HIGH PERFORMANCE RESEARCH COMPUTING

Introduction to AlphaFold for 3D Protein Structure Prediction on Grace

High Performance
Research Computing
DIVISION OF RESEARCH

Fall 2023

AlphaFold for 3D Protein Structure Prediction on Grace

- Resources and Limitations
- Database Files
- Running AlphaFold
  - Google Colab
  - ChimeraX + Google Colab
  - Grace GPU or non-GPU nodes
- Visualization of Results
  - job resource usage
  - view predictions in ChimeraX
  - plotting pLDDT values
- Alternative Workflows

Resource Limitations

- AlphaFold
  - currently AlphaFold can only utilize one GPU
  - about 90% of processing is done on CPU when using DeepMind's workflow
- AlphaFold on HPRC Grace
  - sometimes GPU not detected on certain nodes
- AlphaFold in Google Colab (web browser or ChimeraX app)
  - no guarantee of available resources in Colab
  - runs as a Jupyter notebook on Google Colab cloud servers
    - 12GB RAM max
    - a notebook can run for up to 12 hours per day
      - 24 hours per day with Colab Pro ($9.99/month)
    - not suitable for large predictions

AlphaFold Databases on Grace

/scratch/data/bio/alphafold/2.3.0

```
├── bfd
│   ├── [1.4T]  bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt_a3m.ffdata
│   ├── [1.7G]  bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt_a3m.ffindex
│   ├── [ 16G]  bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt_cs219.ffdata
│   ├── [1.6G]  bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt_cs219.ffindex
│   ├── [304G]  bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt_hhm.ffdata
│   └── [124M]  bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt_hhm.ffindex
├── [4.0K]  example_data
│   ├── [  27]  1L2Y.fasta
│   ├── [ 770]  mmcif_3geh.fa
│   ├── [ 106]  T1083.fasta
│   └── [ 187]  T1083_T1084_multimer.fasta
├── mgnify
│   └── [120G]  mgy_clusters_2022_05.fa
├── params
│   ├── [ 18K]  LICENSE
│   ├── [356M]  params_model_1_multimer_v2.npz
│   ├── [356M]  params_model_1.npz
│   ├── [356M]  params_model_1_ptm.npz
│   ├── [356M]  params_model_2_multimer_v2.npz
│   ├── [356M]  params_model_2.npz
│   ├── [356M]  params_model_2_ptm.npz
│   ├── [356M]  params_model_3_multimer_v2.npz
│   ├── [354M]  params_model_3.npz
│   ├── [355M]  params_model_3_ptm.npz
│   ├── [356M]  params_model_4_multimer_v2.npz
│   ├── [354M]  params_model_4.npz
│   ├── [355M]  params_model_4_ptm.npz
│   ├── [356M]  params_model_5_multimer_v2.npz
│   ├── [354M]  params_model_5.npz
│   └── [355M]  params_model_5_ptm.npz
```

total size: 2.8TB
number of files: 199,000+

```
├── pdb70
│   ├── [ 410]  md5sum
│   ├── [ 53G]  pdb70_a3m.ffdata
│   ├── [2.0M]  pdb70_a3m.ffindex
│   ├── [6.6M]  pdb70_clu.tsv
│   ├── [ 21M]  pdb70_cs219.ffdata
│   ├── [1.5M]  pdb70_cs219.ffindex
│   ├── [3.2G]  pdb70_hhm.ffdata
│   ├── [1.8M]  pdb70_hhm.ffindex
│   └── [ 19M]  pdb_filter.dat
├── pdb_mmcif
│   ├── [9.4M]  mmcif_files
│   └── [143K]  obsolete.dat
├── pdb_seqres
│   └── [230M]  pdb_seqres.txt
├── uniclust30
│   ├── [179G]  UniRef30_2022_02_a3m.ffdata
│   ├── [830M]  UniRef30_2022_02_a3m.ffindex
│   ├── [7.5G]  UniRef30_2022_02_cs219.ffdata
│   ├── [749M]  UniRef30_2022_02_cs219.ffindex
│   ├── [ 42G]  UniRef30_2022_02_hhm.ffdata
│   ├── [ 24M]  UniRef30_2022_02_hhm.ffindex
│   └── [ 379]  UniRef30_2022_02.md5sums
├── uniprot
│   └── [104G]  uniprot.fasta
├── [4.0K]  uniref30
│   ├── [160G]  UniRef30_2021_03_a3m.ffdata
│   ├── [758M]  UniRef30_2021_03_a3m.ffindex
│   ├── [6.7G]  UniRef30_2021_03_cs219.ffdata
│   ├── [683M]  UniRef30_2021_03_cs219.ffindex
│   ├── [ 38G]  UniRef30_2021_03_hhm.ffdata
│   ├── [ 22M]  UniRef30_2021_03_hhm.ffindex
│   └── [ 379]  UniRef30_2021_03.md5sums
└── uniref90
    └── [ 67G]  uniref90.fasta
```

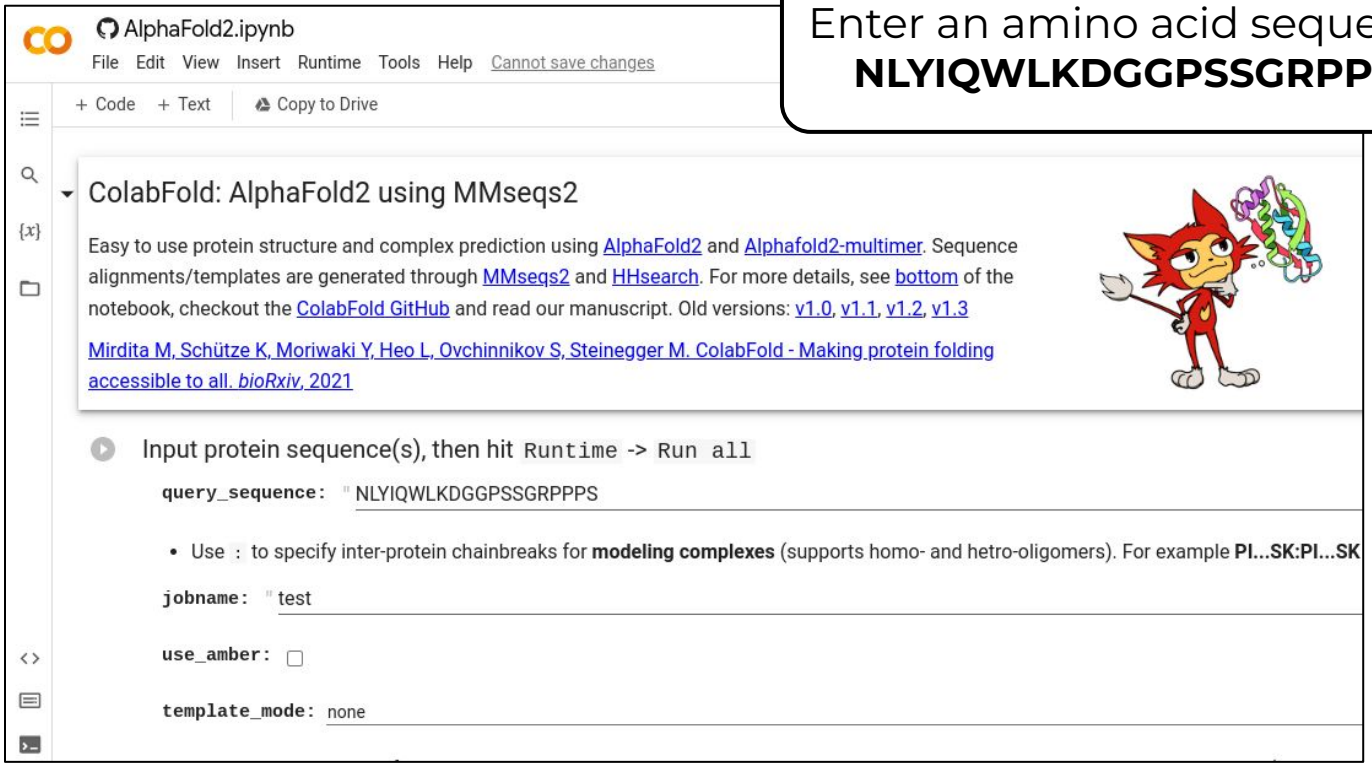
Resources for Running AlphaFold

- Run as a Jupyter Notebook on Google Colab in web browser
- Run as a Jupyter Notebook on Google Colab in ChimeraX
  - ChimeraX Interactive App on HPRC Grace Portal
    - https://portal-grace.hprc.tamu.edu
- Run as a Slurm job script on Grace

https://hprc.tamu.edu/kb/Software/AlphaFold

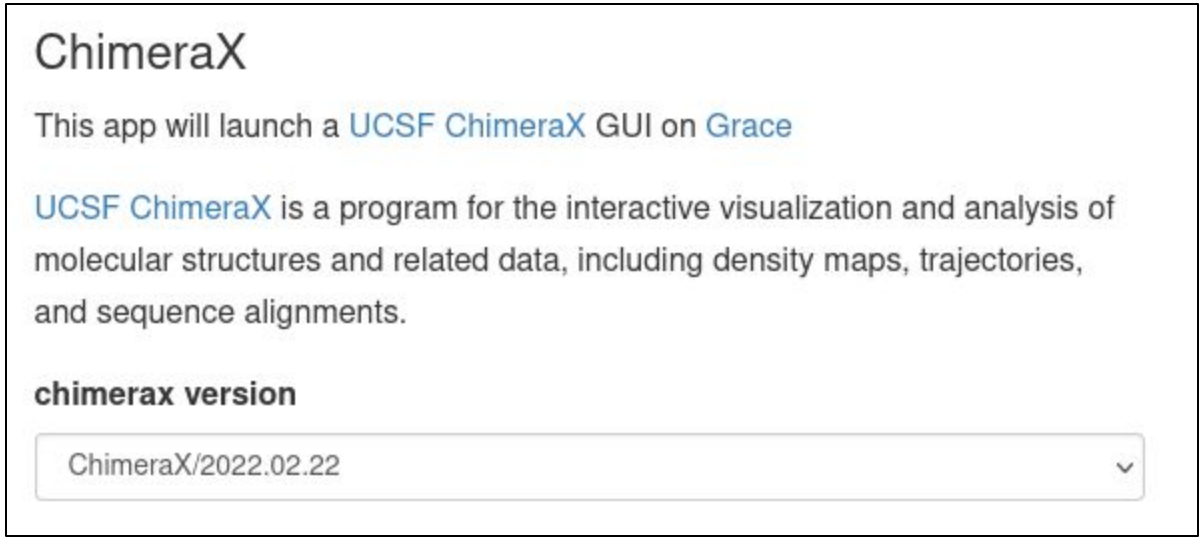
ColabFold AlphaFold2 Jupyter Notebook

Enter an amino acid sequence
NLYIQWLKDGGPSSGRPPPS



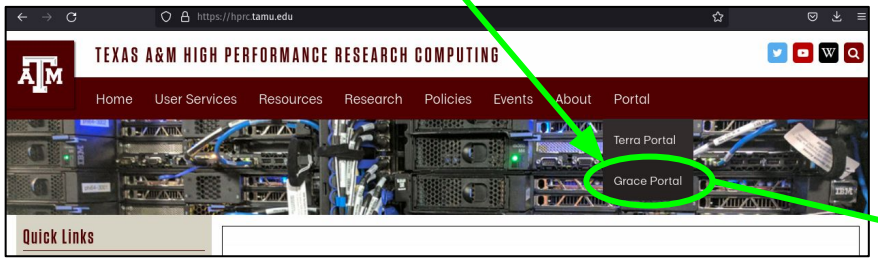
ChimeraX

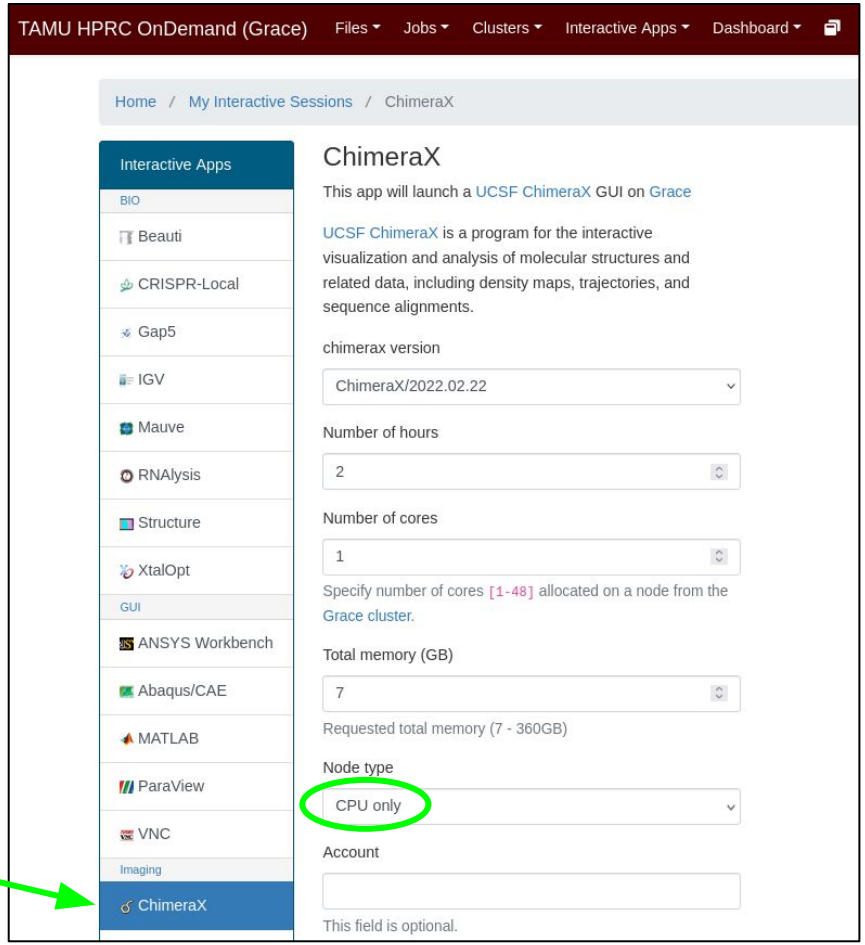
- Can be used to visualize protein structures
- Can be launched using the Grace portal
  - portal-grace.hprc.tamu.edu
- Can be used to run AlphaFold using the daily build version (2022.02.22+)
  - uses Google Colab with limited resources



ChimeraX

- Launch a ChimeraX job on the Grace portal portal-grace.hprc.tamu.edu
  - select Node Type: CPU only
  - AlphaFold runs on Google Colab GPUs so we can use a non-GPU for running ChimeraX
- ChimeraX will be used later for the following
  - run AlphaFold in Google Colab
  - visualize results from an AlphaFold job on Grace





Running AlphaFold on Grace

- Can be run as a job script requesting one GPU
- Shared databases are available: 2.6TB total size

AlphaFold 2.2.3	monomer_ptm 20 aa	multimer 98 & 73 aa
A100	36 minutes	4 hours 49 minutes
A40	35 minutes	-
RTX 6000	33 minutes	4 hours 44 minutes
T4	33 minutes	4 hours 45 minutes
CPU only	40 minutes	6 hours 11 minutes

- Can be run in ChimeraX from the Grace portal
  - using the ChimeraX AlphaFold option
  - all processing done on Google cloud servers

Viewing Maximum Available Resources

The **maxconfig** command will show the recommended Slurm parameters for the maximum available resources (cores, memory, time) per node for a specified accelerator or partition (default Grace partition: long) and show SUs charged

```
[username@grace ~]$ maxconfig

  Grace partitions:   short  medium  long  xlong  vnc  gpu  bigmem  special  gpu-a40
  Grace GPUs in gpu partition:  a100:2  a40:3  rtx:2  t4:4

  Showing max parameters (cores, mem, time) for partition long

  CPU-billing * hours * nodes =   SUs
           48 *   168 *     1 = 8,064

#!/bin/bash
#SBATCH --job-name=my_job
#SBATCH --time=7-00:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=48
#SBATCH --mem=360G
#SBATCH --output=stdout.%x.%j
#SBATCH --error=stderr.%x.%j
```

Viewing Maximum Available Resources

See the recommended Slurm parameters for requesting 1 x A100 GPU with ½ the total CPUs and memory since there are 2 x A100s per node

```
[username@grace ~]$ maxconfig -g a100 -G 1

  Grace partitions:   short  medium  long  xlong  vnc  gpu  bigmem  special  gpu-a40
  Grace GPUs in gpu partition:  a100:2  a40:3  rtx:2  t4:4

  Showing 1/2 of total cores and memory for using 1 x a100 GPU

  (CPU-billing + (GPU-billing * GPU-count)) * hours * nodes = SUs
  (         25 + (         72 *         1)) *    96 *     1 = 9,312

#!/bin/bash
#SBATCH --job-name=my_job
#SBATCH --time=4-00:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=24
#SBATCH --mem=180G
#SBATCH --gres=gpu:a100:1
#SBATCH --output=stdout.%x.%j
#SBATCH --error=stderr.%x.%j
```

Grace SUs Charged: GPU vs CPU-only Jobs

show SU rate for 1 day CPU-only

```
maxconfig -d 1
```

```
CPU-billing * hours * nodes =   SUs
         48 *    24 *     1 =  1,152
```

show SU rate for 1 x A100 GPU for 1 day

```
maxconfig -g a100 -G 1 -d 1
```

```
(CPU-billing + (GPU-billing * GPU-count)) * hours * nodes = SUs
(         25 + (         72 *         1)) *    24 *     1 = 2,328
```

show SU charge rate

```
maxconfig -h
```

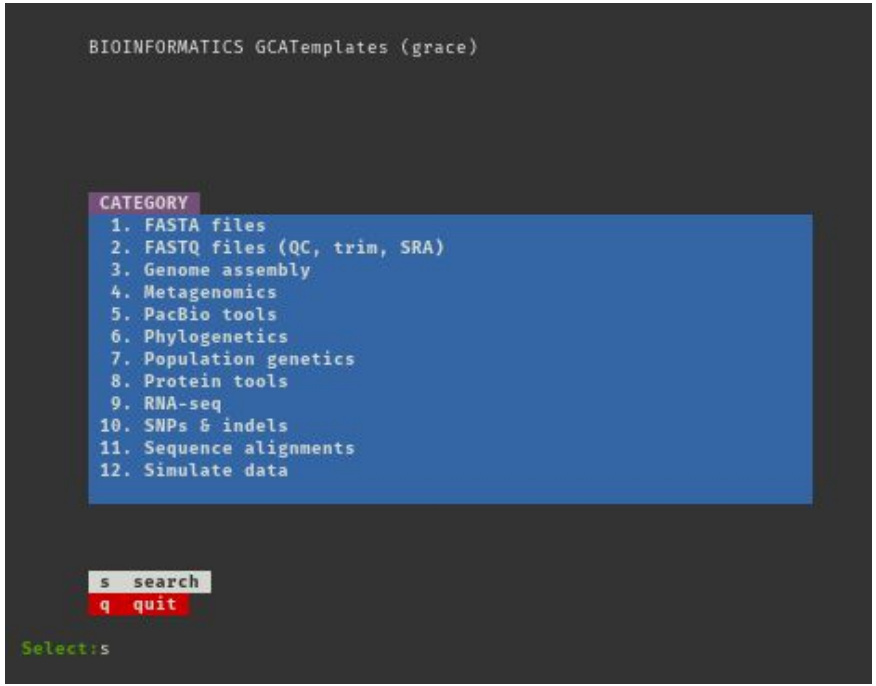
```
SU rate per GPU:
      GPU   SUs per hour
      ----  ------------
      a100        72
      a40         48
      rtx         48
      t4          24
```

Finding AlphaFold template job scripts using GCATemplates on Grace

- Genomic Computational Analysis Templates have example input data so you can run the script for demo purposes

```
mkdir $SCRATCH/af_demo
cd $SCRATCH/af_demo
gcatemplates
```

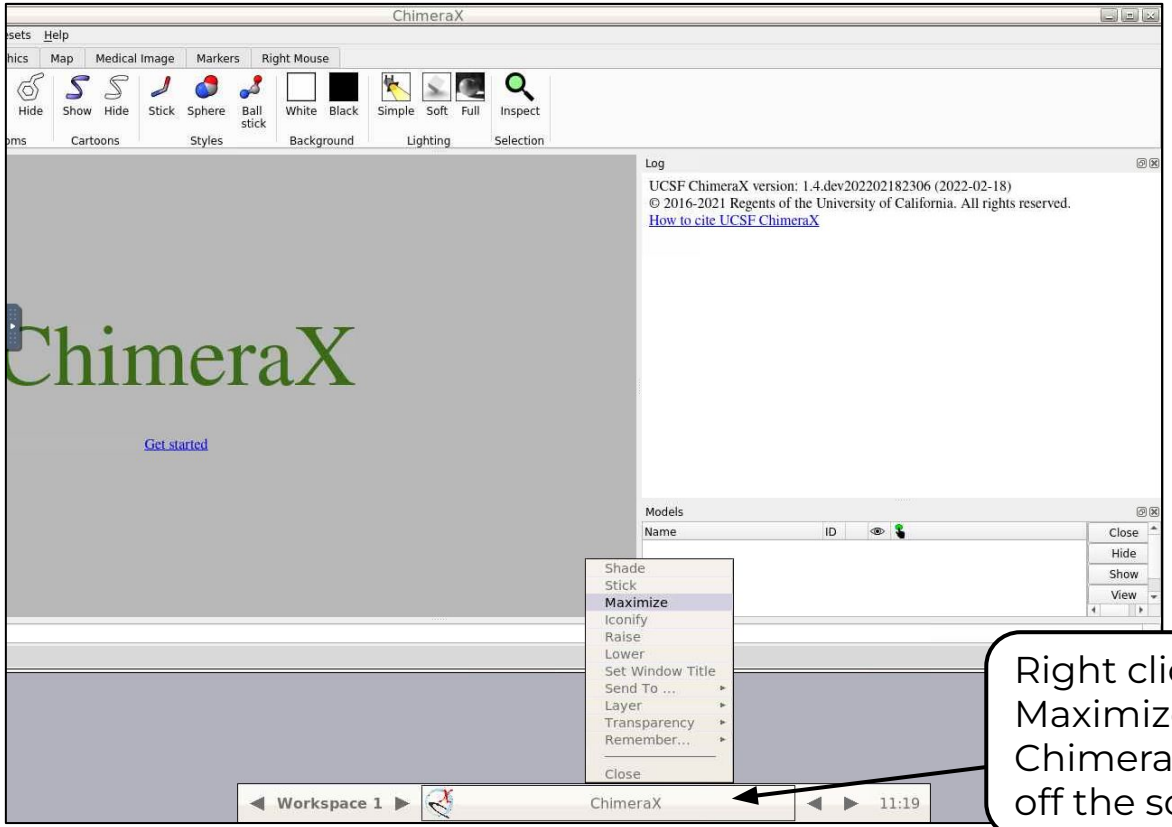
- Type **s** for search then enter **alphafold** to search for the alphafold **2.2.3** template script and select the **reduced_dbs** script
- Review the script and submit the job script which takes about 35 minutes to complete using 1 x A100




```
sbatch run_alphafold_2.2.3_reduced_dbs_monomer_ptm_grace.sh
```

AlphaFold with ChimeraX + Google Colab

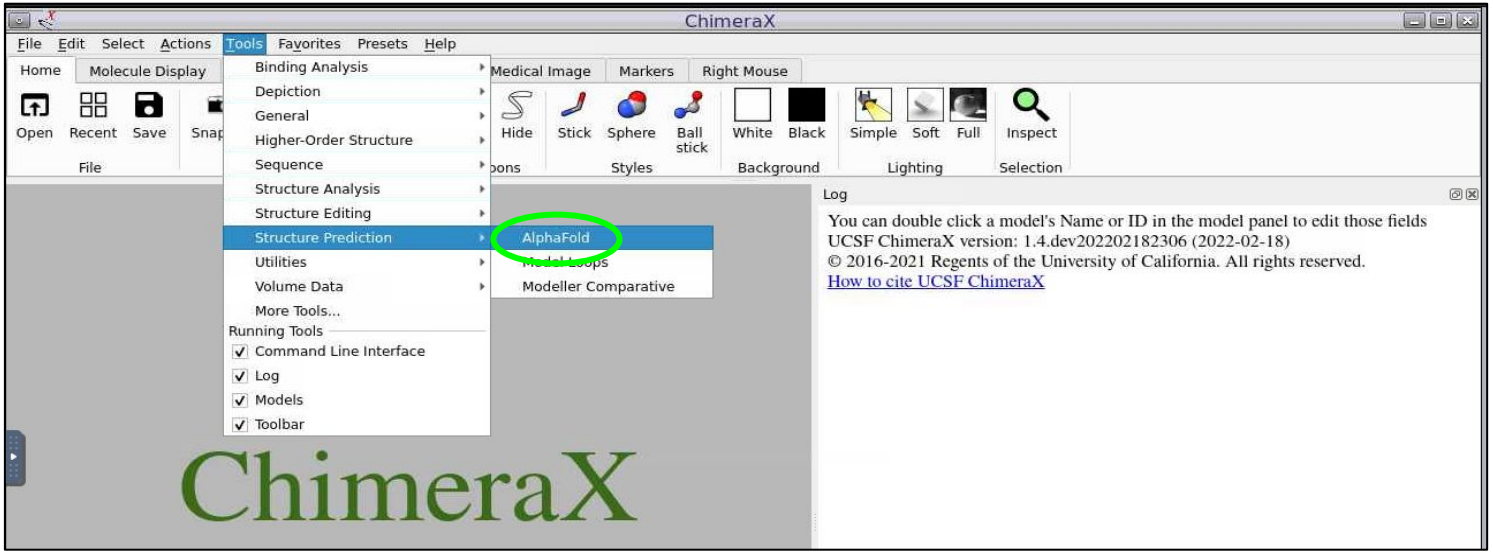
Maximize ChimeraX Window



Right click and select Maximize if the ChimeraX window is off the screen

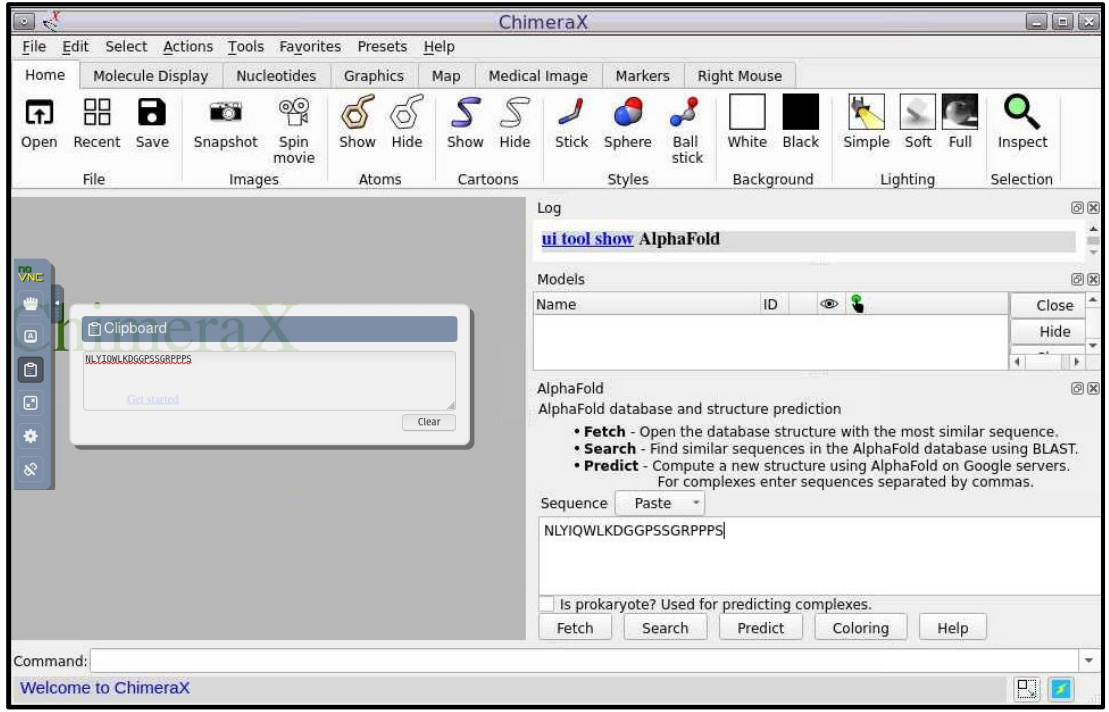
AlphaFold with ChimeraX

- Launch ChimeraX using the HPRC Grace portal
- Select the AlphaFold Structure Prediction option



AlphaFold with ChimeraX

- Enter an amino acid sequence **NLYIQWLKDGGPSSGRPPPS**
  - or paste in Clipboard first then paste in Sequence field
- Click Predict
- A Google Colab page will start and prompt you for your Google login
- Login to your Google account to begin processing
- Prediction completes in about 1 hour
- Not ideal for large prediction jobs



AlphaFold Grace Job Scripts

Example AlphaFold Job Script

- **multimer**
  - dbs in **red** are required for multimer
- AlphaFold can only use one GPU so reserve half the CPU and memory resources so another job can use the other GPU
  - Grace compute nodes have 360GB of available memory and 48 cores

```
#!/bin/bash
#SBATCH --job-name=alphafold        # job name
#SBATCH --time=2-00:00:00           # max job run time dd-hh:mm:ss
#SBATCH --ntasks-per-node=1         # tasks (commands) per compute node
#SBATCH --cpus-per-task=24          # CPUs (threads) per command
#SBATCH --mem=180G                  # total memory per node
#SBATCH --gres=gpu:a100:1           # request 1 A100 GPU
#SBATCH --output=stdout.%x.%j       # save stdout to file
#SBATCH --error=stderr.%x.%j        # save stderr to file

module purge

export SINGULARITYENV_TF_FORCE_UNIFIED_MEMORY=1
export SINGULARITYENV_XLA_PYTHON_CLIENT_MEM_FRACTION=4.0

DOWNLOAD_DIR=/scratch/data/bio/alphafold/2.3.0

# run jobstats in the background (&) to monitor cpu and gpu usage
jobstats &

singularity exec --nv /sw/hprc/sw/containers/alphafold/alphafold_2.3.1.sif \
 python /app/alphafold/run_alphafold.py \
 --use_gpu_relax \
 --data_dir=$DOWNLOAD_DIR \
 --uniref90_database_path=$DOWNLOAD_DIR/uniref90/uniref90.fasta \
 --mgnify_database_path=$DOWNLOAD_DIR/mgnify/mgy_clusters_2018_12.fa \
 --uniclust30_database_path=$DOWNLOAD_DIR/uniclust30/uniclust30_2021_03/UniRef30_2021_03 \
 --bfd_database_path=$DOWNLOAD_DIR/bfd/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt \
 --model_preset=multimer \
 --pdb_seqres_database_path=$DOWNLOAD_DIR/pdb_seqres/pdb_seqres.txt \
 --uniprot_database_path=$DOWNLOAD_DIR/uniprot/uniprot.fasta \
 --template_mmcif_dir=$DOWNLOAD_DIR/pdb_mmcif/mmcif_files \
 --obsolete_pdbs_path=$DOWNLOAD_DIR/pdb_mmcif/obsolete.dat \
 --max_template_date=2022-1-1 \
 --db_preset=full_dbs \
 --output_dir=out_alphafold \
 --fasta_paths=$DOWNLOAD_DIR/example_data/T1083_T1084_multimer.fasta

# run jobstats to create a graph of cpu and gpu usage for this job
jobstats
```

Example AlphaFold Job Script

- **monomer**
  - dbs in **red** required for monomer
- **monomer_ptm**
  - will produce pTM scores that can be graphed using AlphaPickle
- AlphaFold can only use one GPU so reserve half the CPU and memory resources so another job can use the other GPU

```
#!/bin/bash
#SBATCH --job-name=alphafold        # job name
#SBATCH --time=2-00:00:00           # max job run time dd-hh:mm:ss
#SBATCH --ntasks-per-node=1         # tasks (commands) per compute node
#SBATCH --cpus-per-task=24          # CPUs (threads) per command
#SBATCH --mem=180G                  # total memory per node
#SBATCH --gres=gpu:a100:1           # request 1 A100 GPU
#SBATCH --output=stdout.%x.%j       # save stdout to file
#SBATCH --error=stderr.%x.%j        # save stderr to file

module purge

export SINGULARITYENV_TF_FORCE_UNIFIED_MEMORY=1
export SINGULARITYENV_XLA_PYTHON_CLIENT_MEM_FRACTION=4.0

DOWNLOAD_DIR=/scratch/data/bio/alphafold/2.3.0

# run jobstats in the background (&) to monitor cpu and gpu usage
jobstats &

singularity exec --nv /sw/hprc/sw/containers/alphafold/alphafold_2.3.1.sif \
 python /app/alphafold/run_alphafold.py \
 --use_gpu_relax \
 --data_dir=$DOWNLOAD_DIR \
 --uniref90_database_path=$DOWNLOAD_DIR/uniref90/uniref90.fasta \
 --mgnify_database_path=$DOWNLOAD_DIR/mgnify/mgy_clusters_2018_12.fa \
 --uniclust30_database_path=$DOWNLOAD_DIR/uniclust30/uniclust30_2021_03/UniRef30_2021_03 \
 --bfd_database_path=$DOWNLOAD_DIR/bfd/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt \
 --model_preset=monomer \
 --pdb70_database_path=$DOWNLOAD_DIR/pdb70/pdb70 \
 --template_mmcif_dir=$DOWNLOAD_DIR/pdb_mmcif/mmcif_files \
 --obsolete_pdbs_path=$DOWNLOAD_DIR/pdb_mmcif/obsolete.dat \
 --max_template_date=2022-1-1 \
 --db_preset=full_dbs \
 --output_dir=out_alphafold \
 --fasta_paths=$DOWNLOAD_DIR/example_data/T1083.fasta

# run jobstats to create a graph of cpu and gpu usage for this job
jobstats
```

Example AlphaFold Job Script

- **monomer + reduced_dbs**
- dbs in **red** required for monomer + reduced_dbs
- small_bfd_database is a subset of BFD and is generated by taking the first non-consensus sequence from every cluster in BFD
- AlphaFold can only use one GPU so reserve half the CPU and memory resources so another job can use the other GPU

```
#!/bin/bash
#SBATCH --job-name=alphafold        # job name
#SBATCH --time=2-00:00:00           # max job run time dd-hh:mm:ss
#SBATCH --ntasks-per-node=1         # tasks (commands) per compute node
#SBATCH --cpus-per-task=24          # CPUs (threads) per command
#SBATCH --mem=180G                  # total memory per node
#SBATCH --gres=gpu:a100:1           # request 1 A100 GPU
#SBATCH --output=stdout.%x.%j       # save stdout to file
#SBATCH --error=stderr.%x.%j        # save stderr to file

module purge

export SINGULARITYENV_TF_FORCE_UNIFIED_MEMORY=1
export SINGULARITYENV_XLA_PYTHON_CLIENT_MEM_FRACTION=4.0

DOWNLOAD_DIR=/scratch/data/bio/alphafold/2.3.0

# run jobstats in the background (&) to monitor cpu and gpu usage
jobstats &

singularity exec --nv /sw/hprc/sw/containers/alphafold/alphafold_2.3.1.sif \
 python /app/alphafold/run_alphafold.py \
 --use_gpu_relax \
 --data_dir=$DOWNLOAD_DIR \
 --uniref90_database_path=$DOWNLOAD_DIR/uniref90/uniref90.fasta \
 --mgnify_database_path=$DOWNLOAD_DIR/mgnify/mgy_clusters_2022_05.fa \
 --small_bfd_database_path=$DOWNLOAD_DIR/small_bfd/bfd-first_non_consensus_sequences.fasta \
 --model_preset=monomer \
 --pdb70_database_path=$DOWNLOAD_DIR/pdb70/pdb70 \
 --template_mmcif_dir=$DOWNLOAD_DIR/pdb_mmcif/mmcif_files \
 --obsolete_pdbs_path=$DOWNLOAD_DIR/pdb_mmcif/obsolete.dat \
 --max_template_date=2023-1-1 \
 --db_preset=reduced_dbs \
 --output_dir=out_alphafold \
 --fasta_paths=$DOWNLOAD_DIR/example_data/1L2Y.fasta

# run jobstats to create a graph of cpu and gpu usage for this job
jobstats
```

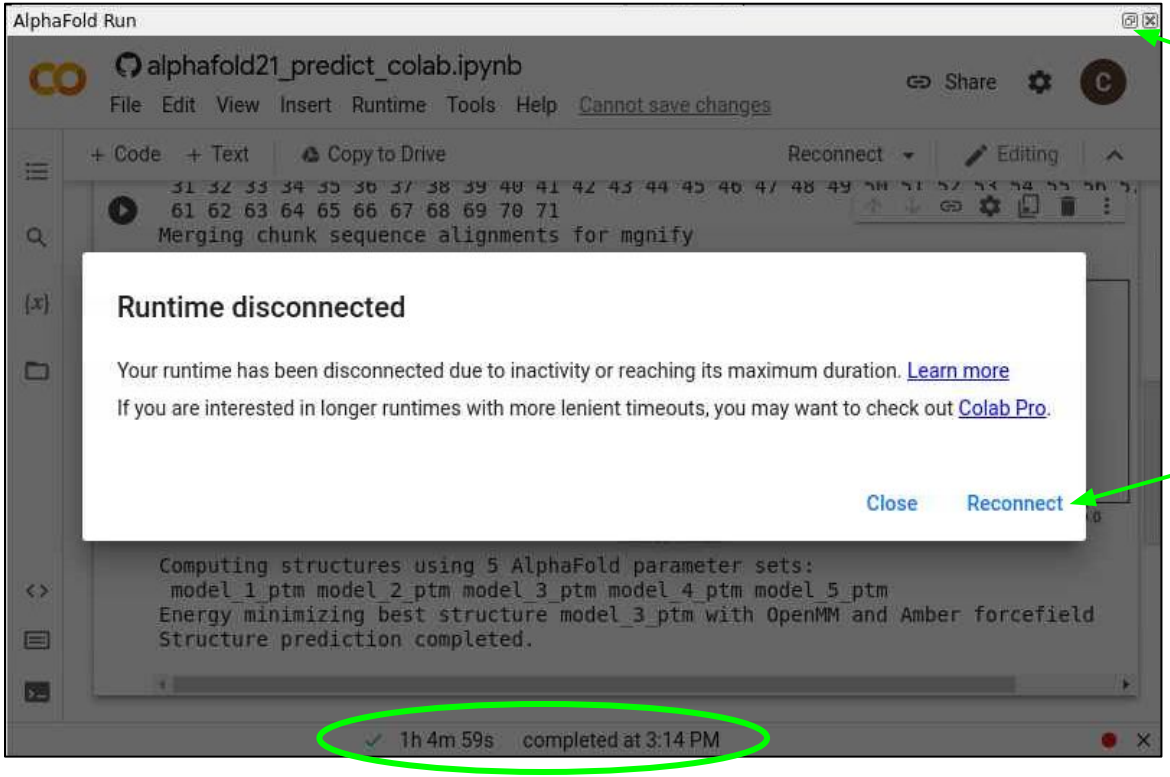
Unified Memory

```
#!/bin/bash
#SBATCH --job-name=alphafold          # job name
#SBATCH --time=2-00:00:00             # max job run time dd-hh:mm:ss
#SBATCH --ntasks-per-node=1           # tasks (commands) per compute node
#SBATCH --cpus-per-task=24            # CPUs (threads) per command
#SBATCH --mem=180G                    # total memory per node
#SBATCH --gres=gpu:a100:1             # request 1 A100 GPU
#SBATCH --output=stdout.%x.%j         # save stdout to file
#SBATCH --error=stderr.%x.%j          # save stderr to file

export SINGULARITYENV_TF_FORCE_UNIFIED_MEMORY=1
export SINGULARITYENV_XLA_PYTHON_CLIENT_MEM_FRACTION=4.0
```

- unified memory can be used to request more than just the total GPU memory for the JAX step in AlphaFold
  - A100 GPU has 40GB memory
  - GPU total memory (40) * `XLA_PYTHON_CLIENT_MEM_FRACTION` (4.0)
  - `XLA_PYTHON_CLIENT_MEM_FRACTION` default = 0.9
- this example script has 160 GB of unified memory
  - 40 GB from A100 GPU + 120 GB DDR from motherboard

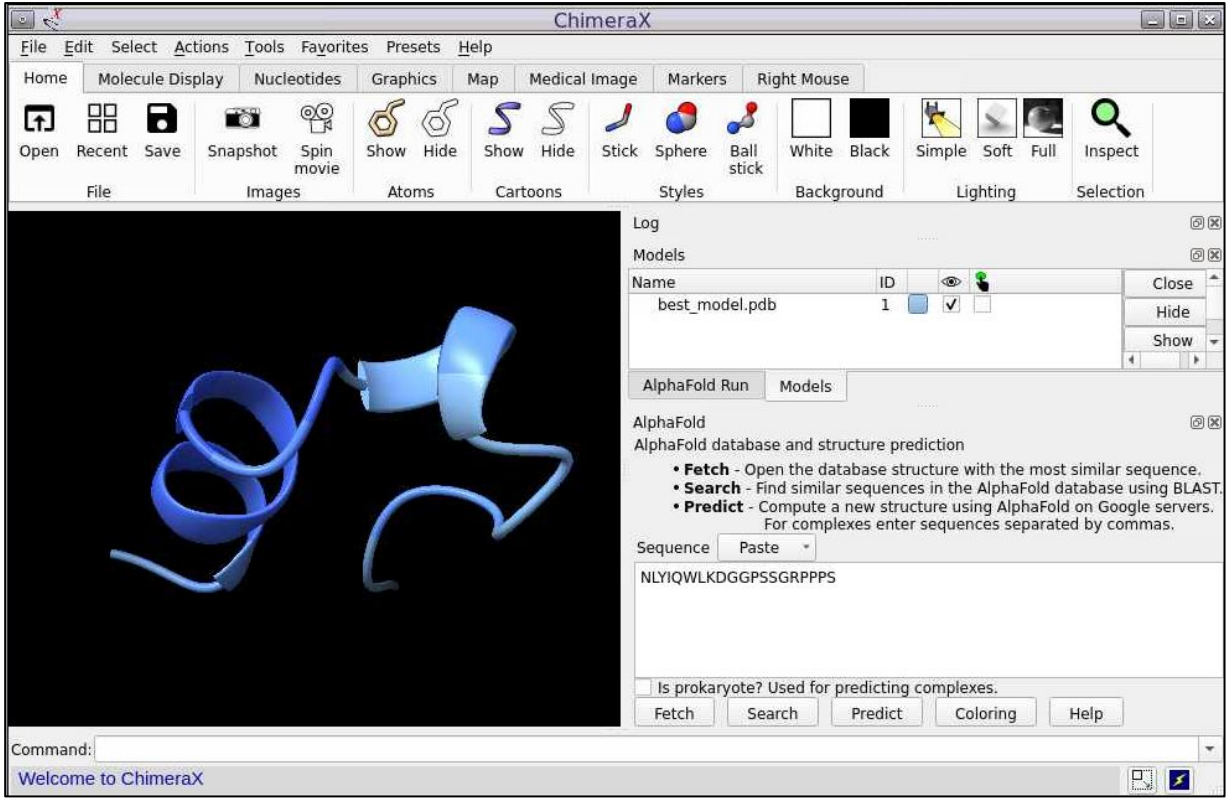
AlphaFold Results Visualization

Visualize Results with ChimeraX



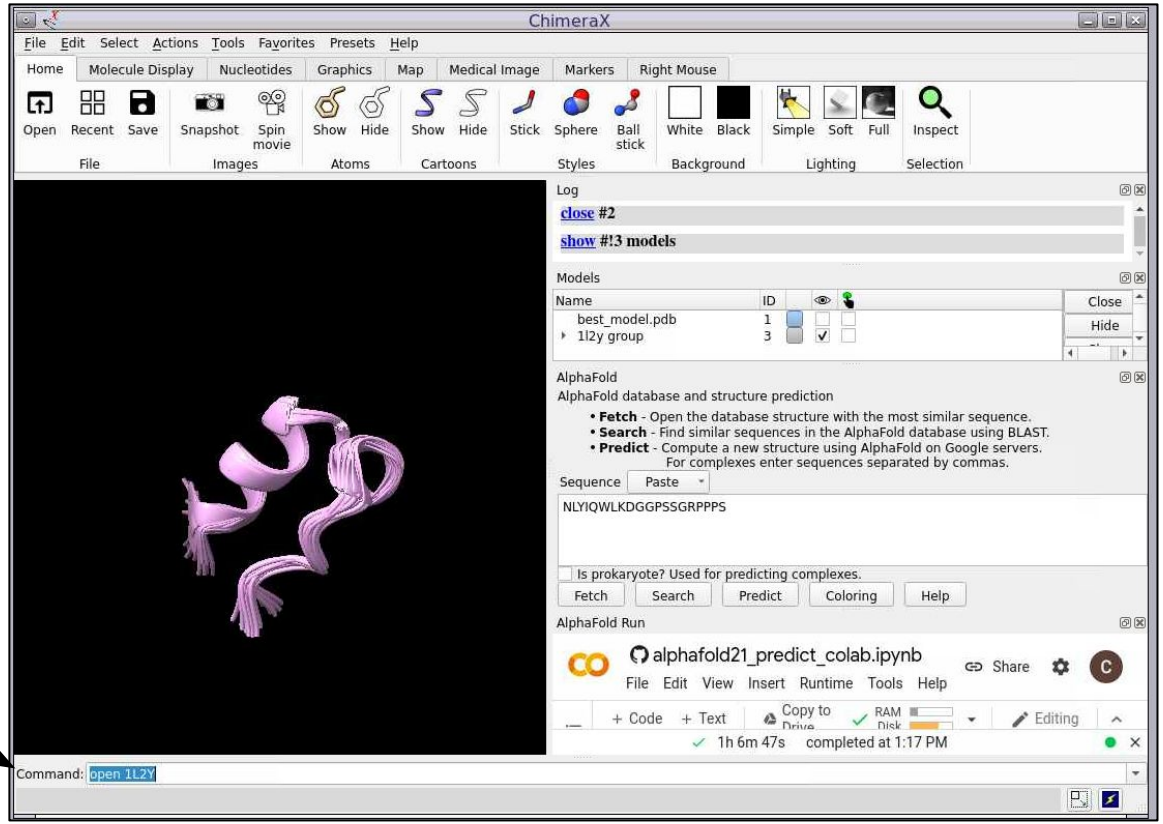
Click the minimize button to return to ChimeraX or click Reconnect then minimize

Visualize AlphaFold Google Colab Results with ChimeraX

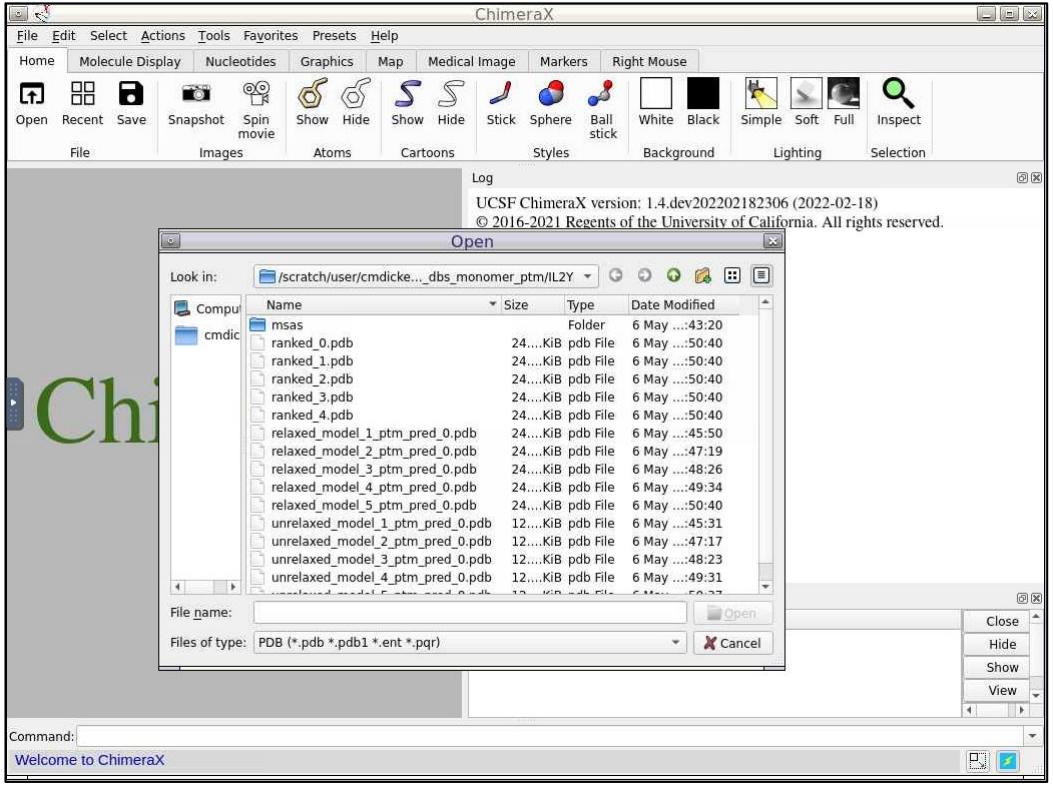


View PDB Structure if Available

type ***open*** and the protein name ***1L2Y***



Visualize AlphaFold Grace Results with ChimeraX



AlphaFold Confidence Metrics

AlphaPickle for Visualization of Confidence Scores

AlphaPickle can be used to create graphs for pLDDT and PAE scores

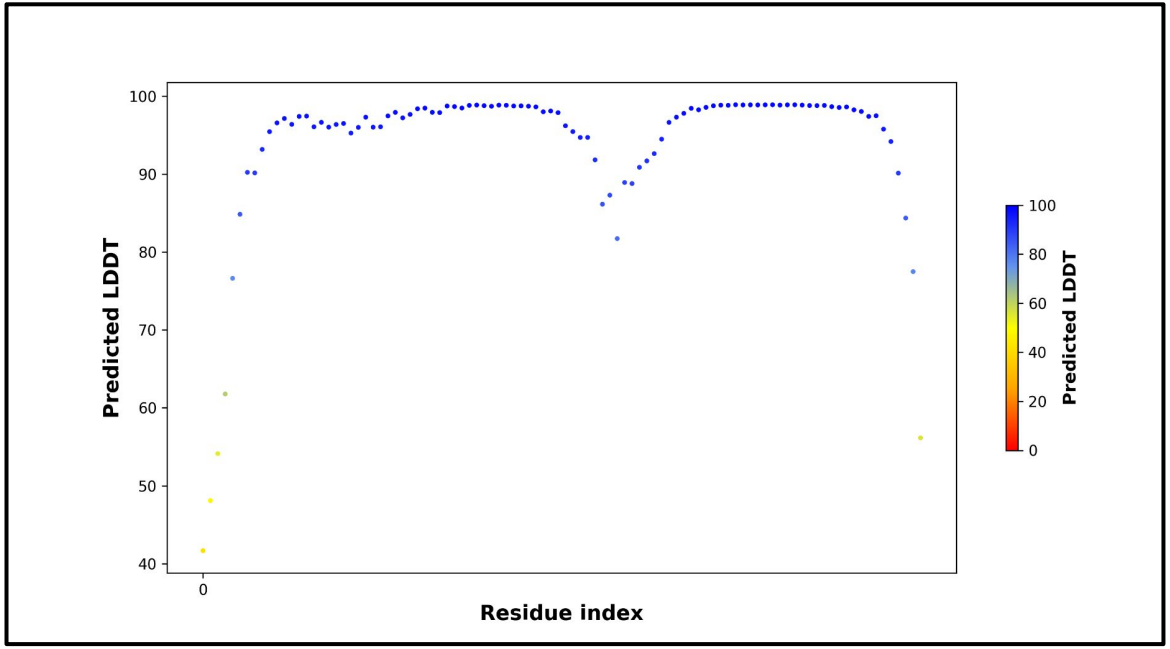
- graphing PAE scores is only available for the **monomer_ptm** and **multimer** model presets
- load the AlphaPickle module at the beginning of the job script
- run AlphaPickle at the end specifying the output directory used in the run_alphafold.py command

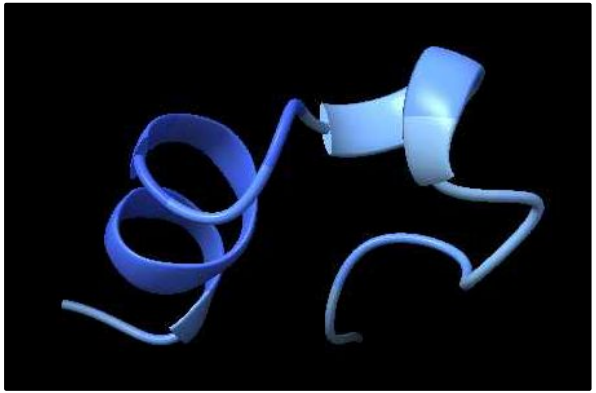
- pLDDT: scale from 0 - 100 of per-residue estimate of prediction confidence
- PAE: Predicted Alignment Error

https://github.com/mattarnoldbio/alphapickle

```
#!/bin/bash
#SBATCH --job-name=alphafold        # job name
#SBATCH --time=2-00:00:00           # max job run time dd-hh:mm:ss
#SBATCH --ntasks-per-node=1         # tasks (commands) per compute node
#SBATCH --cpus-per-task=24          # CPUs (threads) per command
#SBATCH --mem=180G                  # total memory per node
#SBATCH --gres=gpu:a100:1           # request 1 A100 GPU
#SBATCH --output=stdout.%x.%j       # save stdout to file
#SBATCH --error=stderr.%x.%j        # save stderr to file

module purge
module load GCC/10.2.0  CUDA/11.1.1  OpenMPI/4.0.5  AlphaPickle/1.4.1

export SINGULARITYENV_TF_FORCE_UNIFIED_MEMORY=1
export SINGULARITYENV_XLA_PYTHON_CLIENT_MEM_FRACTION=4.0
DOWNLOAD_DIR=/scratch/data/bio/alphafold/2.3.0
# run jobstats in the background (&) to monitor cpu and gpu usage
jobstats &

singularity exec --nv /sw/hprc/sw/containers/alphafold/alphafold_2.3.1.sif \
 python /app/alphafold/run_alphafold.py \
 --use_gpu_relax \
 --data_dir=$DOWNLOAD_DIR \
 --uniref90_database_path=$DOWNLOAD_DIR/uniref90/uniref90.fasta \
 --mgnify_database_path=$DOWNLOAD_DIR/mgnify/mgy_clusters_2022_05.fa \
 --small_bfd_database_path=$DOWNLOAD_DIR/small_bfd/bfd-first_non_consensus_sequences.fasta \
 --model_preset=monomer_ptm \
 --pdb70_database_path=$DOWNLOAD_DIR/pdb70/pdb70 \
 --template_mmcif_dir=$DOWNLOAD_DIR/pdb_mmcif/mmcif_files \
 --obsolete_pdbs_path=$DOWNLOAD_DIR/pdb_mmcif/obsolete.dat \
 --max_template_date=2023-1-1 \
 --db_preset=reduced_dbs \
 --output_dir=out_alphafold_2.3.1 \
 --fasta_paths=$DOWNLOAD_DIR/example_data/1L2Y.fasta

# graph pLDDT and PAE .pkl files; fasta sequence name will be part of the output directory name
run_AlphaPickle.py -od out_alphafold_2.3.1/1L2Y

# run jobstats to create a graph of cpu and gpu usage for this job
jobstats
```

Visualize AlphaFold pLDDT Scores



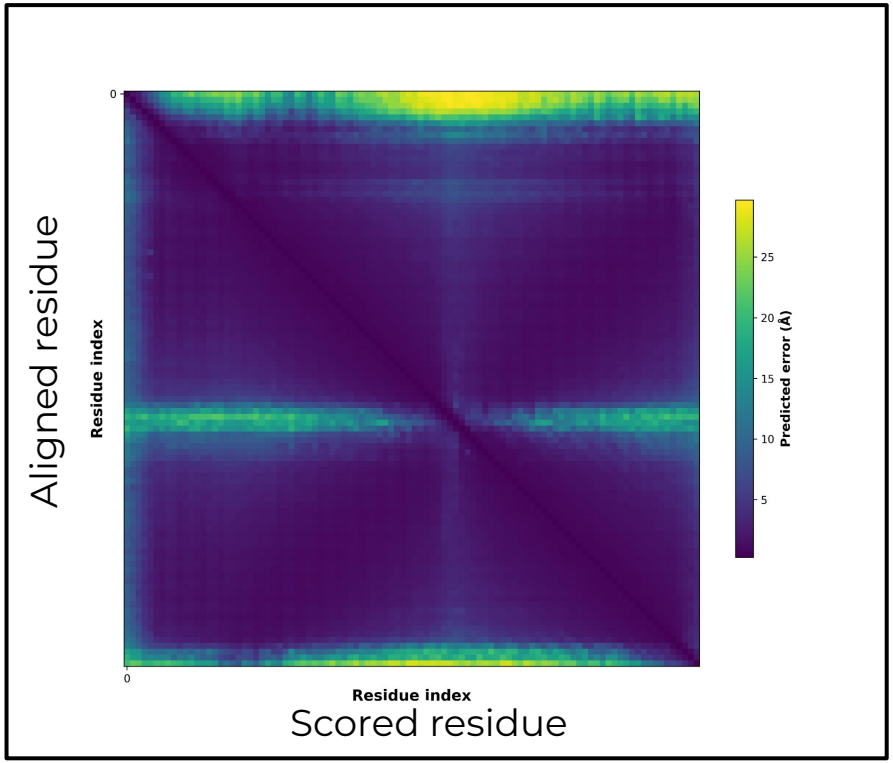


> 90 = Very high
70 - 90 = Confident
50 - 70 = Low
< 50 = Very low

```
eog out_alphafold_2.3.1/1L2Y/ranked_0_pLDDT.png
```

you may get different results compared to the image above when using reduced_dbs

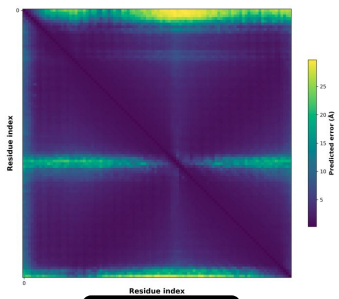
Visualize AlphaFold PAE Results (monomer_ptm)



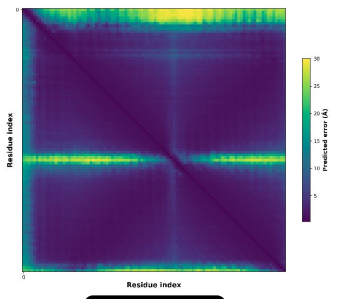
- Low Predicted Aligned Error (PAE) value has higher confidence of accuracy
- Must use monomer_ptm or multimer as model_preset to create PAE image
- The colour at position (x, y) indicates AlphaFold's expected position error at residue x, when the predicted and true structures are aligned on residue y.

```
eog out_alphafold_2.3.1/1L2Y/ranked_0_PAE.png
```

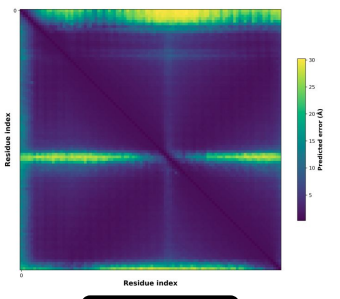
Evaluating Models



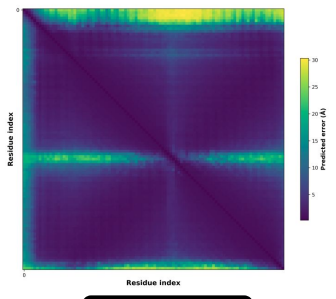
ranked_0



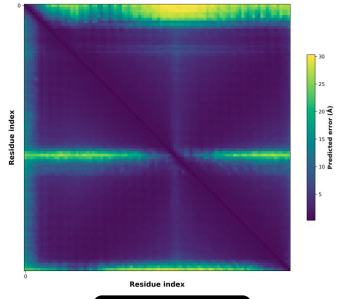
ranked_1



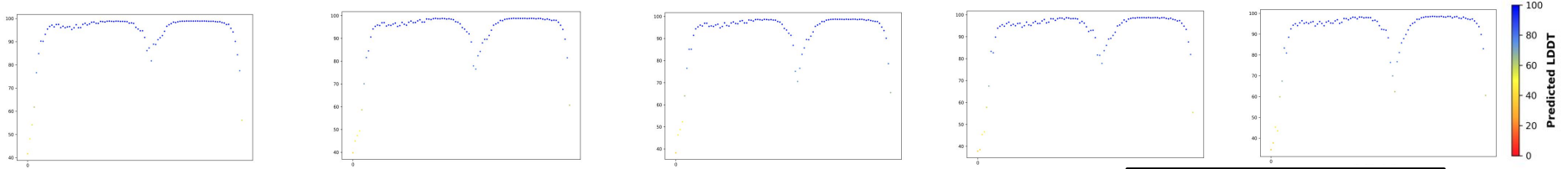
ranked_2



ranked_3



ranked_4



see which model has the top rank based on pLDDT score

```
cat out_alphafold_2.3.1/IL2Y/ranking_debug.json
```


"plddts": {
  "model_1_ptm_pred_0": 94.16585478746399,
  "model_2_ptm_pred_0": 94.64120852328334,
  "model_3_ptm_pred_0": 89.94980057627299,
  "model_4_ptm_pred_0": 77.53515668415058,
  "model_5_ptm_pred_0": 88.40610380463586
},
"order": [
  "model_2_ptm_pred_0",
  "model_1_ptm_pred_0",
  "model_3_ptm_pred_0",
  "model_5_ptm_pred_0",
  "model_4_ptm_pred_0"
]

ranked_0

ranked_4

AlphaFold Job Resource Monitoring

Review CPU usage for a Job

The **seff** command displays CPU and memory resource usage and efficiency

```
seff 8862586
```

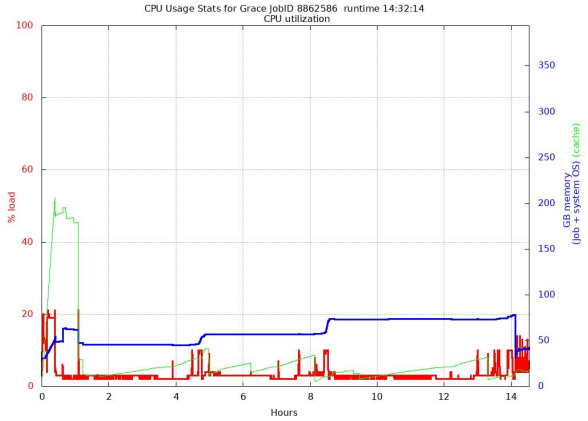
```
Job ID: 8862586
Cluster: grace
User/Group: username/username
State: COMPLETED (exit code 0)
Nodes: 1
Cores per node: 24
CPU Utilized: 08:19:34
CPU Efficiency: 2.39% of 14-12:55:36 core-walltime
Job Wall-clock time: 14:32:19
Memory Utilized: 51.43 GB
Memory Efficiency: 28.57% of 180.00 GB
```

usage stats are not accurate until the job is complete

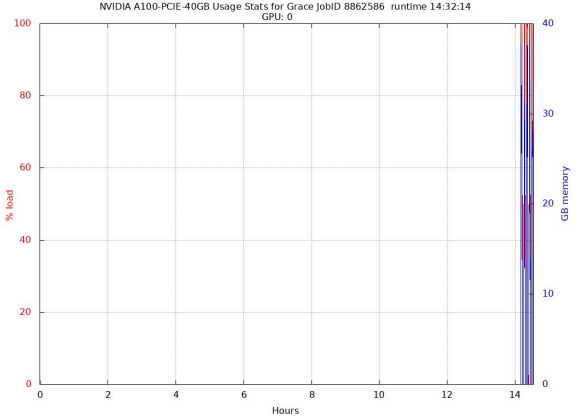
Review GPU and CPU usage for a Job

The **jobstats** command monitors GPU and CPU resource usage and create graphs

```
#!/bin/bash
#SBATCH --job-name=my_gpu_job
#SBATCH --time=1-00:00:00
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=24
#SBATCH --mem=180G
#SBATCH --gres=gpu:a100:1
#SBATCH --output=stdout.%x.%j
#SBATCH --error=stderr.%x.%j

module purge

# run jobstats in the background &
# to monitor resource usage
jobstats &

my_alphafold_command

# run jobstats to create a graph
# of cpu and gpu usage for this job
jobstats
```



eog stats_cpu.8862586.png



eog stats_gpu.8862586.png

- CPU stats are only accurate for jobs using the entire compute node resources (CPUs, memory)
- GPU stats are accurate if using fewer than max CPUs and memory

https://hprc.tamu.edu/kb/Software/useful-tools/jobstats

AlphaFold Workflow Alternatives

ParallelFold

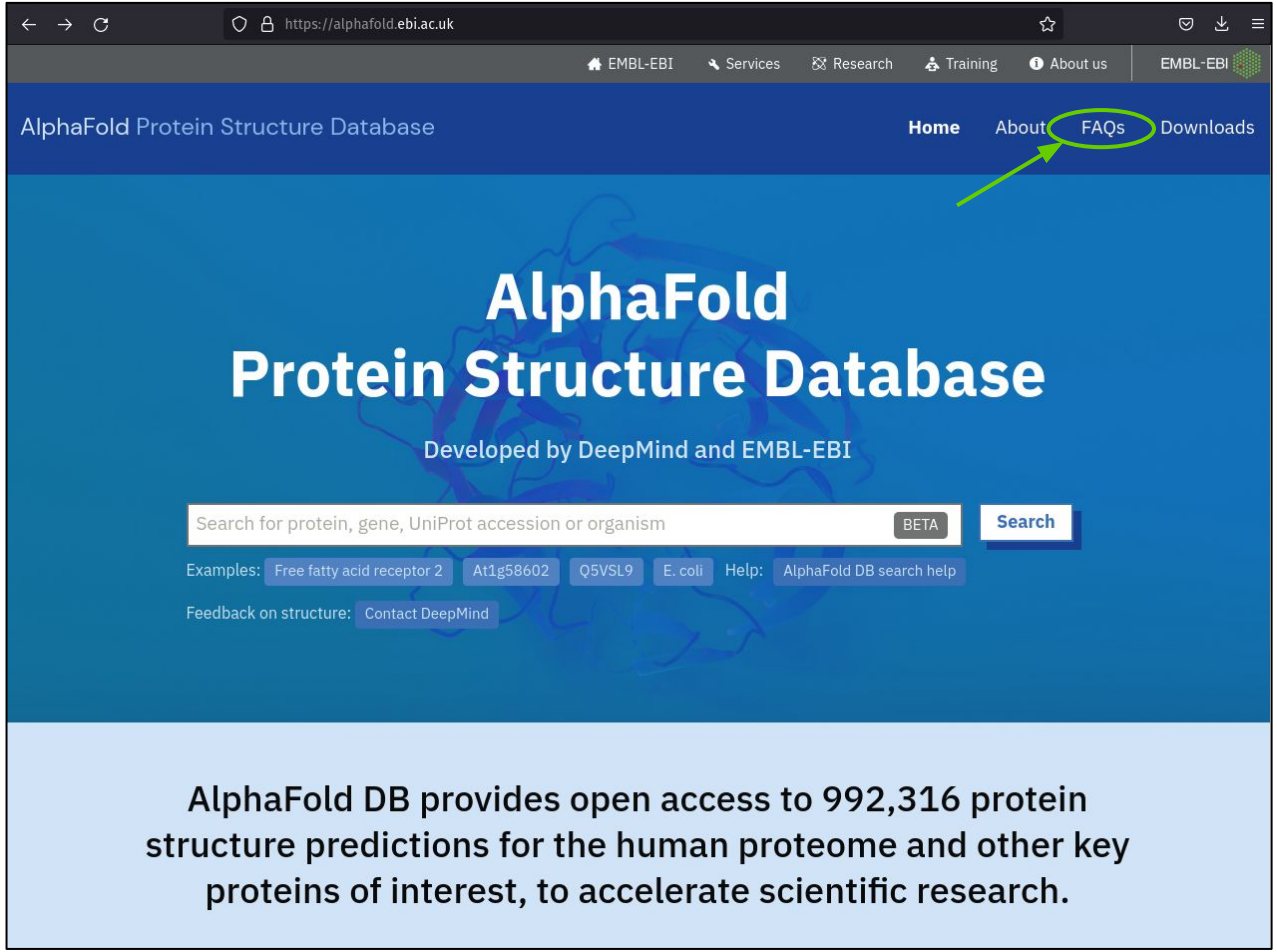
- ParallelFold (ParaFold) breaks the AlphaFold workflow into two steps
  - processing of the three CPU steps in parallel
  - processing of the GPU step
- The parallel portion for CPU steps is not implemented yet resulting in similar or longer runtimes than the DeepMind approach
  - the first three CPU steps, jackhammer, jackhammer and HHblits are supposed to run as three separate processes in parallel but currently this parallelization step is not implemented yet
  - when the CPU processing is parallelized, it will require 21 cores
    - 8 cores for each of the two jackhammer steps
    - 5 cores for the HHblits step.
- The AlphaFold CPU steps may be implemented in parallel soon

https://github.com/Zuricho/ParallelFold

Databases and References

DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) have partnered to create AlphaFold DB to make these predictions freely available to the scientific community.

Search for your protein to see if it the structure has already been predicted using AlphaFold



References

Article | Open Access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper, Richard Evans, ... Demis Hassabis + Show authors

Nature **596**, 583–589 (2021) | Cite this article

Article | Open Access | Published: 22 July 2021

Highly accurate protein structure prediction for the human proteome

Kathryn Tunyasuvunakool, Jonas Adler, ... Demis Hassabis + Show authors

Nature **596**, 590–596 (2021) | Cite this article

Arnold, M. J. (2021) AlphaPickle doi.org/10.5281/zenodo.5708709

HPRC Resources

- Free Help
  - Send an email to help@hprc.tamu.edu if you have any questions regarding Bioinformatics tools usage on HPRC clusters or to schedule a Zoom or in-person visit
    - First spend some time investigating the error
      - read log files, stdout file, stderr file, tool manual
      - Google search
      - Google user groups: many are software specific
    - Include details about your issue
      - The JobID
      - Which cluster or which Galaxy you are using
      - Which software you are using
      - Which modules you have loaded
      - Commands you used in your job script
      - Error messages you are seeing
- HPRC NGS data analysis tools Documentation
  - https://hprc.tamu.edu/kb/Software/Bioinformatics

Let us know when the issue has been resolved so we can close the helpdesk ticket