

---

# AlphaFold on BioHPC

— GUI with Astrocyte and CLI with SLURM

Peng Lian

2023-05-17

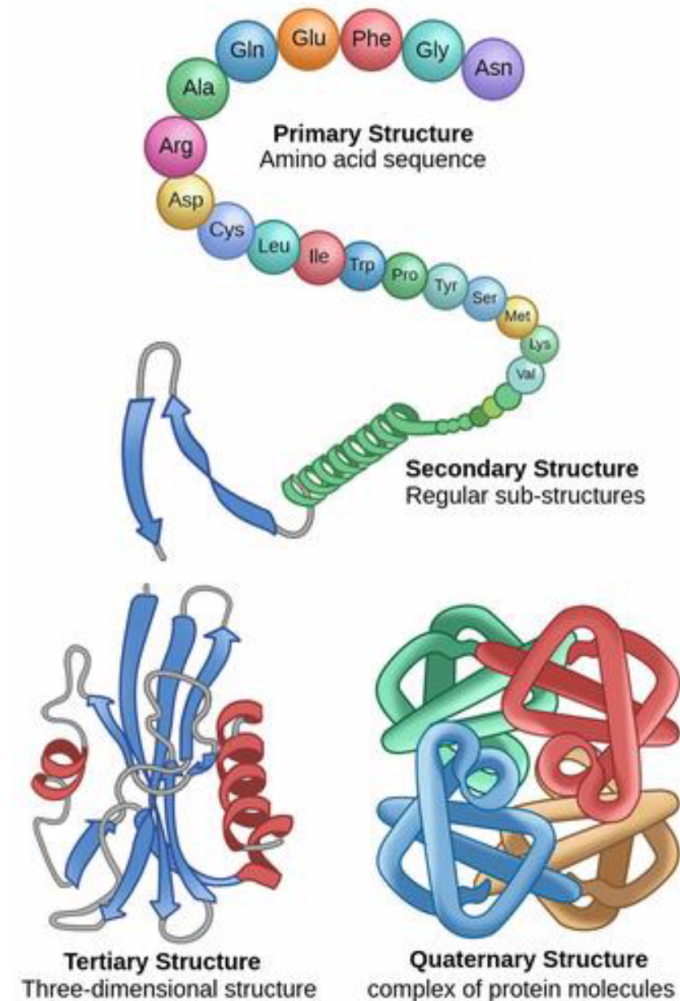
# Outline

---

- Protein Structure
- **Critical Assessment of protein Structure Prediction (CASP)**
- AlphaFold source code on Github
- AlphaFold database
- AlphaFold new feature: Multimer
- Run AlphaFold with Astrocyte AlphaFold Workflow (GUI)
- Run AlphaFold with SLURM (CLI)

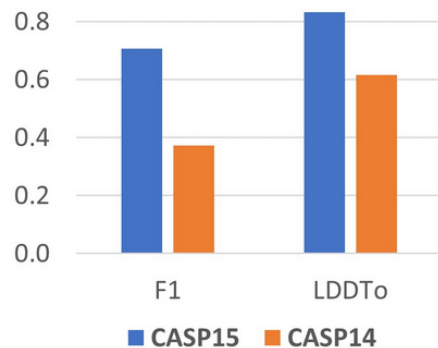
# Protein Structure

- Protein folding problem
  - Sequence of amino acids do not show how they fold into shape
- Why is protein folding important?
  - Protein structure determines the function
  - Structure based drug design
- Methods to obtainer protein structure
  - X-ray crystallography
  - Nuclear Magnetic Resonance (NMR)
  - Cryo-Electron Microscopy (CryoEM)
  - Artificial Intelligence (AI)

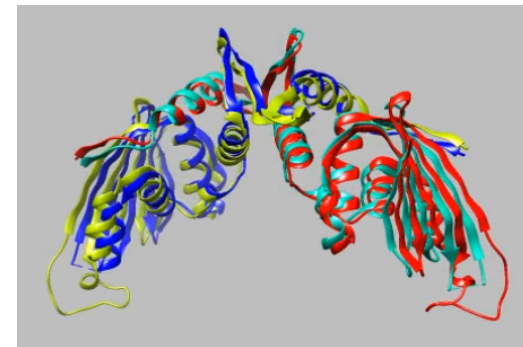


# CASP

- Critical Assessment of protein Structure Prediction. <https://predictioncenter.org>
- Community-wide, worldwide experiment for protein structure prediction
- Take place every two years since 1994
- How to play:
  - Target proteins: recently solved but hold by PDB
  - No one knows the structure of the target proteins
  - Participants predict the structures with their algorithms
  - Evaluation with scores such as GDT-TS (global Distance Test – Total Score), describing percentage of well-modeled residues in the model with respect to the target



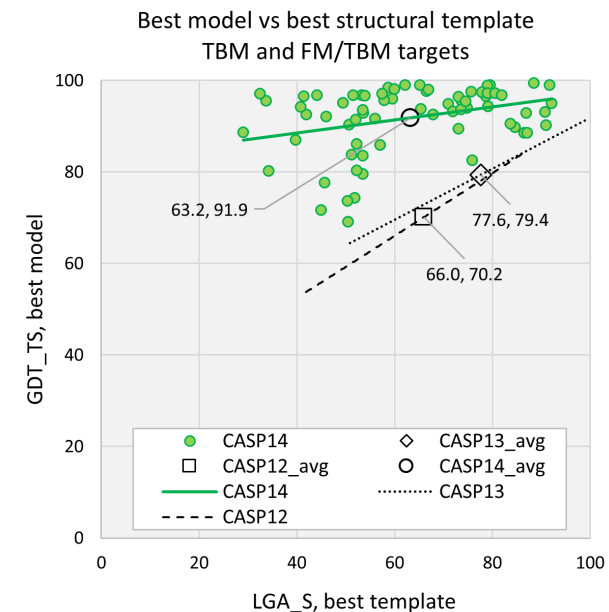
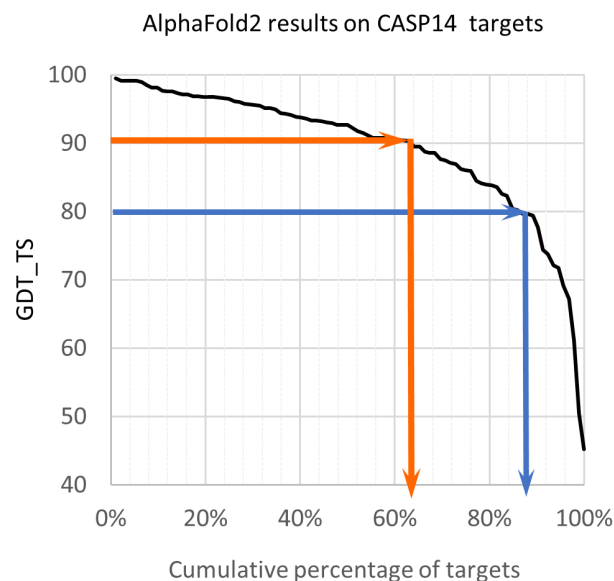
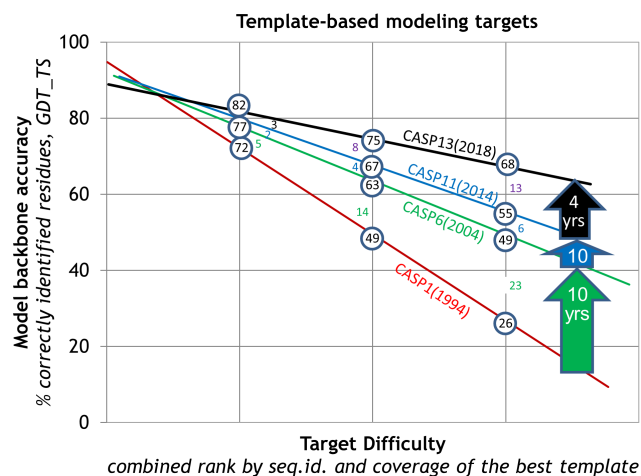
Scores



CASP15: T1113o  
F1=92.2; LDDTo=0.913

# AlphaFold in CASP

- Over the course of CASP, template-based modeling get enormous improvement
- The 2014-2018 model accuracy improvement doubled that of 2004-2014
- CASP14 marked an extraordinary increase in the accuracy of the computed three-dimensional protein structures with the emergence of the advanced deep learning method AlphaFold2
- AlphaFold2 proved to be competitive with the experimental accuracy
- The accuracy of CASP14 models is significantly higher than the corresponding average of previous two CASPs



# AlphaFold Code on Github

## 1. Deepmind <https://github.com/deepmind/alphafold>

- Docker:
  - Use “root” account to run docker
  - On BioHPC, users are not allowed to use “root” privilege to run Docker.
- Databases:
  - full database (2.2TB) /project/apps\_database
  - reduced database (415GB)
- Running Alphafold: [run\\_docker.py](#)

```
1 python3 docker/run_docker.py --fasta_paths=T1050.fasta --max_template_date=2020-05-14
```

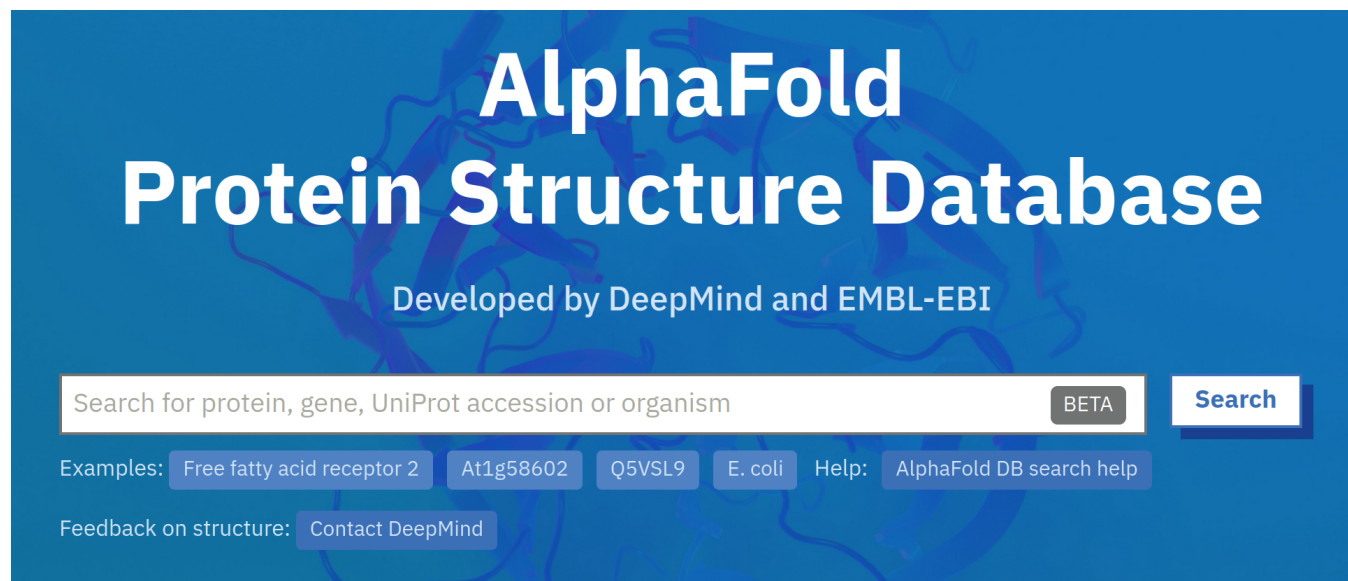
## 2. AlphaFold non Docker [https://github.com/kalininalab/alphafold\\_non\\_docker](https://github.com/kalininalab/alphafold_non_docker)

- Use Conda environment instead of Docker
- Databases: full database and reduced database, same as above
- Installed as BioHPC modules
- Running Alphafold: [run\\_alphafold.sh](#) calls [run\\_alphafold.py](#) (Deepmind)

```
1 run_alphafold.sh -d /project/apps_database/alphafold/database_full -o /your/path/to/dummy_test/ \  
2 -m model_1 -f /your/path/to/query.fasta -t 2020-05-14
```

# AlphaFold Databases

- <https://alphafold.ebi.ac.uk>
- Collaboration between Google Deepmind and EMBL-EBI
- 214,683,829 structures available on the AlphaFold DB website
- Including 48 complete proteomes
- An additional 3,095 structures are included in the human proteome, covering sequences longer than our usual length limit split into fragments.



The screenshot shows the AlphaFold Protein Structure Database homepage. The background is a dark blue with a faint protein structure. The main title "AlphaFold Protein Structure Database" is in large white font. Below it, it says "Developed by DeepMind and EMBL-EBI". There is a search bar with the placeholder text "Search for protein, gene, UniProt accession or organism" and a "BETA" label. To the right of the search bar is a "Search" button. Below the search bar, there are examples: "Examples: Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli Help: AlphaFold DB search help". At the bottom, there is a "Feedback on structure: Contact DeepMind" link.

# Run AlphaFold with Astrocyte (GUI)

Steps to run Astrocyte AlphaFold Workflow:

1. Log in <https://astrocyte.biohpc.swmed.edu>
2. Create a project
3. Upload your sequence file (.fasta file)
4. Run the Astrocyte AlphaFold workflow by selecting parameters. e.g. monomer or multimer
5. Download or online analyze results



# Parameter Forms for Submission

## Parameters

Project

Project 1905: alphafold

Name for this run

A file contains one fasta sequence for monomer OR multiple sequences for multimer prediction (required)

5HT1A.fasta  
 GPHR\_heterodimer.fasta  
 GPHR\_heterodimer-1382.fasta

Choose preset model configuration - the monomer model, the monomer model with extra ensembling, monomer model with pTM head, or multimer model (default: 'monomer') (required)

monomer

This is the path to the database folder on BioHPC (required)

/project/apps\_database/alphafold\_2.1.1/database\_full

Choose preset MSA database configuration - smaller genetic database config (reduced\_dbs) or full genetic database config (full\_dbs) (default: 'full\_dbs') (required)

Full Databases

Maximum template release date to consider (ISO-8601 format - i.e. YYYY-MM-DD). Important if folding historical test sets (required)

2021-12-06

Optional for multimer system, not used by the single chain system. A boolean specifying true where the target complex is from a prokaryote, and false where it is not, or where the origin is unknown. This value determine the pairing method for the MSA (default: 'None')

Unknown

Run multiple JAX model evaluations to obtain a timing that excludes the compilation time, which should be more indicative of the time required for inferencing many proteins (default: 'False') (required)

Do NOT use benchmark

Enable NVIDIA runtime to run with GPUs (default: True) (required)

Use GPU

Comma separated list of devices to pass to 'CUDA\_VISIBLE\_DEVICES' (default: 0) (required)

Use GPU 0

Run Workflow

# Astrocyte AlphaFold Versions

<https://astrocyte.biohpc.swmed.edu/workflow/48/view>

## Published Versions

Version	Git Tag	
astrocyte_alphafold - 1.0.0	<p><b>Astrocyte AlphaFold Workflow</b> This workflow is based on AlphaFold 2.1.1 that supports multimer. It selects SLURM partition automatically.</p> <p><b>Author:</b> Peng Lian, Xiaochu Lou <b>Contact:</b> <a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></p>	<p><a href="#">▶ Run this Version</a></p> <p><a href="#">📖 Documentation</a></p> <p><a href="#">⚙️ Developer Information</a></p>
astrocyte_alphafold - 0.0.3	<p><b>Astrocyte AlphaFold Workflow</b> This workflow is based on AlphaFold 2.1.1 that supports multimer</p> <p><b>Author:</b> Peng Lian, Xiaochu Lou, Yingfei Chen <b>Contact:</b> <a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></p>	<p><a href="#">▶ Run this Version</a></p> <p><a href="#">📖 Documentation</a></p> <p><a href="#">⚙️ Developer Information</a></p>
astrocyte_alphafold - 0.0.2	<p><b>Astrocyte AlphaFold Workflow</b> This is a workflow based on AlphaFold 2.0</p> <p><b>Author:</b> Peng Lian, Erand Smakaj, Xiaochu Lou, Yingfei Chen <b>Contact:</b> <a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></p>	<p><a href="#">▶ Run this Version</a></p> <p><a href="#">📖 Documentation</a></p> <p><a href="#">⚙️ Developer Information</a></p>
astrocyte_alphafold - 0.0.1	<p><b>Astrocyte AlphaFold Workflow</b> This is a workflow based on AlphaFold 2.0</p> <p><b>Author:</b> Peng Lian, Xiaochu Lou, Yingfei Chen <b>Contact:</b> <a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></p>	<p><a href="#">▶ Run this Version</a></p> <p><a href="#">📖 Documentation</a></p> <p><a href="#">⚙️ Developer Information</a></p>

## Test Versions

Version	Git Tag	
astrocyte_alphafold - 1.1.2 (test)	<p><b>Astrocyte AlphaFold Workflow</b> This workflow is based on AlphaFold 2.1.1 that supports multimer. It selects SLURM partition automatically.</p> <p><b>Author:</b> Peng Lian, Xiaochu Lou <b>Contact:</b> <a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></p>	<p><a href="#">▶ Run this Version</a></p> <p><a href="#">📖 Documentation</a></p> <p><a href="#">⚙️ Developer Information</a></p>
astrocyte_alphafold - 1.0.3 (test)	<p><b>Astrocyte AlphaFold Workflow</b> This workflow is based on AlphaFold 2.1.1 that supports multimer. This test version uses large memory GPU.</p> <p><b>Author:</b> Peng Lian, Xiaochu Lou, Yingfei Chen <b>Contact:</b> <a href="mailto:biohpc-help@utsouthwestern.edu">biohpc-help@utsouthwestern.edu</a></p>	<p><a href="#">▶ Run this Version</a></p> <p><a href="#">📖 Documentation</a></p> <p><a href="#">⚙️ Developer Information</a></p>

# Astrocyte AlphaFold Results

## Workflow Output / Visualization

You can **download** an archive file containing all output of the workflow, or **export** it directly to a location on the BioHPC cluster storage for further work.

If you wish to use the output file(s) as input for other runs, select to output to the **incoming** directory.

*Note - Mac OSX cannot extract zip files >4GB. A tar file download will be added shortly.*

Download Workflow Output:

 Download as .zip file

Export Output:

/project/apps/astrocyte/astrocyte\_outgoing/s190450/workflow\_4859\_output

 Export

The **Visualization App** (vizapp) allows you to explore the results of your workflow on the web. Use the buttons below to start/stop and connect to a vizapp session. It takes 30s for the vizapp to start, or longer if there is a queue on the BioHPC cluster. Please stop the vizapp when you are finished using it, as it occupies a slot on the BioHPC cluster.






**Vizapp Status:**

 Start Vizapp

## Output Browser

Click the 'Generate Direct Link' button to obtain a direct web link you can use with external tools, such as the UCSC Browser, that need to access the file directly. These links are valid for 24 hours.

 Current Directory: (/)

	<a href="#">5HT1A.fasta</a>		
	<a href="#">.keep</a>	(0 bytes)	 Generate Direct Link
	<a href="#">5HT1A.fasta_run_alphaFold.log</a>	(34.0 KB)	 Generate Direct Link

# Astrocyte AlphaFold Online Analysis

BioHPC protein viewer

load file <

Open PDB

structure <

surface <

ligand <

selection <

label <

contact <

stage <

snapshot <

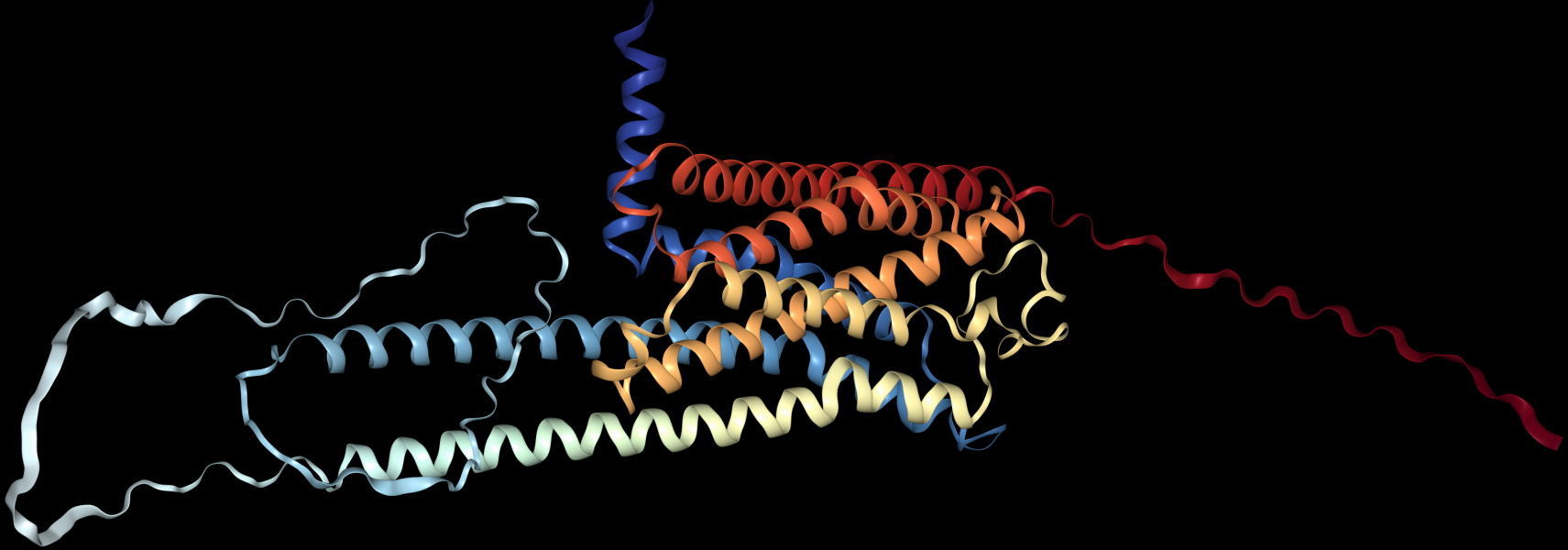
Animation

None Spin Rock

Sequence

Fullscreen

Sequence +



Showing file: output/5HT1A.fasta/5HT1A/ranked\_0.pdb

# Run AlphaFold with SLURM (CLI)

## GPU nodes on BioHPC

Partition	CPU/Node	Memory/Node	GPU/Node	GPU Memory	Nodes
GPU	32	256GB	1 K20/K40	6GB/12GB	71
GPU <sub>p</sub> 4	72	384GB	1 P4	8GB	15
GPU <sub>p</sub> 40	72	384GB	1 P40	24GB	16
GPU <sub>p</sub> 100	56	256GB	2 P100	16GB	11
GPU <sub>v</sub> 100s	72	384GB	1 V100S	32GB	33
GPU <sub>4v</sub> 100	72	384GB	4 V100S	32GB	11
GPU <sub>A</sub> 100	72	1.5TB	1 A100	40GB	16

## Check nodes availability

```

1 sinfo -p GPUv100s
2
3 PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
4 GPUv100s   up      infinite    25   alloc NucleusC[036-052,056-057,059-064]
5 GPUv100s   up      infinite     8    idle NucleusC[053-055,065-069]

```

# Run AlphaFold on BioHPC

Multiple methods are available to run AlphaFold

- WebGPU session: <https://portal.biohpc.swmed.edu/terminal/webgui>
  - 20 hour limit
  - Graphical interface
  - Better for short time jobs
- Submit a SLURM job: <https://portal.biohpc.swmed.edu/sbatch/#/script>
  - Better for long time jobs

For best performance we recommend TurboVNC for webGUI and webGPU sessions, and the NICE DCV client for webWinDCV sessions.

TurboVNC Client Download: [\[Windows\]](#) [\[Mac OSX\]](#) [\[Linux 64-bit .deb\]](#) [\[Linux 64-bit .rpm\]](#) (Version 2.0.91)

NICE DCV Client Download: [\[Windows\]](#) [\[Mac OSX\]](#) [\[Linux .tar.gz\]](#)

Launch a new interactive / visualization job

Note that a session may take time to start if there are no nodes currently free in the cluster. Jobs run for a maximum of 20 hours.

**Job type\***

WebGPUv100s - Tesla V100 GUI + single GPU, high performance ▾

Your session will start immediately, nodes are available.

Launch Job

# Example Commands

```
1 # Load the AlphaFold module
2 module load alphafold/2.1.1
3
4 # Run the command
5 run_alphafold -d /project/apps_database/alphafold_2.1.1/database_full \
6               -o /PATH/TO/YOUR_OUTPUT_DIR \
7               -m multimer \
8               -f /PATH/TO/YOUR_FASTA_FILE/multimer.fasta \
9               -c reduced_dbs \
10              -t 2020-05-14
```

## Note:

- Use same version of alphafold module and alphafold database
- Use 'reduced\_dbs' for testing and 'full\_dbs' for production
- Turn off GPU for extremely long sequences (CPU memory >> GPU memory)
- Provide a correct list of GPU devices
- Model option needs to be agree with the sequences in the fasta file

# Command Usage

## Usage:

```
1 /cm/shared/apps/alphafold/2.1.1/alphafold/run_alphafold <OPTIONS>
```

## Options:

-d <data_dir>	Path to directory of supporting data
-o <output_dir>	Path to a directory that will store the result
-f <fasta_path>	Path to a FASTA file containing sequence. If a FASTA file contains multiple sequences, then it will be folded as a multimer
-t <max_template_date>	Maximum template release date to consider (ISO-8601 format - i.e. YYYY-MM-DD). Important if folding historical test sets
-g <use_gpu>	Enable NVIDIA runtime to run with GPUs (default: true)
-n <openmm_threads>	OpenMM threads (default: all available cores)
-a <gpu_devices>	Comma separated list of devices to pass to 'CUDA_VISIBLE_DEVICES' (default: 0)
-m <model_preset>	Choose preset model configuration - the monomer model, the monomer model with extra ensembling, monomer model with pTM head, or multimer model (default: 'monomer')
-c <db_preset>	Choose preset MSA database configuration - smaller genetic database config (reduced_dbs) or full genetic database config (full_dbs) (default: 'full_dbs')
-p <use_precomputed_msas>	Whether to read MSAs that have been written to disk. WARNING: This will not check if the sequence, database or configuration have changed (default: 'false')
-l <is_prokaryote>	Optional for multimer system, not used by the single chain system. A boolean specifying true where the target complex is from a prokaryote, and false where it is not, or where the origin is unknown. This value determine the pairing method for the MSA (default: 'None')
-b <benchmark>	Run multiple JAX model evaluations to obtain a timing that excludes the compilation time, which should be more indicative of the time required for inferencing many proteins (default: 'false')



# Input FASTA file

- Monomer

```
1 >sequence_1
2 S..E..Q..U..E..N..C..E.....
```

- Multimer (Homomer)

```
1 >sequence_1
2 S..E..Q..U..E..N..C..E.....
3 >sequence_2
4 S..E..Q..U..E..N..C..E.....
```

- Multimer (Heteromer)

```
1 >sequence_1
2 S..E..Q..U..E..N..C..E...A.....
3 >sequence_2
4 S..E..Q..U..E..N..C..E...B.....
5 >sequence_3
6 S..E..Q..U..E..N..C..E...C.....
7 >sequence_4
8 S..E..Q..U..E..N..C..E...D.....
```

# Preset MSA Database

This is the support database of the AlphaFold module. It's big and it's growing. The full database is more than 2.2 TB and the reduced database is more than 615 GB for version 2.1.1.

Module	Database Version	Database Path
alphafold/2.1.1	2.1.1	/project/apps_database/alphafold_2.1.1
alphafold/2.2.0	2.2.0	/project/apps_database/alphafold_2.2.0
alphafold/X.Y.Z	X.Y.Z	/project/apps_database/alphafold_X.Y.Z

**Note:** Please use the same version of the database as the alphafold module!

# Acknowledgement

---

- Thank all BioHPC team members for their support.
- Please acknowledge our contribution by adding the following sentence to your paper:

This research was supported in part by the computational resources provided by the BioHPC supercomputing facility located in the Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center.

# Questions?

**Thanks for your  
attention!**