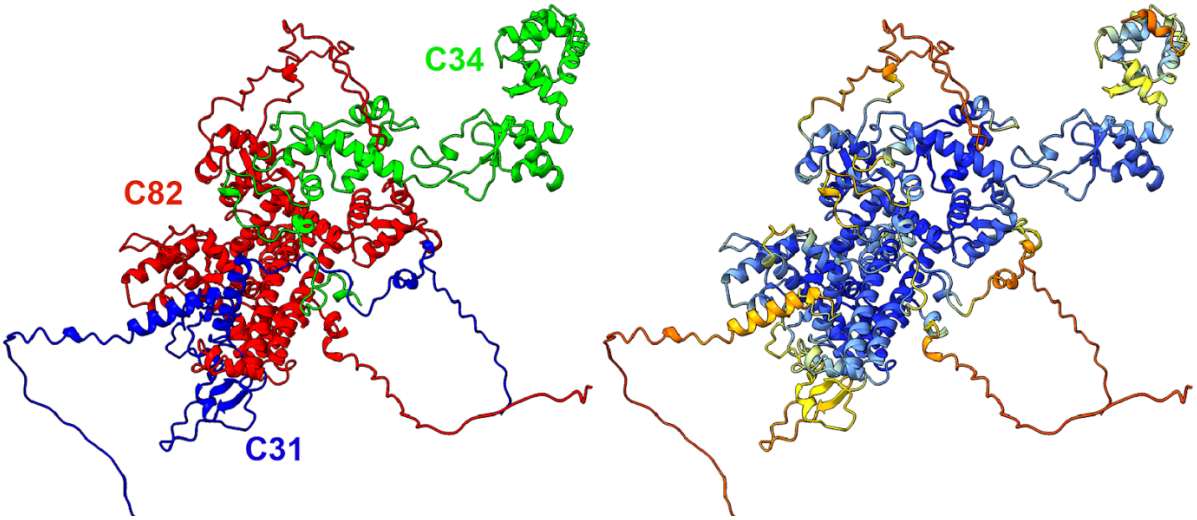
EMBO Practical integrative structural biology course, 2022

Hosted at: European Molecular Biology Laboratory Hamburg, Notkestrasse 85, D-22607 Hamburg, Germany

Tutor: Agnieszka Obarska

Modeling C31, C34 and C82 using AlphaFold



About the Pol III complex

Pol III is a 17-subunit enzyme that transcribes tRNA genes. Its architecture can be subdivided into a core, stalk, heterodimer, and heterotrimer of C82, C34, and C31 subunits (see Figure). During the tutorials, we can focus on modeling the positioning of the C82/C34/C31 heterotrimer subunits relative to the others. The structure of Pol III is quite well characterized, with multiple cryo-EM structures of Pol III published.

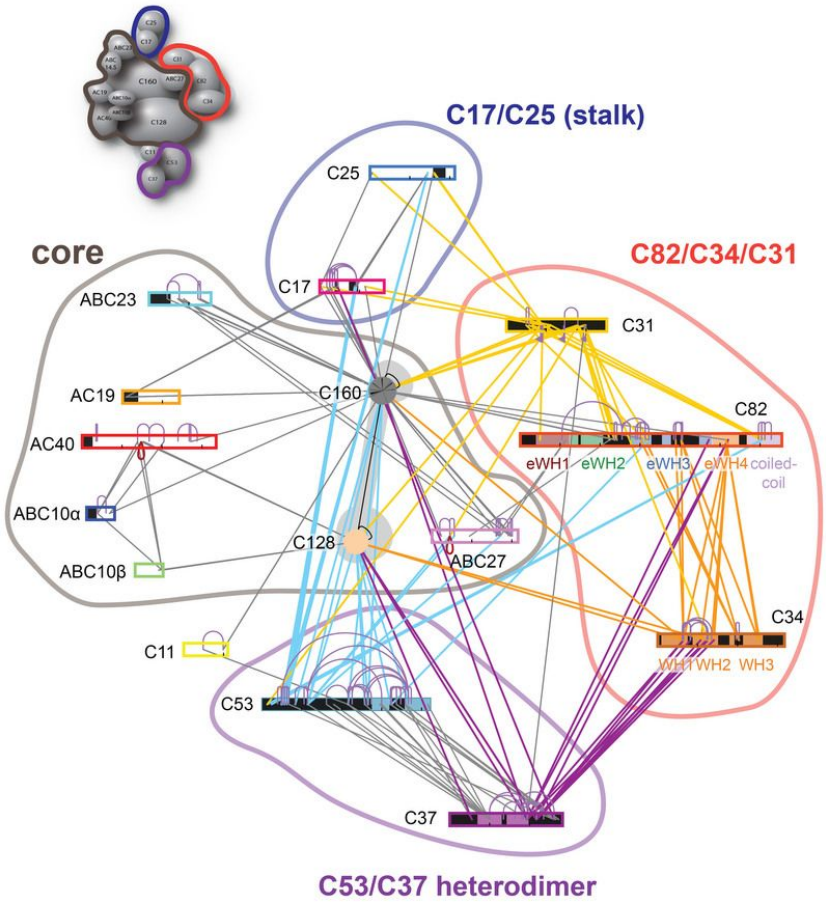


Figure 1. Pol III subunits are shown as rectangular bars except for C160 and C128, which are shown as ovals for the sake of clarity. Inter-links are shown as lines connecting the protein bars, while intra-links are shown as curves. Inter-links to C31 are colored yellow, to C34 - gold, to C37 – violet, to C53 - cyan. The remaining interlinks are colored gray. Domains of C82 and C34 discussed in this work are indicated. The figure was created with xiNET. Figure reproduced from https://www.nature.com/articles/nmeth.3838

Structural modeling

Structural modeling allows **predicting the three-dimensional structure** of proteins using computational methods. The most accurate modeling program is **AlphaFold**, which allows building a model of the target protein based on a sequence alone and optionally using other structures as templates.

Overview

During this practical session, **we will use AlphaFold2 to model Pol III subunits alone and in complexes**.
UniProt ids of Pol III subunits: C31: P17890, C34: P32910, C82: P32349

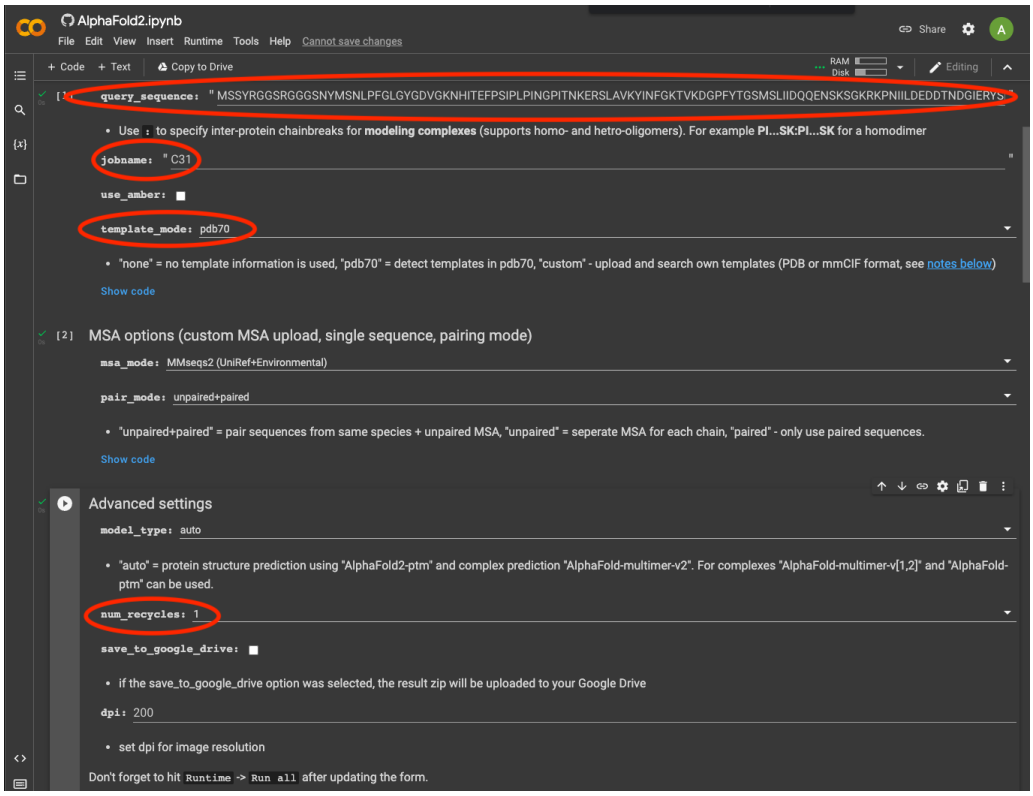
1. We will run modeling of C31 alone using public web server implementation of AlphaFold - **ColabFold**
2. We will run modeling of C31-C82 complex using public web server implementation of AlphaFold - **ColabFold**
3. We will run modeling of C31 using local **AlphaFold Monomer** on computer cluster
4. We will run modeling of C31-C82 complex using local **AlphaFold Multimer** on computer cluster

    - We will talk about how AlphaFold works, what different scores mean etc.
    - We will explore the AlphaFold database of pre-calculated models
    - We will check ChimeraX functions for exploring AlphaFold models

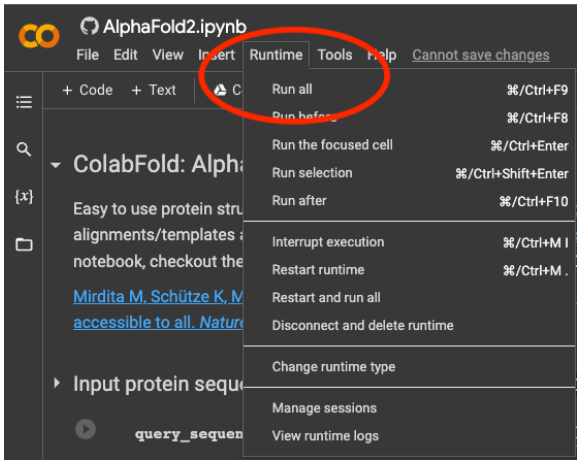
Modeling of C31 using ColabFold:

1. Go to ColabFold: https://github.com/sokrypton/ColabFold
2. Select “AlphaFold2_mmseqs2”
3. Replace the default query sequence with the sequence of C31:
    1. Go to UniProt: https://www.uniprot.org/
    2. Paste UniProt id of C31: P17890
    3. Copy sequence of C31

4. If you want to use templates in modeling set template_model: pdb70, if not leave it “none”.
5. Change num_recycles to 1 in the “Advanced settings” to speed up calculations (normally keep it default!).



6. Go to Runtime-> Run all

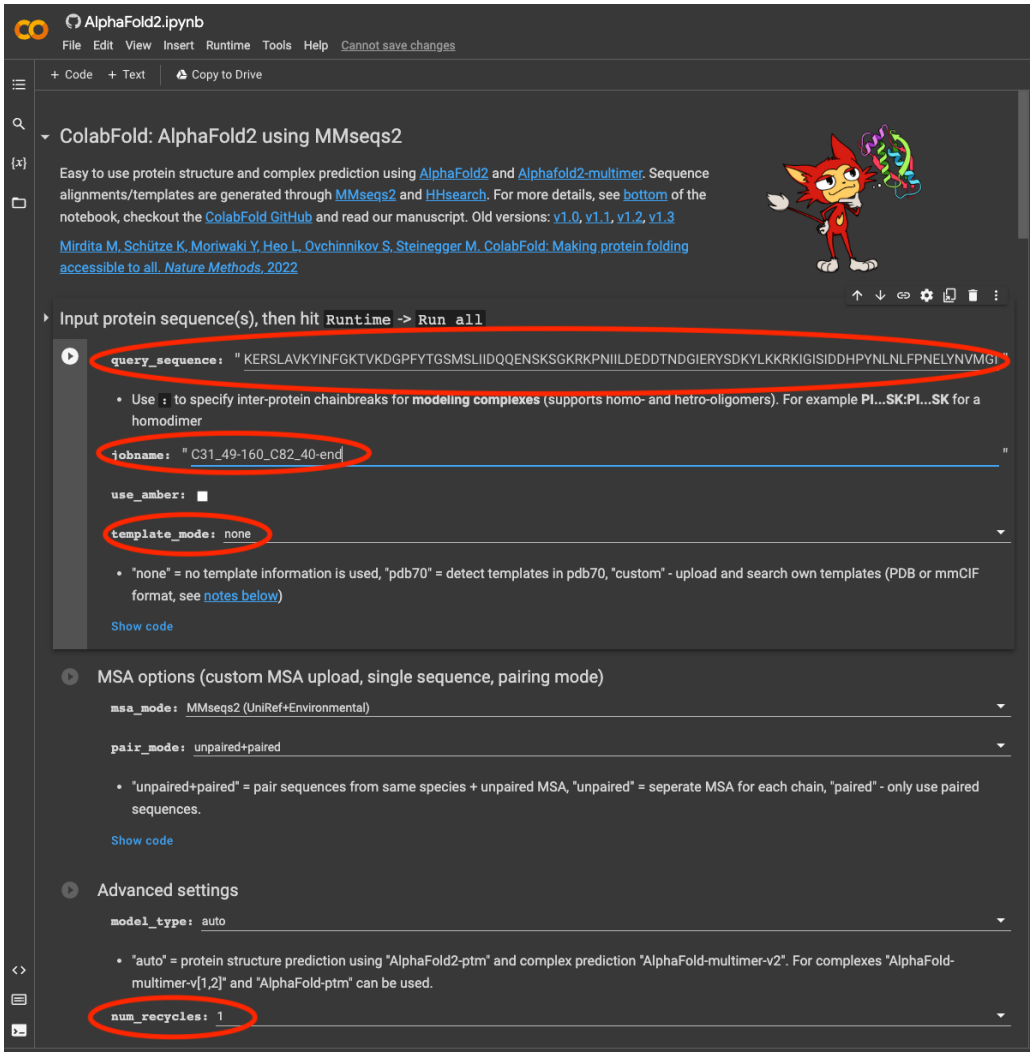


Modeling of C31-C82 complex using ColabFold

1. Go to ColabFold: https://github.com/sokrypton/ColabFold
2. Select “AlphaFold2_mmseqs2”
3. Replace the default query sequence with the sequence of C31 (aa. 49-160) and C82 (40-end) separated by “:”

```
KERSLAVKYINFGKTVKDGPFYTGSMSLIIDQQENSKSGKRKPNIILDEDDTNDGIERYSDKYLKKRKIGISIDDHPYNLNLFPNELYNVM
GINKKKLLAISKFNNADDVF:TLNPDLFLYKELVKAHLGERAASVIGMLVALGRLSVRELVEKIDGMDVDSVKTTLVSLTQLRCVKYLQE
TAISGKKTTYYYYNEEGIHILLYSGLIIDEIITQMRVNDEEEHKQLVAEIVQNVISLGSLTVEDYLSSVTSDSMKYTISSLFVQLCEMGYLIQI
SKLHYTPIEDLWQFLYEKHYKNIPRNSPLSDLKKRSQAKMNAKTDFAKIINKPNELSQILTVDPKTSLRIVKPTVSLTINLDRFMKGRRSK
QLINLAKTRVGSVTAQVYKIALRLTEQKSPKIRDPLTQTGLLQDLEEAKSFQDEAELVEEKTPGLTFNAIDLARHLPAELDLRGSLLSRKPS
DNKKRSGSNAAASLPSKKLKTEDGFVIPALPAAVSKSLQESGDTQEEDEEEEDLDADTEDPHSASLINSHLKILASSNFPFLNETKPGVY
YVPYSKLMPVLKSSVYEYVIASTLGPSAMRLSRCIRDNKLVSEKIINSTALMKEKDIRSTLASLIRYNSVEIQEVPRTADRSASRAVFLFRCK
ETHSYNFMRQNLEWNMANLLFKKEKLKQENSTLLKKANRDDVKGRENELLLPSELNQLKMVNERELNVFARLSRLLSLWEVFQMA
```

4. Change num_recycles to 1 in the “Advanced settings” to speed up calculations (normally keep it default!).

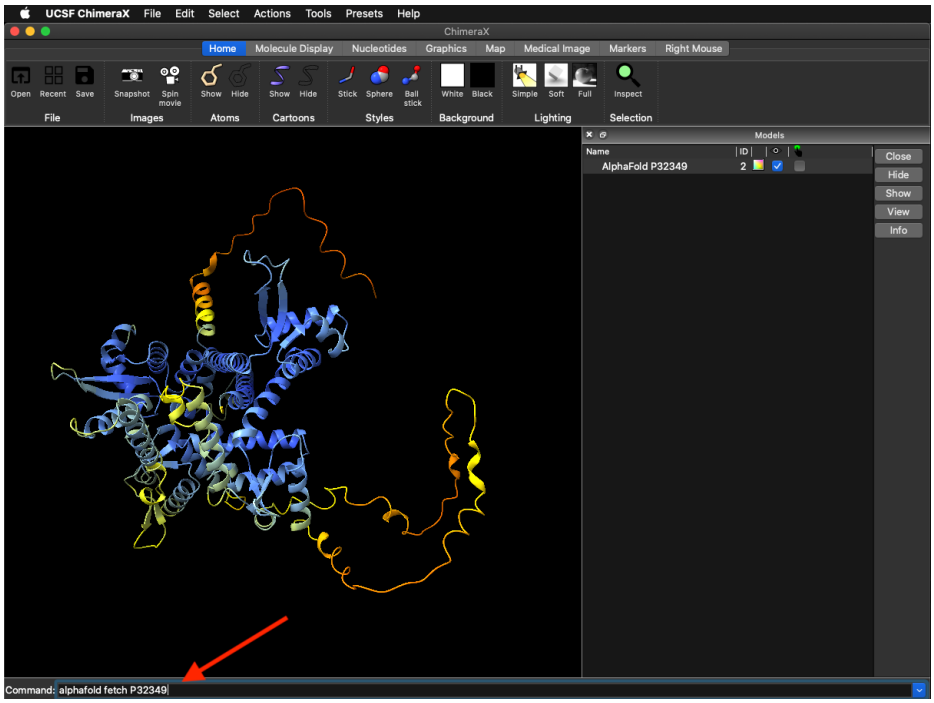


5. Go to Runtime-> Run all

Inspect model in ChimeraX

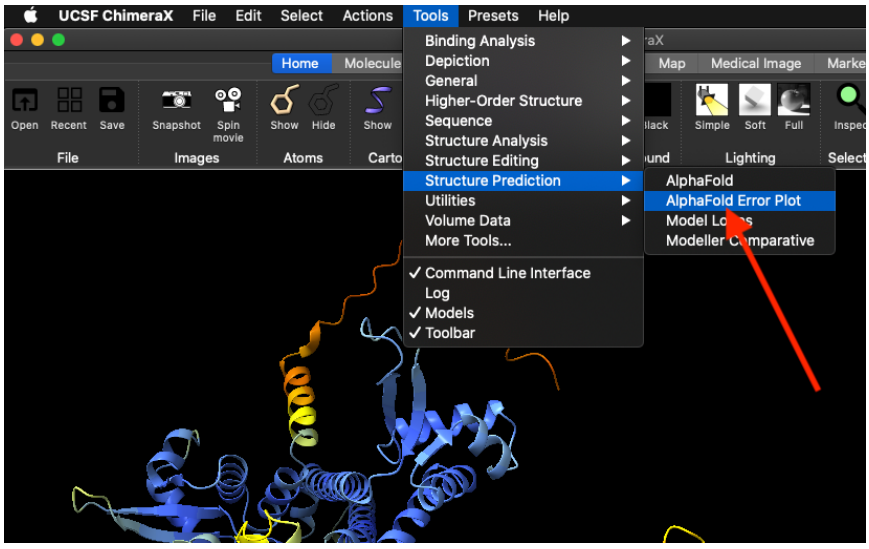
1. Open Chimera X.
2. Open model of C82 from AlphaFold database. In the command line, type:

```
alphafold fetch P32349
```

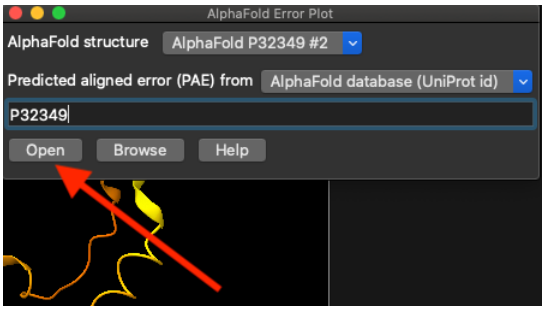


It is colored by pLDDT scores (AlphaFold scheme).

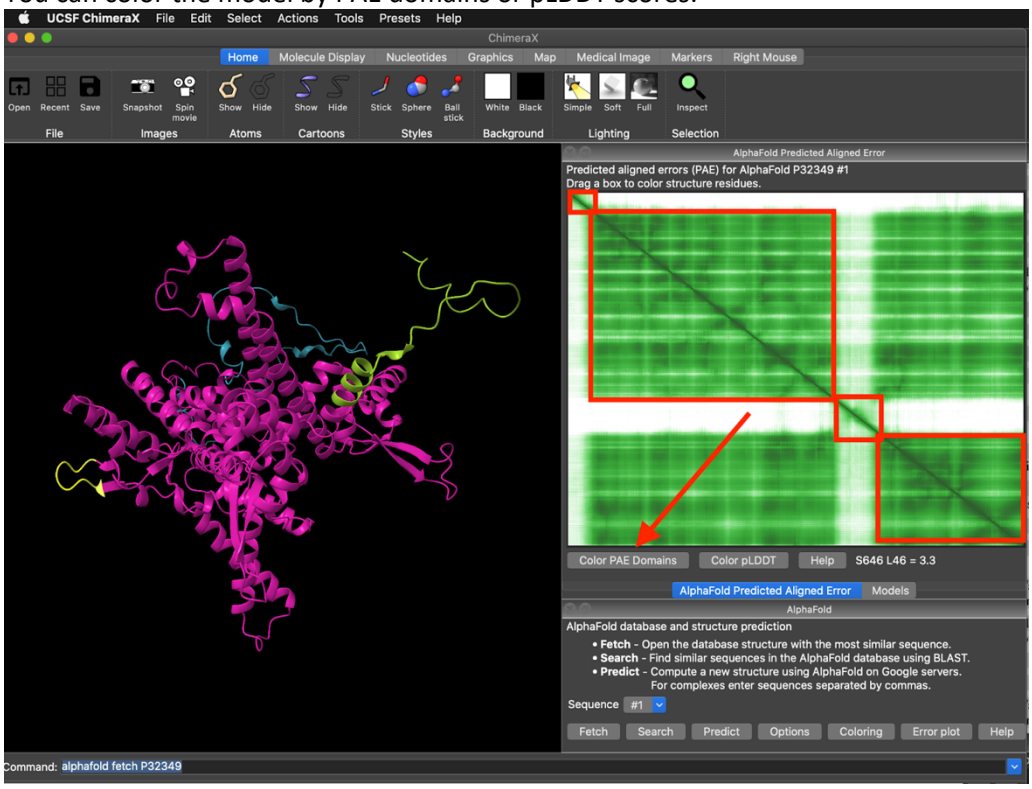
3. Open PAE plot, select from ChimeraX menu:
Tools → Structure Prediction → AlphaFold Error Plot



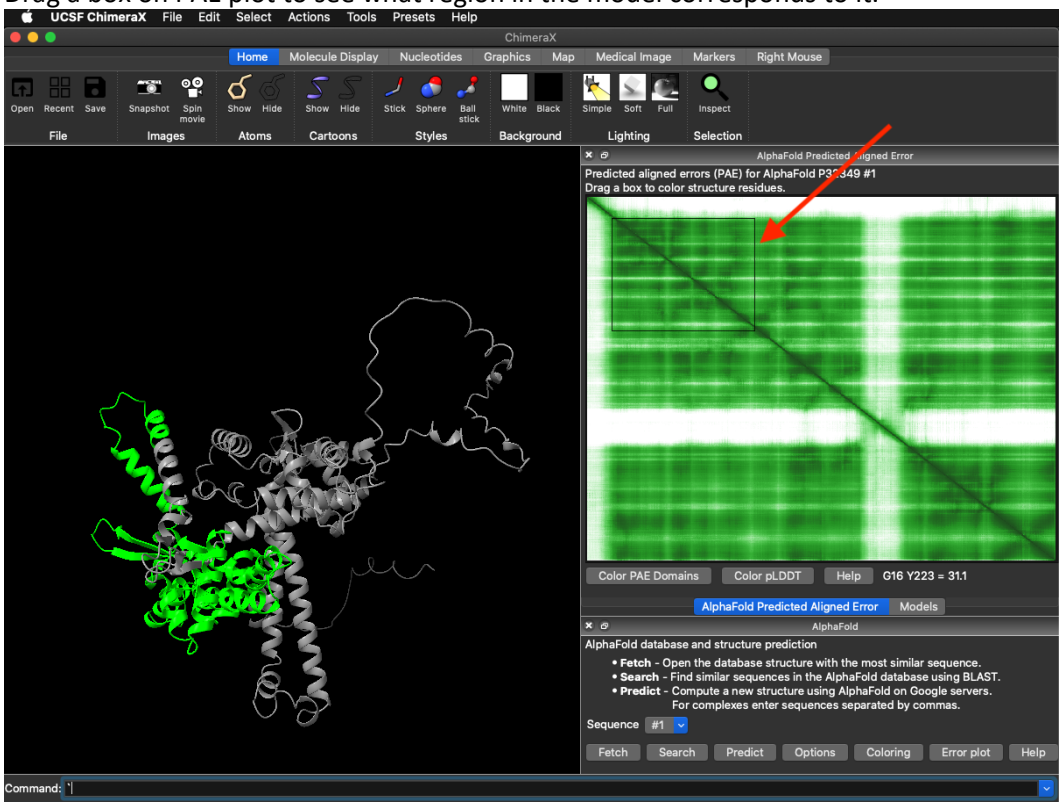
and open PAE from AlphaFold database:



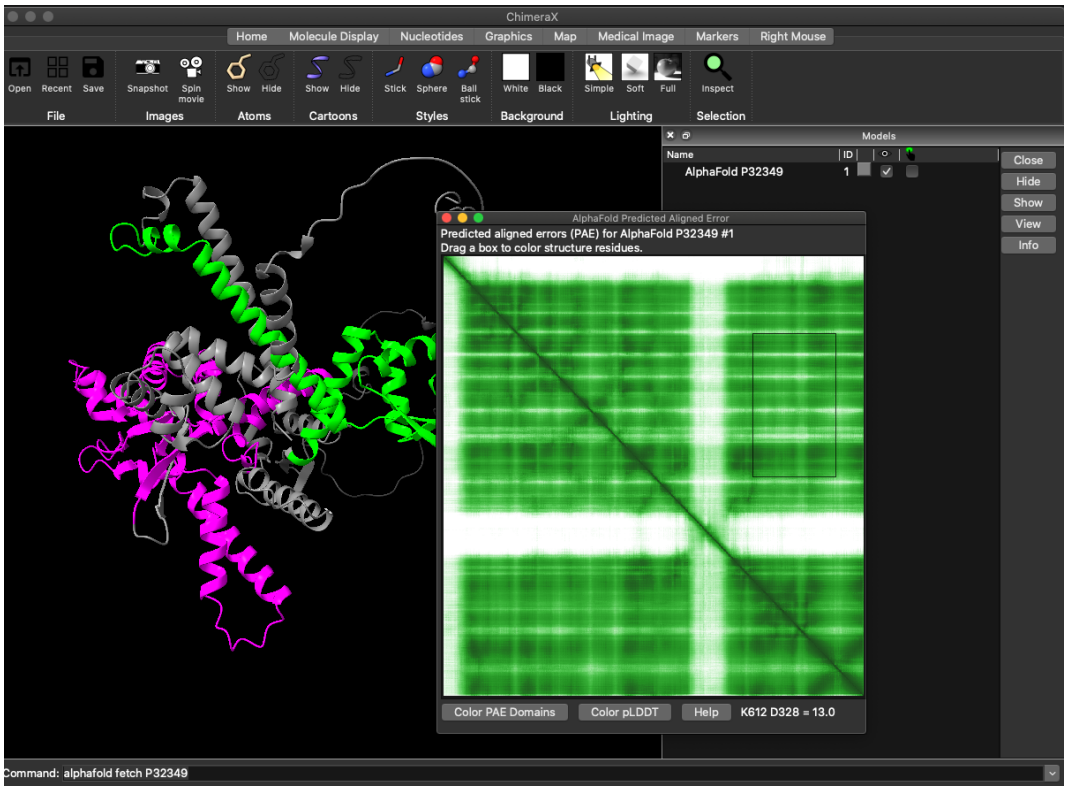
You can color the model by PAE domains or pLDDT scores:



Drag a box on PAE plot to see what region in the model corresponds to it.



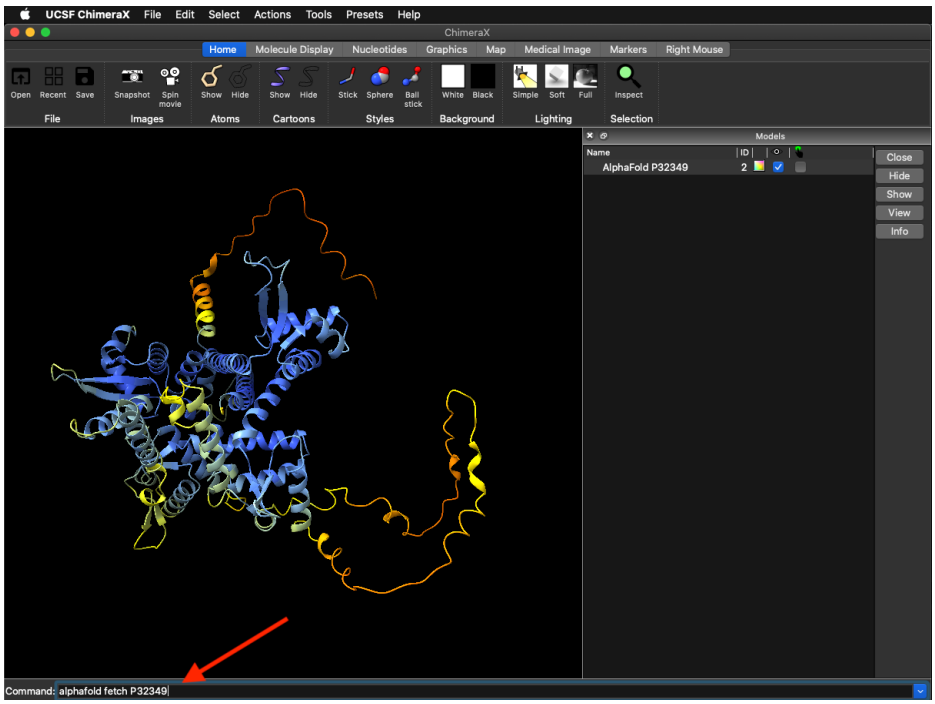
what does it tell you? Is the orientation between the domains confident?



Analyze ColabFold results

1. Download results, unzip, and open model ranked as 1 in ChimeraX.
2. To color by pLDDT scores (AlphaFold scheme). In the command line, type:

```
color bfactor palette alphafold
```



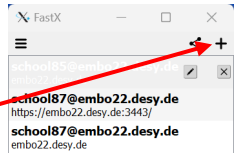
3. To color by chain, type:

```
rainbow chain
```

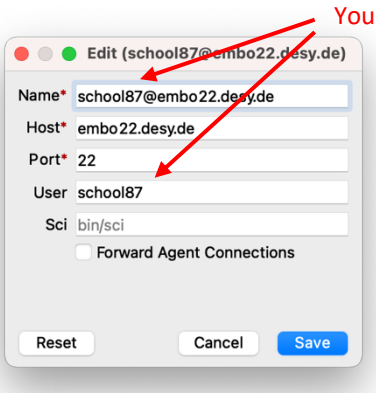
4. Open png files with MSA, PAE… . What do they tell you?

Run the original version of AlphaFold monomer on computer cluster

Remote desktop connection access



1. Start FastX2
2. Click the little plus in the top right corner
3. Fill in like this replacing your user name

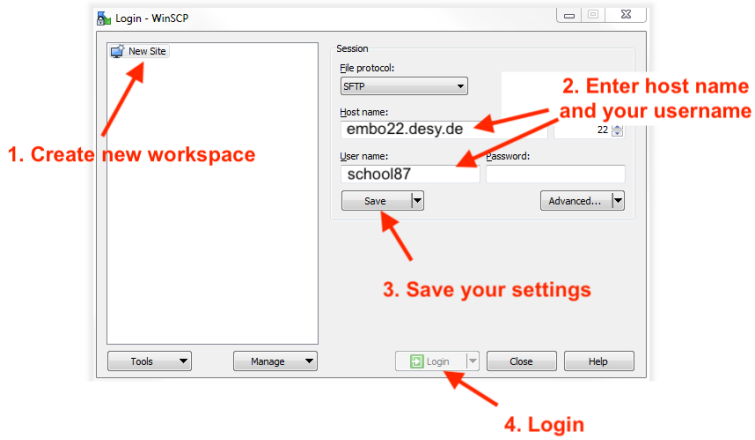


4. Click the little plus in the top right corner

5. and open xterm

To copy files to your local computer use WinSCP

1. Start WinSCP
2. Enter Hostname: embo22.desy.de
3. Enter your username and password.



Run AlphaFold Monomer on cluster

1. In your home create a directory AlphaFold, and a directory where you put the sequences you want to model

```
mkdir AlphaFold
cd AlphaFold
mkdir sequences
```

2. In the "sequences" directory, create fasta file with the sequence you want to model

```
>C31
MSSYRGGSRGGGSNYMSNLPFGLGYGDVGKNHITEFPSIPLPINGPITNKERSLAVKYINFGKTVKDGPFYTGSMSLIIDQQENSKS
GKRKPNIILDEDDTNDGIERYSDKYLKKRKIGISIDDHPYNLNLFPNELYNVMGINKKKLLAISKFNNADDVFTGTGLQDENIGLSML
AKLKELAEDVDDASTGDGAAKGSKTGEGEDDDLADDDFEEDEDEEDDDDYNAEKYFNNGDDDDYGDEEDPNEEAAF
```

3. Copy script to run AlphaFold monomer on desy cluster to your AlphaFold directory:

```
cd ~/AlphaFold
cp /beegfs/desy/group/school/data/AlphaFold/AlphaFold/AlphaFold_monomer.sh .
```

4. Open script with the text editor you like and edit script to your needs.

```
#!/bin/bash
#SBATCH --partition=allgpu
#SBATCH --constraint='A100|V100'
#SBATCH --time=0-12:00
#SBATCH --job-name=C31
#SBATCH --output=C31.out

unset LD_PRELOAD

source /etc/profile.d/modules.sh
module purge
module load maxwell cuda/11.3

# alphafold basics
export PATH=/software/alphafold/2.1.1L/envs/af2.1/bin:$PATH
export TF_FORCE_UNIFIED_MEMORY=1

python3 /software/alphafold/2.1.1L/alphafold/run_alphafold.py \
  --data_dir=/beegfs/desy/group/it/ReferenceData/alphafold \
  --uniref90_database_path=/beegfs/desy/group/it/ReferenceData/alphafold/uniref90/uniref90.fasta \
  --bfd_database_path= \
    /beegfs/desy/group/it/ReferenceData/alphafold/bfd/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt \
  --template_mmcif_dir=/beegfs/desy/group/it/ReferenceData/alphafold/pdb_mmcif/mmcif_files \
  --obsolete_pdbs_path=/beegfs/desy/group/it/ReferenceData/alphafold/pdb_mmcif/obsolete.dat \
  --pdb70_database_path=/beegfs/desy/group/it/ReferenceData/alphafold/pdb70/pdb70 \
  --mgnify_database_path=/beegfs/desy/group/it/ReferenceData/alphafold/mgnify/mgy_clusters.fa \
  --uniclust30_database_path= \
    /beegfs/desy/group/it/ReferenceData/alphafold/uniclust30/uniclust30_2018_08/uniclust30_2018_08 \

  --model_preset=monomer \
  --max_template_date=2013-01-01 \
  --output_dir=/home/$USER/AlphaFold/results/ \
  --fasta_paths=/beegfs/desy/group/school/data/AlphaFold/AlphaFold/sequences/C31.fasta
```

sbatch options

AlphaFold command

paths to AlphaFold databases

- Here you can change the output directory
- And path to the sequence you want to model
- by setting max_template_data you can exclude some templates from modeling

5. Run script & check the job

```
sbatch AlphaFold_monomer.sh
squeue -u username
```

Run AlphaFold Multimer on cluster

1. In the sequence directory, create fasta file with the sequences (or fragments of sequence) you want to model in complex

```
>C31
KERSLAVKYINFGKTVKDGPFYTGSMSLIIDQQENSKSGKRKPNIILDEDDTNDGIERYSDKYLKKRKIGISIDDHPYNLNLFPNELYNVMGINKKKLLAISKFNNADDVF
>C82
TLNPDLFLYKELVKAHLGERAASVIGMLVALGRLSVRELVEKIDGMDVDSVKTTLVSLTQLRCVKYLQETAISGKKTTYYYYNEEGIHILLYSGLIIDEIITQMRVNDEEEH
KQLVAEIVQNVISLGSLTVEDYLSSVTSDSMKYTISSLFVQLCEMGYLIQISKLHYTPIEDLWQFLYEKHYKNIPRNSPLSDLKKRSQAKMNAKTDFAKIINKPNELSQILT
VDPKTSLRIVKPTVSLTINLDRFMKGRRSKQLINLAKTRVGSVTAQVYKIALRLTEQKSPKIRDPLTQTGLLQDLEEAKSFQDEAELVEEKTPGLTFNAIDLARHLPAELDL
RGSLLSRKPSDNKKRSGSNAAASLPSKKLKTEDGFVIPALPAAVSKSLQESGDTQEEDEEEEDLDADTEDPHSASLINSHLKILASSNFPFLNETKPGVYYVPYSKLMPVL
KSSVYEYVIASTLGPSAMRLSRCIRDNKLVSEKIINSTALMKEKDIRSTLASLIRYNSVEIQEVPRTADRSASRAVFLFRCKETHSYNFMRQNLEWNMANLLFKKEKLKQE
NSTLLKKANRDDVKGRENELLLPSELNQLKMVNERELNVFARLSRLLSLWEVFQMA
```

2. Copy script to run AlphaFold monomer on desy cluster

```
cp /beegfs/desy/group/school/data/AlphaFold/AlphaFold/AlphaFold_multimer.sh .
```

3. Open script with the text editor you like and edit script to your needs.

```
#!/bin/bash
#SBATCH --partition=allgpu
#SBATCH --constraint='A100|V100'
#SBATCH --time=0-12:00
#SBATCH --job-name=C31_C82
#SBATCH --output=C31_C82.out

unset LD_PRELOAD

source /etc/profile.d/modules.sh
module purge
module load maxwell cuda/11.3

# alphafold basics
export PATH=/software/alphafold/2.1.1L/envs/af2.1/bin:$PATH
export TF_FORCE_UNIFIED_MEMORY=1
echo $USER

python3 /software/alphafold/2.1.1L/alphafold/run_alphafold.py \
 --data_dir=/beegfs/desy/group/it/ReferenceData/alphafold \
 --uniref90_database_path=/beegfs/desy/group/it/ReferenceData/alphafold/uniref90/uniref90.fasta \
 --bfd_database_path= \
  /beegfs/desy/group/it/ReferenceData/alphafold/bfd/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt \
 --template_mmcif_dir=/beegfs/desy/group/it/ReferenceData/alphafold/pdb_mmcif/mmcif_files \
 --obsolete_pdbs_path=/beegfs/desy/group/it/ReferenceData/alphafold/pdb_mmcif/obsolete.dat \
 --mgnify_database_path=/beegfs/desy/group/it/ReferenceData/alphafold/mgnify/mgy_clusters.fa \
 --uniclust30_database_path= \
  /beegfs/desy/group/it/ReferenceData/alphafold/uniclust30/uniclust30_2018_08/uniclust30_2018_08 \
 --uniprot_database_path=/beegfs/desy/group/it/ReferenceData/alphafold/uniprot/uniprot.fasta \
 --pdb_seqres_database_path=/beegfs/desy/group/it/ReferenceData/alphafold/pdb_seqres/pdb_seqres.txt \

 --model_preset=multimer \
 --max_template_date=2013-01-01 \
 --output_dir=/home/$USER/AlphaFold/results/ \
 --fasta_paths=/beegfs/desy/group/school/data/AlphaFold/AlphaFold/sequences/C31_C82.fasta
```

sbatch options

AlphaFold command

paths to AlphaFold databases

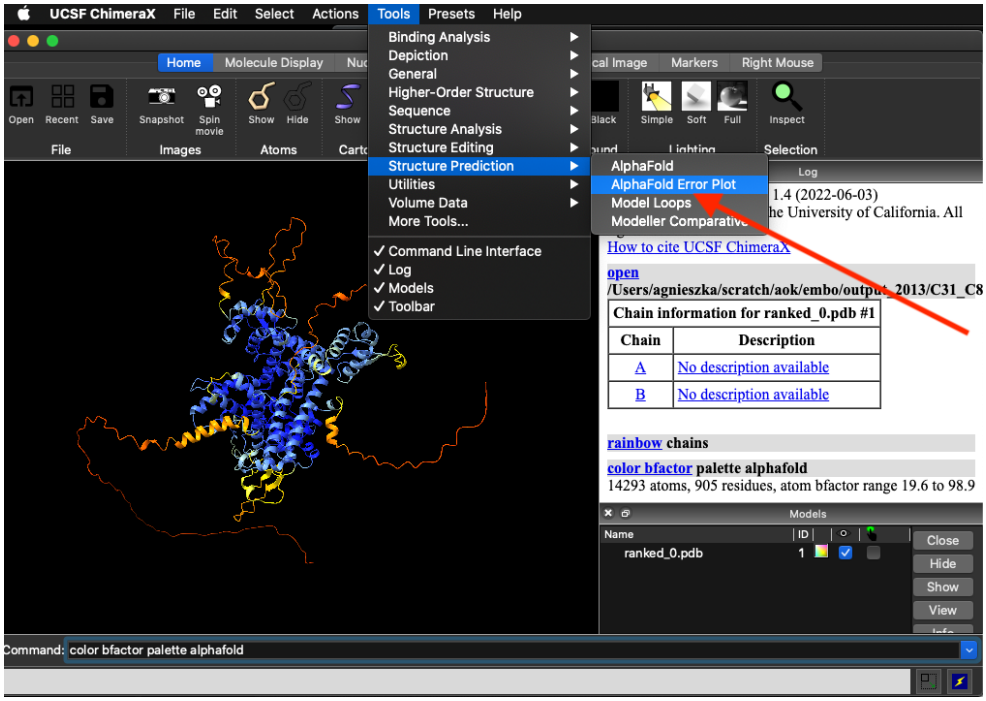
- Here you can change the output directory
- And path to the sequence you want to model
- by setting max_template_data you can exclude some templates from modeling

4. Run script and check your job

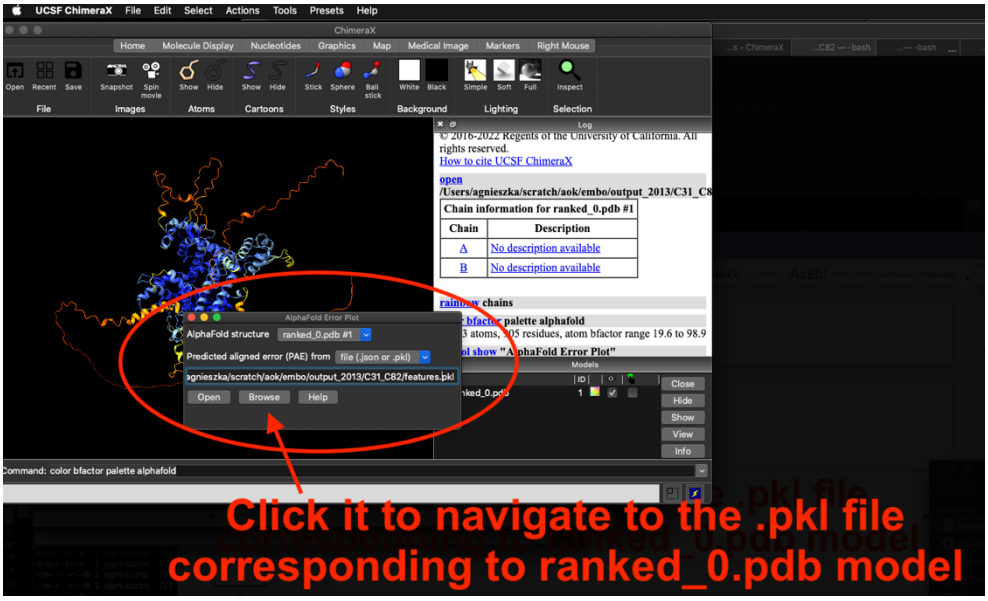
```
sbatch AlphaFold_multimer.sh
squeue -u username
```

Analyze models in ChimeraX

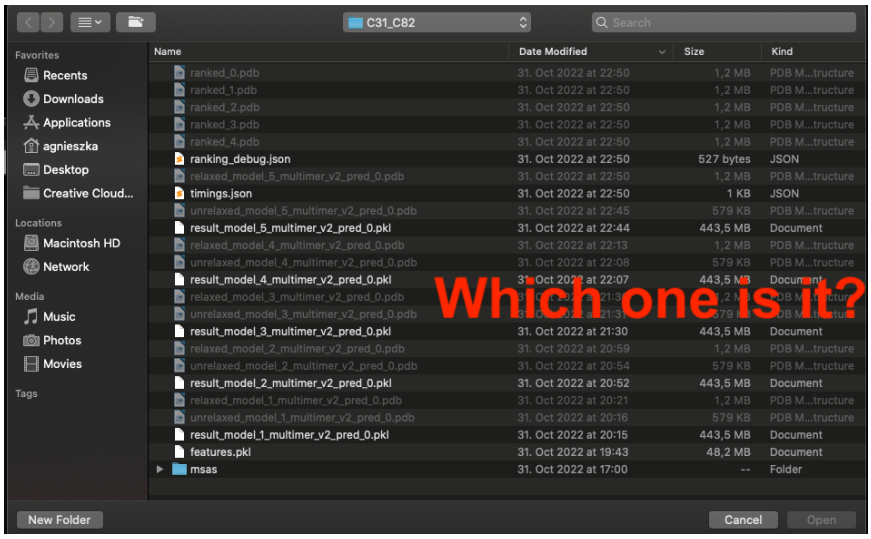
1. Open ChimeraX
2. Open ranked_0.pdb. File → Open
3. Open PAE plot



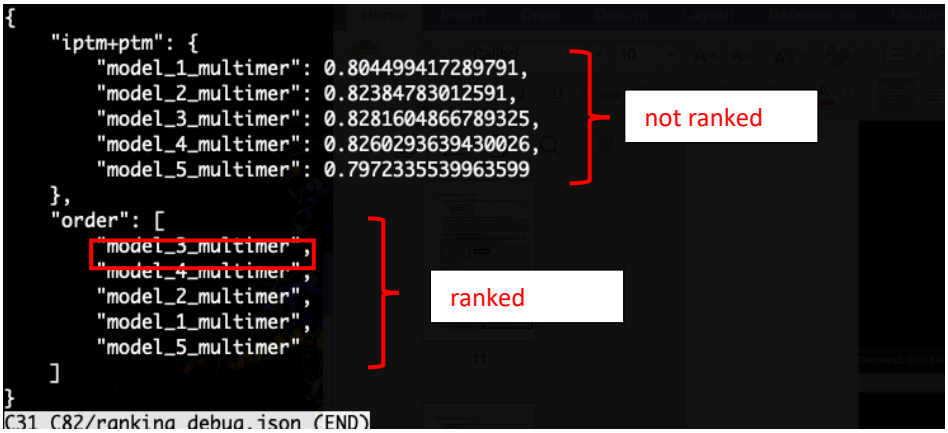
You will see another window



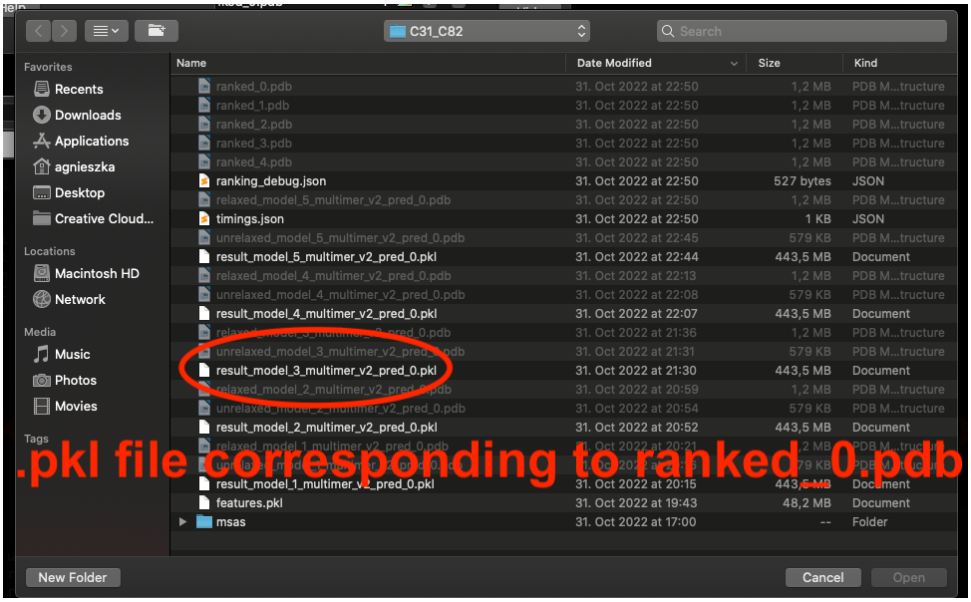
But how do we know which one it is?



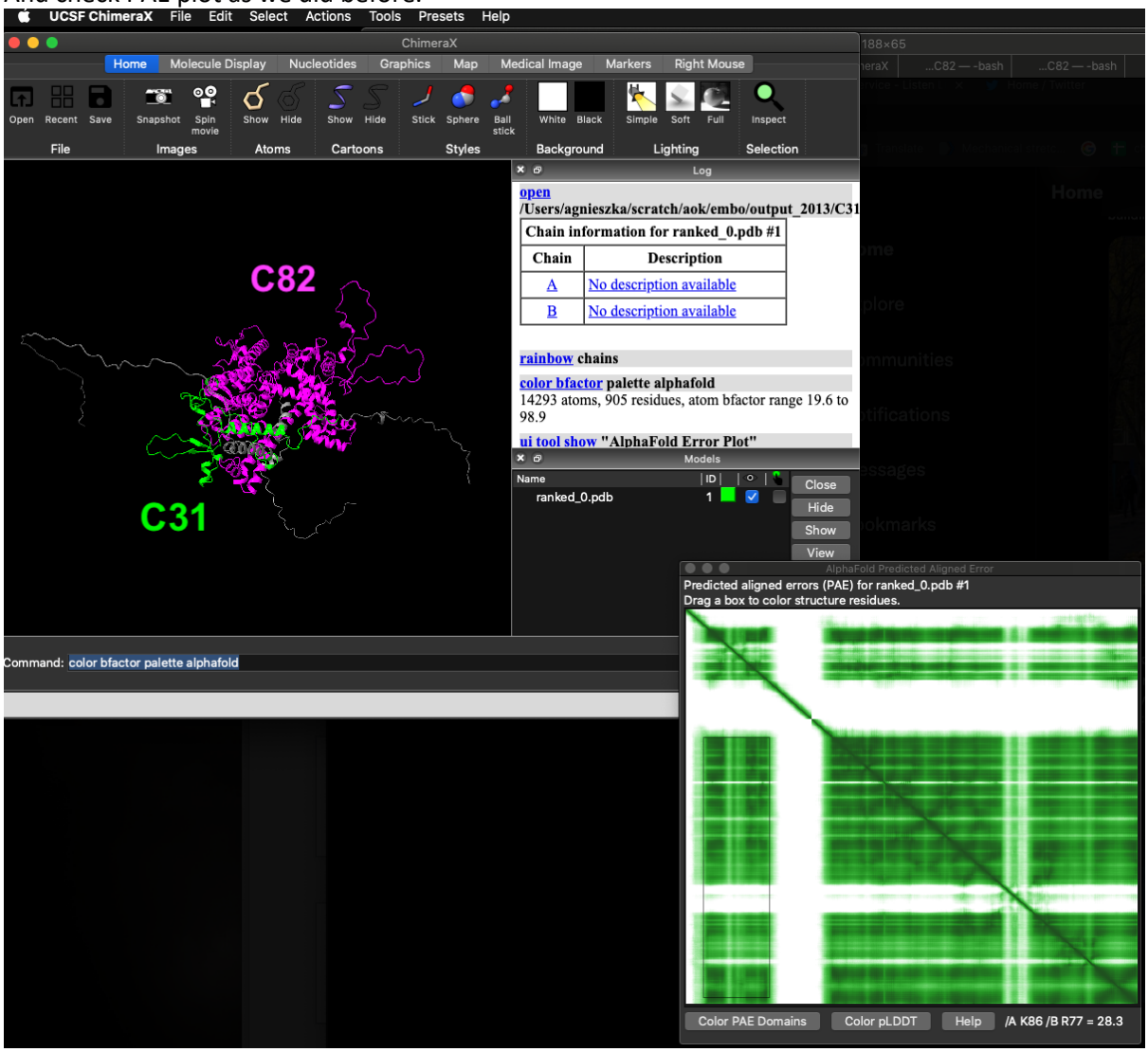
To find out, you have to open ranking_debug.json file to see what model was ranked as 0.



Now open corresponding .pkl file



And check PAE plot as we did before.

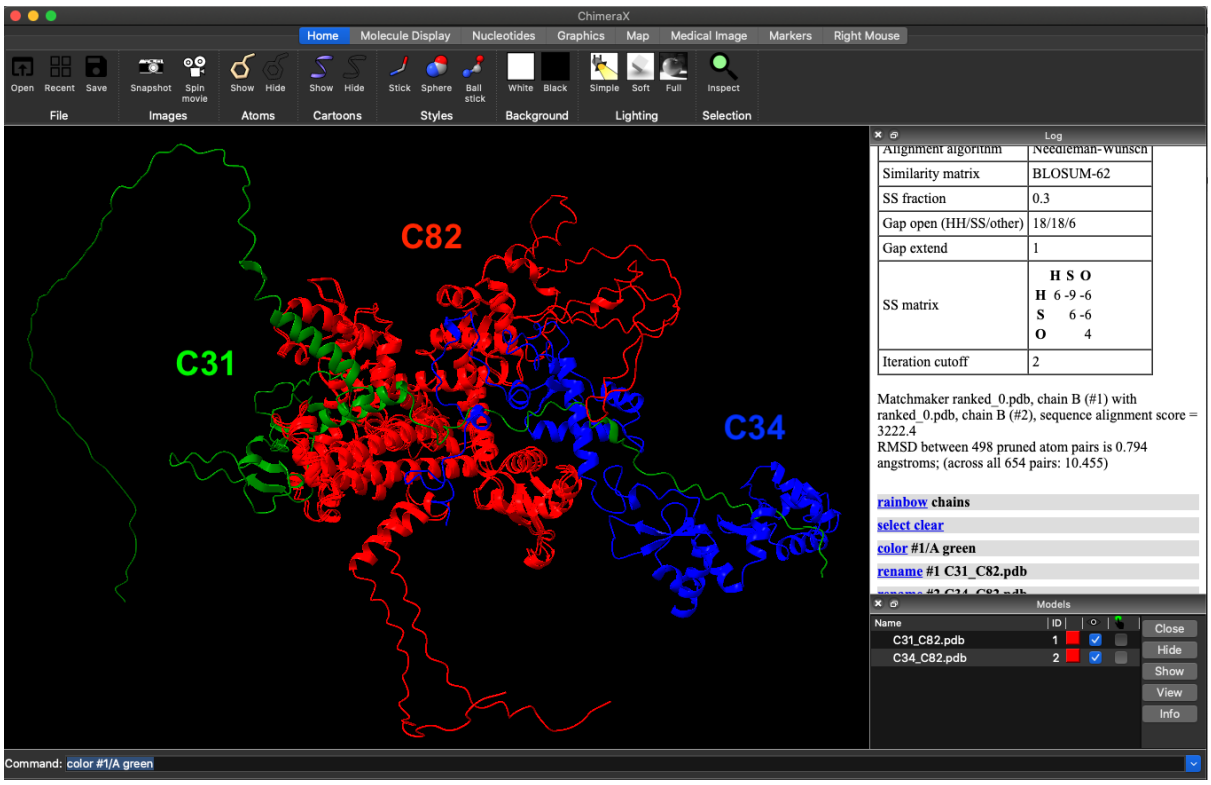


Superpose models of individual subcomplexes

Pre-calculated AlphaFold models of C31, C34, C82 subunits as well as C31-C34, C31-C82, and C34-C82 complexes you can find here **/beegfs/desy/group/school/data/AlphaFold/AlphaFold/results**

1. Open ChimeraX
2. Open models of C31-C82 and C34-C82 dimers
3. Superpose them by typing in the command line and color one of the C31 or C34 subunits e.g. green

```
mm #2 to #1
color #1/A green
```



Do they make sense together?
Try to model trimer C31-C34-C82.

If you want to learn more about using AlphaFold within ChimeraX:
https://www.youtube.com/watch?v=oxblwn0_PMM