UT Southwestern Medical Center
Lyda Hill Department of Bioinformatics

BioHPC

Alphafold – Google Deepmind's protein prediction platform

[web] portal.biohpc.swmed.edu
[email] biohpc-help@utsouthwestern.edu

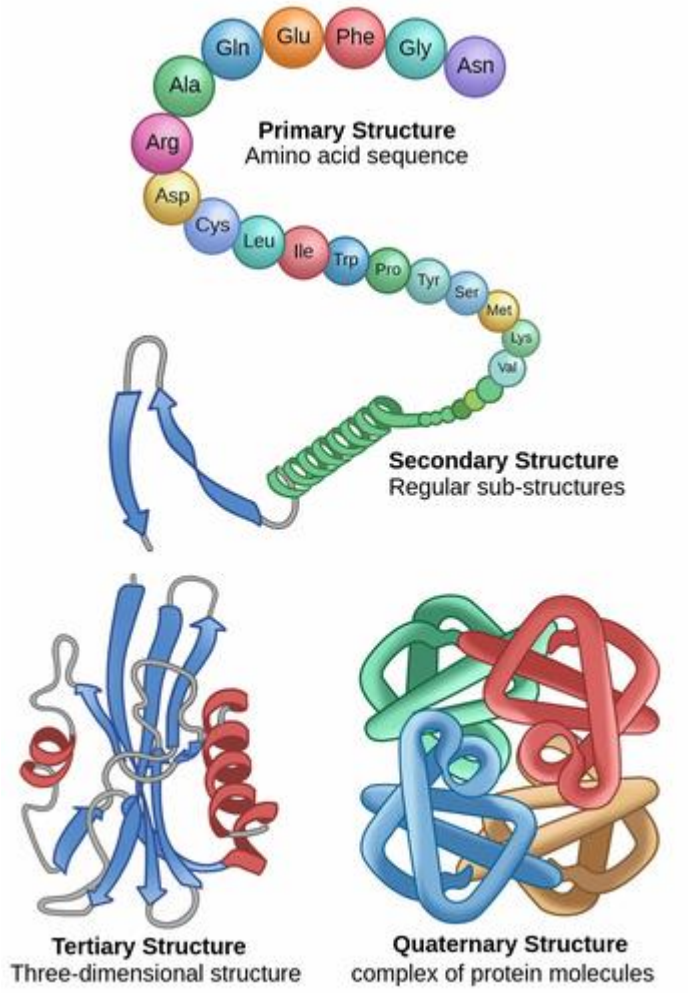
2021-09-15

Outline

- Protein structure
- CASP, the Critical Assessment of protein Structure Prediction
- Alphafold paper
- Alphafold source code on git hub
- Alphafold database
- Alphafold limitation
- Demo: how to run Alphafold on BioHPC

Protein structure

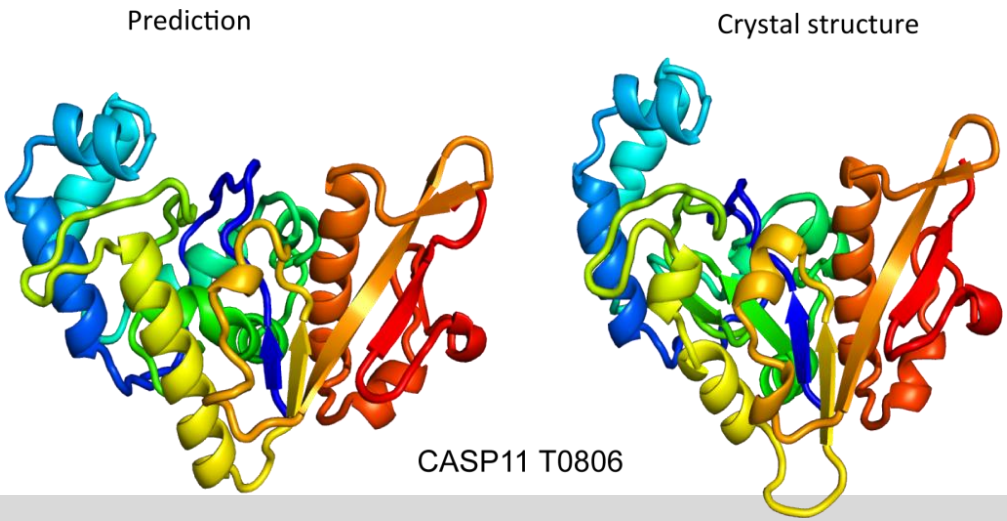
- Protein structure
- Protein folding problem
  - Sequence of amino acids do not show how they fold into shape
- Why is protein folding important?
  - Protein structure dictates its function
  - Scientists can develop drugs based on known protein structures
- Methods:
  - X-ray crystallography
  - NMR (Nuclear Magnetic Resonance)
  - CryoEM
  - AI (Artificial Intelligence)



Ref to
https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery

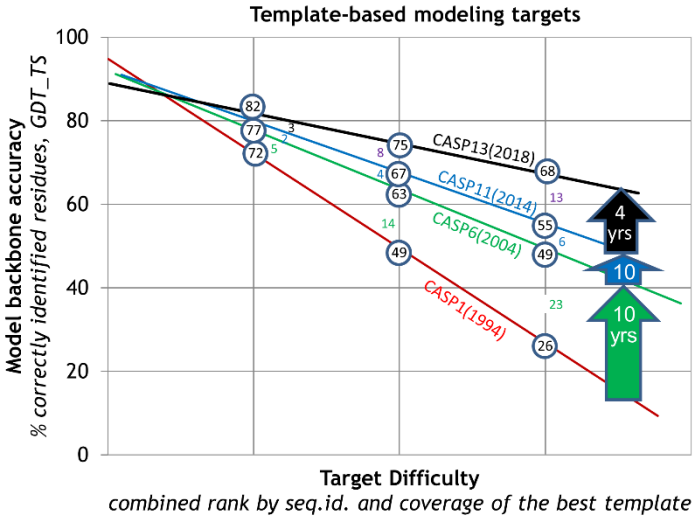
CASP

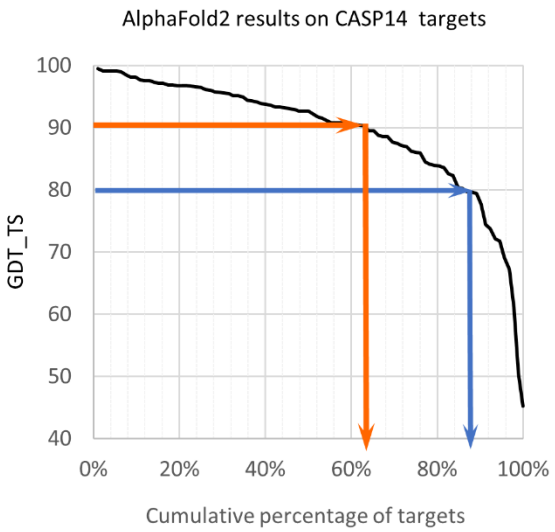
- CASP, Critical Assessment of protein Structure Prediction
- Community-wide, worldwide experiment for protein structure prediction
- Take place every two years since 1994
- Problem:
  - Target proteins: structures just been solved and hold by Protein Data Bank
  - Neither predictors nor the organizers and assessors know the structures of the target proteins
  - Evaluation: GDT-TS (global Distance Test – Total Score) describing percentage of well-modeled residues in the model with respect to the target

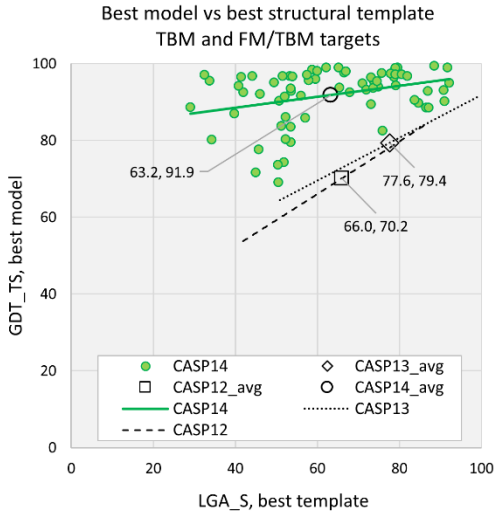


Alphafold in CASP

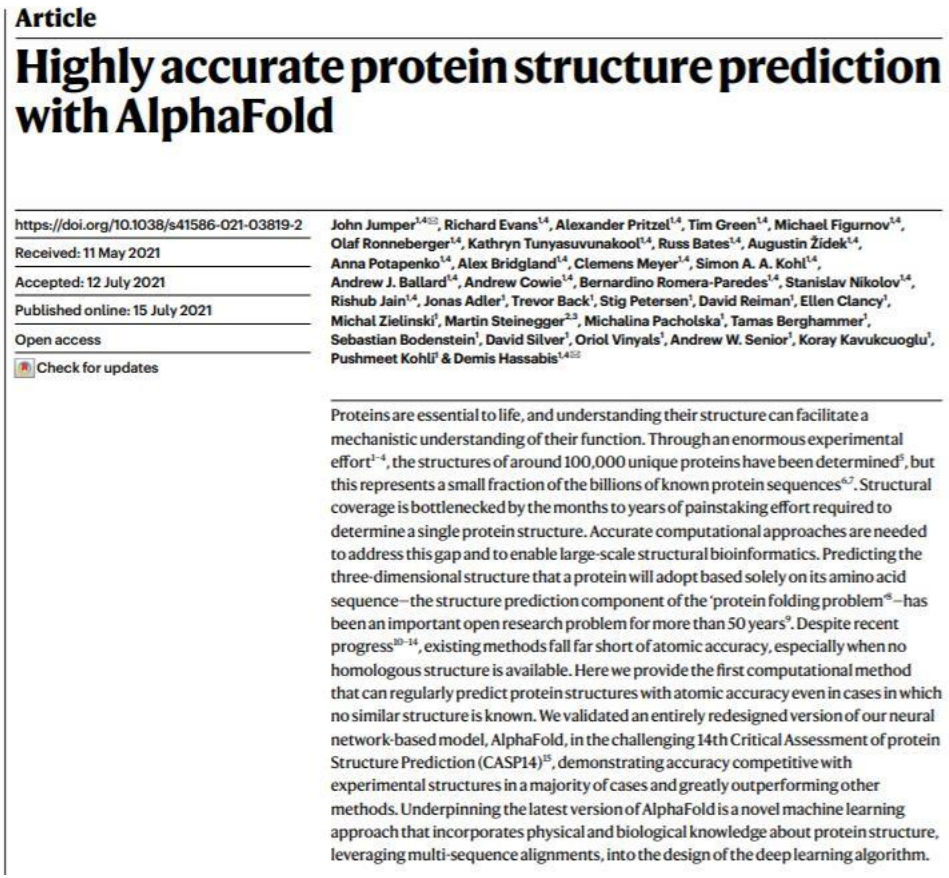
- Over the course of CASP, template-based modeling get enormous improvement
- The 2014-2018 model accuracy improvement doubled that of 2004-2014
- CASP14 marked an extraordinary increase in the accuracy of the computed three-dimensional protein structures with the emergence of the advanced deep learning 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 paper



Article

Highly accurate protein structure prediction with AlphaFold

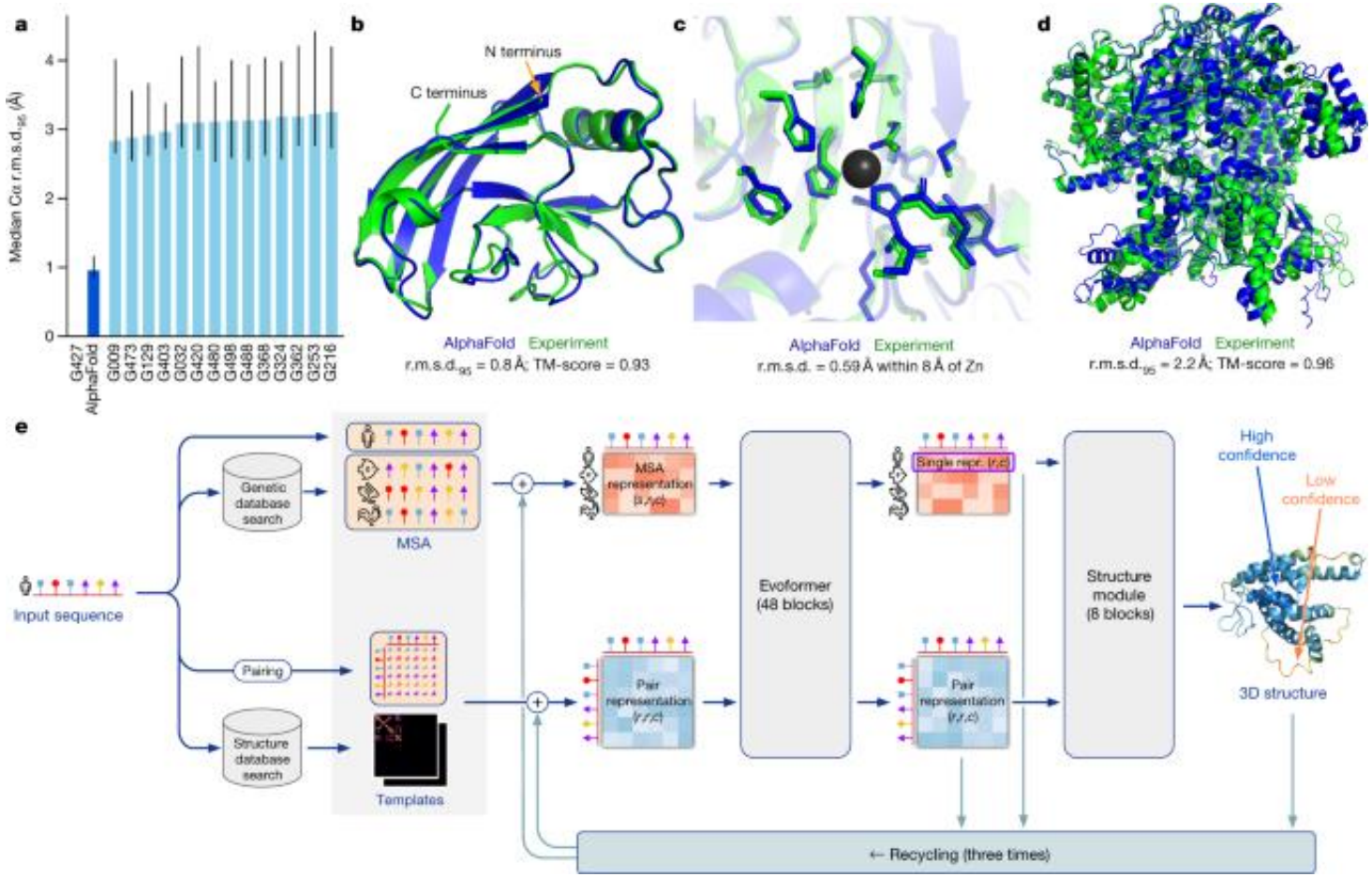
https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4✉], Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4✉]

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort[1–4], the structures of around 100,000 unique proteins have been determined[5], but this represents a small fraction of the billions of known protein sequences[6,7]. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the 'protein folding problem'[8]—has been an important open research problem for more than 50 years[9]. Despite recent progress[10–14], existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)[15], demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.

Alphafold paper

- 

Alphafold code on github

A) Deepmind https://github.com/deepmind/alphafold

- Docker:
  - Use “root” (administrative account) to run docker
  - On HPC, users are not allowed to use “root” to run docker
- Databases:
  - full database (2.2TB) /project/apps_database/alphafold/database_full
  - reduced database (415GB)
- Running Alphafold: run_docker.py

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

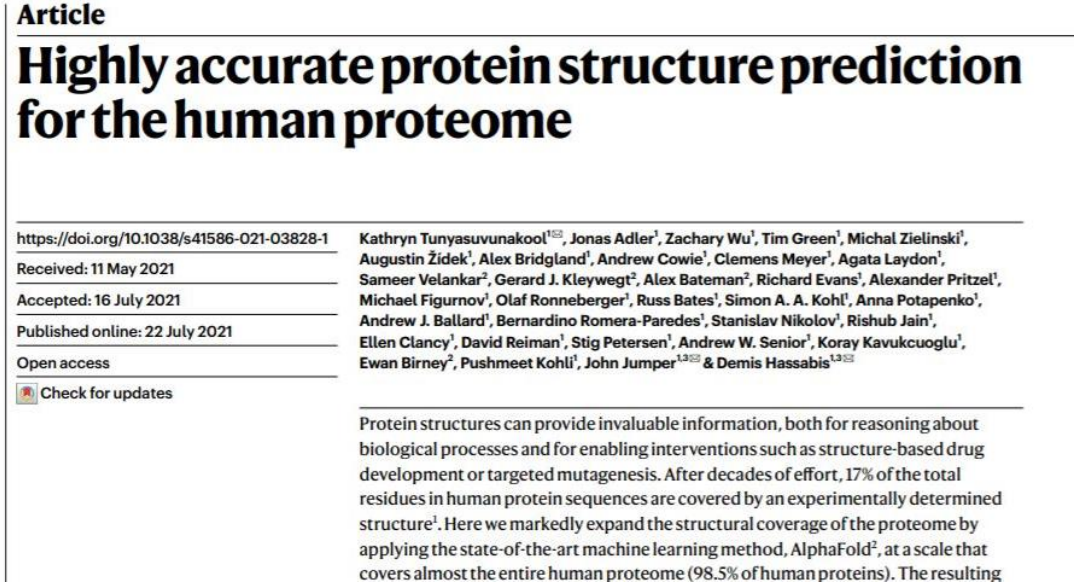
B) Alphafold non docker https://github.com/kalininalab/alphafold_non_docker

- Conda environment instead of docker
- Download Alphafold git repo
- Databases: full database and reduced database, same as above
- Install as a BioHPC module
- Running Alphafold: run_alphafold.sh calls run_alphafold.py (Deepmind)

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

Alphafold database

- https://alphafold.ebi.ac.uk/
- collaboration between Deepmind and EMBL-EBI
- Human proteome (98.5%): 58% of residues with a confident prediction, of which a subset (36% of all residues) have very high confidence
- Will cover over 100 million proteins in UniRef90 in the coming months



Article

Highly accurate protein structure prediction for the human proteome

https://doi.org/10.1038/s41586-021-03828-1
Received: 11 May 2021
Accepted: 16 July 2021
Published online: 22 July 2021
Open access
Check for updates

Kathryn Tunyasuvunakool[1], Jonas Adler[1], Zachary Wu[1], Tim Green[1], Michal Zielinski[1], Augustin Žídek[1], Alex Bridgland[1], Andrew Cowie[1], Clemens Meyer[1], Agata Laydon[1], Sameer Velankar[2], Gerard J. Kleywegt[2], Alex Bateman[2], Richard Evans[1], Alexander Pritzel[1], Michael Figurnov[1], Olaf Ronneberger[1], Russ Bates[1], Simon A. A. Kohl[1], Anna Potapenko[1], Andrew J. Ballard[1], Bernardino Romera-Paredes[1], Stanislav Nikolov[1], Rishub Jain[1], Ellen Clancy[1], David Reiman[1], Stig Petersen[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Ewan Birney[2], Pushmeet Kohli[1], John Jumper[1,3] & Demis Hassabis[1,3]

Protein structures can provide invaluable information, both for reasoning about biological processes and for enabling interventions such as structure-based drug development or targeted mutagenesis. After decades of effort, 17% of the total residues in human protein sequences are covered by an experimentally determined structure[1]. Here we markedly expand the structural coverage of the proteome by applying the state-of-the-art machine learning method, AlphaFold[2], at a scale that covers almost the entire human proteome (98.5% of human proteins). The resulting

Tunyasuvunakool, Kathryn, et al. "Highly accurate protein structure prediction for the human proteome." Nature 596.7873 (2021): 590-596.

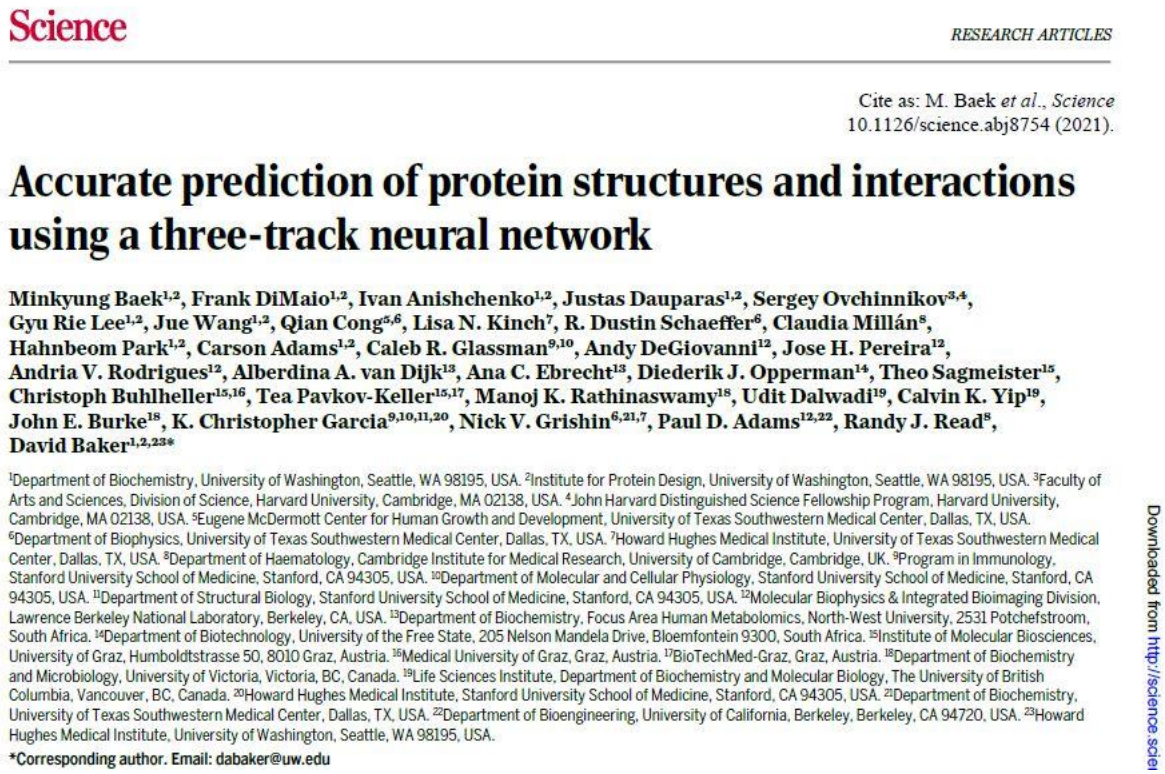
Alphafold limitation

- Current use cases: predicting the structure of a single protein chain with a naturally occurring sequence
- Limitations:
  - Multi-chain prediction (complex)
  - Regions that are intrinsically disordered or unstructured in isolation
  - AlphaFold has not been validated for predicting the effect of mutations
  - Where a protein is known to have multiple conformations AlphaFold usually only produces one of them
  - AlphaFold does not predict the positions of any non-protein components found in experimental structures (such as cofactors, metals, ligands, ions, DNA/RNA, or post-translational modifications)

Ref to https://alphafold.ebi.ac.uk/faq#faq-8

Also…

- RoseTTAfold from Baker's lab in University of Washington in Seattle



Science

RESEARCH ARTICLES

Cite as: M. Baek *et al.*, *Science*
10.1126/science.abj8754 (2021).

Accurate prediction of protein structures and interactions using a three-track neural network

Minkyung Baek[1,2], Frank DiMaio[1,2], Ivan Anishchenko[1,2], Justas Dauparas[1,2], Sergey Ovchinnikov[3,4], Gyu Rie Lee[1,2], Jue Wang[1,2], Qian Cong[5,6], Lisa N. Kinch[7], R. Dustin Schaeffer[6], Claudia Millán[8], Hahnbeom Park[1,2], Carson Adams[1,2], Caleb R. Glassman[9,10], Andy DeGiovanni[12], Jose H. Pereira[12], Andria V. Rodrigues[12], Alberdina A. van Dijk[13], Ana C. Ebrecht[13], Diederik J. Opperman[14], Theo Sagmeister[15], Christoph Buhlheller[15,16], Tea Pavkov-Keller[15,17], Manoj K. Rathinaswamy[18], Udit Dalwadi[19], Calvin K. Yip[19], John E. Burke[18], K. Christopher Garcia[9,10,11,20], Nick V. Grishin[6,21,7], Paul D. Adams[12,22], Randy J. Read[8], David Baker[1,2,23]*

[1]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA 98195, USA. [3]Faculty of Arts and Sciences, Division of Science, Harvard University, Cambridge, MA 02138, USA. [4]John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA 02138, USA. [5]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX, USA. [6]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [7]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA. [8]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. [9]Program in Immunology, Stanford University School of Medicine, Stanford, CA 94305, USA. [10]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA 94305, USA. [11]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA 94305, USA. [12]Molecular Biophysics & Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [13]Department of Biochemistry, Focus Area Human Metabolomics, North-West University, 2531 Potchefstroom, South Africa. [14]Department of Biotechnology, University of the Free State, 205 Nelson Mandela Drive, Bloemfontein 9300, South Africa. [15]Institute of Molecular Biosciences, University of Graz, Humboldtstrasse 50, 8010 Graz, Austria. [16]Medical University of Graz, Graz, Austria. [17]BioTechMed-Graz, Graz, Austria. [18]Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada. [19]Life Sciences Institute, Department of Biochemistry and Molecular Biology, The University of British Columbia, Vancouver, BC, Canada. [20]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, CA 94305, USA. [21]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, TX, USA. [22]Department of Bioengineering, University of California, Berkeley, Berkeley, CA 94720, USA. [23]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.

***Corresponding author. Email: dabaker@uw.edu**

Baek, Minkyung, et al. "Accurate prediction of protein structures and interactions using a three-track neural network." Science 373.6557 (2021): 871-876.

GPU nodes on BioHPC

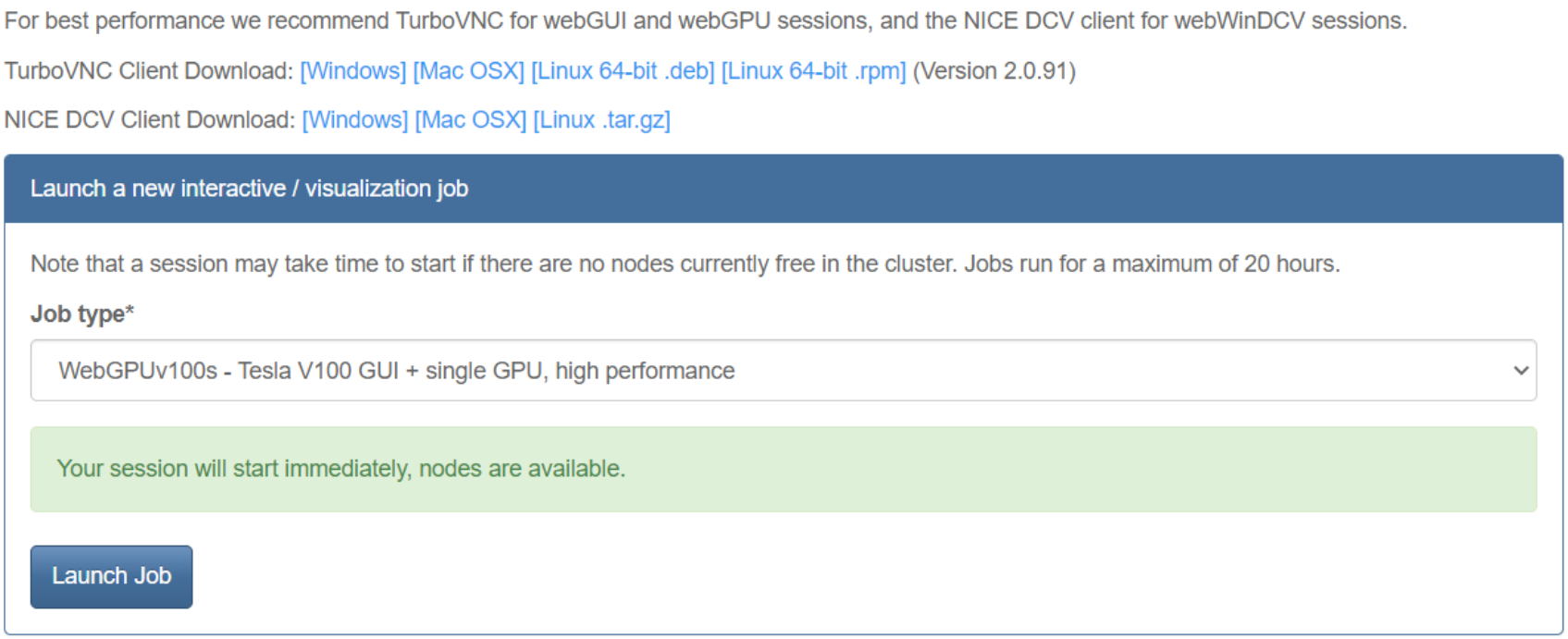
GPU Partition	Number of CPU/Node	Memory Per Node	Number of GPU/Node	GPU Memory	Number of nodes
GPU	32	256GB	1 K20/K40	6GB/12GB	8
GPUp4	72	384GB	1 P4	8GB	16
GPUp40	72	384GB	1 P40	24GB	16
GPUp100	56	256GB	2 P100	16GB	12
GPUv100s	72	384GB	1 V100S	32GB	32
GPU4v100	72	384GB	4 V100S	32GB	12
GPUA100	72	1.5TB	1 A100	40GB	16

Check node availability

```
[s179389@Nucleus005 ~]$ sinfo -p GPUp4
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
GPUp4        up   infinite      5  alloc NucleusC[002,012-013,016-017]
GPUp4        up   infinite     11   idle NucleusC[003-011,014-015]
```

Run alphaFold on BioHPC__webGPU

Launch a webGPU session through: https://portal.biohpc.swmed.edu/terminal/webgui/



Run alphaFold on BioHPC__webGPU

```
[s179389@NucleusC038 alphafold_test]$ ls input_fasta/
query_2.fasta  query.fasta
[s179389@NucleusC038 alphafold_test]$ ls output_alphafold/
test1  test2
[s179389@NucleusC038 alphafold_test]$ module load alphafold/2.0
        This package provides an implementation of the inference pipeline of AlphaFold v2.0, it's
currently running using conda env on biohpc
This application runs under a conda env alphafold/2.0_non_docker.
To run alphafold2:
run_alphafold <option>
To get a full discription of option and paramets:
run_alphafold --help
[s179389@NucleusC038 alphafold_test]$ run_alphafold -d
/project/apps_database/alphafold/database_full  -o ./output_alphafold/test1/  -m
model_1,model_2  -f ./input_fasta/query.fasta -t 2020-05-14
……
I0914 14:22:35.268986 46912496447232 run_alphafold.py:280] Using random seed
5465313646353885249 for the data pipeline
……
```

Run alphaFold on BioHPC__webGPU

Usage: run_alphafold.sh <OPTIONS>
Required Parameters:
-d <data_dir> Path to directory of supporting data
-o <output_dir> Path to a directory that will store the results.
-m <model_names> Names of models to use (a comma separated list)
-f <fasta_path> Path to a FASTA file containing one sequence
-t <max_template_date> Maximum template release date to consider (ISO-8601 format - i.e. YYYY-MM-DD). Important if folding historical test sets

Optional Parameters:
-n <openmm_threads> OpenMM threads (default: all available cores)
-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')
-g <use_gpu> Enable NVIDIA runtime to run with GPUs (default: True)
-a <gpu_devices> Comma separated list of devices to pass to 'CUDA_VISIBLE_DEVICES' (default: 0)
-p <preset> Choose preset model configuration - no ensembling and smaller genetic database config (reduced_dbs), no ensembling and full genetic database config (full_dbs) or full genetic database config and 8 model ensemblings (casp14)

https://github.com/kalininalab/alphafold_non_docker

Run alphaFold on BioHPC__genetic database

```
$ run_alphafold -d /project/apps_database/alphafold/database_full -o
/project/biohpcadmin/s179389/alphafold_test/dummy_test_reduced_database/ -m
model_1 -f /project/biohpcadmin/s179389/alphafold_test/example/query.fasta -t 2020-
05-14 -p reduced_dbs
```

```
/project/apps_database/alphafold/
└── database_full
    ├── bfd
    ├── mgnify
    ├── params
    ├── pdb70
    ├── pdb_mmcif
    ├── scripts
    ├── small_bfd
    ├── uniclust30
    └── uniref90
```

Run alphaFold on BioHPC__input &output

```
/project/biohpcadmin/s179389/alphafold_test
├── input_fasta
│   ├── query_2.fasta
│   └── query.fasta
└── output_alphafold
    ├── query
    └── query_2
```

```
/project/biohpcadmin/s179389/alphafold_git/dummy_test/
└── query
    ├── features.pkl
    ├── msas
    │   ├── bfd_uniclust_hits.a3m
    │   ├── mgnify_hits.sto
    │   ├── pdb70_hits.hhr
    │   └── uniref90_hits.sto
    ├── ranked_0.pdb
    ├── ranking_debug.json
    ├── relaxed_model_1.pdb
    ├── result_model_1.pkl
    ├── test0914.py
    ├── timings.json
    └── unrelaxed_model_1.pdb
```

Run alphaFold on BioHPC__SLURM Job

submit a SLURM job through: https://portal.biohpc.swmed.edu/sbatch/#/script

```
module load alphafold/2.0

# COMMAND GROUP 1


run_alphafold -d /project/apps_database/alphafold/database_full \
        -o /project/biohpcadmin/s179389/alphafold_test/dummy_test_slrum/ \
        -m model_1,model_2 \
        -f /project/biohpcadmin/s179389/alphafold_test/example/query.fasta \
        -t 2020-05-14

# END OF SCRIPT
```

Run alphaFold on BioHPC__SLURM Job__Multiple sequence prediction

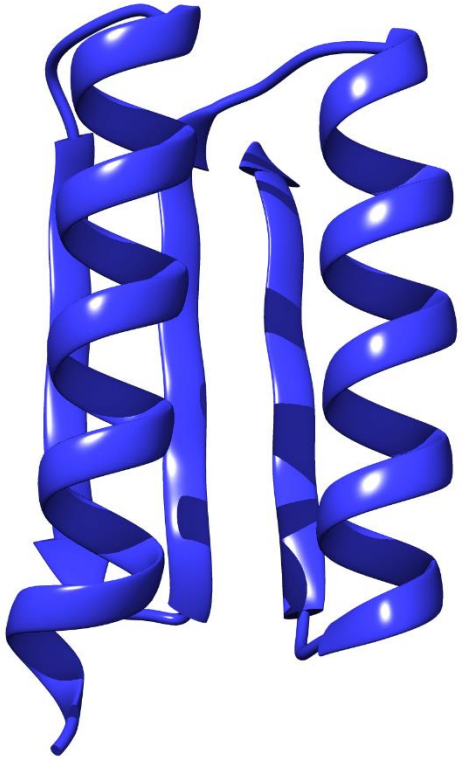
submit a SLURM job through: https://portal.biohpc.swmed.edu/sbatch/#/script

```
module load alphafold/2.0

# COMMAND GROUP 1
run_alphafold -d /project/apps_database/alphafold/database_full \
        -o /project/biohpcadmin/s179389/alphafold_test/dummy_test_slrum/ \
        -m model_1,model_2 \
        -f /project/biohpcadmin/s179389/alphafold_test/example/query_2.fasta \
        -t 2020-05-14 &

run_alphafold -d /project/apps_database/alphafold/database_full \
        -o /project/biohpcadmin/s179389/alphafold_test/dummy_test_slrum/ \
        -m model_1,model_2 \
        -f /project/biohpcadmin/s179389/alphafold_test/example/query.fasta \
        -t 2020-05-14 &
wait
# END OF SCRIPT
```

Run alphaFold on BioHPC__test



Query: 70 AA

Query_2: 350 AA

UT Southwestern Medical Center
Lyda Hill Department of Bioinformatics

BioHPC

Questions? Comments? Collaborations?

Email: biohpc-help@utsouthwestern.edu

Thanks!

Run alphaFold on BioHPC__test

Run alphaFold on BioHPC__test

```
>dummy_sequence
GWSTELEKHREELKEFLKKEGITNVEIRIDNGRLEVRVEGGTERLKRFLEELRQKLEKK
GYTVDIKIE


>sample sequence consisting of 350 residues
MTANHLESPNCDWKNNRMAIVHMVNVTPLRMMEEPRAAVEAAFEGIMEPAVVGDMVEYWN
KMISTCCNYYQMGSSRSHLEEKAQMVDRFWFCPCIYYASGKWRNMFLNILHVWGHHHYPR
NDLKPCSYLSCKLPDLRIFFNHMQTCCHFVTLLFLTEWPTYMIYNSVDLCPMTIPRRNTC
RTMTEVSSWCEPAIPEWWQATVKGGWMSTHTKFCWYPVLDPHHEYAESKMDTYGQCKKGG
MVRCYKHKQQVWGNNHNESKAPCDDQPTYLCPPGEVYKGDHISKREAENMTNAWLGEDTH
NFMEIMHCTAKMASTHFGSTTIYWAWGGHVRPAATWRVYPMIQEGSHCQC
```