مركز خليفة للتقانات الحيوية والهندسة الوراثية
Khalifa Center for Genetic Engineering & Biotechnology

AI-guided structural modelling workshop
Genomes to structures

Yu Sugihara, Andrés Posbeyikian, James Canham, Sophien Kamoun

6th of May, 2025

Session 1. Running AlphaFold 3 on server

Objectives: To learn how to run AF3 on the server, modify modelling parameters (*i.e.* stoichiometries). Learn to interpret confidence metrics, and download modelling results.

Exercise 1.1. Modelling individual structures of proteins

In this exercise, you will model the structure of the partial sequence of the rice immune receptor protein, Pikp-1.

To begin, navigate to https://alphafoldserver.com/, create an account, and submit your first job:

>Pikp-1_186-486
GLKQKIVIKVAMEGNNCRSKAMALVASTGGVDSVALVGDLRDKIEVVGYGIDPIKLISALRK
KVGDAELLQVSQANKDVKETTPMLAPVKSICEFHKVKTVCILGLPGGGKTTVARELYDALG
THFPCRVFVSVSPSSSPSPNLTKTLADIFAQAQLGVTDTLSTSYGGSGTGRALQQHLIDNIS
AFLLNKKYLIVIDDIWHWEEWEVIRKSIPKNDLGGRIIMTTRLNSIAEKCHTDDNDVFVYEVG
DLDNNDALSLSWGIATKSGAGNRIGTGEDNPCYDIVNMCYGMPLALIWLSSAL

a) What does **pTM** stand for? What is the **pTM** score for the prediction?

b) What can you say about the *local* measure of confidence (**pLDDT**)? Is it high all over the model? Are there any amino acids predicted with higher confidence than others?

c) What can you say about the predicted aligned error (**PAE**) plot? Are there any regions where the relative positions are unreliable?

Exercise 1.2. Modelling a multimeric complex

In this exercise, you will explore the oligomeric state predictions for a given protein. Now, you will model the **HMA domain of Pikp-1** protein used in **Exercise 1.1** in dimeric (n=2), trimeric (n=3), and tetrameric (n=4) states.

Your task is to evaluate the confidence in the relative orientation of the modeled polypeptide chains. For this, you will need to interpret the **PAE plots** of the predictions. You will also complete the table below with predicted **pTM** and **ipTM** scores, and based on this assessment, determine the likely oligomeric state of the protein (monomer, dimer, trimer, or tetramer).

>Pikp-HMA
GLKQKIVIKVAMEGNNCRSKAMALVASTGGVDSVALVGDLRDKIEVVGYGIDPIKLISALRK
KVGDAELLQVSQANKD

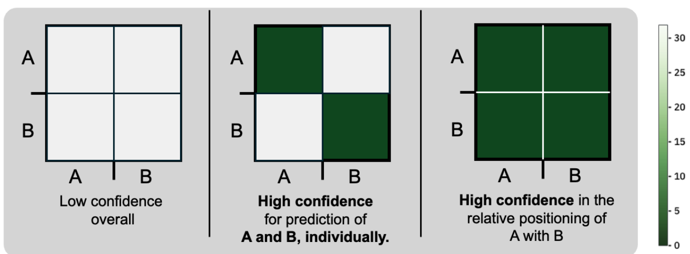
a) What does **ipTM** stand for?

b) Complete the table below with your modelling results:

Oligomeric state	ipTM	pTM
Monomer (n=1)	-	
Dimer (n=2)		
Trimer (n=3)		
Tetramer (n=4)		

c) What does the predicted alignment error (**PAE)** look like for the different oligomeric states? Is there high confidence in the relative positioning of the protomers, across all cases tested?

Hint: Remember what we learned about predicted alignment error interpretation.



Exercise 1.3. Protein-Protein interaction prediction

Next, let's model three different heteromeric complexes between **pathogen proteins** and the rice **Pikp-1** domains. Your goal here is to analyze the AF3 models and make a hypothesis about which effector may interact with the plant host protein, based on the modelling confidence metrics alone.

>Pikp-1_186-486
GLKQKIVIKVAMEGNNCRSKAMALVASTGGVDSVALVGDLRDKIEVVGYGIDPIKLISALRK
KVGDAELLQVSQANKDVKETTPMLAPVKSICEFHKVKTVCILGLPGGGKTTVARELYDALG
THFPCRVFVSVSPSSSPSPNLTKTLADIFAQAQLGVTDTLSTSYGGSGTGRALQQHLIDNIS
AFLLNKKYLIVIDDIWHWEEWEVIRKSIPKNDLGGRIIMTTRLNSIAEKCHTDDNDVFVYEVG
DLDNNDALSLSWGIATKSGAGNRIGTGEDNPCYDIVNMCYGMPLALIWLSSAL

Below are the amino acid sequences of three pathogen proteins, secreted by the fungus *Pyricularia oryzae* (causal agent of the rice blast disease):

>Pathogen_protein_1 (AVR-PikD)
ETGNKYIEKRAIDLSRERDPNFFDHPGIPVPECFWFMFKNNVRQDAGTCYSSWKMDMKV
GPNWVHIKSDDNCNLSGDFPPGWIVLGKKRPGF

>Pathogen_protein_2 (AVR-Pib)
ATQVTILKKGERITWVEVPKGESREFNIRGKYFTVSVSDDGTPSISGSKYTVE

>Pathogen_protein_3 (AVR-Pii)
LPTPASLNGNTEVATISDVKLEARSDTTYHKCSKCGYGSDDSDAYFNHKCN

a) Based on these modelling results, which pathogen effector(s) is most likely to interact with the rice Pikp-1? What confidence metrics allow you to conclude this?

b) Based on the modelling results, which Pikp-1 domain contributes to binding to the pathogen protein?

c) What kind of experiments would you perform to test this hypothesis?

Session 2: Analysing AF3 outputs using ChimeraX

Objectives: The main goal of this session is to become familiarized with AF3 output analysis, using commands in the molecular visualization program **UCSF ChimeraX**.

Exercise 2.1. Introduction to ChimeraX basic commands.

Let's get started with a few basic commands. In this exercise you must open a protein model from the Protein Data Bank (PDB), make chain and residue selections, and color different regions. To begin, **run this command** in the ChimeraX console:

```
>open 7B1I
```

In the log, click on [more info…], to look at the metadata associated with 7B1I:

a) What's the **title** of the paper associated with 7B1I ?
b) What was the **experimental method** used for its resolution?

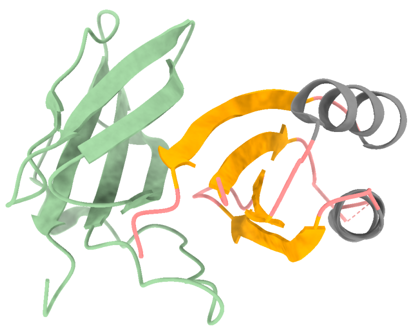
Navigate to the 'Molecule Display' panel, and click on the 'chain' icon .

Now click on the 'Sequence' icon .

c) Which chain in the model is longer, chain A or B?

Now, let's color all alpha-helices gray. To do this, in the sequence panel for chain B, **click on one of the yellow sequence rectangles** (corresponding to α-helices). This will select all α-helices on chain B. Now, in the top navigation panel, click on **Actions → Color → Gray**.

Now do the same for the β-sheets, and color them orange. Your model should now be looking like this:



Now, let's do something similar, using the ChimeraX console.

```
>sel #1/c
```

Note that the **#** denotes the model ID (we only have one open at the moment, so there is only model **'1'**), and the slash / is used to specify the chain ID. Next, color this selection gray.

```
>color sel gray
```

The selection and coloring steps can be condensed into a single line:

```
>color #1/c #2B89AD
```

Here we are using the hex-code for a teal blue (#2B89AD). This is convenient when you need to use a color that is not in the default options.

We can specify a range or a list of residues in the console, using a colon ' : ' after the chain ID.

`>sel #1/b:1-10` (to select residues 1 to 10 from chain B)
`>sel #1/b:1,5,20` (to select residues 1, 5 and 20 from chain B)
`>sel #1/b:1-10,66-73` (to select two ranges of residues)

Here is a cheat-sheet of the hierarchical specifiers that ChimeraX uses:

Symbol	Reference Level	Definition	Examples
#	model	model number assigned to the data in ChimeraX (hierarchical, with positive integers separated by dots: N, N.N, N.N.N, *etc.*)	**#1** **#1.3**
/	chain	chain identifier (case-insensitive unless both upper- and lowercase chain IDs are present)	**/A**
:	residue	residue number *OR* residue name (case-insensitive)	**:51** **:glu**
@	atom	atom name (case-insensitive)	**@ca**

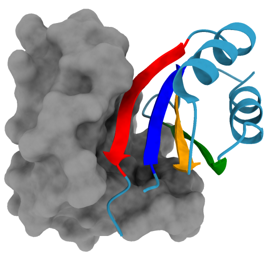
Now let's color each beta-sheet of the protein:

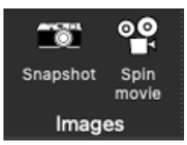
```
>color #1/b:3-9 blue           (β-sheet no. 1)
>color #1/b:30-36 green        (β-sheet no. 2)
>color #1/b:42-47 orange       (β-sheet no. 3)
>color #1/b:66-73 red          (β-sheet no. 4)
```

Finally, let's show the AVR-Pik protein (chain C) in **volumetric** representation, by showing its surface.

```
>surface #1/c
```

Find an angle that shows the binding of **AVR-Pik** to **OsHIPP19**, and take a picture and a spin movie by navigating to the Home panel, and images subpanel.





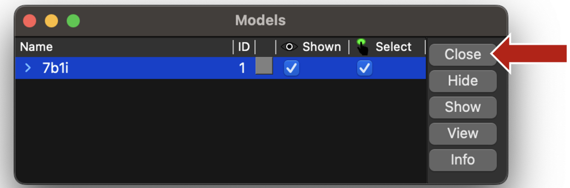
By default, the picture and the video should be saved to your Desktop, but you can check where they were saved by typing pwd (print working directory) in the ChimeraX console, and checking the directory which will be printed in your `log`.

ChimeraX doesn't save your progress automatically, so to save a session, you can type out the command.

Note: The proper direction of the '/' bars will depend on your operating system.

`>save ~/Users/{your-username-here}/Desktop/`**`ChimeraX_Exercise1.cxs`**

Now that you have saved your session and completed the exercise, select the `7B1I` model in the Models panel, and close it.



Exercise 2.2. Visualizing AF3 models in ChimeraX with their confidence metrics

In this section, we will visualize AlphaFold 3 (**AF3**) prediction results in **ChimeraX**, along with their associated confidence metrics. We will color residues by **pLDDT** scores and display predicted aligned error (**PAE**) values for contacting residues, as pseudobonds.

As an example, we will use a prediction generated in **Exercise 1.3**. We will open the model for **Pikp-1_186–486** in complex with **Pathogen_Protein_1**. To begin, download the modelling results from the AlphaFold 3 web server by clicking the **Download** button. You should obtain a ZIP file named, for example:

`fold_pikp_1_186_486_with_pathogen_protein_1.zip`

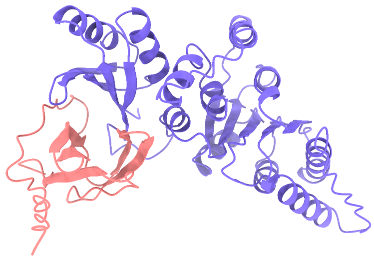
Unzip this file to extract the contents. You should see 17 files, including:

- `[...]_model_0.cif`: top-ranked model, containing the predicted atomic coordinates.
- Several `.cif` files corresponding to other ranked models.
- `.json` files containing the **PAE** data (error estimates) for each predicted model.

Today, we will analyze the top-ranked model—`model_0.cif`.
Drag this file into ChimeraX to load the structure.

Now, color the model by chains,

```
>color #1 bychain
```



And show the sequences of each chain by clicking on the 'Sequence' icon . From looking at the sequences, it's evident that the shorter chain B corresponds to the effector **'Pathogen_Protein_1'**, and the longer chain A corresponds to **'Pikp-1_186–486'**.

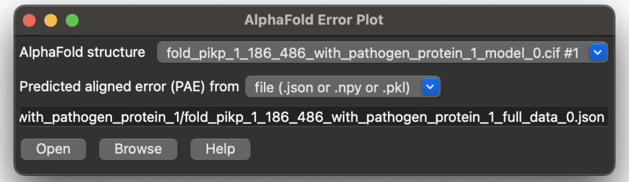
```
chain A   1 GLKQKIVIKVAMEGNNCRSKAMALVASTGGVDSVALVGDLRDKIEVVGYG
chain A  51 IDPIKLISALRKKVGDAELLQVSQANKDVKETTPMLAPVKSICEFHKVKT
chain A 101 VCILGLPGGGKTTVARELYDALGTHFPCRVFVSVSPSSSPSPNLTKTLAD
chain A 151 IFAQAQLGVTDTLSTSYGGSGTGRALQQHLIDNISAFLLNKKYLIVIDDI
chain A 201 WHWEEWEVIRKSIPKNDLGGRIIMTTRLNSIAEKCHTDDNDVFVYEVGDL
chain A 251 DNNDALSLSWGIATKSGAGNRIGTGEDNPCYDIVNMCYGMPLALIWLSSA
chain A 301 L
```

```
chain B  1 ETGNKYIEKRAIDLSRERDPNFFDHPGIPVPECFWFMFKNNVRQDAGTCYSSWKM
chain B 56 DMKVGPNWVHIKSDDNCNLSGDFPPGWIVLGKKRPGF
```

Next, we will load the **confidence metrics** associated with this AlphaFold model. To do this:

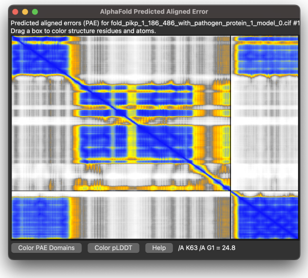
1. In the top menu, go to: `Tools > Structure Prediction > AlphaFold Error Plot`

2. A new window will open with a dropdown to select the model:



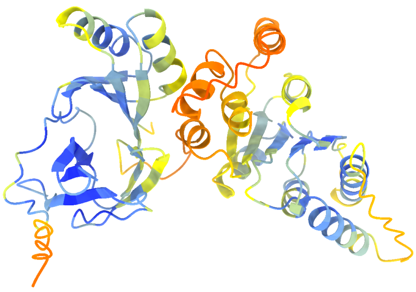
3. In most cases, ChimeraX will automatically find the correct `.json` file.

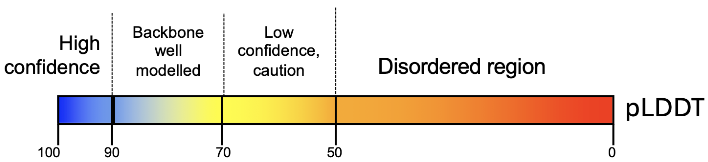
If not, click **Browse** and manually select the JSON file for your top-ranked model (e.g. `fold_pikp_1_186_486_with_pathogen_protein_1_full_data_0.json`), and click **Open**.



You should now see the **PAE** data associated with your model. This is a very important window because it shows the error data associated with your AlphaFold model, and allows you to color the structure by **pLDDT** scores, or by **PAE domains**.

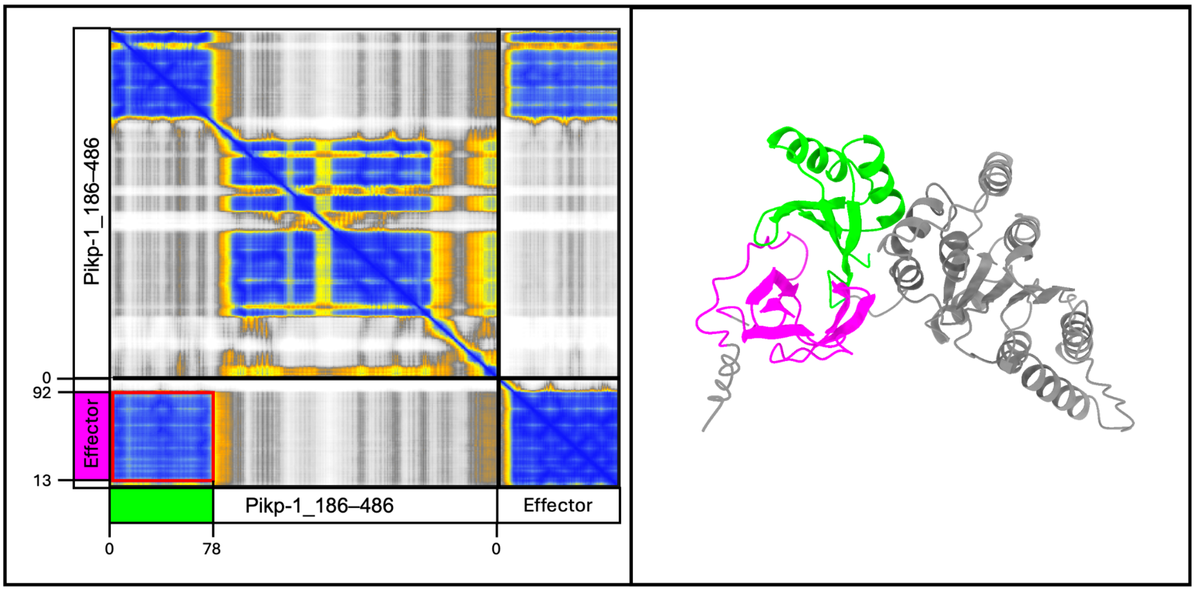
Let's color our prediction by **pLDDT**. You should have something that looks like this:





A) What can you say about the pLDDT scores of the residues predicted at the interaction **interface** between the two proteins?

The PAE plot we opened for this model is interactive, meaning that if you make a selection in the PAE plot, it will be reflected in your 3D model:

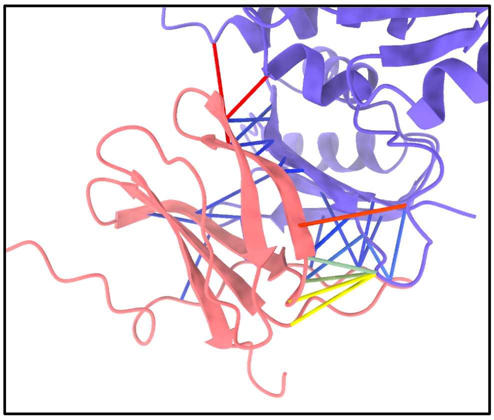


As you can see here, selecting the high-confidence region in the marginal of the PAE plot leads to the highlighting in **pink** and **green** on the 3D structure, of the N-terminal end of **Pikp-1_186-486**, and almost all of the **Pathogen_Protein_1** (effector). The way in which we interpret this is that AF3 is confident about the relative positioning of the N-term residues of **Pikp-1_186-486** in relation to all of the residues in the effector protein.

Another way of visualizing the PAE data of predicted contacts is by using the following command in the ChimeraX terminal:

```
>alphafold contacts #1/a to #1/b distance 3.5
```

This command will only work if you have already loaded the PAE data into ChimeraX. Notice that you can specify two chains for which you want to find contacts, and a **distance** threshold which we will set to 3.5 Ångstroms.



Any pairs of residues that are within a distance of 3.5Å will be shown with a line or 'pseudo-bond' drawn, connecting them. And this pseudobond will be colored based on the AlphaFold predicted alignment error (**PAE**) metric, using ChimeraX paecontacts palette.

Predicted Aligned Error (PAE)

B) Can you tell which β-sheet on the side of the effector protein is predicted to have more high-confidence (i.e. low PAE) contacts with Pikp, at a distance equal or lower to 3.5 Å?

The command we just used can also be modified to search specific ranges or lists of residues, following the same selection algebra we saw in **Exercise 2.1**

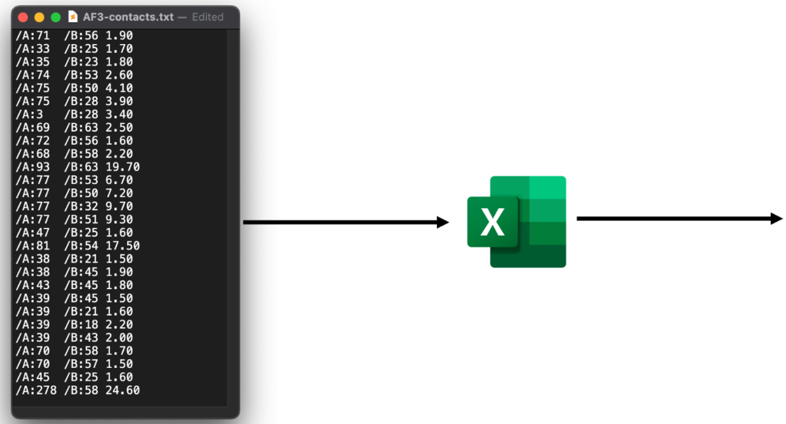
C) If we run the command below, we will be plotting contacts between **which residues,** and at what distances?

```
>alphafold contacts #1/b:20 to #1/a distance 4.5
```

Exercise 2.3. Exporting information from a protein-protein interface in ChimeraX

Now, let's see how we can export the predicted contacts from **Exercise 2.2** from **ChimeraX** into another programme, like **Excel**, for downstream analysis. We can export a list of these residues to a text file by adding the following at the end of our **`alphafold contacts`** command:

```
> alphafold contacts #1/a to #1/b distance 3.5 outputFile AF3-contacts.txt
```




```
/A:71  /B:56 1.90
/A:33  /B:25 1.70
/A:35  /B:23 1.80
/A:74  /B:53 2.60
/A:75  /B:50 4.10
/A:75  /B:28 3.90
/A:3   /B:28 3.40
/A:69  /B:63 2.50
/A:72  /B:56 1.60
/A:68  /B:58 2.20
/A:93  /B:63 19.70
/A:77  /B:53 6.70
/A:77  /B:50 7.20
/A:77  /B:32 9.70
/A:77  /B:51 9.30
/A:47  /B:25 1.60
/A:81  /B:54 17.50
/A:38  /B:21 1.50
/A:38  /B:45 1.90
/A:43  /B:45 1.80
/A:39  /B:45 1.50
/A:39  /B:21 1.60
/A:39  /B:18 2.20
/A:39  /B:43 2.00
/A:70  /B:58 1.70
/A:70  /B:57 1.50
/A:45  /B:25 1.60
/A:278 /B:58 24.60
```

Position A (Pikp-1_186-486)	Position B (Pathogen_Protein_1)	PAE
/A:38	/B:21	1.5
/A:39	/B:45	1.5
/A:70	/B:57	1.5
/A:39	/B:21	1.6
/A:45	/B:25	1.6
/A:47	/B:25	1.6
/A:72	/B:56	1.6
/A:33	/B:25	1.7
/A:70	/B:58	1.7
/A:35	/B:23	1.8
/A:43	/B:45	1.8
/A:38	/B:45	1.9
/A:71	/B:56	1.9
/A:39	/B:43	2
/A:39	/B:18	2.2
/A:68	/B:58	2.2
/A:69	/B:63	2.5
/A:74	/B:53	2.6
/A:3	/B:28	3.4
/A:75	/B:28	3.9
/A:75	/B:50	4.1
/A:77	/B:53	6.7
/A:77	/B:50	7.2
/A:77	/B:51	9.3
/A:77	/B:32	9.7
/A:81	/B:54	17.5
/A:93	/B:63	19.7
/A:27	/B:58	24.6

This results in a text file containing a list, where every row shows a pair of residues involved in a contact at a distance lower than our 3.5 Å distance cutoff, and with a given PAE score. This data can be further processed in Excel to remove low confidence contacts (*e.g.* **`/A:81 /B:54 17.50`**).

> **Confident** PAE scores tend to be in the range of **0-10**, with **10-15 being a gray zone**, and anything **above 15** being of **poor confidence**.

A) Study the predicted interface of **Pathogen_Effector_1** and the rice **Pikp-1_186-486**, and generate a list of candidate residues at the interface that you would mutate in either protein to abolish their interaction.

B) What key experiments would you conduct to test these designed mutants in the lab? Briefly describe how you would design the experiments, and which controls you would incorporate.

Session 3: Multiple protein structure alignment using FoldMason

Objectives: The main goal of this session is to learn how to analyze structurally similar proteins, using **TM-align, Foldseek**, and **FoldMason**.

Exercise 3.1. Superimposing structures in ChimeraX

Let's download pathogen proteins from the URL (XXX), and drag and drop the PDB files in **Dataset 1** **i**nto ChimeraX to visualize them. Then, you can type the commands below to superimpose these structures:

```
>match #2 to #1
>match #3 to #1
```

a) What do these structures look like? Do you think all three structures are similar or different?

b) Which two structures look the most similar among the three?

Exercise 3.2. Scoring protein structure similarities using TM-align

Please visit the **RCSB.org website** (https://www.rcsb.org/alignment) and upload the PDB files. Complete the table below with **RMSD** and **TM-scores** for the different query structures (AVR-Pia and AVR-PikD). The structure uploaded at the top will be used as the reference structure.

a) Complete the table below with your TM-align results:

Reference	Subject	RMSD	TM-score
AVR-Pia	AVR-PikD		
AVR-Pia	AVR-Mgk1		
AVR-PikD	AVR-Pia		
AVR-PikD	AVR-Mgk1		

b) Based on the RMSD and TM-scores, which structure—AVR-Mgk1 or AVR-PikD—is more similar to AVR-Pia?

Hint: The lower the RMSD, the better the structural alignment between the two proteins. TM-scores below 0.2 usually indicate that the proteins are unrelated, while scores above 0.5 generally suggest that the proteins share the same fold.

c) When you swapped the reference and subject structures, how did the RMSD and TM-scores change?

Exercise 3.3. Searching structurally similar proteins using Foldseek

Please visit the **Foldseek website** (https://search.foldseek.com/search) and upload the PDB file of AVR-PikD (Dataset 1). Since you have just learned about TM-align, let's click Databases & search settings and set TM-align as Mode. Click SEARCH and search the structurally similar proteins in the database.

a) Can you find similar structures in the hit proteins? Do structurally similar proteins have high sequence identities?

Hint: Click Alignment bottom and remember RMSD and TM-scores!

b) What kind of species are predominant in AFDB50? Can you see any species which have structurally similar proteins but are taxonomically far from dominant species?

Exercise 3.4. Multiple protein structure alignment using FoldMason

Please visit the **FoldMason website** (https://search.foldseek.com/foldmason) and upload 80 PDB files in Dataset 2 to FoldMason. All the PDB files are structurally similar to that of AVR-PikD which is also known as MAX fold structure. Click ALIGN to start structure-based alignment for all the 80 PDB in FoldMason.

a) Which protein is structurally most similar to AVR-PikD based on the results?

b) If you leverage the programs like Foldseek and FoldMason, what kind of project could you envision?