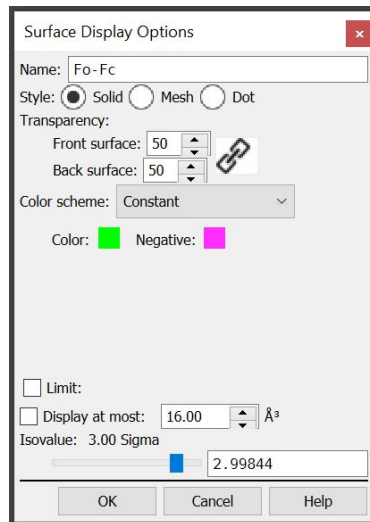
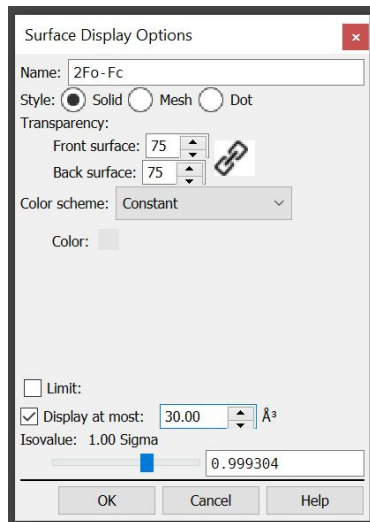




# M How to view the electron density?

Style changes made to the electron density maps:

- Coloured the 2Fo-Fc map grey, set the style as solid, and the transparency of front/back surfaces to 75.
- Coloured the negative Fo-Fc map magenta, set the style as solid, and the transparency of front/back surfaces to 50.

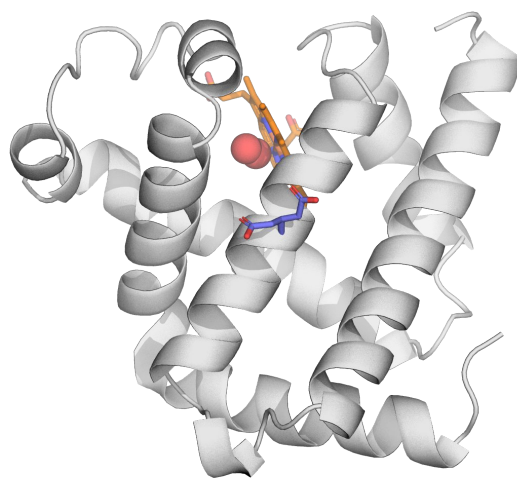


# X-ray quality checks

## ❑ Any alternate conformations of protein residues or ligands?

At times, residues within the crystal can be observed in two or more distinct conformations. Such cases are recorded in the **occupancy** field of the PDB file.

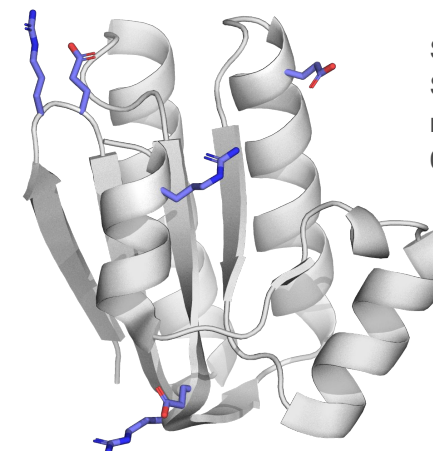
Occupancy reports the **fraction of molecules** which contain that specific conformation. For residues with a single conformation, the occupancy is 1.



Alternate locations of Glu109 in myoglobin (PDB ID: 1A6M).

## ❑ Were some parts of the protein modelled in (0-occupancy atoms)?

In some cases, crystallographers model in parts of the protein (usually loops) that are not seen clearly in the electron density. Such atoms have an occupancy of 0.



Several side chains in bacterial SpoIIAA protein have been modelled in with the occupancy of 0 (PDB ID: 1H4X).

Make sure none of the relevant parts of your target have 0-occupancy atoms as their positions are not reliable.



# X-ray quality checks

## ❑ What about the missing atoms/residues?

If parts of the protein are **flexible** and have space to move within the crystal lattice, it is usually not possible to reliably model them due to very weak electron density.

To avoid possible misinterpretations, crystallographers deposit structures to the PDB without these parts included. In the majority of cases, these are **N- and C-terminal tails or loops**.

You can easily find what's missing from the model in the header of PDB files under **REMARK 465** (missing residues) and **REMARK 470** (residues with missing heavy atoms).

```
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     LEU A   73
REMARK 465     ARG A   74

REMARK 470 MISSING ATOM
REMARK 470 THE FOLLOWING RESIDUES HAVE MISSING ATOMS (M=MODEL NUMBER;
REMARK 470 RES=RESIDUE NAME; C=CHAIN IDENTIFIER; SSEQ=SEQUENCE NUMBER;
REMARK 470 I=INSERTION CODE):
REMARK 470   M RES CSSEQI  ATOMS
REMARK 470     PRO A   1    CG  CD
REMARK 470     ILE A  14    CG1 CG2 CD1
```

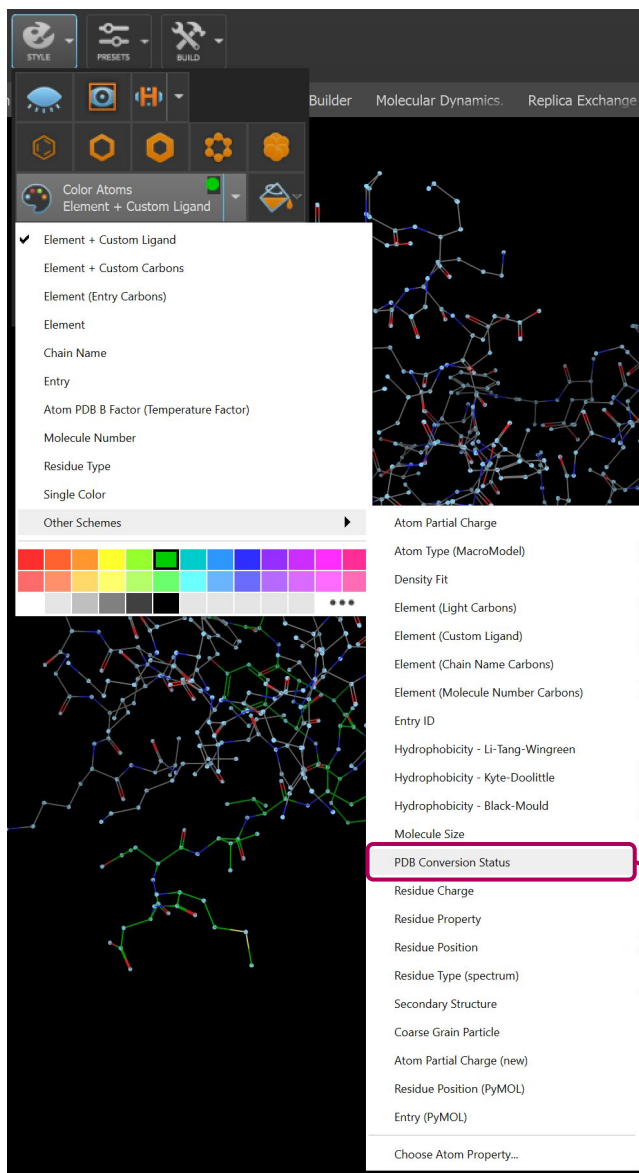
## To model or not to model missing parts?

Unfortunately, there are no clear-cut rules and it's up to the modeller to estimate whether the addition of the missing parts will cause more harm than good on the project.

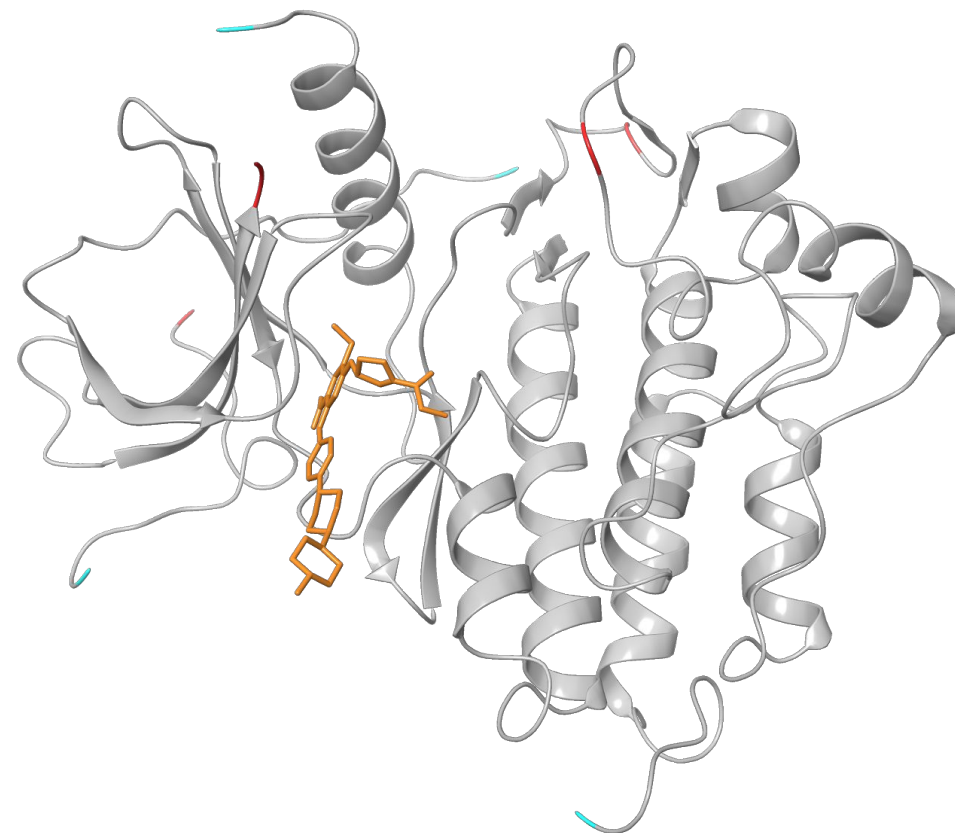
### Some tips:

- Adding missing loops around binding sites when running docking experiments with a rigid receptor is likely to cause artifacts.
- Any calculations based on MD simulations cannot run if there are any of the heavy atoms missing.
- If you decide not to add the missing loops back in, it's always a good idea to cap the terminal residues to avoid placing charges incorrectly.

# M How to see the location of missing atoms/residues?



- Non-standard residues connected by geometry and/or CONECT records.
- Standard residue, but with missing atoms.
- Adjacent residue is missing.
- Standard residue with unrecognized atom names connected by geometry.
- Residue with an alternate location indicator.
- Standard residues connected by standard templates.



*colour residues based on the  
PDB Conversion Status  
colour scheme*

# M How to see the location of missing atoms/residues?

The screenshot displays the Multiple Sequence Viewer/Editor interface. The main window shows a sequence alignment between P00533 (EGFR\_HU...) and 5Y9T (A). The sequences are displayed in rows, with missing residues highlighted in darker red. The interface includes a menu bar (File, Edit, Select, View, Align), a toolbar with search and alignment options, and a status bar at the bottom showing '2 SEQUENCES selected' and '1 STRUCTURES in Workspace'.

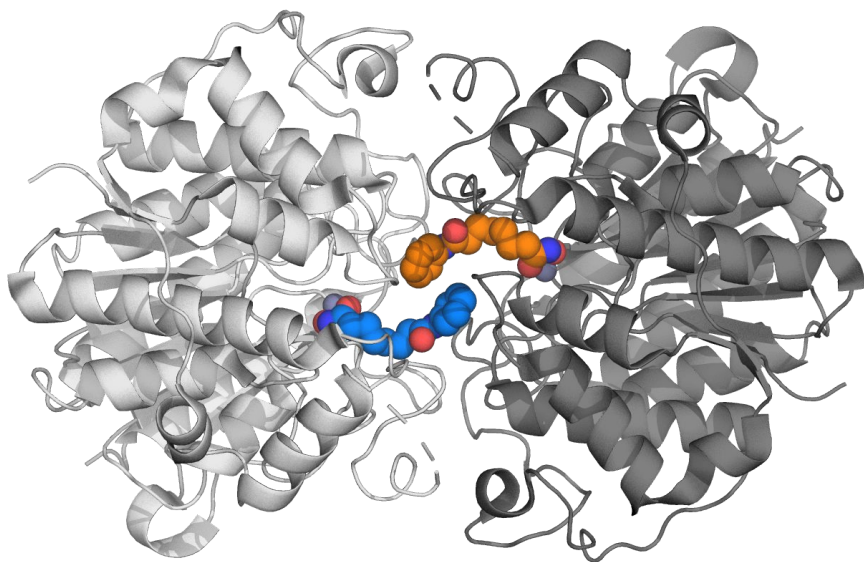
Sequence	630	640	650	660	670	680
5Y9T A	H	P	N	C	T	Y
P00533 EGFR_HU...	H	P	N	C	T	Y
5Y9T A	L	Q	E	R	E	L
P00533 EGFR_HU...	L	Q	E	R	E	L
5Y9T A	K	E	L	R	E	A
P00533 EGFR_HU...	K	E	L	R	E	A
5Y9T A	D	N	I	G	S	Q
P00533 EGFR_HU...	D	N	I	G	S	Q
5Y9T A	Y	H	A	E	G	G
P00533 EGFR_HU...	Y	H	A	E	G	G
5Y9T A	E	R	L	P	Q	P
P00533 EGFR_HU...	E	R	L	P	Q	P
5Y9T A	T	D	S	N	F	Y
P00533 EGFR_HU...	T	D	S	N	F	Y

- **Multiple sequence viewer** allows you to easily compare the sequence of your structure to the canonical sequence from UniProt.
- Note that the missing residues are shown by darker colour shades.
- Such visualisations help you to quickly gauge which missing regions could be difficult to model back in.

# X-ray quality checks

## ❑ Could the crystal contacts have caused some artifacts?

Proteins in the crystal lattice are much more **tightly packed** than in solution. Such packing can stabilise conformational states that wouldn't necessarily be adopted in solution.



For example, in the crystal structure of the HDAC8 protein complexed with SAHA (PDB ID: 1T69), the ligand directly interacts with both the protein and the ligand of a crystal mate which brings the validity of its binding pose into question.

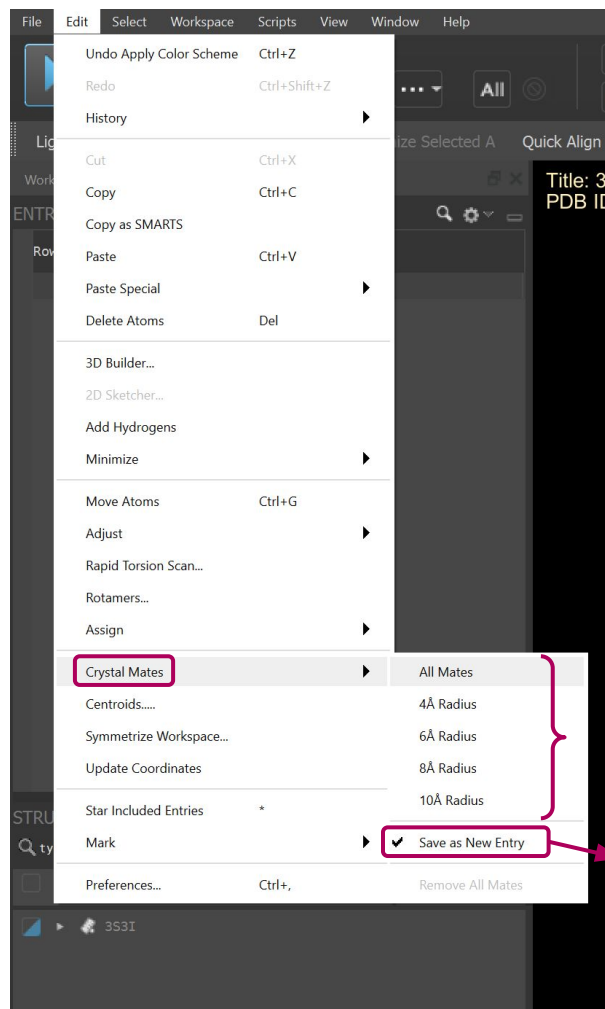
It is highly recommended to generate **crystal symmetry mates** that can be found at 5 Å of the asymmetric unit cell and check whether any pertinent parts of the protein or ligand can be found at the **protein-protein interfaces**.

The majority of modern visualisation softwares can easily generate crystal symmetry mates at any desired distance range or even create whole **unit cells** or **supercells**.

**Note:** Crystal symmetry mates are generated based on the input coordinates and the symmetry operators. If you at any point alter the original coordinates (e.g. by aligning the model to a reference structure), the created symmetry mates will be **incorrect**.



# M How to check the crystal contacts?



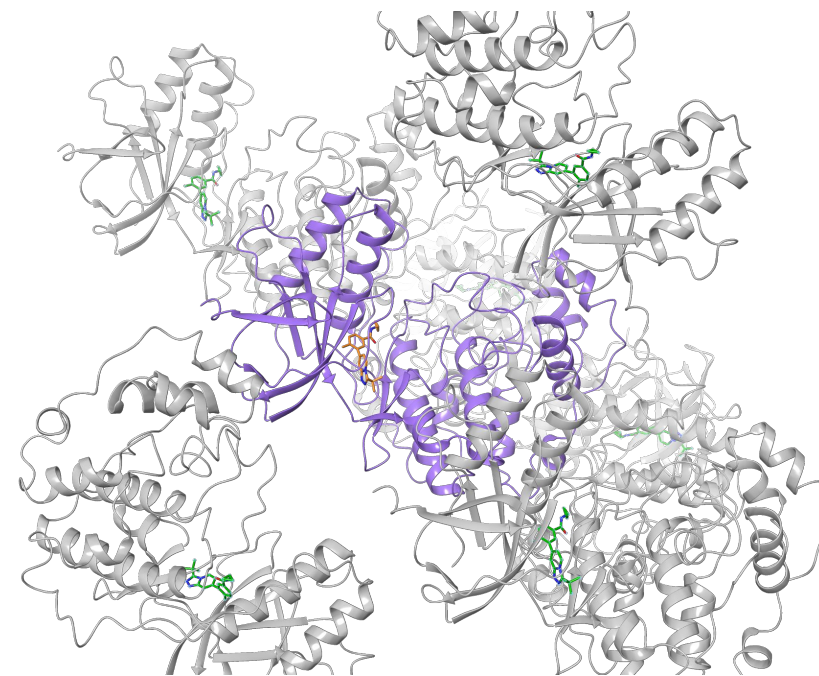
The structure must contain the **CRYST1** section with the **space group information** so that the appropriate symmetry operations can be applied to create **crystal mates**.

CRYST1    45.280    84.960    123.510    90.00    90.00    90.00    P 21 21 21    4

*lengths of unit cell edges (Å)*      *unit cell angles (°)*      *space group*      *# of ASUs*

*choose whether you want to add a layer of residues within a certain distance radius or whole molecules*

*save them as new entries (default) or include them in the existing one*



The asymmetric unit (ASU) is coloured purple (with ligand in orange), while its crystal mates are coloured grey (with ligands in green) (PDB ID: 3S3I).

# The final checklist

## Function-related checks

- What's the subcellular location of the protein?
- Is the protein a monomer or a multimer?  
If a multimer, is it a homomer or a heteromer?
- Is the protein known for multiple conformational states?
- What about atypical chemical forms?
- Maybe there are some PTMs?
- Are any metals involved?
- Does the protein bind any other cofactors?

## Sequence-related checks

- Is the whole protein there? Any missing (sub)domains?
- Are you working with the correct sequence?
- Are there any "extras", e.g. signalling peptides or expression tags?
- Are there any homologues?

## X-ray related checks

- Is the resolution high enough?
- Are the R and  $R_{\text{free}}$  factors low?
- What are the B-factors like?
- What about the RSCC values?
- Are there any geometric outliers or clashes?
- What were the experimental conditions like?
- Are there any parts of the structure that don't fit the electron density well?
- Any alternate conformations of protein residues or ligands?
- Were some parts of the protein modelled in (0-occupancy atoms)?
- What about the missing atoms/residues?
- Could the crystal contacts have caused some artifacts?

# Recommended reading

The **FEBS**  
Journal

REVIEW ARTICLE |  Free Access

## Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures

Alexander Wlodawer, Wladek Minor, Zbigniew Dauter, Mariusz Jaskolski

First published: 06 December 2007 | <https://doi.org/10.1111/j.1742-4658.2007.06178.x> | Citations: 134

	<b>HHS Public Access</b> Author manuscript <i>Postepy Biochem.</i> Author manuscript; available in PMC 2017 September 24.
---	---

Published in final edited form as:  
*Postepy Biochem.* 2016 ; 62(3): 242–249.

### The young person's guide to the PDB\*

Wladek Minor<sup>1,8</sup>, Zbigniew Dauter<sup>2</sup>, and Mariusz Jaskolski<sup>3,4</sup>

<sup>1</sup>Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22908, USA <sup>2</sup>Macromolecular Crystallography Laboratory, National Cancer Institute, Argonne National Laboratory, Argonne, IL 60439, USA <sup>3</sup>Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University, Poznan, Poland <sup>4</sup>Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

- There are also quite a few slide decks online from Gerard J. Kleywegt aimed at non-crystallographers with lots of illustrative examples.

# Protein Preparation Workflow



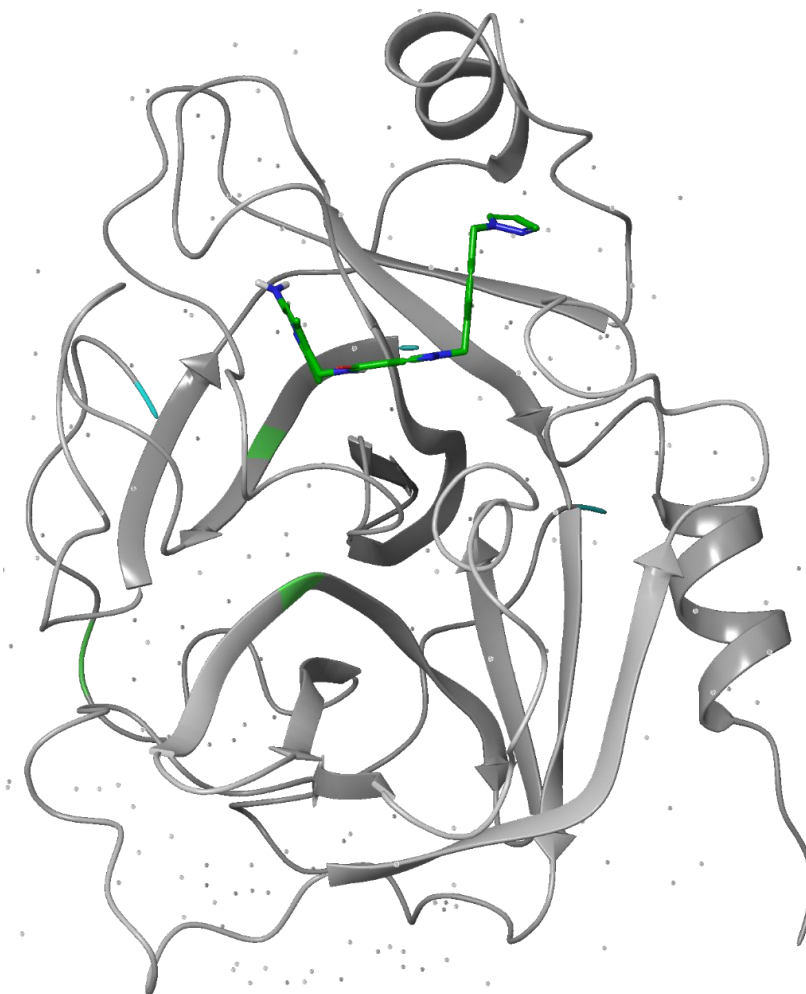


# M How to prepare your protein?

Tasks -> Browse -> Protein Preparation and Refinement -> Protein Preparation Workflow  
(can also be found in the Favourites toolbar under Protein Preparation)

You can use the **PDB Conversion Status** colour scheme to highlight possible issues:

- Non-standard residues connected by geometry and/or CONECT records.
- Standard residue, but with missing atoms.
- Adjacent residue is missing.
- Standard residue with unrecognized atom names connected by geometry.
- Residue with an alternate location indicator.
- Standard residues connected by standard templates.



Protein Preparation Workflow (Interactive) - 5TZ9

Preparation Workflow | Diagnostics | Substructures

INTERACTIVE

Click a button to run a step. Wait for each process to finish and ensure the correct entry is in the Workspace before requesting the next step.

- 1. Specify Protein**  
Source: **Workspace** Get PDB...  
Entry: 5TZ9 (1)  
Review Structure Global Settings ▾
- 2. Preprocess**  
 Cap termini  Fill in missing side chains More Options ▾  
Preprocess Default actions selected (6)
- 3. Diagnose and Analyze**  
*Run diagnostics and review the structure in the secondary tabs.*  
Check Structure
- 4. Optimize H-bond Assignments**  
*Optimize to address any overlapping hydrogens.* Settings ▾  
Optimize - OR - Assign with Constraints...
- 5. Minimize and Delete Waters**  
*Run a restrained minimization, then optionally delete specified waters.* Settings ▾  
Clean Up

Workflow group: 5TZ9-proteinprep\_1

# M How to prepare your protein?

Check here **before and after preprocessing** to see:

- Unknown atom types.
- Missing side-chain atoms
- Overlapping atoms (typically hydrogens which will be corrected in the H-bond optimisation stage).
- Alternate positions (where you can choose which one you'd like to proceed with).

Protein Preparation Workflow (Interactive) - 5TZ9 - 2-preproc...  
Preparation Workflow Diagnostics Substructures  
Check Workspace Entry Entry: 5TZ9 - 2-preprocessed (3)  
One issue was found. See Reports for more information about the protein.

Valences Missing Overlapping Alternates Reports

View: Steric Clashes  
Select: Average B-factors  
Gamma Atom B-Factor  
Peptide Planarity  
Sidechain Planarity  
Improper Torsions  
C-alpha Stereochemistry  
Missing Atoms  
Protein Reliability Report...  
Ramachandran Plot...

	Distance	Min Allowed	Delta
E:ACE 390: CH3 - E:PHE 524: CA	2.452	3.4	0.948
E:ACE 390: CH3 - E:PHE 524: C	1.783	3.42	1.637
E:ACE 390: CH3 - E:PHE 524: O	2.441	3.2	0.759
E:ACE 390: C - E:PHE 524: C	1.773	3.25	1.477
E:ACE 390: C - E:PHE 524: O	2.593	3.4	0.807
E:ACE 390: C - E:ASP 577: OD2	2.595	3.42	0.825
E:ACE 390: O - E:PHE 524: O	2.126	3.2	1.074
E:ACE 390: O - E:CYS 574: CB	2.855	3.44	0.585
E:ACE 390: O - E:THR 413: C	1.815	3.22	1.405
E:ACE 390: O - E:THR 413: O	2.597	3.22	0.623
E:ACE 390: O - E:THR 413: O	1.79	3	1.21
E:ACE 390: O - E:THR 413: O	2.64	3.2	0.56
E:ACE 390: O - E:THR 413: O	2.598	3.22	0.621

Export...  All reports Structure average: N/A  
< Workflow Substructures >

**Protein Reports** allows you to check and export all the possible geometric outliers, steric clashes, missing atoms, and unusually large B-factors, while the **Ramachandran Plot** allows for a quick visual of backbone dihedrals.

# M How to prepare your protein?

The PPW protocol is composed of 3 steps:

- **Preprocess**
- Optimize H-Bond Assignments
- Minimize

Assigns **bond orders** to all bonds in the structure based on a range of factors, including connectivity, bond length, bond angles and dihedral angles. For HET groups, it first checks Chemical Components Dictionary using SMARTS patterns. **Always check** the HET group assignments.

**Hydrogens are typically missing** in X-ray structures and must be added for any further calculation. It's usually a good idea to remove the original hydrogens as they're often added incorrectly and might cause compatibility issues with other suite applications.

**Force fields treat metal compounds as ionic** rather than covalent, so it's necessary to replace the existing bonds to metals with **zero-order bonds** (to keep the molecule intact) and correct the formal charge on the metal and the neighboring atoms to treat the bonds as ionic.

Forms a bond between **sulphur atoms** that are within 3.2 Å of each other. **Renames the CYS residues to CYX** if the bond is added. **Always check** which bonds have been formed and whether they are biologically relevant.

Protein Preparation Workflow (Interactive) - 5TZ9 - 2-preproc... INTERACTIVE

Click a button to run a step. Wait for each process to finish and ensure the correct entry is in the Workspace before requesting the next step.

1. Specify Protein  
Source: **Workspace** (Get PDB...)  
Entry: 5TZ9 - 2-preprocessed (3) (Review Structure) (Global Settings)

2. Preprocess (✓)  
 Cap termini  Fill in missing side chains (Preprocess) (More Options)

3. Diagnose and Analyze  
Run diagnostics and review the structure (Check Structure)

4. Optimize H-bond Assignments  
Optimize to address any overlapping hydrogen bonds (Optimize)

5. Minimize and Delete Waters  
Run a restrained minimization, then delete the specified waters. (Clean Up)

More Options:

- Align to:  First selected entry  PDB: [ ]
- Assign bond orders  Using CCD database
- Replace hydrogens
- Create  Zero-order bonds to metals
- Disulfide bonds
- Antibody annotation scheme: Kabat
- Renumber residues to match scheme
- Add terminal oxygens to protein chains
- Convert selenomethionines to methionines
- Delete waters beyond hets: 8.00 Å
- Fill in missing loops (using Prime)
- Generate het states (with Epik): pH: 7.4 +/- 2.0 (Max states to process automatically: 1)

Workflow group: 5TZ9-proteinprep\_1

It's **critical to carefully check** where the missing atoms and residues are and whether modelling them in incorrectly can have jeopardise further calculations.

Typically, **missing side chains** tend to be on the protein surface, so adding them back won't create issues. If they're in the ligand binding site, then you need to make sure they're correctly rebuilt.

If you've decided to **rebuild loops**, it's probably better to use more sophisticated tools to get a suitable conformation, such as Prime Homology Modelling, Prime Refine Loops, or even Molecular Dynamics.

If you've decided to leave the gaps in, make sure you **cap the termini**, so that no charges are introduced at wrong locations.

Water molecules are **better left untouched** during the preparation process as their presence is usually beneficial for advanced calculations, such as FEP+ and MD simulations. They can be easily removed at a later stage.

Runs **Epik** to generate probable ionization and tautomeric states in the specified pH range for all HET groups, as well as states prepared for binding to metals if the HET group is coordinated to a metal. Think carefully of the **pH range** you'd like to work with and **always check** if the results make sense. If still in doubt, you can always use **Jaguar** for more precise calculations.



# M How to prepare your protein?

Substructures tab allows you to **easily select and visualise** individual chains, water molecules, and HET groups and decide whether you want to keep them in the system.

## HET table

For each of the HET groups, you can (and should) **review all the Epik-generated states** and decide which one you'd like to proceed with. The one with the lowest state penalty is automatically selected.

When you select a HET state, its **Epik state penalty** is shown together with the hbond count and total charge.

Protein Preparation Workflow (Interactive) - 5TZ9 - 2-preprocessed

Preparation Workflow Diagnostics Substructures

Reload from Workspace Entry: 5TZ9 - 2-preprocessed (3)

Choose items below to view in Workspace, copy, or delete. Select

**Ligands, Metals, Other.** The Lig column shows detected ligands. To change the classification, visit the [Ligand Detection...](#) settings, then click *Reload from Workspace* above.

The *Preprocess* step may generate multiple states for your ligands. The (likely) most favorable state will be checked by default. Optionally choose a different state to keep.

Lig	Chain	Res Name + #	S1	S2
X	E	7SD 701	<input checked="" type="checkbox"/>	<input type="checkbox"/>

**Waters:**

Chain	Res Name + #
E	HOH 801
E	HOH 802
E	HOH 803
E	HOH 804

**Chains:** [Expand to PDB chain](#)

Chain	Type
E	Protein

1 item selected Clear Copy to New Entry Delete from Entry

Prepare Selected Only... < Diagnostics Workflow >

State penalty: 0.70 kcal/mol; H-Bond count: 0; Q: +1

water table

chain table



# M How to prepare your protein?

The PPW protocol is composed of 3 steps:

- Preprocess
- **Optimize H-Bond Assignments**
- Minimize

Protonation states of protein residues are determined with **PROPKA** by default.

This **empirical method** typically works well, but it fails to correctly determine the  $pK_a$  in certain situations:

- **Missing/erroneous solvation** can result in a rather different environment for the residue which can be then assigned a rare protonation state (e.g. uncharged Lys).
- **Small geometrical inaccuracies** can also lead to different assessments of the area.
- **Catalytic residues** can adopt atypical protonation/tautomeric states as part of their mechanism of action which PROPKA might not assign correctly.

**Always check PROPKA's results** to make sure the results are as expected. If in doubt, try to **prepare several structures** and compare the results.

Protein Preparation Workflow (Interactive) - 5TZ9 - 2-preprocessed

Preparation Workflow | Diagnostics | Substructures

Click a button to run a step. Wait for each process to finish and ensure the correct entry is in the Workspace before requesting the next step. **INTERACTIVE**

1. Specify Protein  
Source: **Workspace**   
Entry: 5TZ9 - 2-preprocessed (3)

2. Preprocess  
 Cap termini  Fill in missing side chains   
 8 actions selected

3. Diagnose and Analyze

4. Optimize H-bond Assignments  
Optimize to address any overlapping hydrogens.   
 - or -

5. Minimize and Delete Waters  
Run a restrained minimization, then optionally delete specified waters.

Workflow group: 5TZ9 - 2-preprocessed-proteinprep\_1

You can (and should) use the Interactive Optimizer to **manually review and adjust** the residues.

Interactive H-bond Optimizer

Analysis  
pH: 7.4  Include current orientations  
 Use PROPKA  Label pKas  Only treat Workspace selection  
  Use crystal symmetry

View all species 300 species total  
 View cluster: 5 77 clusters total  
Cluster 5

Display result: -1 Score: N/A

#	Lock	Species	State
1	<input type="checkbox"/>	E:ASN 474	No Flip
2	<input type="checkbox"/>	E:ASN 481	No Flip
3	<input type="checkbox"/>	E:HIS 434	HID
4	<input type="checkbox"/>	E:HIS 472	HID
5	<input type="checkbox"/>	E:HIS 482	HID
6	<input type="checkbox"/>	E:THR 55...	Initial
7	<input type="checkbox"/>	E:SER 57...	Initial
8	<input type="checkbox"/>	E:SER 59...	Initial

Pick to locate species

*performs H-Bond assignment automatically and creates a new entry*

*performs the restrained minimization and creates a new entry*

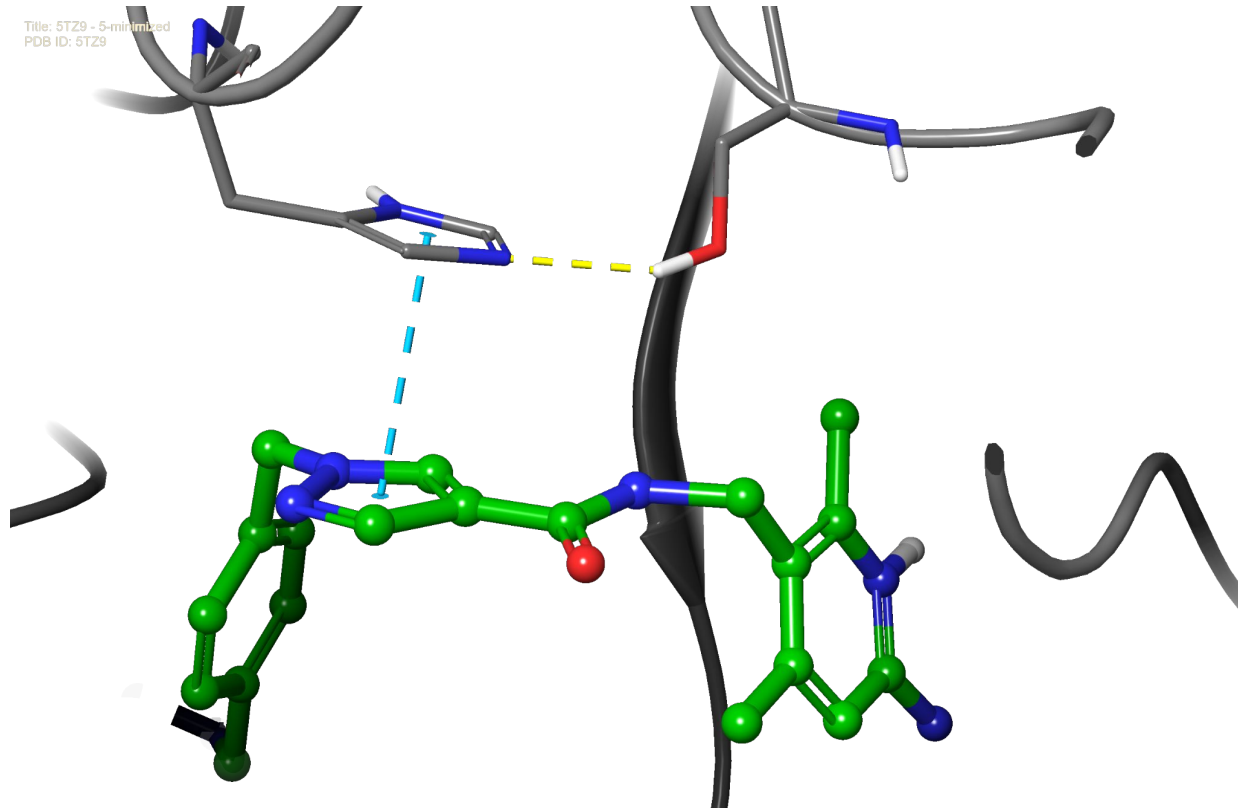
# M Further post-PPW activities to explore

Depending on the envisioned modelling task, you might want to consider running additional analyses:

- **Comparing multiple prepared structures** - If you're having doubts about the "correctness" of your prepared structure, preparing several others could help you identify potential problems.
- **Metal coordination** - Make sure to check whether the metal cofactors have **complete coordination shells** around them. If any of the coordination sites are vacant, stability issues might arise in subsequent calculations and you should consider whether a water molecule or other complexing group should be bonded to the metal using zero-order bonds.
- **Additional hydration** - Positions of water molecules in crystal structures are often unreliable which is why it's advisable to re-evaluate them, especially those in the ligand binding sites. Schrödinger's **WaterMap** application can be used to highlight regions of questionable solvation and, occasionally, areas where solvation is clearly missing from the original structure.
- **Manual loop rebuilding** - It's always worth the effort to try to rebuild any missing loops using Schrödinger's **Prime** application, as you will get a better idea of what possible effects such modelled features could have on your system.
- **Stability assessment** - Running a molecular dynamics simulation can often help estimate the overall stability of the system.
  - For example, if you've decided **not to rebuild loops** and just cap the termini, you can check if the regions around the chain breaks are stable throughout the simulation.
  - You can also check whether a ligand changes its binding pose in a **protein-ligand complex** which could indicate the original pose was stable due to the crystal environment.
  - In general, MD simulation should allow the protein to relax parts involved in crystal contacts. However, **large conformational changes** could also hint at issues with the original structure or even the force field.

# Molecular Dynamics helps identify stable states

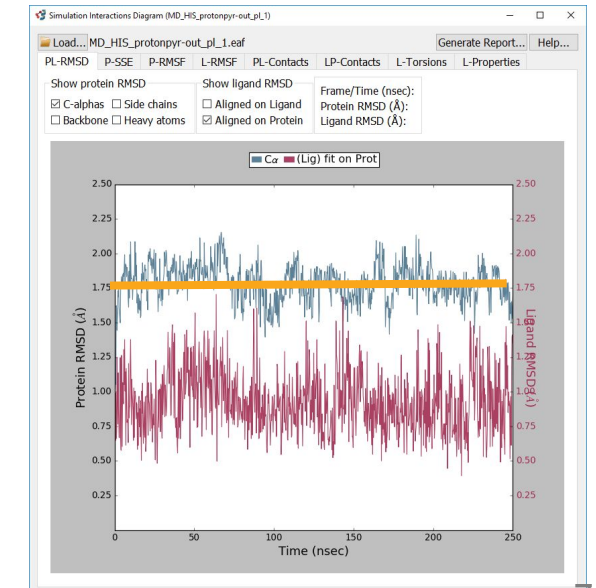
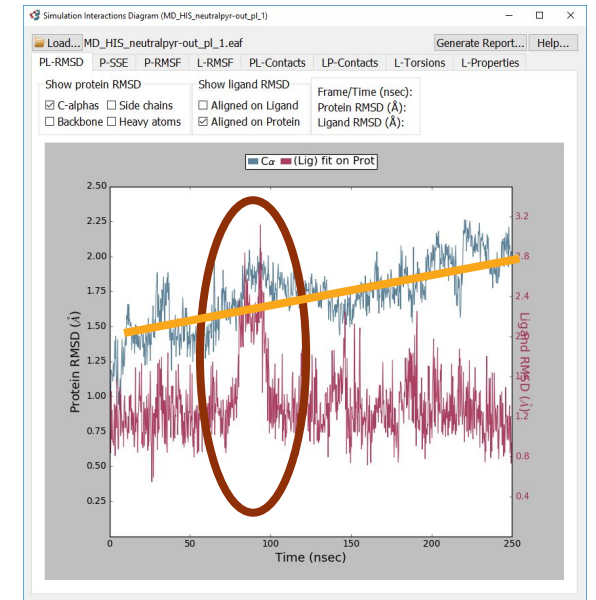
Title: 5TZ9 - 5-minimized  
PDB ID: 5TZ9



## HIS/N

- Increasing  $C\alpha$  RMSD
- Ligand changes binding mode

 HIS/N+1



# The final checklist

## Function-related checks

- What's the subcellular location of the protein?
- Is the protein a monomer or a multimer?  
If a multimer, is it a homomer or a heteromer?
- Is the protein known for multiple conformational states?
- What about atypical chemical forms?
- Maybe there are some PTMs?
- Are any metals involved?
- Does the protein bind any other cofactors?

## Sequence-related checks

- Is the whole protein there? Any missing (sub)domains?
- Are you working with the correct sequence?
- Are there any "extras", e.g. signalling peptides or expression tags?
- Are there any homologues?

## X-ray related checks

- Is the resolution high enough?
- Are the R and  $R_{\text{free}}$  factors low?
- What are the B-factors like?
- What about the RSCC values?
- Are there any geometric outliers or clashes?
- What were the experimental conditions like?
- Are there any parts of the structure that don't fit the electron density well?
- Any alternate conformations of protein residues or ligands?
- Were some parts of the protein modelled in (0-occupancy atoms)?
- What about the missing atoms/residues?
- Could the crystal contacts have caused some artifacts?

## Preparation-related checks

- Are the bond orders assigned to HET groups correct?
- Are the protonation/tautomer states of HET groups reasonable?
- Any titratable residues of interest?
- Does the hydrogen bond network (incl waters) make sense?
- Have residues/loops been rebuilt correctly?
- Have non-standard residues or PTMs been treated correctly?
- Is the resulting structure stable?





**Schrödinger**

**See you after  
lunch!**