Protein Bioinformatics

Yazhini Hong Su Michel van Kempen
Alexandra Kolodyazhnaya Amirhossein Hajialiasgary Najafabadi
Johannes Söding

November 01, 2023

Contents

Protein structure prediction

Proteins are molecular machines that carry out almost all cellular functions in cells. Proteins perform their function with the help of a 3D structure that is determined by their amino acid sequences. Protein structure modeling is the process of predicting the 3-D structure of a protein from its amino acid sequence.

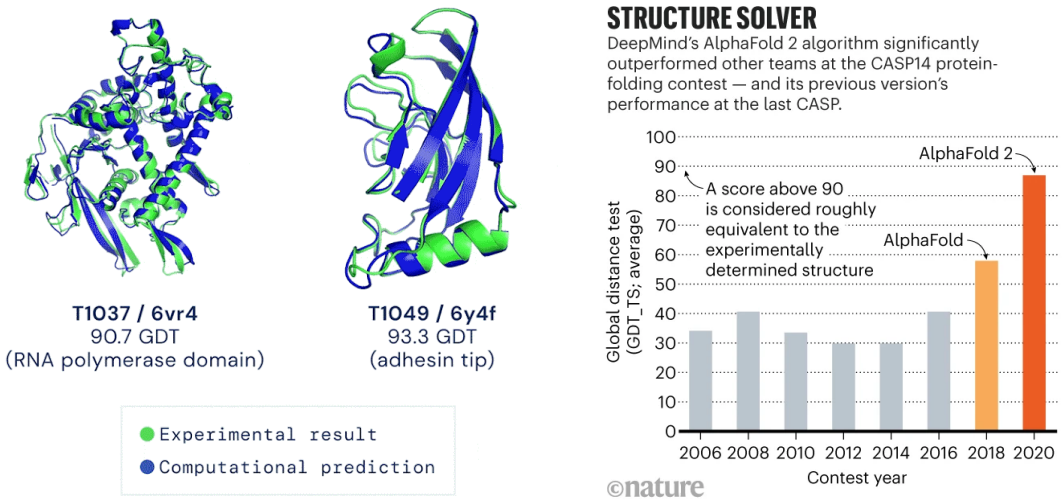
In this section you will learn how to:

1. Predict the 3-D structure of a single-chain protein with ColabFold.
2. Predict the 3-D structure of a two-chain protein complex with ColabFold.
3. Assess the quality of predicted structures.

Have fun!

1.1 AlphaFold: AI-based protein structure prediction tool

AlphaFold is an artificial intelligence (AI) system developed by DeepMind that predicts a protein's 3D structure from its amino acid sequence. It emerged as the top performer at CASP13 in 2018, and its successor, AlphaFold2, continued this success at CASP14 in 2020, consistently delivering accuracy that rivals experimental methods.



(a) (b)

1.2 Prediction of a single-chain protein structure using ColabFold

In this section, we will work with a single-chain protein (UniProt id: I1EYW3) sequence from the *Amphimedon queenslandica (Sponge)* organism.

ColabFold:

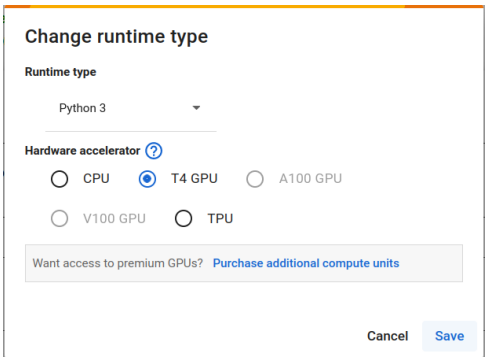
ColabFold is an easy-to-use, Google Colab-based implementation of the AlphaFold2 structure prediction suite. ColabFold [1] makes use of both to offer a simple, user-friendly, and fast tool to predict 3-D structures of proteins. Google Colab offers free CPU and, importantly, free GPU resources for running Jupyter Notebooks.

Tips for Colab:
- You can show/hide the code with **View** → **Show/hide code**, or click on the ▷ button left from the code cell.

1. Open the ColabFold Notebook[1] in Google Colab and sign in with your Google account. The usage of Google Colab is free but requires a Google account.

[1] https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

2. A GPU is required for the structure prediction, configure the notebook to use a GPU: **Runtime → Change runtime type**



3. Step-by-step instructions on how to run AlphaFold using ColabFold:

   - First paste the sequence into the field `query_sequence` and type a jobname. You can give any job name as you prefer. We use "test" here.

     ```
     >tr|I1EYW3|I1EYW3_AMPQE 40S ribosomal protein S12 OS=Amphimedon
     MAAGDDSSQGMKLKEAMKEVLKESLKHDGLARGLREAVKALDKRQAYLCIVAKNCSEAGY
     LRLVEALCKEHQISLLKVEDKEELGEWVGLCKIDKDGKPRKIVKCSCVVVKDIGTDTEAW
     STVQEYIKTQTAAAV
     ```

   - Then select the `num_relax`. If you want to use Amber force fields to "relax" the predicted structure, you can enable this option.

   - Then select the `template_mode`.

     After these three basic steps, you can submit your job by hitting **Runtime → Run all** following the default settings for the remaining configuration. Alternatively, you have the option to proceed and customize the remaining settings.

   - Then select `msa_mode`. This option allows you to specify which MSA database to create the MSA.

   - Then select `pair_mode`. This option controls MSA pairing.

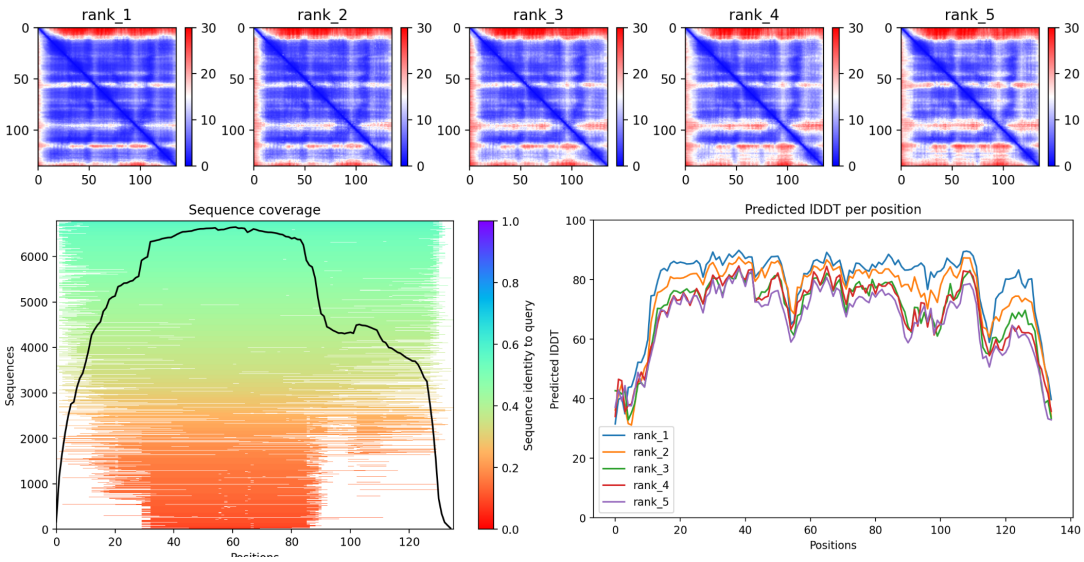
   - Then select `model_type`.

   - Then select `num_recycles`.

     By default, AlphaFold2 predicts five different structures.

   - Hit **Runtime → Run all** to start the prediction (This will take a few minutes...).

4. The prediction results can be visualized with the plots below. The five predicted models are ranked by confidence from high (rank 1) to low (rank 5).



? How to judge an AlphaFold2 model?

? How reliable is your AlphaFold2 model?

? How good is the input MSA?

5. Check the predicted 3-D structure (rank 1). Have fun playing with the cartoon view (ribbon representation).

Note: Further instructions for how to use ColabFold, descriptions about the results, and acknowledgments can be found at the bottom of the Colab page.

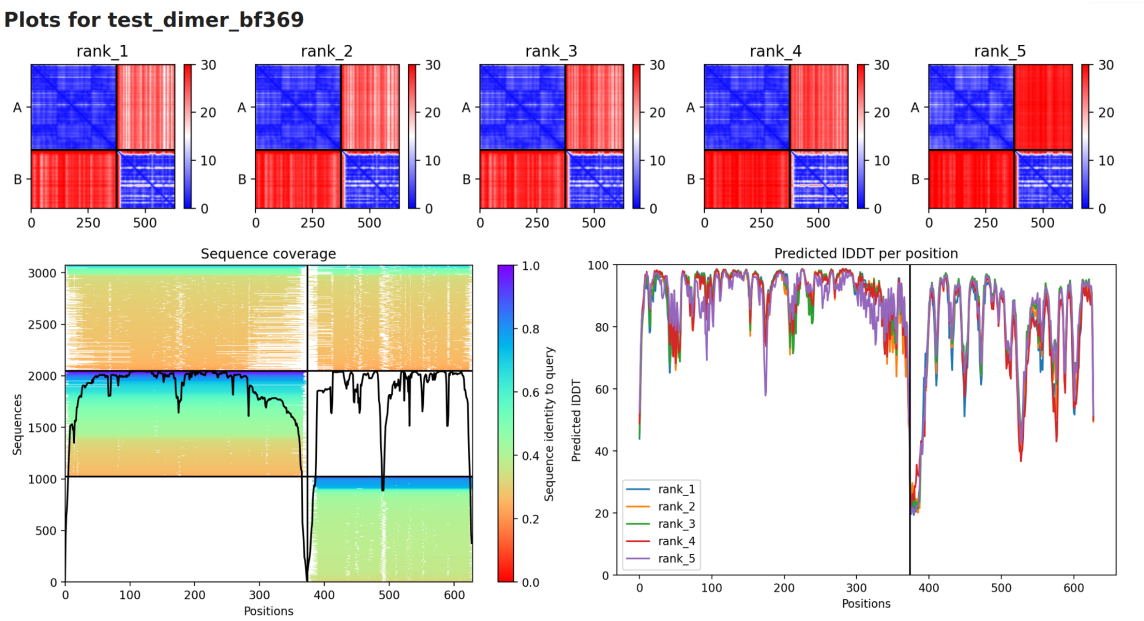
1.3 Protein complex prediction with ColabFold

AlphaFold Multimer is an extension of AlphaFold2 that has been specifically built to predict protein-protein complexes. Here, we use ColabFold to predict the structure of a two-chain protein complex (PDB id: 6QF7).

Since the prediction will take a lot of time, we have provided the prediction results `Tutorialsession1_dimer_input.txt` and `Tutorialsession1_dimer.results.zip` for download. We run ColabFold with default parameters for this job.

Tips: A fasta protein sequence file containing your multiple sequences is required. Since this is a multimer, please include all sequences you would like to fold together. You need to delimit different chains with the `:` character.

Prediction results:

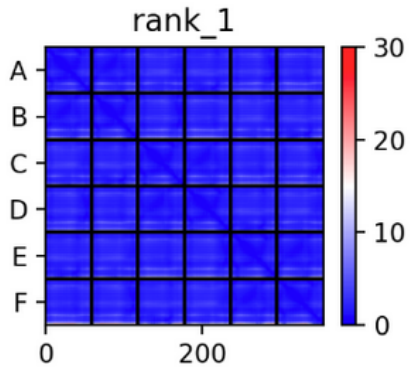


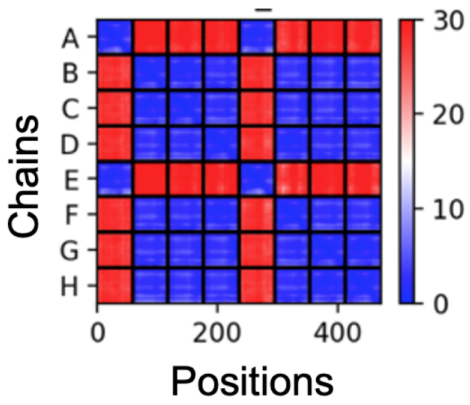
? What does the pLDDT tell us?

? Do high pLDDT values within all domains mean that AlphaFold is confident in their relative positions?

? What does the PAE plot here tell us?

? Is the PAE plot symmetrical? Why?




Chains
A
B
C
D
E
F
G
H
0
200
400
Positions
30
20
10
0

What is pLDDT?

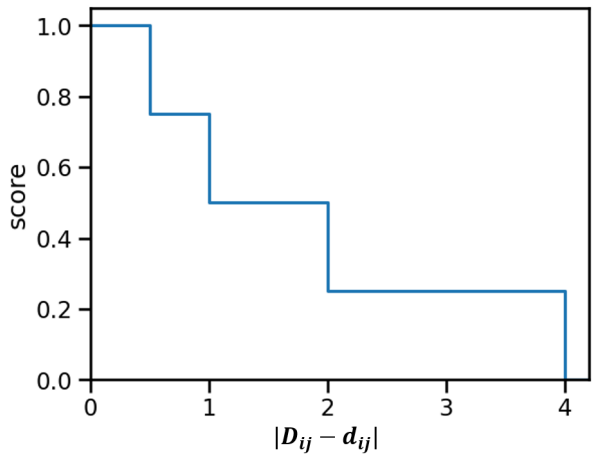
pLDDT is AlphaFold2 per-residue prediction of its lDDT-Cα scores.

The lDDT-Cα is calculated as follows:

$$\text{lDDT} = \frac{100}{L} \sum_{i=1}^{L} \frac{1}{N_i} \sum_{j,|i-j|\geq r, D_{ij}<15} \text{step\_function}(|D_{ij} - d_{ij}|) \tag{1.1}$$

$$\text{step\_function}(x) = 0.25(\mathbb{1}_{x<0.5} + \mathbb{1}_{x<1.0} + \mathbb{1}_{x<2.0} + \mathbb{1}_{x<4.0}) \tag{1.2}$$

$$N_i = \sum_{j,|i-j|\geq r, D_{ij}<15} 1 \tag{1.3}$$



where D_{ij} denotes the distance at Cα atoms between amino acid residues i and j within the ground truth structure, d_{ij} denotes the distance between pairs of these residues within the predicted structure, L is the length of the sequence, the filtering condition for atom pairs is $D_{ij} < 15$Å and $|i-j| \geq r$, r is a minimum sequence separation parameter, and t is the tolerance threshold. The final LDDT score is the average of four fractions computed using the thresholds 0.5Å, 1Å, 2Å and 4Å.

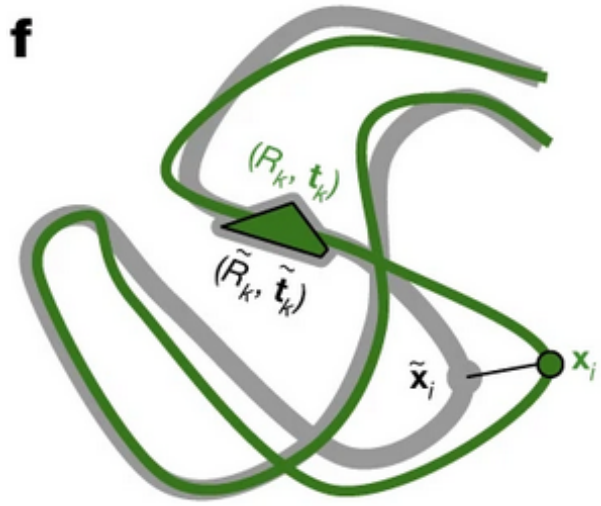
lDDT is a metric ranging from 0 to 100. Roughly, lDDT measures the percentage of correctly predicted inter-atomic distances. It rewards locally correct structures, and getting individual domains right.

pLDDT behaves similarly, as a measure of local confidence. It ranges from 0 to 100 (100 is most confident).

Note: In the training step, pLDDT is calculated based on the ground truth structure. For evaluation, pLDDT is predicted with neural networks.

What is the PAE?

Predicted Aligned Error (PAE) is AlphaFold's prediction of its position error at residue x if the predicted and the true structures were aligned on residue y.



For each alignment, defined by aligning the predicted frame ($(R_k,\ t_k)$; green) to the corresponding true frame (grey), AlphaFold2 computes the distance of all predicted atom positions x_i from the true atom positions.

The PAE aims to measure confidence in the relative positions of pairs of residues. Mainly used to assess relative domain positions, but applicable whenever pairwise confidence is relevant.

The PAE is displayed as a 2D plot.

Protein structure search

In this section, we will work with an uncharacterized protein structure (predicted with ColabFold) from the freshwater demosponge *Spongilla lacustris* [2]. You will learn how to:

- Find similar protein structures with Foldseek.
- Utilizing the protein structure repositories Protein Data Bank (PDB), AlphaFold Database, and AlphaFold Clusters.

2.1 Remote homology detection using Foldseek

Our goal is to explore our protein of interest by seeking its homologous counterparts. We can achieve this through sequence or structure searches. While sequence searches are standard, structure searches excel at uncovering distant homologies, when protein sequences have diverged. As our sponge protein has limited matches in sequence databases, we'll employ a structure search in the AlphaFold database.

? Why are sequence-based methods not sufficient to annotate all proteins?

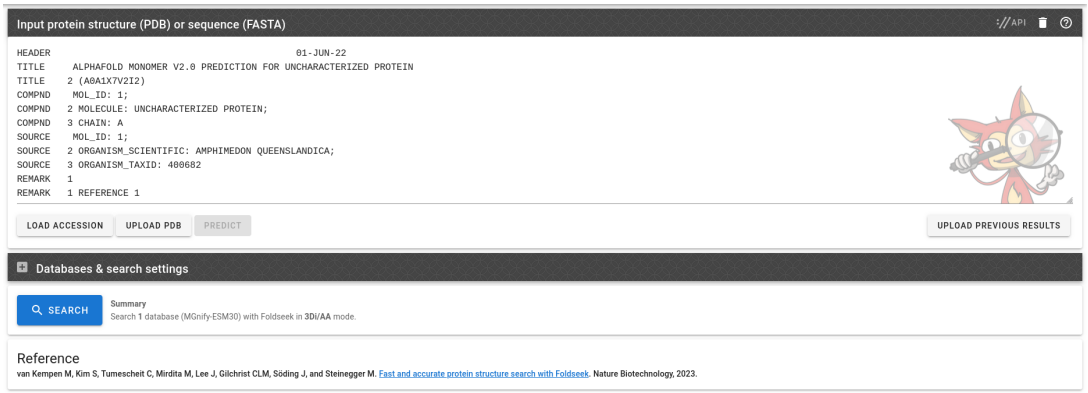
As the protein structure determines its function, and as the structure can also be better conserved than its sequence, the idea is to search with the protein structure instead of its sequence. Your task is to discover structurally similar proteins with annotations that may help to gain insights into our protein.

Foldseek
"Foldseek enables fast and sensitive comparisons of large structure sets. It reaches sensitivities similar to state-of-the-art structural aligners while being at least 20,000 times faster. To facilitate access to Foldseek, we developed a user-friendly webserver optimized to quickly return results for single queries." [3]

Step-by-step instructions on how to use Foldseek:

1. **Upload Your Predicted Structure:**
   Begin by taking the provided structure obtained from `https://wwwuser.gwdg.de/~compbiol/molbio_course/2023/sponge_protein.pdb` and upload it to Foldseek search.foldseek.com).

2. **Set Databases & Search Settings:**
   By default, Foldseek searches through all available databases. However, you can refine your search by selecting specific databases or taxonomic filters. In this context, limit your search to the `PDB` (containing all experimentally solved structures) and `AlphaFold/UniProt50` (the largest AlphaFold database with structures for all sequences clustered by 50% sequence identity).



3. **Initiate the Search:**
   Start the search by clicking the 

(**SEARCH**) button. Please note that this process takes a couple of seconds.

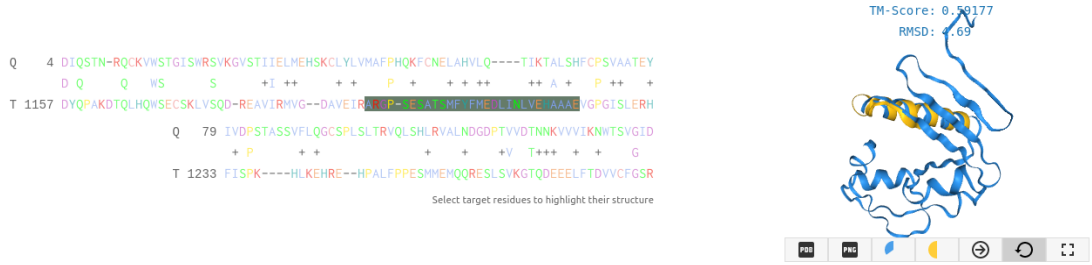
4. **Examine the Search Results:**
   The results page displays matches to the different databases, sorted by structural similarity. Hits are ranked based on an alignment score, which considers the sequence alignment, TM-score, and LDDT score. For each match in a database the page provides the probability of a match being homologous (`Prob.`), an E-value describing the expected number of matches by pure chance at the given database size and alignment score, and the sequence identity, which is the fraction of identical amino acids in the alignment. Notably, finding matches with low sequence identities (below 20%) would be challenging with standard sequence alignment methods [4].

AF-Q5F5L9-F1-model_v4	RNA polymerase-binding transcri...	Neisseria gonorrhoeae FA 1090	1.00	41.4	6.76e-2	1–41
AF-Q93GN5-F1-model_v4	Conjugative transfer	Salmonella enterica subsp. enter...	0.99	30	3.54e-1	1–40
AF-Q9HVK7-F1-model_v4	DksA C4-type domain-containing...	Pseudomonas aeruginosa PAO1	0.98	42.4	4.44e-1	1–33
AF-A0A0H3GKN8-F1-model_v4	Phage/conjugal plasmid C-4 type...	Klebsiella pneumoniae subsp. pn...	0.94	35.8	9.41e-1	1–39
AF-P41039-F1-model_v4	Uncharacterized protein YbiI	Escherichia coli K-12	0.93	37.1	7.51e-1	1–35
AF-Q8ZQN5-F1-model_v4	Putative DnaK suppressor protein	Salmonella enterica subsp. enter...	0.91	34.1	7.51e-1	1–41
AF-P44221-F1-model_v4	Uncharacterized protein HI_1497	Haemophilus influenzae Rd KW20	0.87	37.5	1.72e+0	2–33
AF-O32I79-F1-model_v4	DksA C4-type domain-containing...	Shigella dysenteriae Sd197	0.85	37.1	1.18e+0	1–35

? How would the E-Value change for a match when the database size is reduced by half?

5. **Analyze the Alignment Visualization:**
   Click the button, to view the alignment of a match. Examine the hit through the 3-D viewer, and assess the structural similarity. The Root Mean Square Deviation

(RMSD) describes the deviation between the aligned parts in the superposition of the two structures. The TM-score, similarly, is based on the superposition, but goes on step further and calculates a structural similarity score between 0 and 1, where a TM-score of 0.5 marks the threshold of homology.



? Are the RMSD or the TM-score reliable indicators of structural homology, or where do they fail (look for an example)?

6. **Access Database Entries:**
   Each hit links to its corresponding database entry.

2.2 Protein structure databases

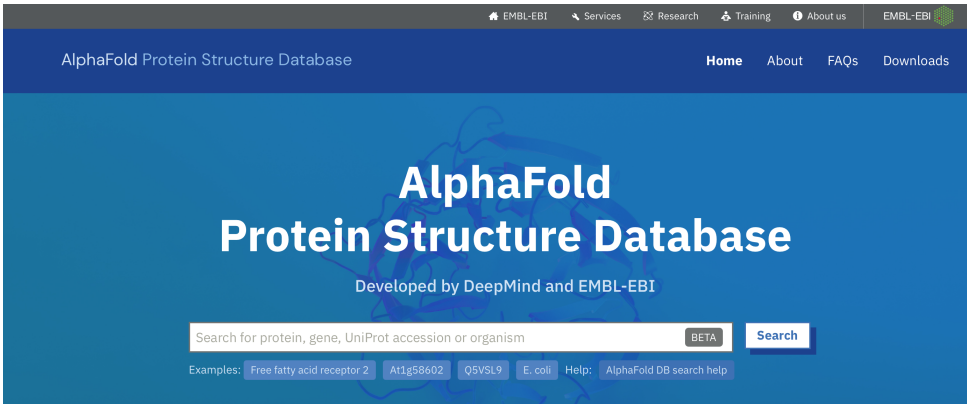
2.2.1 Protein Data Bank (PDB)

The Protein Data Bank (PDB) is a repository exclusively for experimental protein structures, housing around 0.18 million structures. Up until 2021, it used to be the largest protein structure database. Notably, nearly all experiments have been performed with its proteins and therefore most literature resource are linked to the PDB.

? Why should we consider the matches for our protein in the PDB as potentially unreliable, leading to concerns about the transferability of their annotations?

2.2.2 AlphaFold Database (AFDB)

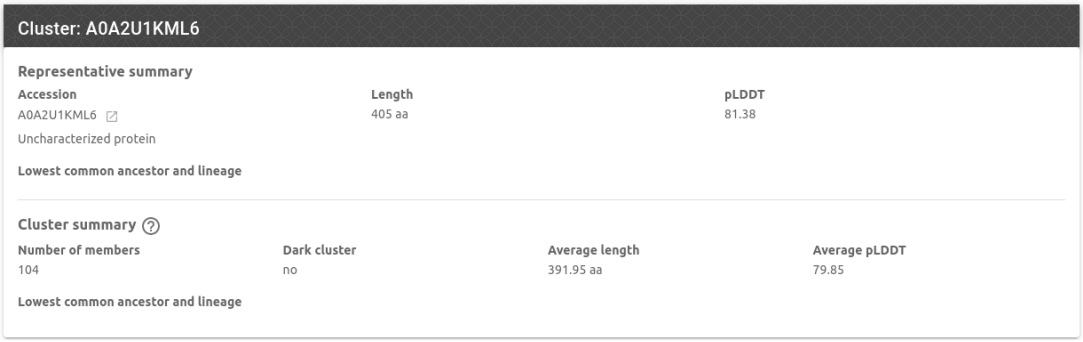
EMBL-EBI and DeepMind have together developed a database for protein structure models predicted by AlphaFold (`https://alphafold.ebi.ac.uk`). Currently, it has the 3-D models for the complete human proteome and 47 other reference organisms such as *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Danio rerio*, and *Rattus norvegicus*. It also contains predictions for most UniProt sequences, resulting in more than 200 million entries. You can retrieve predicted protein 3-D structures using keywords such as protein name, Gene ID, Source Organism, and UniProt ID.



? Do the matches for our protein in the AFDB exhibit higher quality compared to the PDB?

2.2.3 AlphaFold Database clusters

Barrio-Hernandez et al. [5] utilized Foldseek to cluster the complete AlphaFold Database into 2.27 million clusters based on structural similarity. These clusters are available for exploration at `https://cluster.foldseek.com`. Users can simply input their protein of interest to access information about its structural neighbors and potentially uncover new insights into its biological role.



? Identify a suitable match for our protein within the AFDB and search with its UniProt ID in the cluster database. Does this provide us with additional information?

Protein structure analysis

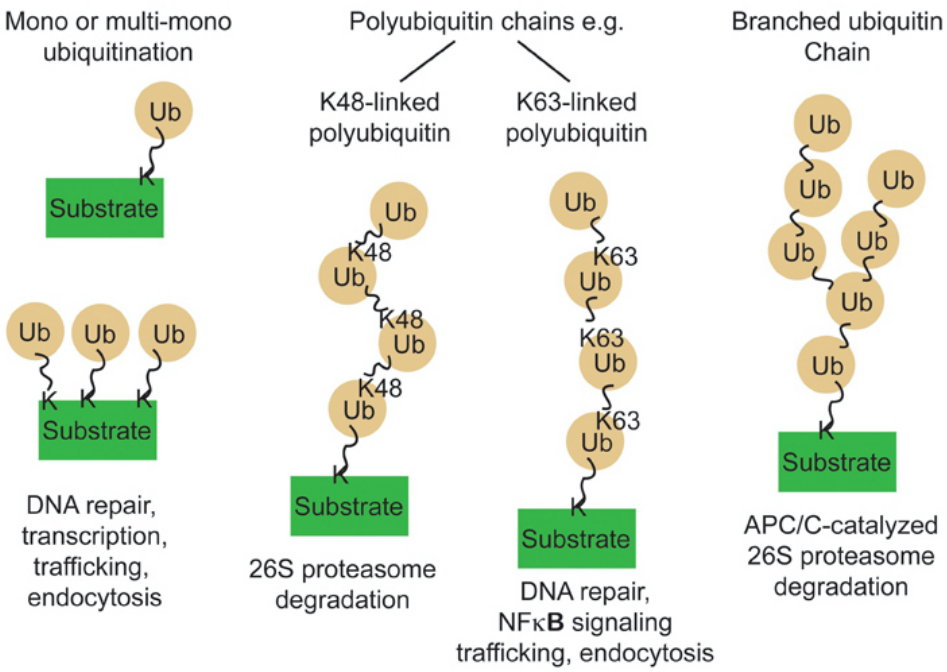
The protein function is mediated at a specific location in the 3D structure (e.g. active site, interface region). Due to functional constraints, these regions are more conserved in evolution than other parts of the protein. In this section, we will analyse protein structure, focusing on functional regions and evolutionary conservation patterns.

All files required for this session are provided in `Tutorialsession3_files.zip`.

Let's look at two proteins from *Amphimedon queenslandica* (Sponge) that mediate protein degradation.

Protein 1: Polyubiquitin-B (Ub, UniProt ID: A0A1X7V2I2). It binds to proteins to be targeted for degradation.

Protein 2: E3 ubiquitin-protein ligase NEDD4-like (UniProt ID:A0A1X7UV05). It interacts with Polyubiquitin-B to facilitate the ubiquitination of target proteins during the E1-E2-E3 Ub conjugation cascade.



Ubiquitin conjugation cascade system [6].

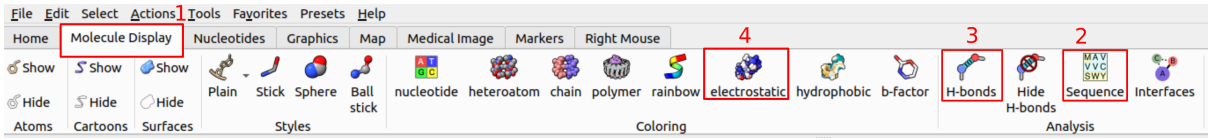
Download AlphaFold predicted 3D structures for both proteins. *(Alternatively, find them*

in the MolBioTutorial_session3 folder shared with you Protein1_AF-A0A1X7V2I2-F1-model_v4_polyUbiquitinB.pdb and Protein2_AF-A0A1X7UV05-F1-model_v4_E3ligase.pdb)

3.1 Visualization

To view and analyze structures, we will use ChimeraX. (freely downloadable from here or here for MacOS)

Explore options in *Molecular Display* panel to understand the topology of the Polyubiquitin-B structure.



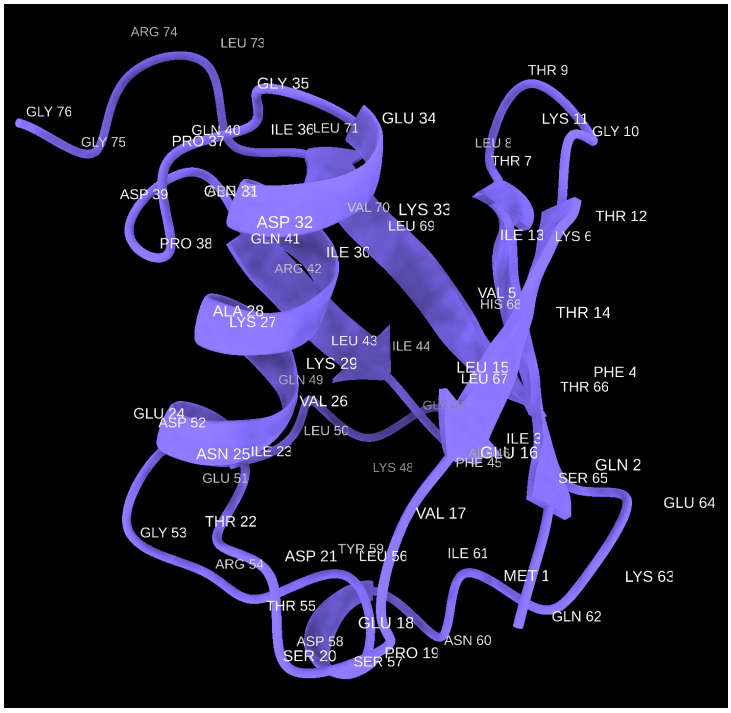
- **How many structural domains do you see in Polyubiquitin-B and what is the length (in aa) of a single domain?**
- **How many H-bonds are formed?**
- **Which surface region(s) is/are negatively charged?**

Hint: red - negatively charged; blue - positively charged

We will focus on single-domain for in-depth analysis. Select a single domain (range 1-76aa) using the sequence panel or use command **select /A:1-76** and save it as a separate PDB file. **Home -> Save -> enable "Save selected atoms only"**.

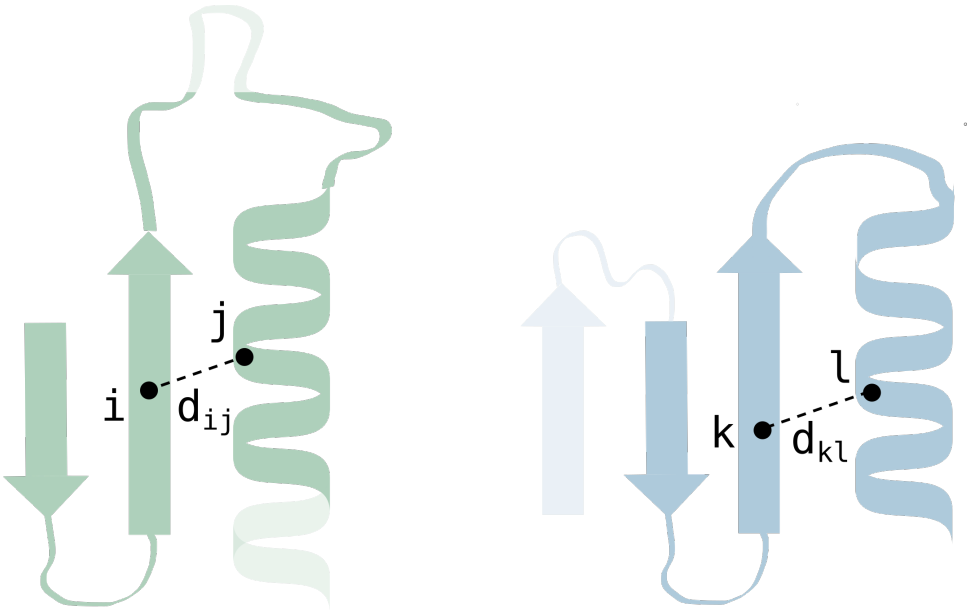
3.2 Residue-level examination

Open a single-domain file that was saved in the above step (or use the shared Polyubiquitin-B_singledomain.pdb) in ChimeraX. You can view residue type and numbering on the 3D structure. **Actions -> Label -> Residues -> Name and Number**.



Inter-residue interactions: Interaction between residues governs protein folding and the formation of 3D structure. As interacting residues form geometrical constraints for 3D structure, they either evolve at a slow rate (well-conserved) or co-evolve to preserve their interactions.

Proteins with similar 3D structures have similar inter-atomic interaction patterns. This principle is applied in Foldseek for structure similarity search.



Let's examine inter-atomic interactions in Polyubiquitin-B domain. Set the view **Molecule Display -> Atoms (Show)** to display atoms as *Stick* or *Ball stick* representation. Next, use **Tools -> Structure Analysis -> Distances** panel or **distance** function in the command line. To select an atom, press **Ctrl** + **Shift** + **Enter**. (command line: eg.,

distance /A:510@NZ /A:424@CB)

Exercise:

Obtain inter-atomic distance between the following residue pairs

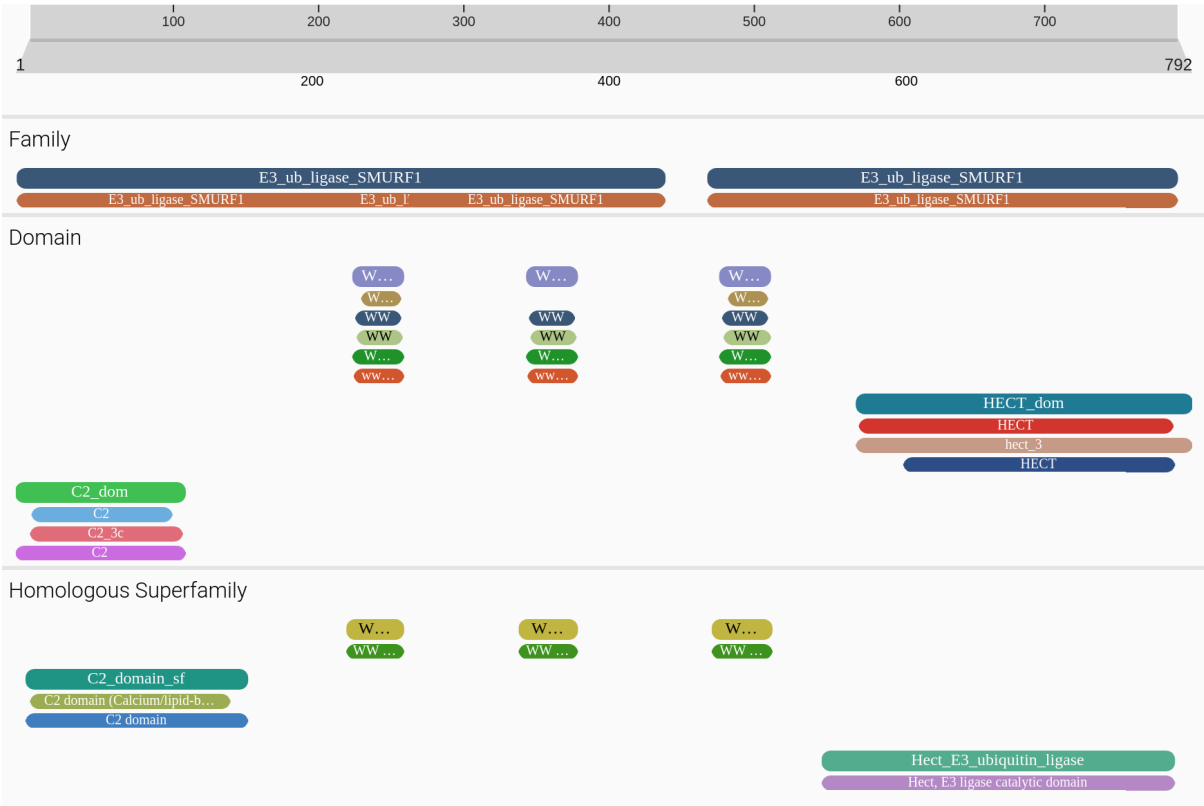
1. GLU 18 (OE1) - LYS 33 (NZ) (residue_type position atom_type)
2. ILE 36 (CD1) - LEU 71 (CD2)
3. LYS 27 (NZ) - ASP 52 (OD2)

? Which residue pair forms a hydrogen bond?

Distance < 4Å

3.3 Conservation

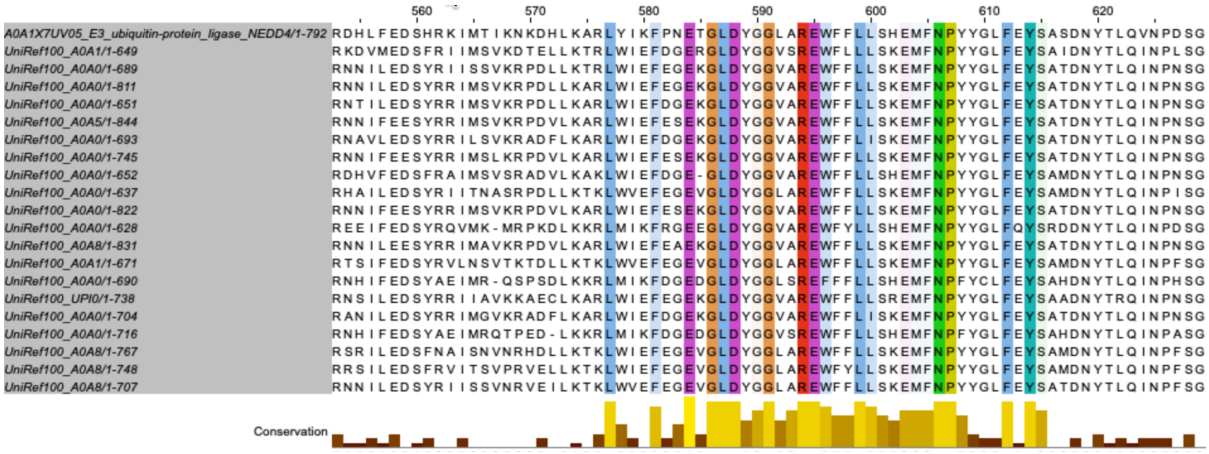
The rate of amino acid evolution varies within proteins. Let's look at the E3 ubiquitin-protein ligase NEDD4-like. This protein has two E3 ubiquitin-protein ligase, SMURF1 type families. This region accepts ubiquitin from an E2 ubiquitin-conjugating enzyme and subsequently transfers the ubiquitin to targeted substrate proteins (IPR024928). Also, the protein comprises WW domains (IPR001202), C2 domain (IPR035892) and HECT domain (IPR000569). Together, they help in ligase function and are better conserved than other regions in the protein.



? How do you find conservation of residues in protein?

Multiple sequence/structure alignment

Launch Jalview with *MSA_E3ligaseNEDD4like.fasta* file to view multiple sequence alignment of E3 ubiquitin-protein ligase NEDD4-like. `https://www.jalview.org/jalview-js/JalviewJS/`



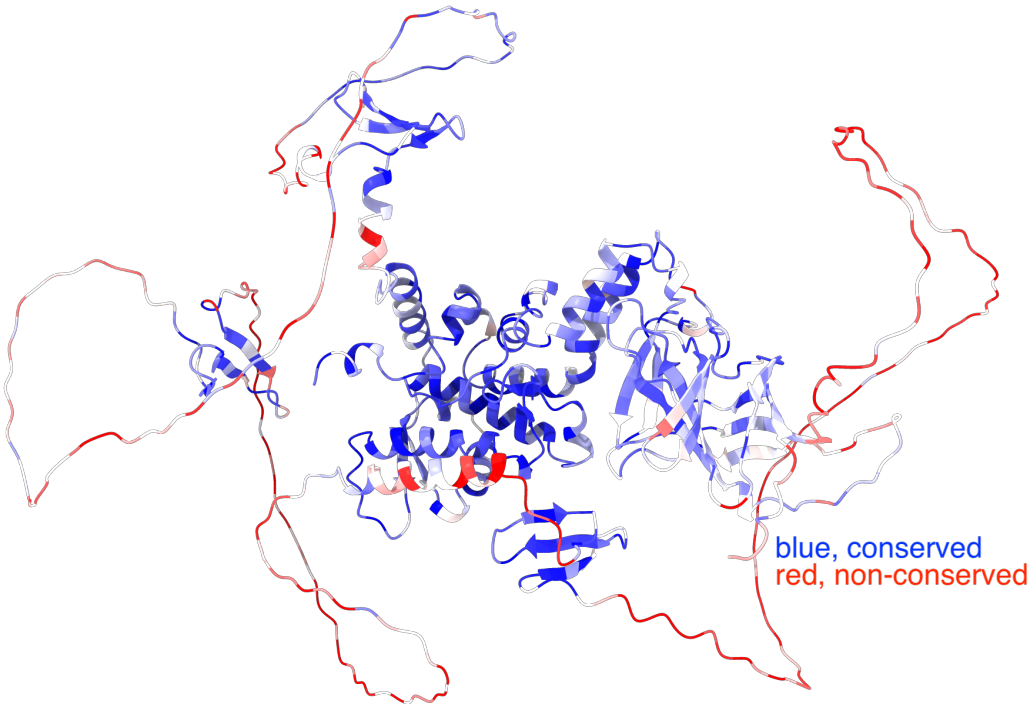
In Jalview output, **Conservation** indicates physicochemical properties conserved at a given position.

Consensus indicates the most frequent amino acid at the position.

Quality score is an ad-hoc measure for the likelihood of observing mutation at the position. A high score indicates no mutation or observed mutations are favourable.

Occupancy indicates how many sequences have an aligned amino acid at the position.

Conservation of key functional residues that mediate ubiquitination: Open Protein2_AF-A0A1X7UV05-F1-model_v4_E3ligase.pdb in chimeraX and set view by **Molecule Display -> Coloring (by b-factor)** to colour residues by their residue conservation.



In this protein, there are 7 key residues that perform ubiquitination by forming contacts with E2 ligase. Some of them are VAL689, TYR691, ILE692, LEU700 and TYR736. By looking at colouring by conservation, you see they are well conserved in evolution.

Find out two more functional residues in the following sets by examining the conservation score as coloured.

```
Set 1: THR683, PHE685, LYS690
Set 2: ASP727, ARG738, LEU740
```

? Which one in each set is a functional residue?

Set 1: PHE685
Set 2: LEU740

3.4 Analysis of interface regions

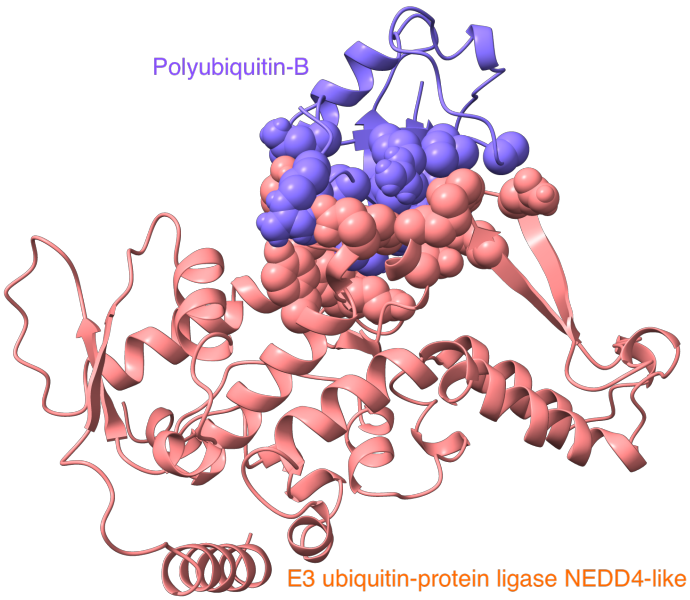
In this part, we will focus specifically on regions that form physical interactions between Polyubiquitin-B and E3 ubiquitin-protein ligase NEDD4-like proteins.

To analyse the interface, we need a protein-protein complex structure. As of now, we only have separate AlphaFold predicted structures for these proteins.

? Any thoughts on how to get protein-protein complex structure?

Protein docking

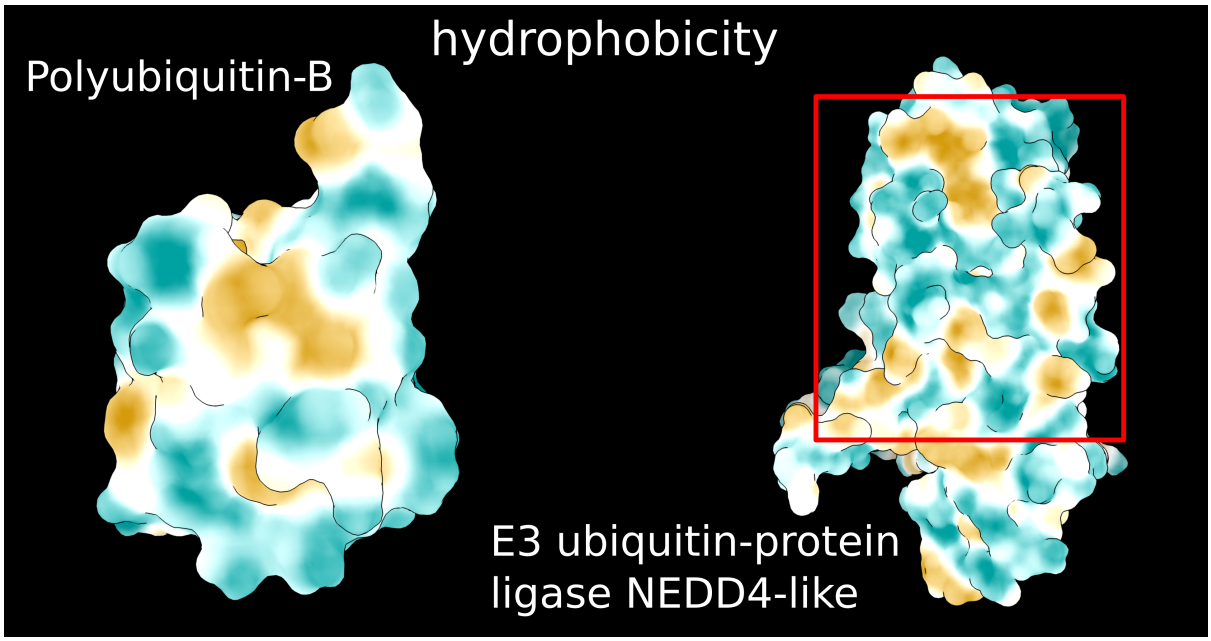
Use pre-generated complex structure named *E3ligase_polyubiquitinB_complex.pdb* given to you in `Tutorialsession3_files` folder. For simplicity, a truncated version covering the C-terminal region (521-792) of E3 ubiquitin-protein ligase NEDD-like protein is used as it is the region that interacts with Polyubiquitin-B.



? How many interface residues are at the interface of two proteins?

Hint: Explore **Molecule Display -> Interfaces**.

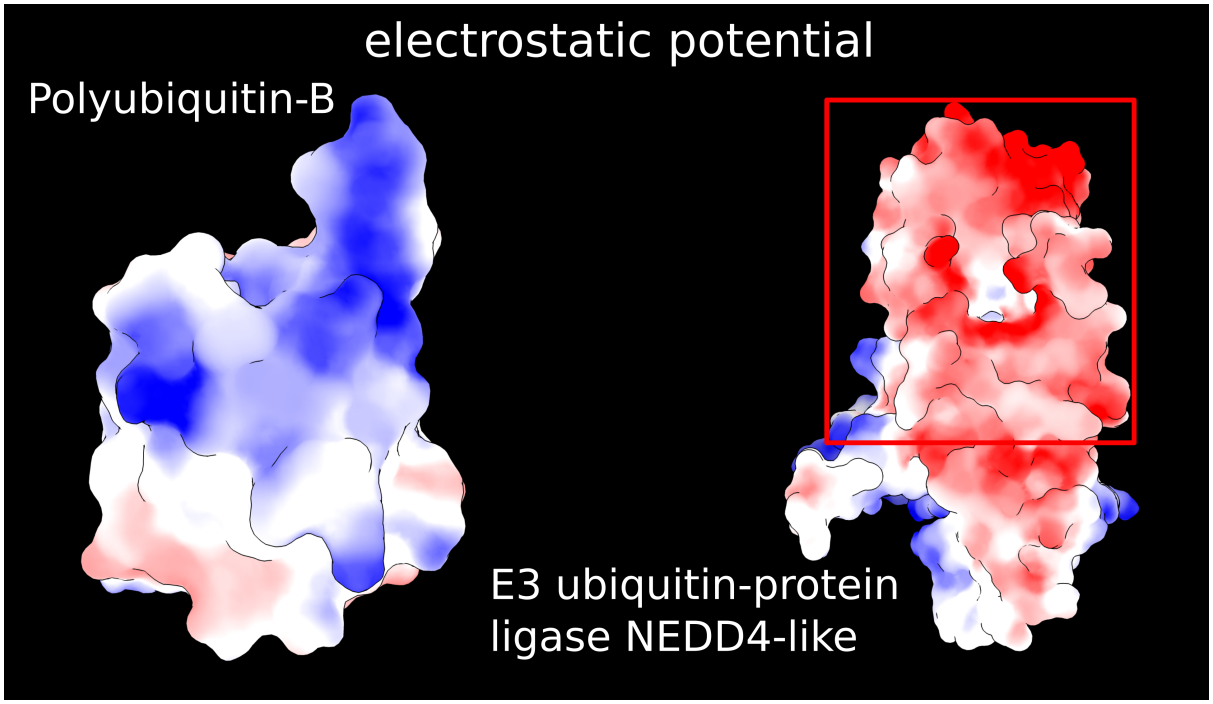
In general, the interface region comprises hydrophobic patches and has opposite-charged residues in complementary positions.



cyan - hydrophilic; goldenrod - hydrophobic

Red box locates the interface region within E3 ubiquitin-protein ligase NEDD4-like protein.

? Whose interface region is more negatively charged among two proteins?



red - negatively charged; blue - positively charged

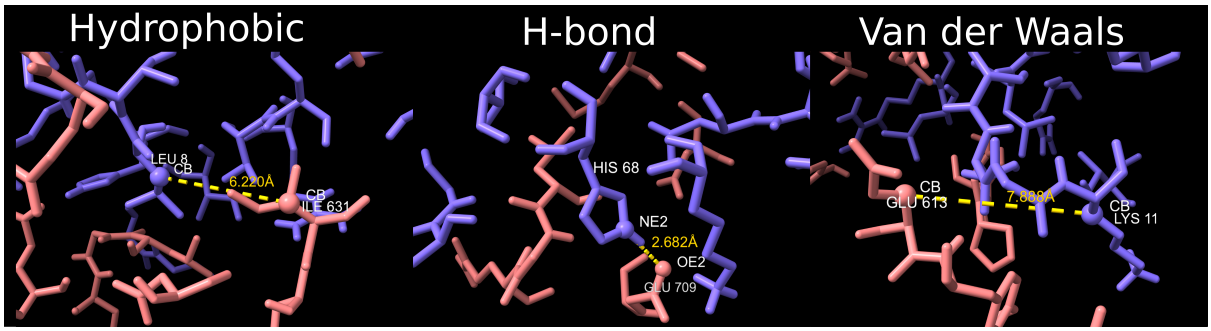
? Can you guess another feature of the interface?

Hint: explore 3D structure

Shape

Different interaction types can be formed at the interface (**Optional**).

For convenience, select only interface residues and save them as a separate .pdb file (or use interface_residues.pdb). You will see that different interacting residue pairs form different types of interactions.



Exercise:

Find a couple of more interacting residue pairs that form hydrophobic (distance <7Å) or H-bond (distance <4Å).

Hydrophobic:
ILE 36 - A (CB) TYR 608 - B (CB)
ILE 44 - A (CB) LEU 711 - B (CB)
VAL 70 - A (CB) ILE 631 - B (CB)
LEU 71 - A (CB) TYR 608 - B (CB)
LEU 73 - A (CB) TYR 609 - B (CB)

Salt bridge:
LYS 6 - A (NZ) GLU 709 - B (OE1))
ARG 74 - A (NH1) GLU 633 - B (OE1)

Appendix

4.1 Letter codes for amino acids in a protein chain

A	Alanine	Ala
C	Cysteine	Cys
D	Aspartic Acid	Asp
E	Glutamic Acid	Glu
F	Phenylalanine	Phe
G	Glycine	Gly
H	Histidine	His
I	Isoleucine	Ile
K	Lysine	Lys
L	Leucine	Leu
M	Methionine	Met
N	Asparagine	Asn
P	Proline	Pro
Q	Glutamine	Gln
R	Arginine	Arg
S	Serine	Ser
T	Threonine	Thr
V	Valine	Val
W	Tryptophan	Trp
Y	Tyrosine	Tyr

Bibliography

[1] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, Jun 2022.

[2] Fabian Ruperti, Nikolaos Papadopoulos, Jacob Musser, and Detlev Arendt. Beyond sequence similarity: cross-phyla protein annotation by structural prediction and alignment. *bioRxiv*, 2022.

[3] Michel van Kempen, Stephanie Kim, Charlotte Tumescheit, Milot Mirdita, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *bioRxiv*, 2022.

[4] Burkhard Rost. Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, 12(2):85–94, 02 1999.

[5] Inigo Barrio-Hernandez, Jingi Yeo, Jürgen Jänes, Milot Mirdita, Cameron L M Gilchrist, Tanita Wein, Mihaly Varadi, Sameer Velankar, Pedro Beltrao, and Martin Steinegger. Clustering predicted structures at the scale of the known protein universe. *Nature*, 622(7983):637–645, October 2023.

[6] Lori Buetow and Danny T Huang. Structural insights into the catalysis and regulation of e3 ubiquitin ligases. *Nature reviews Molecular cell biology*, 17(10):626–642, 2016.