Making AlphaFold2 Accessible to Biologists: Streamlining Protein Structure Prediction for Non-Technical Users

Miguel Esteva Avila and Julie Iskander

Walter and Eliza Hall Institute of Medical Research

We’d like to acknowledge the Traditional Owners of the land on which we meet and share our knowledge today, the people of the Kulin nation.

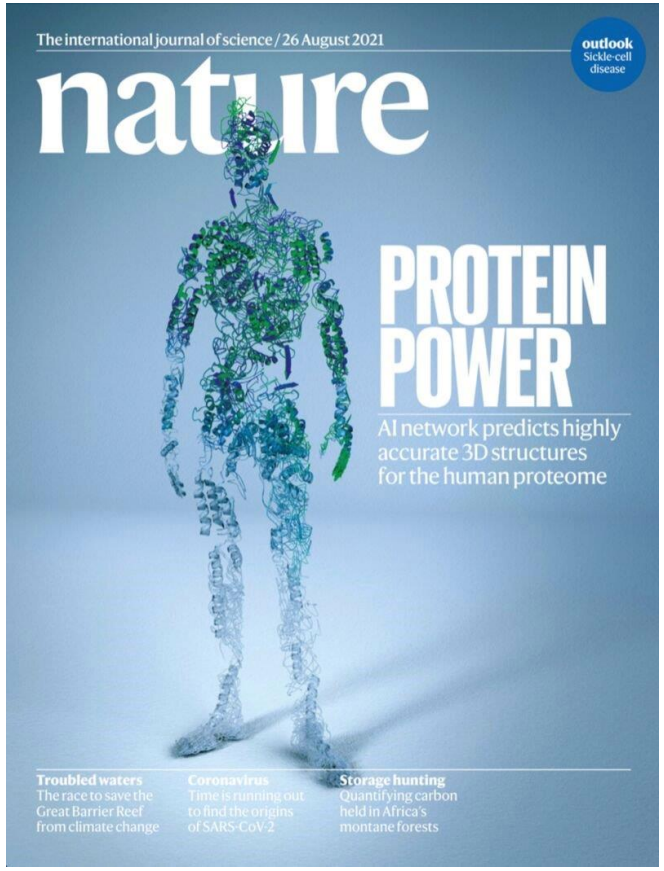
We pay our respects to Elders past, present, and emerging.


WEHI
brighter together
Walter+Eliza Hall
Walter+Eliza Hall
Institute of Medical Research

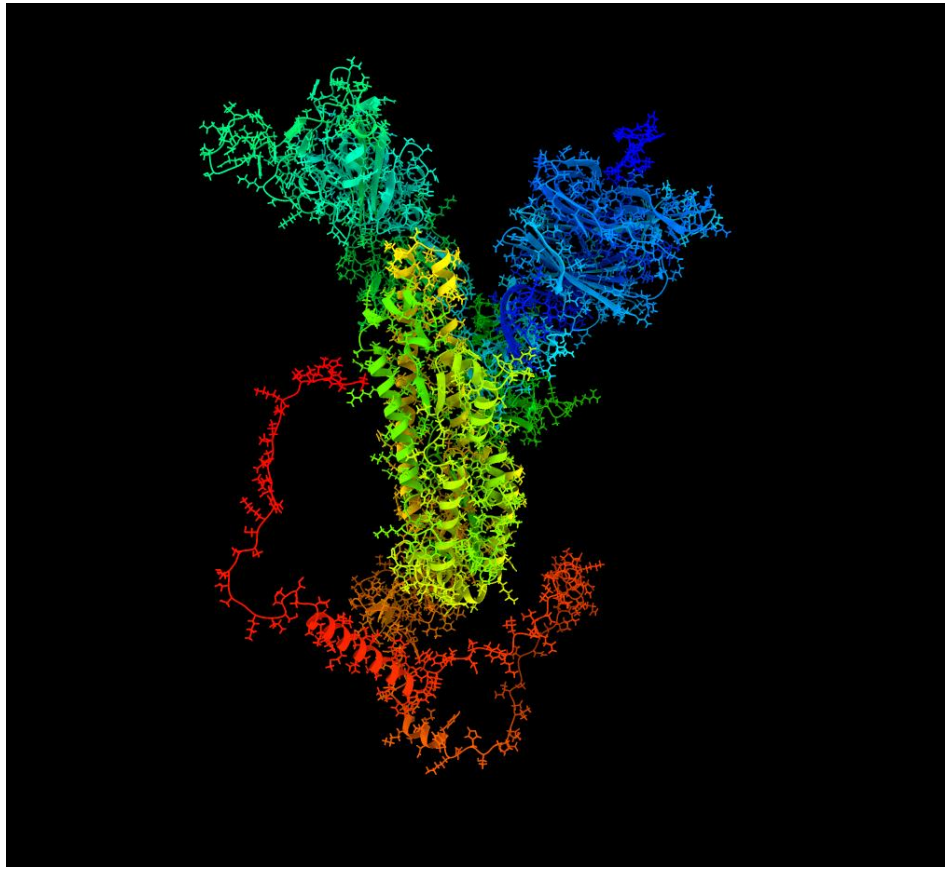
WHY ALPHAFOLD?

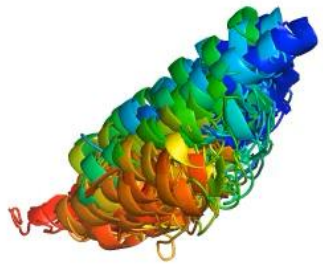
Why AlphaFold?

- For decades, scientists have been trying to find a method to reliably predict a protein's structure just from its sequence of amino acids.
- **"Protein Folding Problem"**
- A complementary alternative to determining it through costly and time-consuming experimentation – could help dramatically accelerate research.



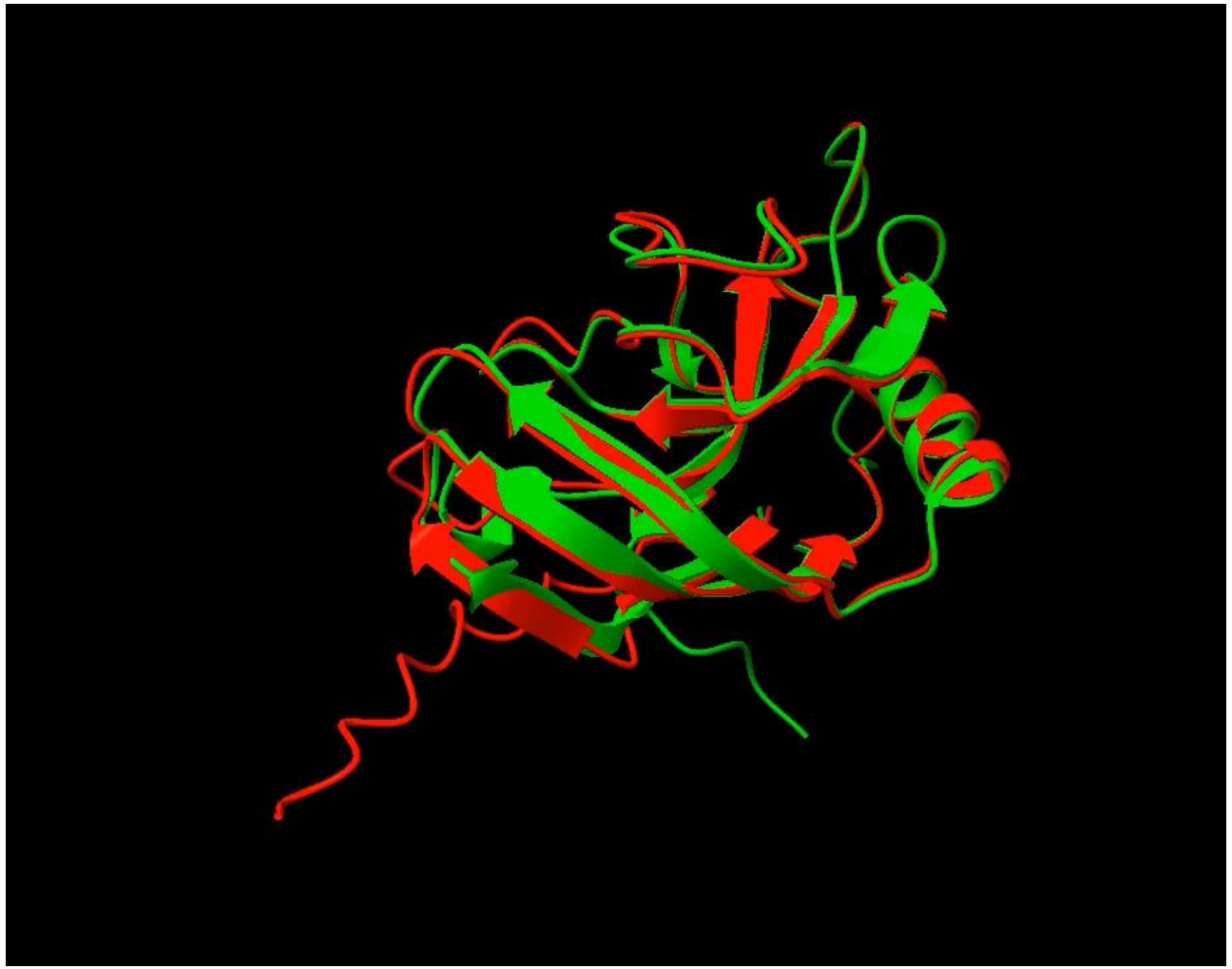
```
>YP_009724390.1 S [organism=Severe acute respiratory syndrome coronavirus 2] [GeneID=43740568]
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHV
SGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPF
LGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPI
NLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYN
ENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASV
YAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIAD
YNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYF
PLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFL
PFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLT
PTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLG
AENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGI
AVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC
LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIG
VTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDI
LSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLM
SFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNT
FVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVA
KNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD
SEPVLKGVKLHYT
```





Recycling iteration 0, block 01
Secondary structure assigned from the final prediction

https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-021-03819-2/MediaObjects/41586_2021_3819_MOESM4_ESM.mp4



6Y4F Green: experimental, Red: AlphaFold2

The 2024 chemistry laureates

The Nobel Prize in Chemistry 2024 was awarded with one half to David Baker "for computational protein design" and the other half jointly to Demis Hassabis and John M. Jumper "for protein structure prediction".

Demis Hassabis and John Jumper have successfully utilised artificial intelligence to predict the structure of almost all known proteins. David Baker has learned how to master life's building blocks and create entirely new proteins.

David Baker, Demis Hassabis and John Jumper. Ill. Niklas Elmehed © Nobel Prize Outreach

https://nobelprize.org/prizes/chemistry/

THE ALPHAFOLD JOURNEY AT WEHI

July 2021

Article | Open access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper ✉, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, ... Demis Hassabis ✉ + Show authors

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

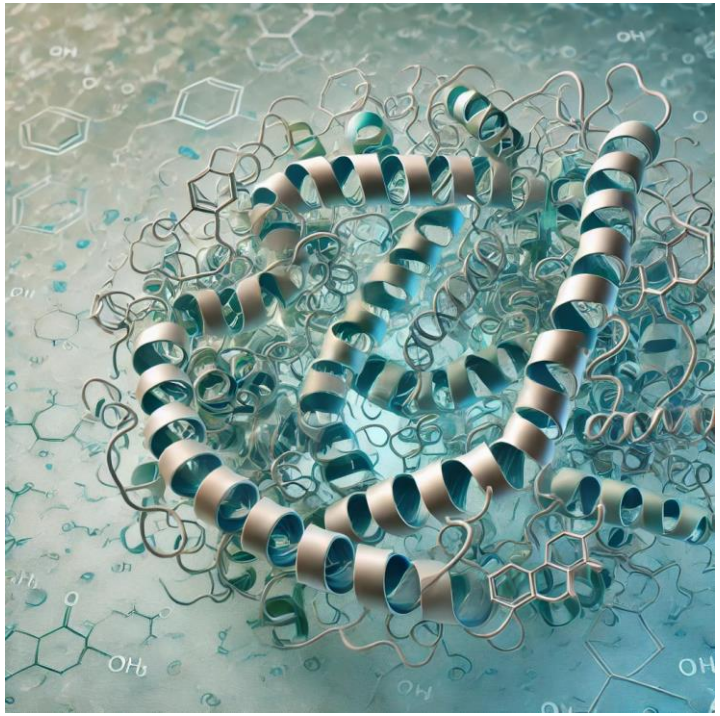
Shortly after

“How can I run AlphaFold?”

”Would it be possible to make AlphaFold available as a module on HPC?”

Challenges running AlphaFold for Biologists

- Complexity and computational requirements of AlphaFold2.
- Lack of technical skills among many biologists for HPC tasks.
- Challenges profiling jobs.
- Complex conda environment.



WEHI
brighter together

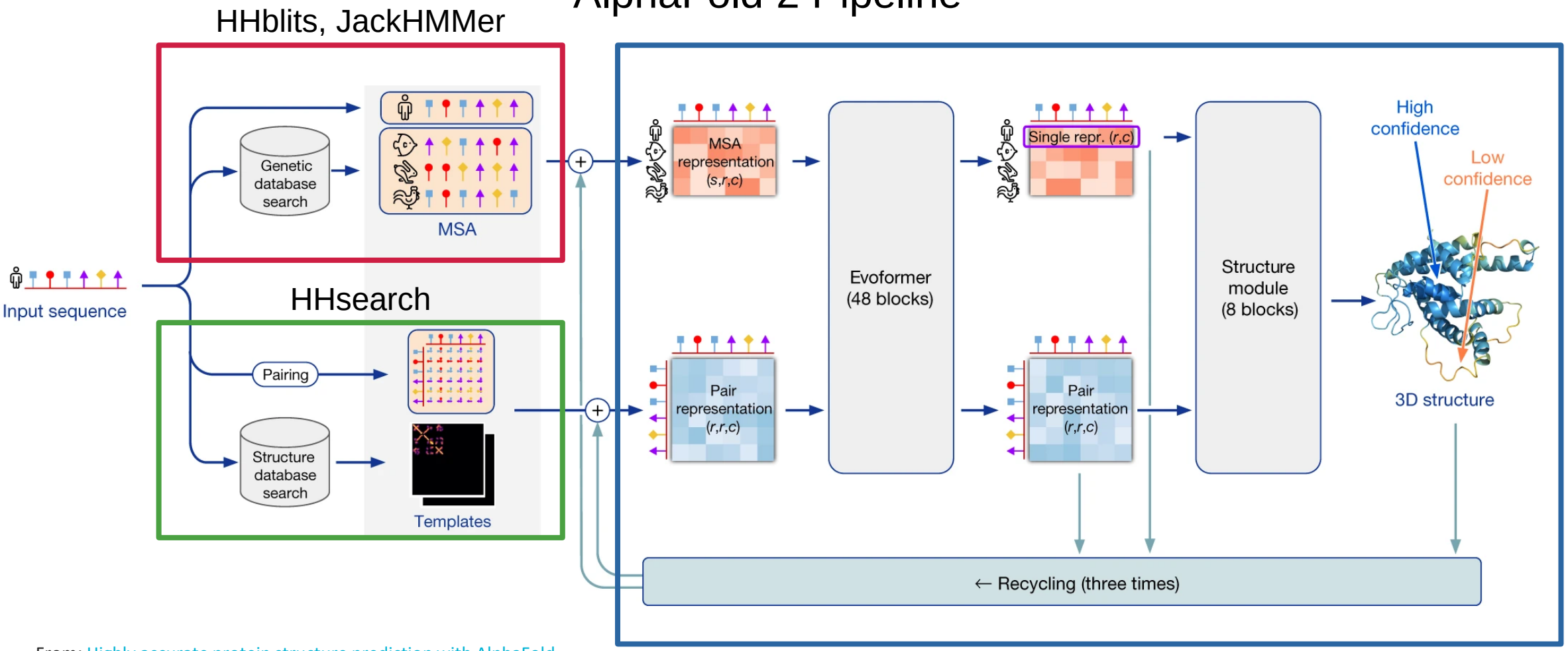
Initial challenges

- Docker was the most robust way to run AlphaFold.
- Required databases have a large storage footprint.

Singularity implementation of AlphaFold2

- AlphaFold2.sif was born!
- Better dependency management.
- Based on original Dockerfile.
- Still challenging for a lot of interested researchers.

AlphaFold 2 Pipeline



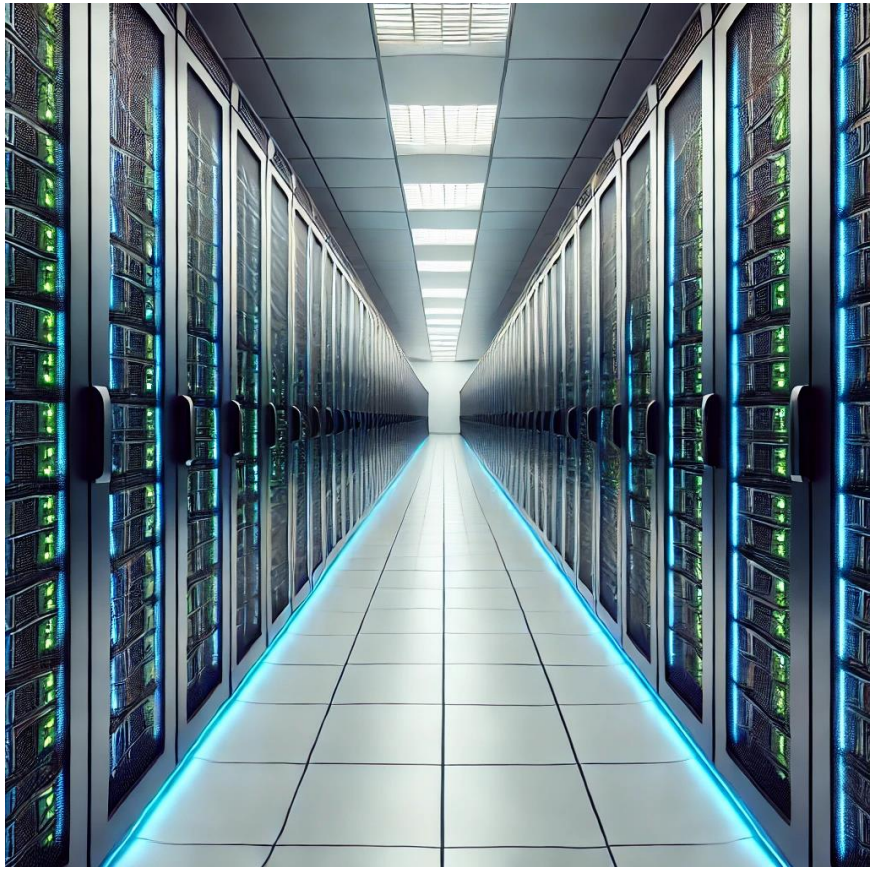
From: [Highly accurate protein structure prediction with AlphaFold]

AlphaFold 2.0.0/1

- Re-runs of the same experiment starting from scratch.
- Whole workflow running as a single job (CPU only jobs using GPU nodes).

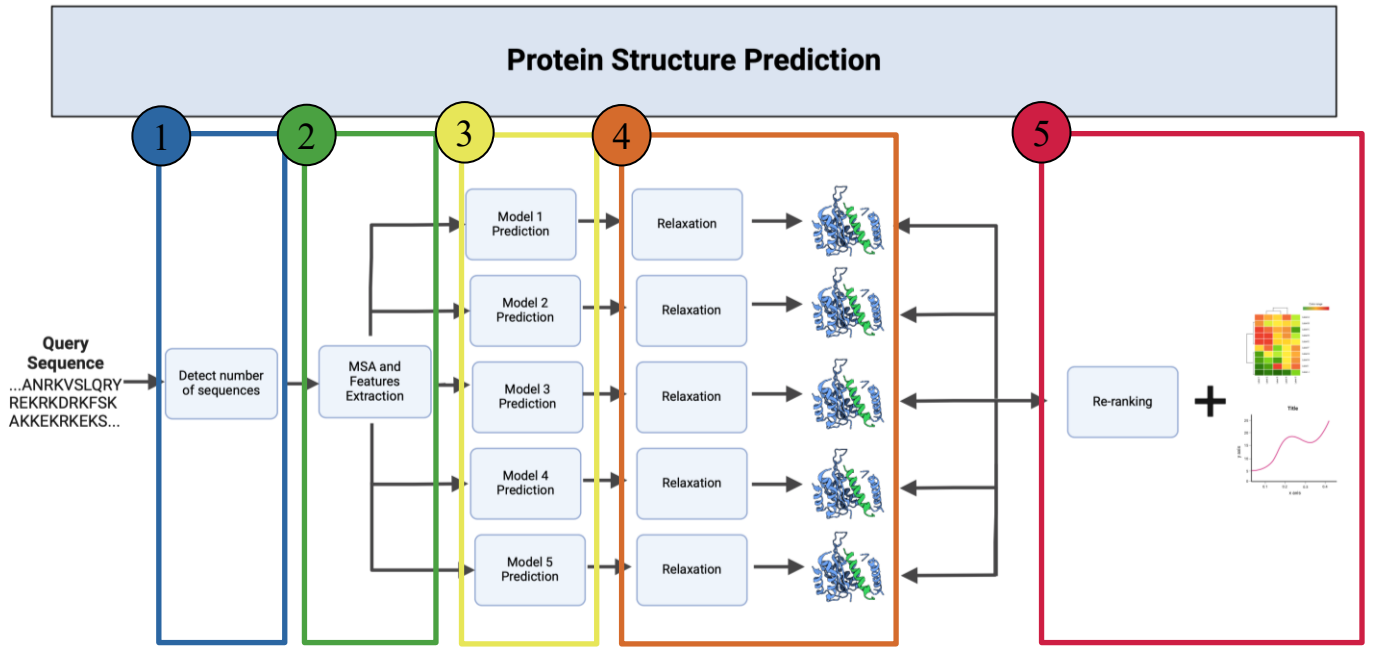
Challenges running AlphaFold on Milton HPC

- Use of command-line.
- Time-limit on GPUs (48 hours).
- Efficient use of resources (GPUs).



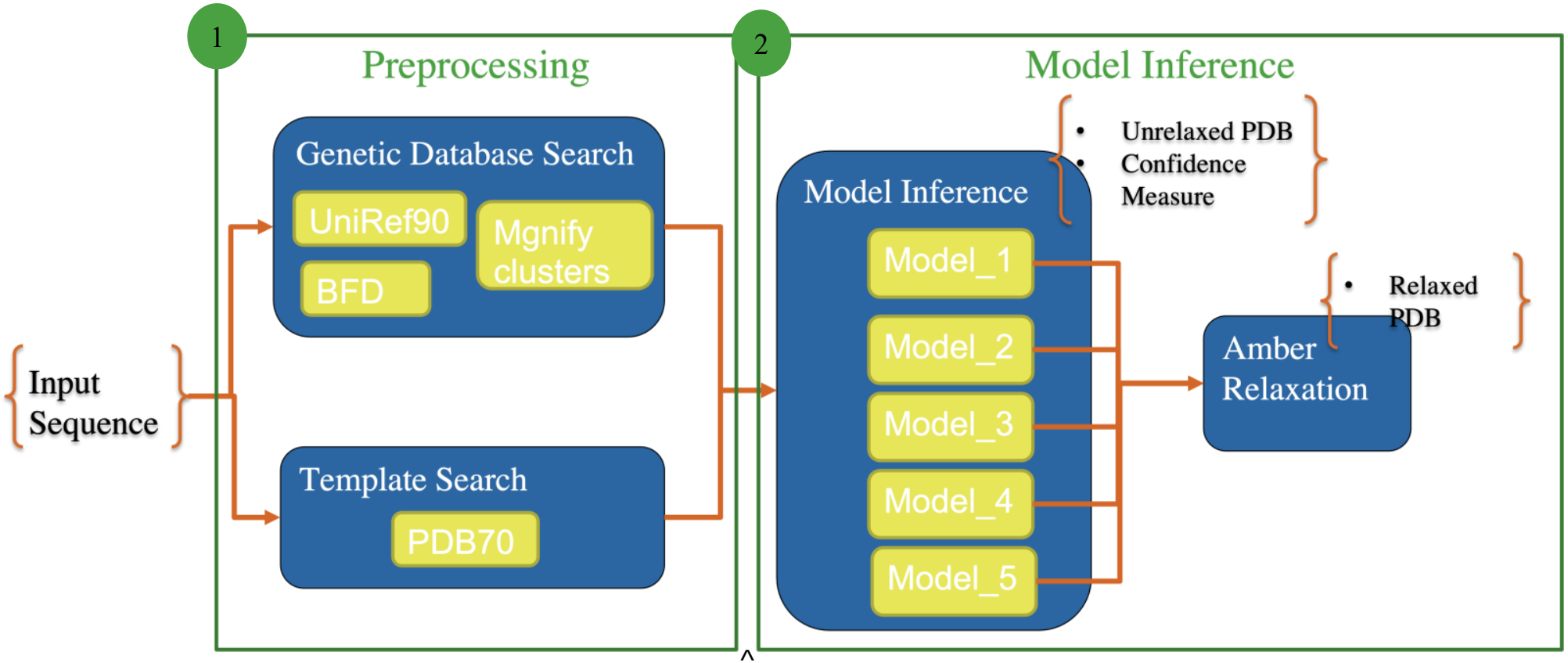
Modified AlphaFold Pipeline

- Modified code to run each step independently, in separate jobs.
- Resource control for MSA, template, and features.
- More efficient resource utilisation.



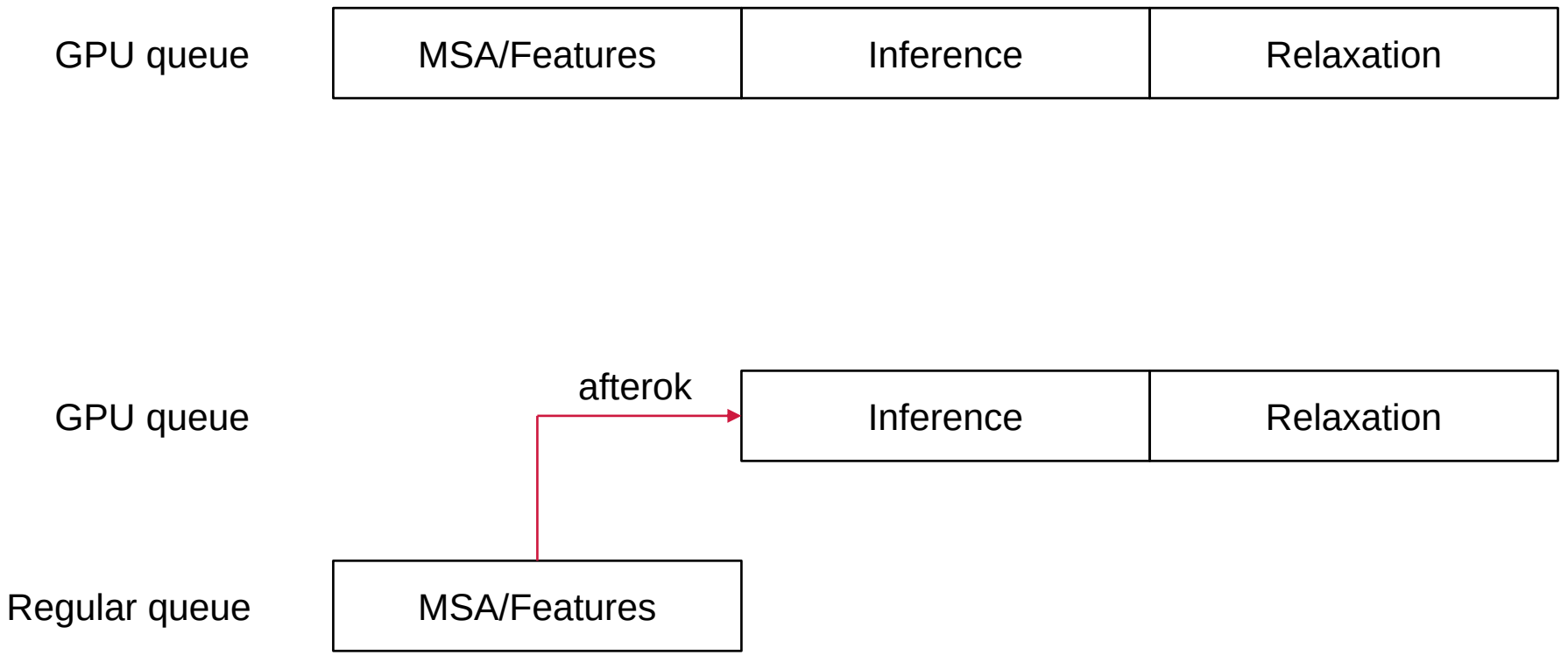
Faster results!

AlphaFold 2 Simplified



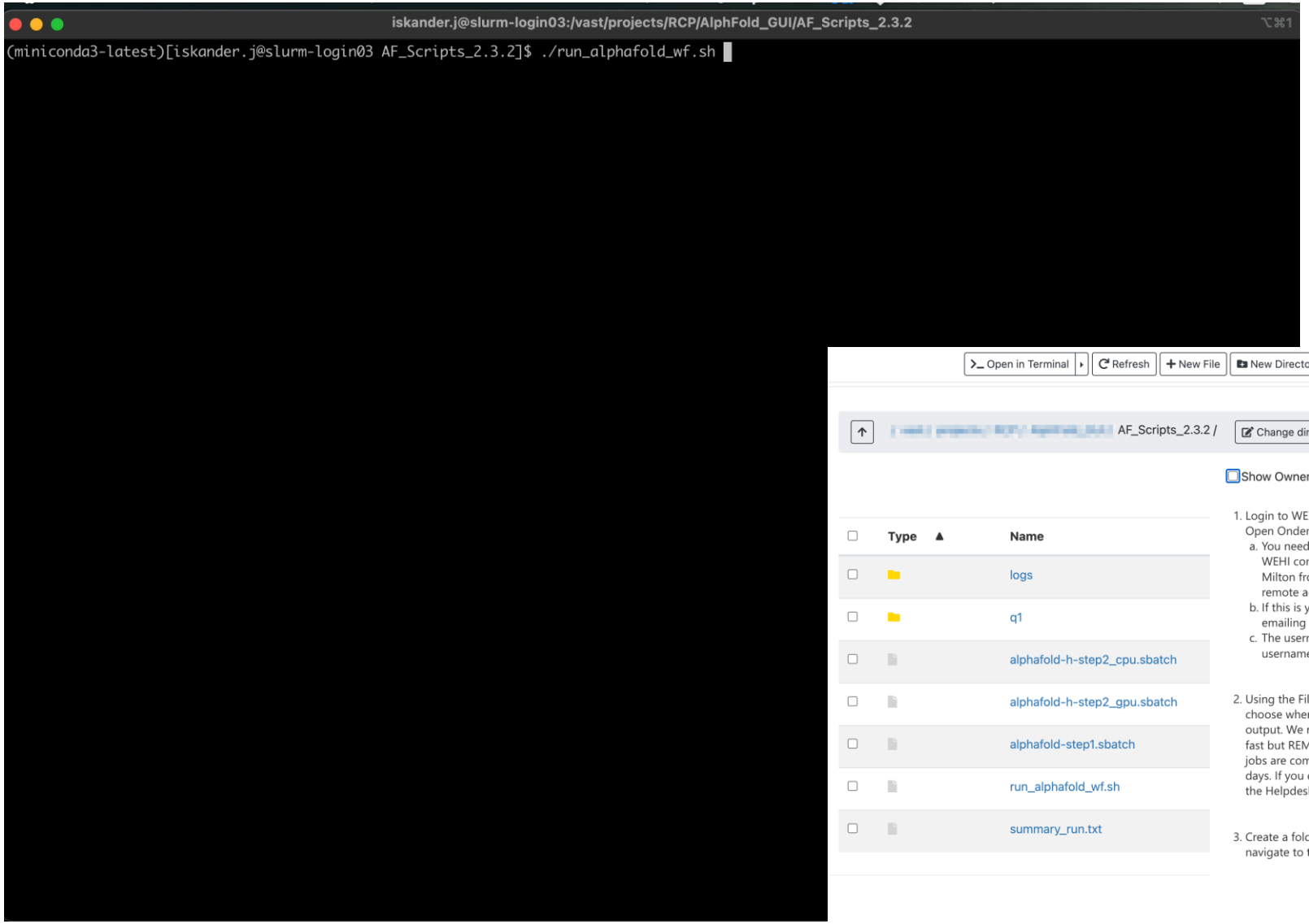
features.pkl

```
# alphafold -h

Usage: alphafold [OPTION: -g -l -c -d -p -f -m -n -i -b -r -x -u -h] -o <output path> -t <max template date> <input FASTA file(s)>...



Apptainer/Singularity is needed to run this script!

OPTION:
  -o  output path            [Required] Path to a directory that will store the results. Will be created if non-existent.
  -g  use GPU                Enable (true)/disable (false) the use of GPU (Default: true).
  -l  GPU(s) to use          Comma separated list of GPU(s) to use e.g. 0,1,2,3. Will be ignored in Slurm.
  -c  custom templates path  Path to a directory containing custom cif templates. (Default: /vast/projects/alphafold/databases/pdb_mmcif/mmcif_files).
  -s  pdb seqres db path     Path to a file containing custom templates. (Default: /vast/projects/alphafold/databases/pdb_seqres/pdb_seqres.txt).
  -d  databases path         Path to the directory containing relevant databases. (Default: /vast/projects/alphafold/databases).
  -p  database preset        Choose db preset model (reduced_dbs|full_dbs). (Default: full_dbs).
  -f  features only          Stop pipeline once feature extraction is complete and features.pkl is created.
  -j  relax only             If set run relaxation only, checks if unrelaxed pdbs exist (Default: false).
  -m  model preset           Model to use (monomer|monomer_casp14|monomer_ptm|multimer). (Default: monomer).
  -n  model indices          Model indices to use (comma-separated list 0,...,4. No spaces). (Default: 0).
  -i  num of predictions     Number of multimer predictions per model (each with a different random seed)
                                 will be generated per model. E.g. if this is 2 and there are 5 models,
                                 then there will be 10 predictions per input. (Default: 5 per model).
                                 Note: this FLAG only applies if model_preset is multimer (-m multimer).
  -b  benchmark mode         Enable benchmark mode by setting -b. (Default: disabled).
                                 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.
  -t  max template date     [Required] Maximum template release date to consider in YYYY-MM-DD format.
  -r  models to relax       The models to run the final relaxation step on. (all|best|none, default: best).
                                 If 'all', all models are relaxed which is time consuming..
                                 If 'best', only the most confident model is relaxed.
                                 If 'none' relaxation is not run (might yield models with stereochemical violations).
  -x  use CPU relax         Relaxation on CPU (will disable GPU relaxation). Set with no arguments.
                                 Relax on GPU can be much faster than CPU, so it is recommended
                                 to set when GPU is not being used or available. (Default: use GPU).
  -u  use precomputed msas  Whether to read MSAs that have been written to disk. (Default: false).
  -h  help                  Print this help menu.

    Environment variables:

       AF_UNIREF30_DB: override HHblits uniref30 database. A directory containing UniRef30_YYYY_MM_* files.
       AF_PDB70_DB: override HHsearch pdb70 database. A directory containing pdb70_* files.
```


GPU queue
MSA/Features
Inference
Relaxation
GPU queue
afterok
Inference
Relaxation
Regular queue
MSA/Features

Simplified AlphaFold

```
#!/bin/bash
#SBATCH --job-name=AlphaFold
#SBATCH --partition=gpuq
#SBATCH --error=AF.err
#SBATCH --output=AF.log
#SBATCH --time=0-05:00:00
#SBATCH --gres=gpu:A100:1
#SBATCH --mem=64G
#SBATCH -c 8
#SBATCH -n 1

module load alphafold/2.3.2
alphafold -m monomer -t $(date "+%Y-%m-%d") -o output T1050.fasta
```



Bash Script Wrapper Step-by-step guide

1. Login to WEHI:
   Open Ondemand (https://ondemand.hpc.wehi.edu.au/)
   a. You need to be connected to the WEHI network by using your WEHI computer onsite or through the VPN. If you are accessing Milton from a non-WEHI device, you need to log in via the remote access portal (https://rap.wehi.edu.au/).
   b. If this is your first time, you need to request access to Slurm by emailing support@wehi.edu.au.
   c. The username and password will be the same as your WEHI username and password.

2. Using the Files Menu (Figure 2), after logging in to Ondemand, choose where you want to create your fasta file and get your output. We recommend using VAST Scratch for that. VAST is fast but REMEMBER to copy the results to another area once the jobs are completed to avoid loss if they are not accessed in 14 days. If you do not have access to VAST, request access through the Helpdesk. In this tutorial, VAST is used (Figure 3).

3. Create a folder for the project (Figure 4) for example, called AF, and navigate to the folder (Figure 5).

Figure 2

Figure 3

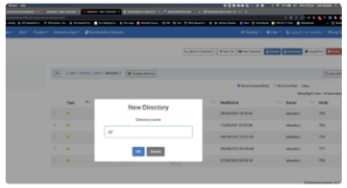


Figure 4

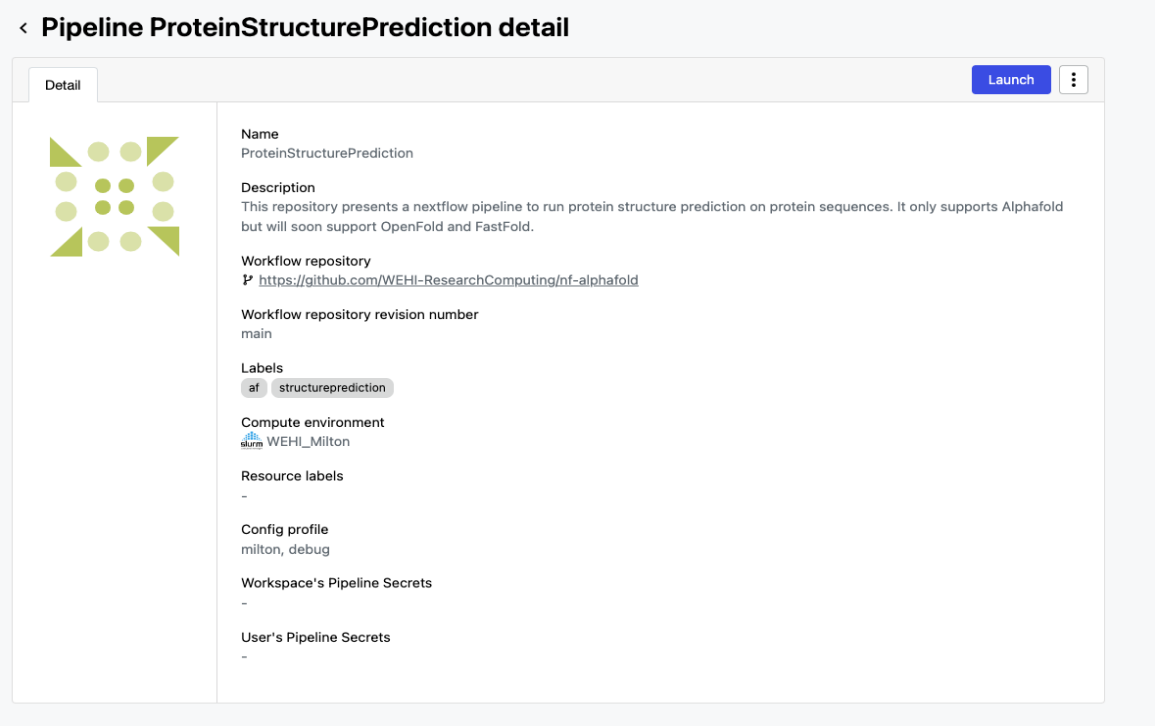
Figure 5

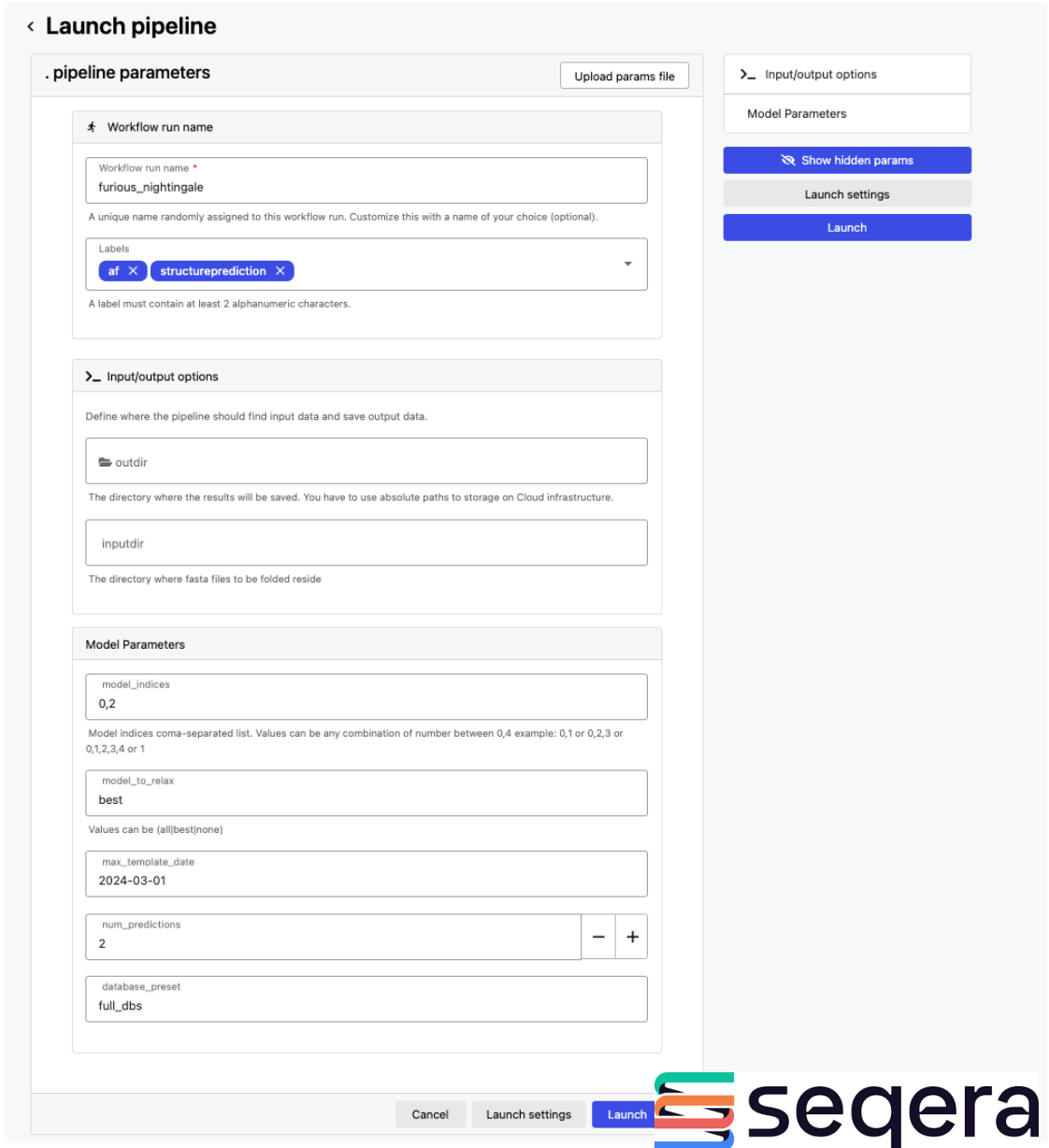
IS THIS ACCESSIBLE ENOUGH?

Challenges

- Folding batches of proteins was still time-consuming and manual.
- Require interacting with command line.
- Troubleshooting meant digging up log files from the filesystem.
- Some biologists found it a hinderance still.

Using Seqera Tower





- WEHI Custom Nextflow Pipeline
- Enhanced Efficiency and Speed

What next?

1

Other folding tools to run in parallel

- OpenFold
- FastFold
- RosettaFold-AA

2

Structural protein structure comparison

- Dalilite
- Foldseek

3

Protein Design Workflows

- RF Diffussion
- ProteinMPNN

Acknowledgements

- RC Team
  - RCP
  - ITS RS Team
- Australia Bio Commons
  - Ziad El Bkhetan
  - Johan Gustafsson
- Structural Biology Division
  - Richard Birkinshaw
  - Joshua Hardy

WEHI
brighter together

Questions