

HIGH PERFORMANCE RESEARCH COMPUTING

ACES: AlphaFold Protein Structure Prediction



High Performance
Research Computing
DIVISION OF RESEARCH

Fall 2024

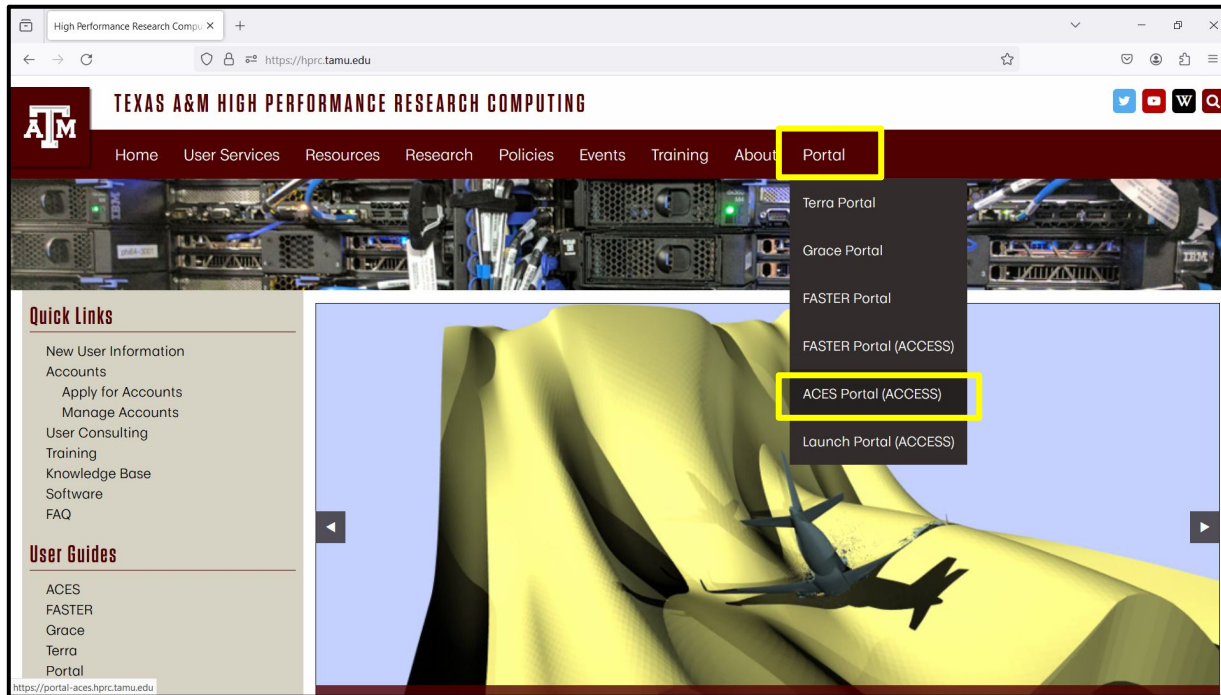


High Performance Research Computing | hprc.tamu.edu | NSF Award #2112356

ACES: AlphaFold Protein Structure Prediction

- ACES Login
- ACES Cluster Utilities
- AlphaFold Job Submission
- AlphaFold Job Scripts using ParaFold
- AlphaFold History
- Selection and Limitations of Resources
- Database Files for Sequence Prediction
- AlphaFold Results Visualization
 - job resource monitoring and usage
 - view predicted structures in Jmol
 - plotting pLDDT values
- AlphaFold 3 Server

Accessing the HPRC ACES Portal



HPRC webpage: hprc.tamu.edu

Accessing ACES via the Portal (ACCESS)



Consent to Attribute Release

TAMU ACES ACCESS OIDC requests access to the following information. If you do not approve this request, do not proceed.

- Your CI/Logon user identifier
- Your name
- Your email address
- Your username and affiliation from your identity provider

ACCESS CI (XSEDE)

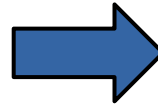
LOG ON

By selecting "Log On", you agree to the [privacy policy](#).

Select an Identity Provider

ACCESS CI (XSEDE)

Select the Identity Provider appropriate for your account.



Log-in using your ACCESS credentials.



If you had an XSEDE account, please enter your XSEDE username and password for ACCESS login.

ACCESS ID

ACCESS Password

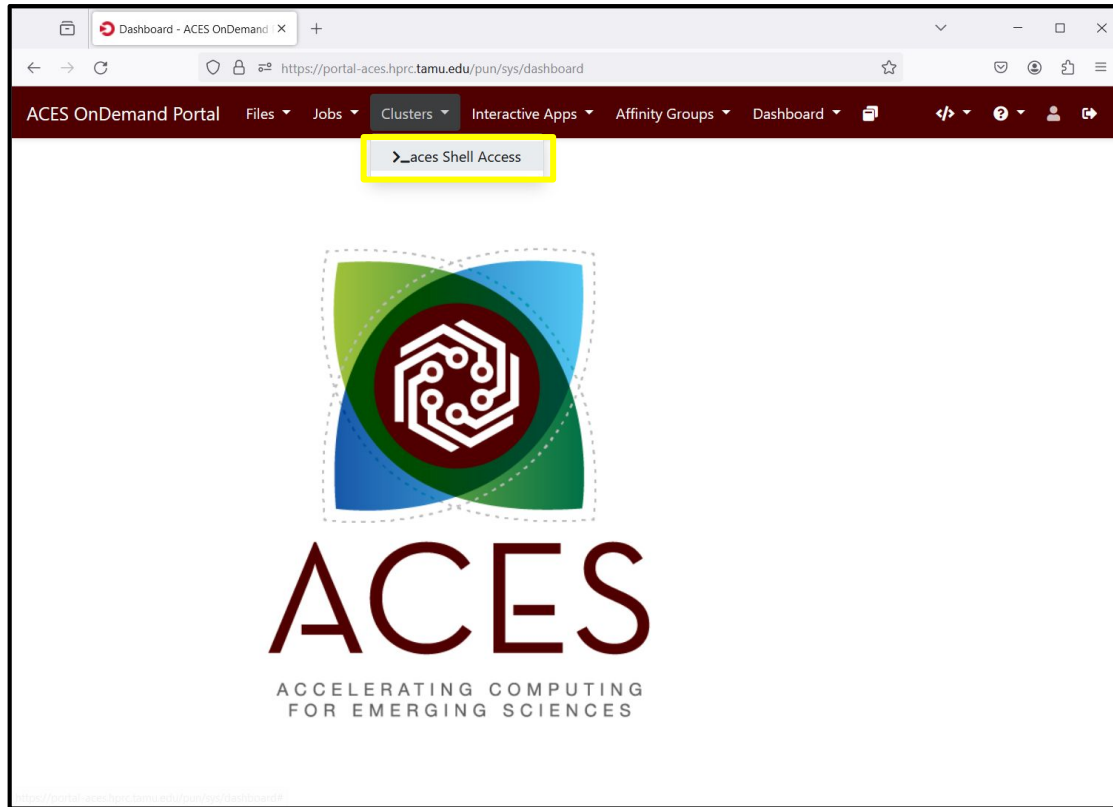
LOGIN

[Register for an ACCESS ID](#)

[Forgot your password?](#)

[Need Help?](#)

Shell Access via the Portal



ACES Cluster Utilities

A number of cluster utilities are available to help you query resources from the command line such as available nodes, GPUs, cores, memory, template job scripts and shared conda and Python environments.

`myjob`

`maxconfig`

`gcatemplates`

`jobstats`

`toolchains`

`gpuavail`

`cpuavail`

`envsavail`

`venvavail`

`maintenance`

use the `-h` or `--help` flag with any utility to see available options

Show Your Job Details using myjob

- The myjob command can be used to see detailed information related to your job.
 - Status (PENDING, RUNNING, COMPLETED, FAILED, ...)
 - Node List
 - Submit time, Start time, End time, Total runtime
 - CPU Efficiency
 - Memory Utilized, Memory Efficiency
- will advise you if your job is PENDING due to a scheduled maintenance.
- will advise you if your job FAILED due to CRLF characters in the job script and provide a link to the HPRC documentation on how to resolve this issue.
- will advise you if your job FAILED due to file or disk quota being reached.
 - will show you the directory in your \$HOME directory that has the most files when \$HOME file quota is reached.

<https://hprc.tamu.edu/kb/Software/useful-tools/myjob>

Show Your Job Details using myjob

```
[userid@aces ~]$ myjob 245660
```

```
    Job ID: 245660
    Cluster: aces
    User/Group: /username
    State: COMPLETED (exit code 0)
    Partition: cpu
    Node Count: 1
    NodeList: ac047
    Cores per node: 10
    CPU Utilized: 00:33:23
    CPU Efficiency: 0.82% of 2-19:49:20 core-walltime
    Submit time: 2024-09-17 15:49:50
    Start time: 2024-09-17 15:49:55
    End time: 2024-09-17 22:36:51
    Job Wall-clock time: 06:46:56
    Memory Utilized: 17.33 GB
    Memory Efficiency: 17.33% of 100.00 GB
    Job Name: parafold-cpu
    Job Submit Directory: /scratch/user/username/af_demo
    Submit Line: sbatch run_parafold_alphafold_2.3.2_monomer_ptm_h100_aces.sh
```

use the -h flag to view usage
`myjob -h`

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 ACES partition: cpu).

```
[username@aces ~]$ maxconfig

ACES partitions:  cpu  gpu  pvc  bittware  d5005  memverge  nextsilicon
ACES GPUs in gpu partition:  a30:2  h100:2  h100:4  pvc:2  pvc:4

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

#!/bin/bash
#SBATCH --job-name=my_job
#SBATCH --time=7-00:00:00
#SBATCH --nodes=1          # max 64 nodes for partition cpu
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=96
#SBATCH --mem=488G
#SBATCH --output=stdout.%x.%j
#SBATCH --error=stderr.%x.%j
```

<https://hprc.tamu.edu/kb/Software/useful-tools/maxconfig>

Viewing Maximum Available Resources

See the recommended Slurm parameters for requesting 1 x H100 GPU with $\frac{1}{4}$ the total CPUs and memory since there are 4 x H100s per node.

```
[username@aces ~]$ maxconfig -g h100 -G 1

ACES partitions:  cpu gpu pvc bittware d5005 memverge nextsilicon
ACES GPUs in gpu partition:  a30:2 h100:2 h100:4 pvc:2 pvc:4

Showing 1/4 of total cores and memory for using 1 x h100 GPU

#!/bin/bash
#SBATCH --job-name=my_job
#SBATCH --time=2-00:00:00
#SBATCH --partition=gpu
#SBATCH --nodes=1      # max 8 nodes for partition gpu
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=24
#SBATCH --mem=125G
#SBATCH --gres=gpu:h100:1
#SBATCH --output=stdout.%x.%j
#SBATCH --error=stderr.%x.%j
```

<https://hprc.tamu.edu/kb/Software/useful-tools/maxconfig>

Checking GPU Configuration & Availability on ACES

- Use the command line (shell) to see the current GPU configuration and availability.
- The GPU configuration can change since ACES is a composable resource cluster.
- If there are no GPUs in the AVAILABILITY output, it means that a GPU job that you submit may take a while to start.
- AlphaFold does not support running on PVC GPUs.

```
[username@aces ~]$ gpuavail
```

CONFIGURATION	
NODE	NODE
TYPE	COUNT
-----	-----
gpu:pvc:4	16
gpu:h100:2	10
gpu:a30:2	2
gpu:h100:4	2
gpu:pvc:2	1

AVAILABILITY					
NODE	GPU	GPU	GPUs	CPUs	GB MEM
NAME	TYPE	COUNT	AVAIL	AVAIL	AVAIL
-----	-----	-----	-----	-----	-----
ac041	h100	4	1	87	421
ac045	h100	2	1	88	422
ac051	pvc	2	2	96	488
ac065	a30	2	2	96	488

<https://hprc.tamu.edu/kb/Software/useful-tools/gpuavail>

Check non-GPU node Availability

Use the `cpuavail` command to see non-GPU nodes readily available for jobs.

```
[username@aces ~]$ cpuavail
```

CONFIGURATION		AVAILABILITY		
NODE	NODE	NODE	CPUs	GB MEM
TYPE	COUNT	NAME	AVAIL	AVAIL
-----		-----		
CPU-only	54	ac006	8	196
GPU	40	ac007	6	86
other	14	ac017	8	88
		ac021	44	4
		ac022	4	190
		ac042	54	214
		ac043	12	92
		ac052	60	244
		ac053	64	248
		ac063	12	228
		ac073	8	88
		ac080	1	121

ACES Cluster maintenance

- You can use the maintenance command to see if there is a scheduled cluster maintenance.

```
[username@aces ~]$ maintenance
```

```
The scheduled 11 hour ACES maintenance will start in:
```

```
3 days 16 hours 41 minutes
```

```
Scheduled jobs will not start if they overlap with this maintenance window.
```

A 7-day job submitted at the time of the above message will remain queued and will not start until after the maintenance is complete.

Submit an AlphaFold Job

Finding AlphaFold template job scripts using GCATemplates on ACES

- Genomic Computational Analysis Templates are job scripts that use examples input data, which you can run for demo purposes.

```
mkdir $SCRATCH/af_demo
```

```
cd $SCRATCH/af_demo
```

```
gcatemplates
```

- Type **s** for search, then enter **alphafold** to search for the alphafold **2.3.2** template script, and select the **parafold** monomer_ptm script.
- Review the script.

```
BIOINFORMATICS GCATemplates (ACES)

CATEGORY
1. FASTQ files (QC, trim, SRA)
2. Protein tools

s search
q quit

Select: s
```

Example AlphaFold (ParaFold) Job Script

```
#!/bin/bash
#SBATCH --job-name=parafold-cpu      # job name
#SBATCH --time=7-00:00:00           # max job run time dd-hh:mm:ss
#SBATCH --ntasks-per-node=1         # tasks (commands) per compute node
#SBATCH --cpus-per-task=48          # CPUs (threads) per command
#SBATCH --mem=244G                   # total memory per node
#SBATCH --output=stdout.%x.%j       # save stdout to file
#SBATCH --error=stderr.%x.%j        # save stderr to file

module purge
module load GCC/11.3.0 OpenMPI/4.1.4 AlphaFold/2.3.2-CUDA-11.8.0
module load ParaFold/2.0-CUDA-11.8.0

ALPHAFOLD_DATA_DIR=/scratch/data/bio/alphafold/2.3.2
protein_fasta=/scratch/data/bio/alphafold/example_data/T1083_T1084_multimer.fasta

# First, run CPU-only steps to get multiple sequence alignments
run_alphafold.sh -d $ALPHAFOLD_DATA_DIR -o pf_output_dir -p multimer -i $protein_fasta -t 2024-1-1 -f

# Second, run GPU steps as a separate job after the first part completes successfully
sbatch --job-name=parafold-gpu --time=2-00:00:00 --ntasks-per-node=1 --cpus-per-task=24 --mem=122G \
--gres=gpu:h100:1 --partition=gpu --output=stdout.%x.%j --error=stderr.%x.%j \
--dependency=afterok:$SLURM_JOBID<<EOF
#!/bin/bash
module purge
module load GCC/11.3.0 OpenMPI/4.1.4 AlphaFold/2.3.2-CUDA-11.8.0
module load ParaFold/2.0-CUDA-11.8.0 AlphaPickle/1.4.1
jobstats -i 1 &
run_alphafold.sh -g -u 0 -d $ALPHAFOLD_DATA_DIR -o pf_output_dir -p multimer -i $protein_fasta -t 2024-1-1
# graph pLDDT and PAE .pkl files
run_AlphaPickle.py -od pf_output_dir/T1083_T1084_multimer
jobstats
EOF
```


Submit and Monitor the Job

- Run the `cpuavail` utility to see cluster usage status.

```
[username@aces ~]$ cpuavail
```

- Edit your job script to use 10 cores and 100 GB memory.
- Submit the job script to the Slurm scheduler.
 - completes in about 3 hours so we will review a completed job

```
[username@aces ~]$ sbatch run_parafold_alphafold_2.3.2_monomer_ptm_h100_aces.sh
```

```
Submitted batch job 245660
```

- Monitor the job status.

```
[username@aces ~]$ squeue --me
```

JOBID	NAME	USER	PARTITION	NODES	CPUS	STATE	TIME	TIME_LEFT	START_TIME	REASON	NODELIST
245660	parafold-cpu	username	cpu	1	10	RUNNING	6.59	6-23:53:01	2024-09-17T15:49	None	ac047

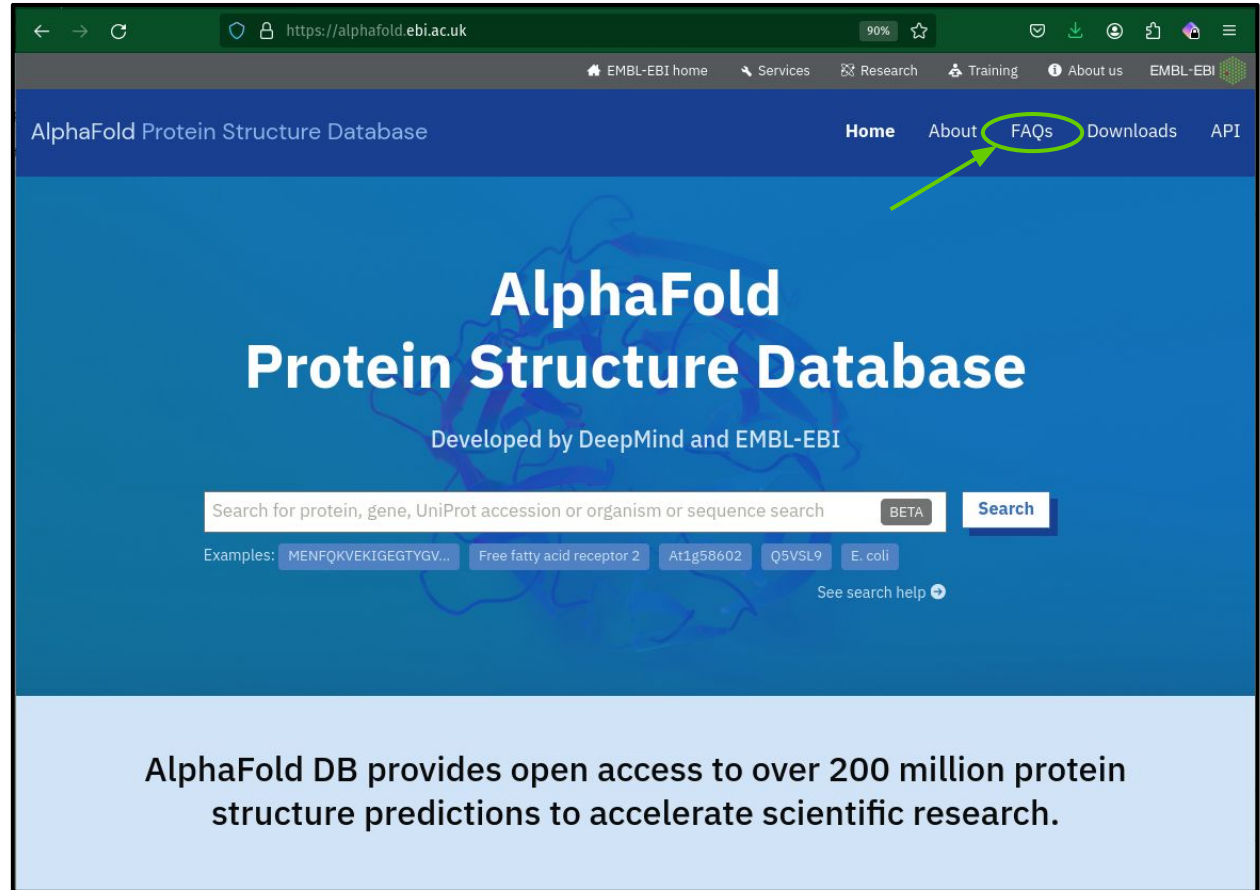
AlphaFold History

- An Artificial Intelligence program developed by DeepMind
- 2018 AlphaFold 1 placed 1st at [CASP 13](#)
- 2020 AlphaFold 1 code released as open source
- 2020 AlphaFold 2 placed 1st at [CASP 14](#)
- 2021 AlphaFold publication in [Nature](#)
 - Highly accurate protein structure prediction with AlphaFold
- 2021 AlphaFold 2 code released as open source on [GitHub](#)
- 2024 AlphaFold 3 available on the deepmind [AlphaFold Server](#)
 - 20 jobs per day allowed for academic researchers

DeepMind and EMBL's European Bioinformatics Institute ([EMBL-EBI](https://www.ebi.ac.uk/)) have partnered to create the AlphaFold Protein Structure [Database](#) to make over 200 million predictions freely available to the scientific community.

Search for your protein to see if the structure has already been predicted using AlphaFold 2.

See the [FAQs](#)



The screenshot shows the AlphaFold Protein Structure Database website. The browser address bar displays <https://alphafold.ebi.ac.uk>. The top navigation menu includes links for Home, About, **FAQs** (highlighted with a green circle and arrow), Downloads, and API. The main heading is "AlphaFold Protein Structure Database" with the subtext "Developed by DeepMind and EMBL-EBI". A search bar is present with the placeholder text "Search for protein, gene, UniProt accession or organism or sequence search" and a "BETA" label. Below the search bar are examples: "MENFQKVEKIGEGTYGV...", "Free fatty acid receptor 2", "At1g58602", "Q5VSL9", and "E. coli". A "See search help" link is also visible. At the bottom of the page, a light blue banner states: "AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research."

Selection and Limitations of Resources

Resource Limitations

- AlphaFold
 - Currently AlphaFold can only utilize one GPU.
 - minimum amino acid length: 16
 - maximum amino acid length:
 - 2,700 proteomes / Swiss-Prot
 - 1,280 all other UniProt
- AlphaFold DeepMind workflow
 - Only about 10% of job runtime is performed on GPU.
- ACES Job Script Configuration
 - In your job script, request only $\frac{1}{2}$ of the cores and memory when using 1 x H100 on a GPU node that has 2 x H100 installed so the other H100 GPU on that node is available for other jobs.

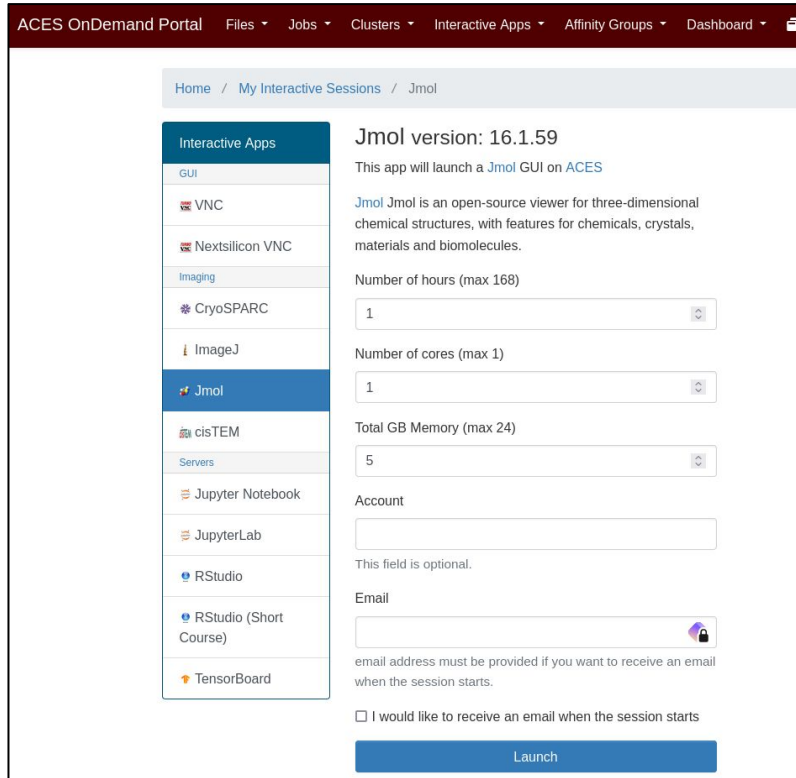
AlphaFold Databases for Structure Predictions on ACES

/scratch/data/bio/alphafold/2.3.2

Database	Size	File Count	monomer	multimer
bfd	1.8T	7	✓	✓
mgnify	120G	2	✓	✓
params	5.3G	17	✓	✓
pdb70	56G	10	✓	-
pdb_mmcif	264G	211,106	✓	✓
pdb_seqres	257M	2	-	✓
uniprot	114G	2	-	✓
uniref30	467G	15	✓	✓
uniref90	77G	2	✓	✓
small_bfd	17G	2	✓	✓
example_data	6K	5	✓	✓
TOTAL	2.9T	211,170		

AlphaFold Results Visualization

Visualize AlphaFold Results with Jmol on the ACES Portal



ACES OnDemand Portal | Files | Jobs | Clusters | Interactive Apps | Affinity Groups | Dashboard

Home / My Interactive Sessions / Jmol

Interactive Apps

- GUI
- VNC
- Nextsilicon VNC
- Imaging
- CryoSPARC
- ImageJ
- Jmol**
- cisTEM
- Servers
- Jupyter Notebook
- JupyterLab
- RStudio
- RStudio (Short Course)
- TensorBoard

Jmol version: 16.1.59
This app will launch a Jmol GUI on ACES

Jmol Jmol is an open-source viewer for three-dimensional chemical structures, with features for chemicals, crystals, materials and biomolecules.


Number of hours (max 168):

Number of cores (max 1):

Total GB Memory (max 24):

Account:

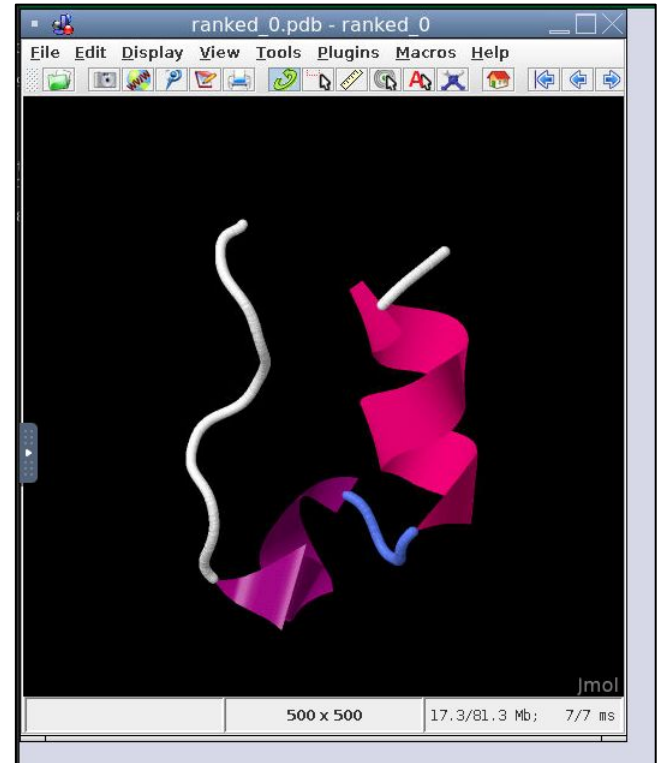
This field is optional.

Email: 

email address must be provided if you want to receive an email when the session starts.

I would like to receive an email when the session starts

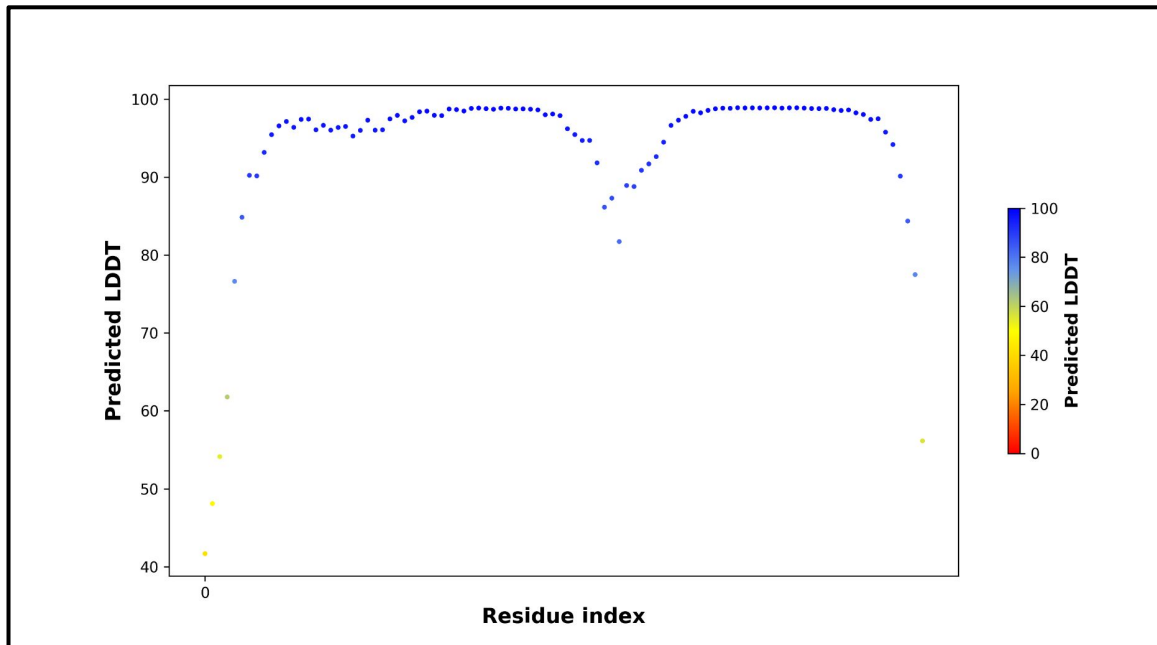
Launch



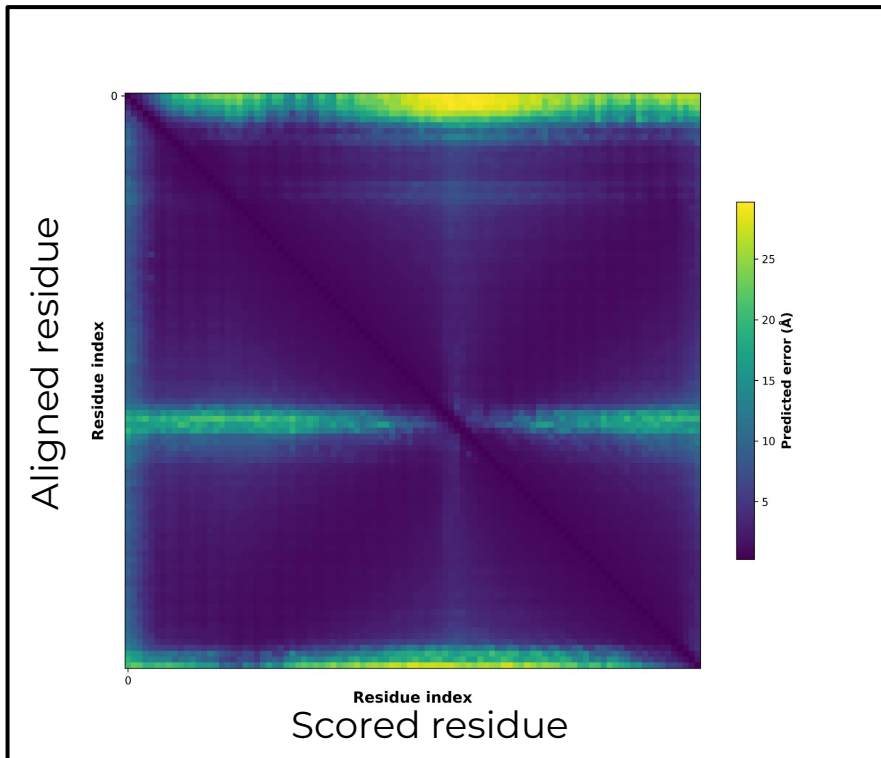
`/scratch/training/alphafold/out_parafold_1L2Y_monomer_ptm_full_dbs/1L2Y/ranked_0.pdb`

AlphaFold Confidence Metrics

Visualize AlphaFold pLDDT Scores



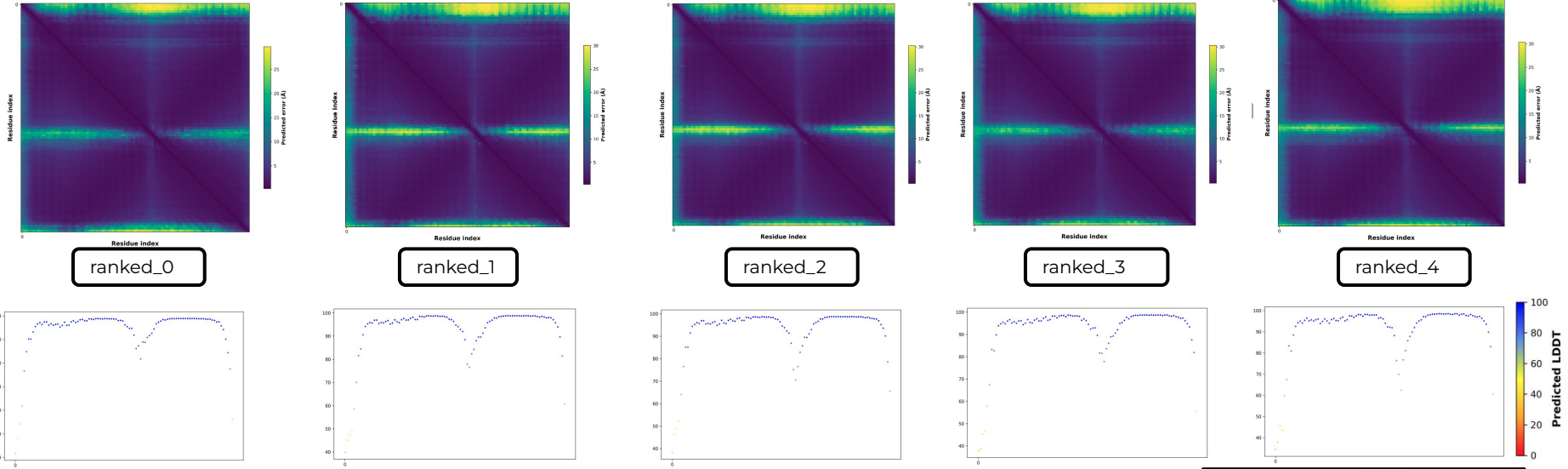
Visualize AlphaFold PAE Results (monomer_ptm)



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

`/scratch/training/alphafold/out_1L2Y_monomer_ptm_reduced_dbs/1L2Y/ranked_0_PAE.png`

Evaluating Models



See which model has the top rank based on pLDDT score.

```
cat /scratch/training/alphafold/out_1L2Y_monomer_ptm_reduced_dbs/1L2Y/ranking_debug.json
```

```
"pLDDTs": {  
  "model_1_ptm_pred_0": 94.14318865567142,  
  "model_2_ptm_pred_0": 94.83124839667653,  
  "model_3_ptm_pred_0": 90.41209232774796,  
  "model_4_ptm_pred_0": 90.73102198891624,  
  "model_5_ptm_pred_0": 92.6319870089253  
},  
"order": [  
  "model_2_ptm_pred_0",  
  "model_1_ptm_pred_0",  
  "model_5_ptm_pred_0",  
  "model_4_ptm_pred_0",  
  "model_3_ptm_pred_0"  
]
```

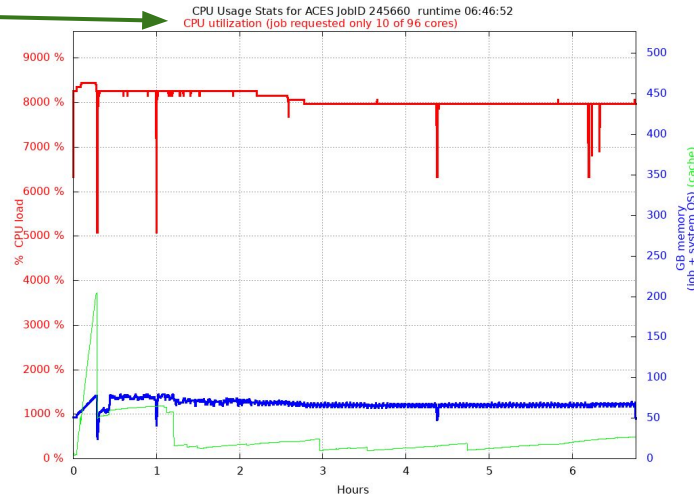
ranked_0

ranked_4

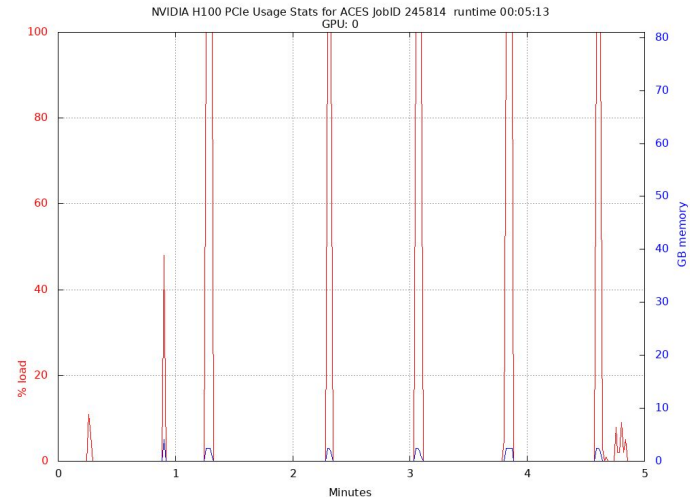
AlphaFold Job Resource Monitoring

Review GPU and CPU usage for a Job

The `jobstats` command monitors GPU and CPU resource usage and create graphs.



stats_cpu.245660.png



stats_gpu.245814.png

view images in Portal Files app

- CPU stats are only accurate for jobs using the entire compute node resources. (CPUs, memory), but we primarily want to make sure GPUs were used.
- GPU stats are accurate if using fewer than max CPUs and memory because your job will be the only job running on the requested GPU.

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

ParaFold Workflow

- The ParaFold module uses the same AlphaFold installation as the AlphaFold module.
- ParaFold divides the AlphaFold workflow into two steps which can be run as two separate jobs:
 - CPU-only: processing the CPU steps to generate multiple sequence alignments
 - GPU: processing the GPU steps to generate predictions
- Test run of multimer (T1083_T1084_multimer.fasta) with full_dbs
- Runtimes for the same job script varied +/- 1 hour; TM-scores also vary

AlphaFold 2.3.2	Runtime	Highest Scoring Model	TM-score**
ParaFold	3 hrs 10 min*	model_1_multimer_v3_pred_4	0.892
DeepMind	2 hrs 45 min	model_1_multimer_v3_pred_1	0.883

* combined time for the separate CPU 3 hour job and GPU 10 min job

** measure of similarity between two protein structures

<https://github.com/Zuricho/ParallelFold>

Graphing Confidence Scores with AlphaPickle

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.

```
module load GCC/11.3.0 OpenMPI/4.1.4 AlphaFold/2.3.2-CUDA-11.8.0
module load ParaFold/2.0-CUDA-11.8.0 AlphaPickle/1.4.1
```

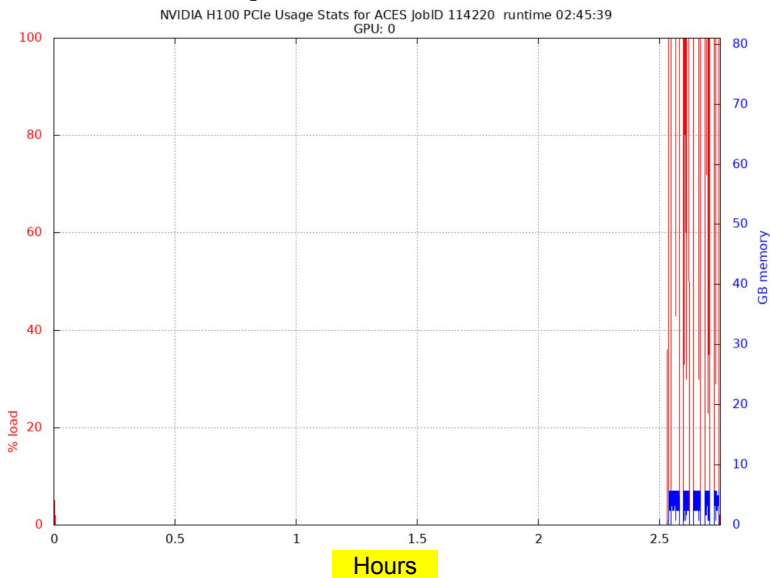
```
run_AlphaPickle.py -od pf_output_dir/T1083_T1084_multimer
```

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

Comparison of DeepMind vs ParaFold Workflows

- AlphaFold DeepMind's workflow (1 CPU+GPU job) vs ParaFold's workflow (1 CPU-only job + 1 GPU job) for the same multimer full_dbs analysis
- The ParaFold workflow significantly reduces GPU idle time

DeepMind Workflow



ParaFold Workflow (GPU job)



The first job of the ParaFold workflow (CPU-only) completed in 3 hours

AlphaFold 3 Server

AlphaFold 3 vs AlphaFold 2

The new AlphaFold model demonstrates substantially improved accuracy over many previous specialized tools: far greater accuracy for protein–ligand interactions compared with state-of-the-art docking tools, much higher accuracy for protein–nucleic acid interactions compared with nucleic-acid-specific predictors and substantially higher antibody–antigen prediction accuracy compared with AlphaFold-Multimer v.2.37,8.

(from Abstract)

Abramson, J., Adler, J., Dunger, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 493–500 (2024).

<https://doi.org/10.1038/s41586-024-07487-w>

AlphaFold 3 Server Terms

- AlphaFold [Server](#) is only available for non-commercial use by individuals and non-commercial organizations (universities, non-profit organizations and research institutes, educational and government bodies), or for journalism.
- You must not use AlphaFold Server or its outputs:
 - in connection with any commercial activities, including research on behalf of commercial organizations;
 - in any automated system that predicts the binding or interaction of the protein with ligands or peptides, such as Glide or AutoDock; or
 - to train machine learning models or related technology for biomolecular structure prediction similar to AlphaFold Server.
- You can publish, share and adapt AlphaFold Server output in accordance with our terms, including the requirement to provide clear notice that ongoing use is subject to [AlphaFold Server Output Terms of Use](#) and of any modifications you make.

AlphaFold Server BETA Server About FAQs

Remaining jobs: 20

Limited to 20 jobs per day

AlphaFold Server allows you to model a structure consisting of many biological molecules

- Remaining jobs refresh each day
- Jobs can be up to 5,000 tokens - see more details on token calculation, accepted formats, seed selection and other features in our FAQ
- Use the entity bar to chemically modify proteins and nucleic acids
- Get in touch with the AlphaFold team if you have any questions

Explore these examples of structures to see it in action - try them out without using your quota until you begin editing!

Protein-RNA-Ion: PDB 8AW3 Protein-Glycan-Ion: PDB 7BBV Protein-DNA-Ion: PDB 7RCE

Ok, got it

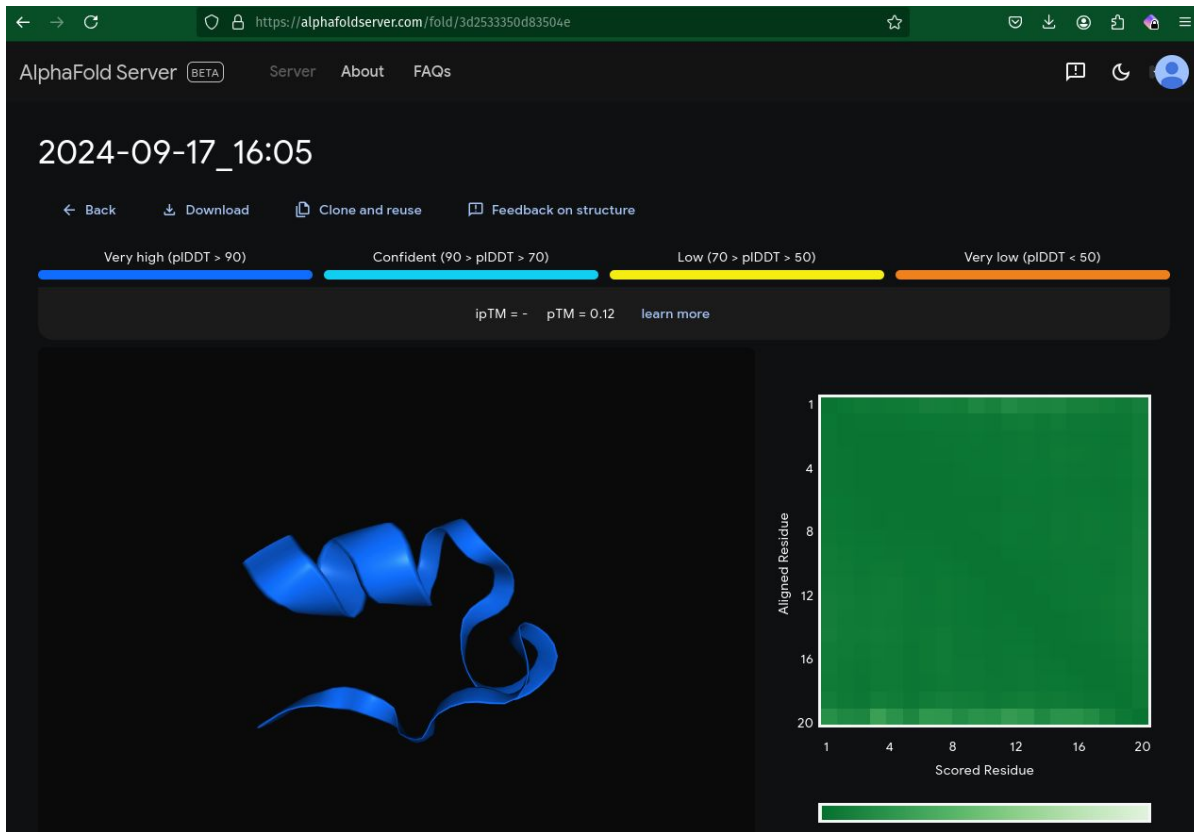
Upload JSON Clear

Molecule type Protein Copies 1

>Paste sequence or fasta Input

Enter 1L2Y sequence:
NLYIQWLKDGGPSSGRPPPS



AlphaFold 3 Server Results



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](#)

Zhong, B, et al. (2021) ParaFold doi.org/10.48550/arXiv.2111.06340

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

ACES Documentation

- ACES KnowledgeBase Documentation hprc.tamu.edu/kb
- ACES User Guide hprc.tamu.edu/kb/User-Guides/ACES
- Email your questions to help@hprc.tamu.edu
 - received emails generate helpdesk tickets

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



**HIGH PERFORMANCE
RESEARCH COMPUTING**
TEXAS A&M UNIVERSITY

<https://hprc.tamu.edu>

HPRC Helpdesk:

help@hprc.tamu.edu

Phone: 979-845-0219

Take our short course survey!



HPRC Survey

https://u.tamu.edu/hprc_shortcourse_survey

Help us help you. Please include details in your request for support, such as, Cluster (ACES, FASTER, Grace, Launch), NetID (UserID), Job information (JobID(s), Location of your jobfile, input/output files, Application, Module(s) loaded, Error messages, etc), and Steps you have taken, so we can reproduce the problem.

