

Predict a molecular inventory of your organism

Proteomics bioinformatics Community Meeting - 2024 Dec 02

Dr Keiran Rowell



Sequence genome

4544 proteins



OpenFold

Democratizing AI for Biology

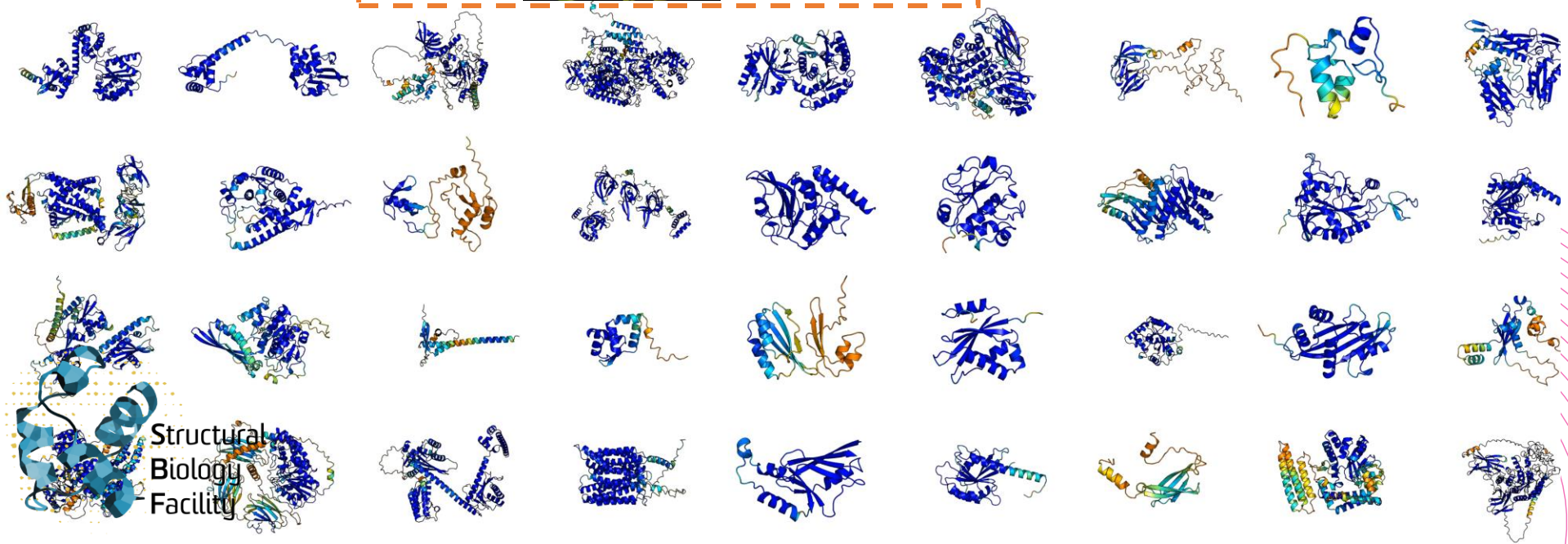
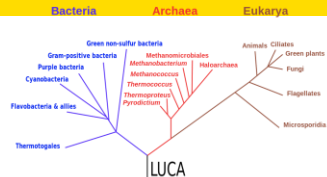
Evo-AI methods

4544 structures

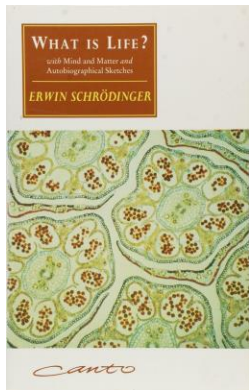


Struct search

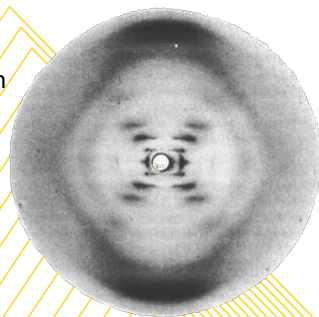
AlphaFoldDB



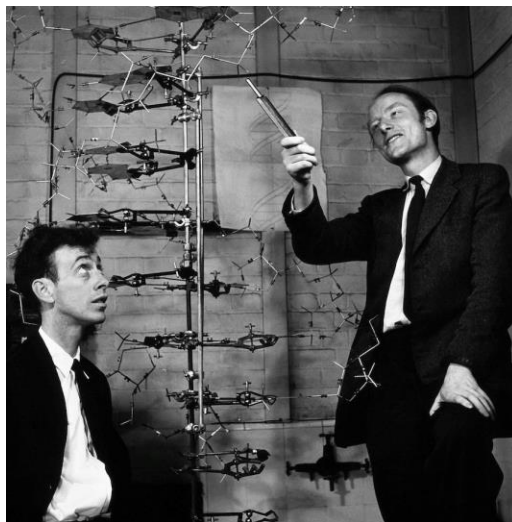
Models in Chemical Biology - function follows chemical form



Schrödinger speculates on the molecules of life
1944



Wilkins & Fraklin's X-ray diffraction images of DNA 1953



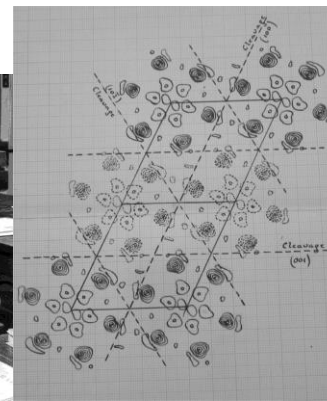
Watson & Crick's DNA double-helix model 1953



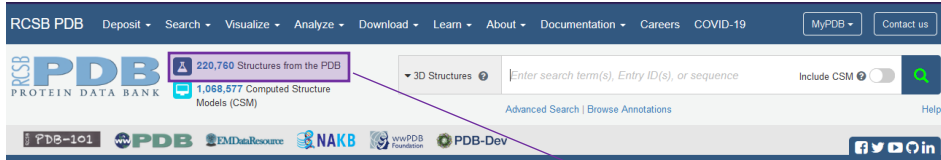
Perutz & Kendrew's model of the 3D structure of a protein (myoglobin)
1957



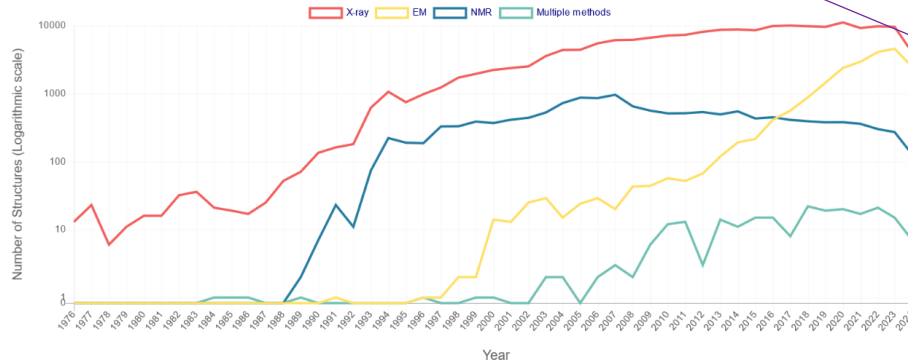
Kathleen Lonsdale resolves the structure of benzene 1924



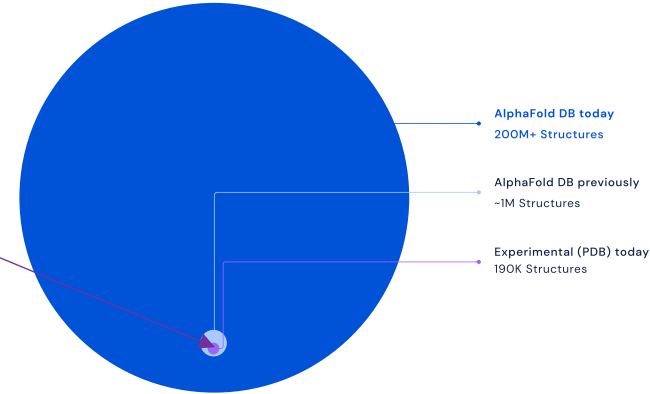
X-rays ruled 20thC – CryoEM & Calcs the 21st



Number of Released PDB Structures per Year



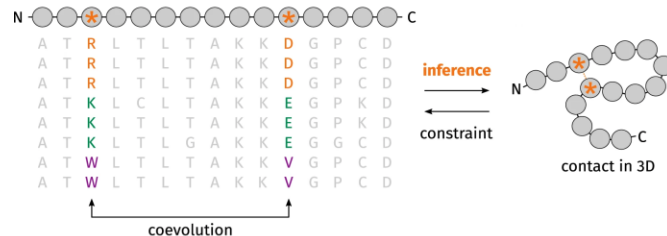
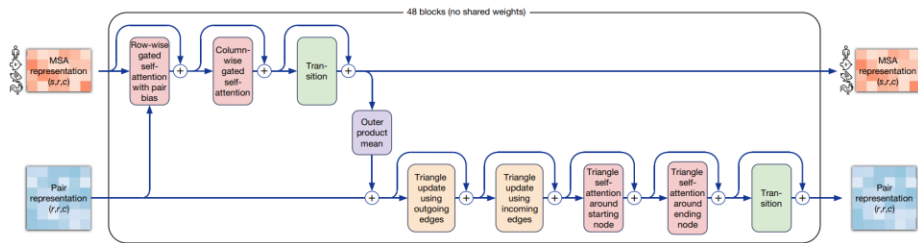
Number of Protein Structures



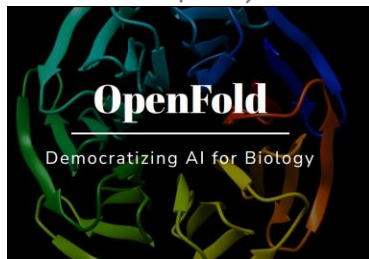
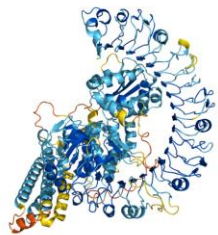
Compute power – Moore's law +
DNA Sequencing – super-Moore's law +
Deep Learning – context scaling laws
=
Comp.Struct.Bio. – **1000x** in 2 years

X'Fold' programs - co-evolution DeepLearning

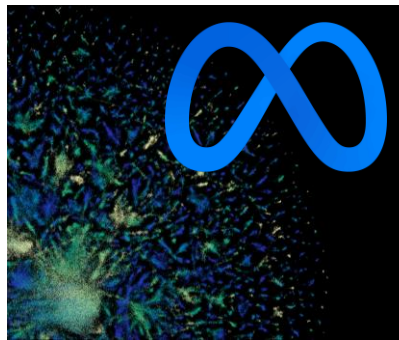
DL: No physics of folding — Transformers — 1D sequence -> 3D structure



'Evo' formers (GPU+CPU+DBs)
(Alpha|Open|RoseTTA|Ab)Fold



Protein-LMs (GPU)
ESMFold, ProtTrans



Struct Alignment (DBs)
Foldseek, Foldmason

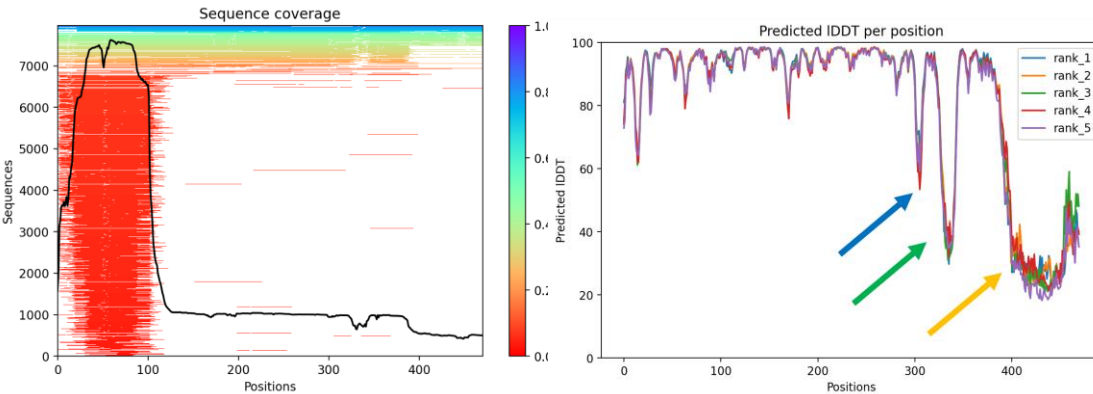


X'Fold' programs - No Physics

Supplementary Videos: Ahdritz, G., Bouatta, N., Floristean, C. *et al.* OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat Methods* **21**, 1514–1524 (2024). <https://doi.org/10.1038/s41592-024-02272-z>

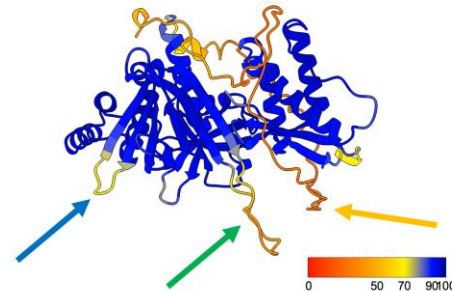
0000

PDB ID: 7RDT



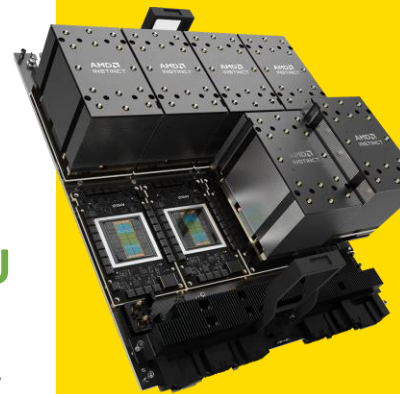
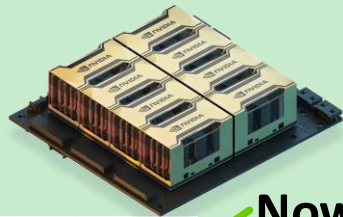
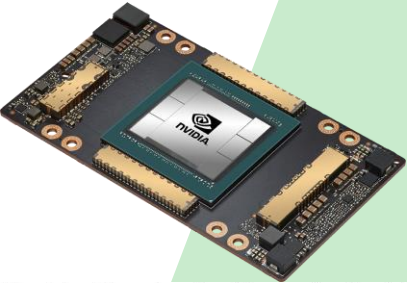
0000

PDB ID: 7B3A

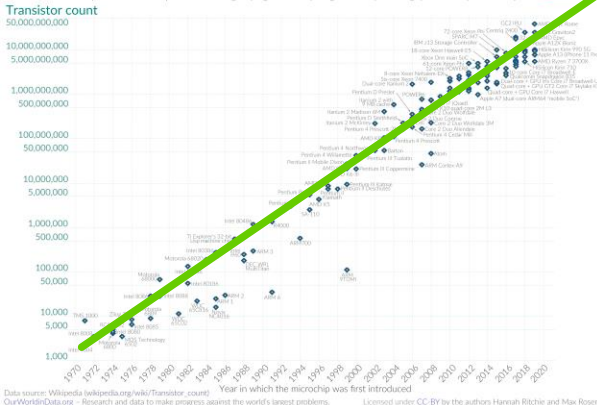


Confidence metrics: Dr Michael Healy, ['AlphaFold2 How-To Guide'](#)

GPUs – Multiply things *really* fast

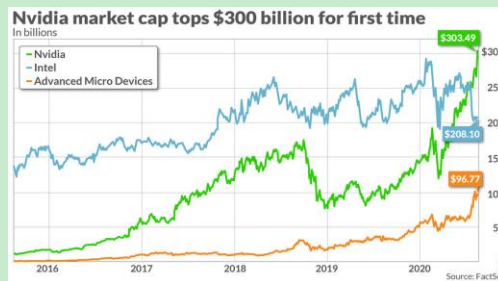


Moore's Law: The number of transistors on microchips doubles every two years. Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.



Now **100 billion** transistors in each GPU

- 100s of trillions multiplies *each second*
- 93 million parameter network less scary



NVIDIA - AMD

- Both design GPUs
- ½ of NVIDIA are software engineers

We can't keep GPUs busy! They are *so fast*

- GPU calculations x50 faster than CPU. 2 hrs vs 5 days
- **ESMFold** is pure GPU. **AlphaFold** GPU+CPU+file retrieval.
 - 613 proteome 8 hrs vs 22 days. 4,622 proteome 10 days vs ?? (2 years)



UNSW
SYDNEY

Where to turn? - *Galaxy*, *Uni* compute cluster, *proteinfold*

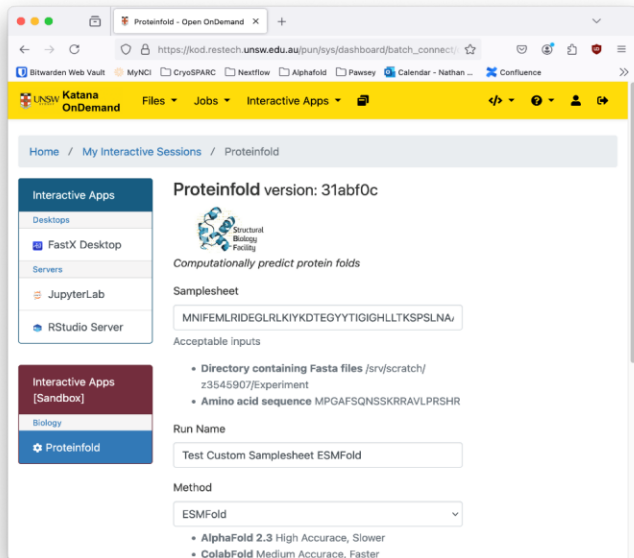


ABOUT ACTIVITIES **SERVICES** TRAINING & EVENTS DOMAINS NEWS CONTACT HELP

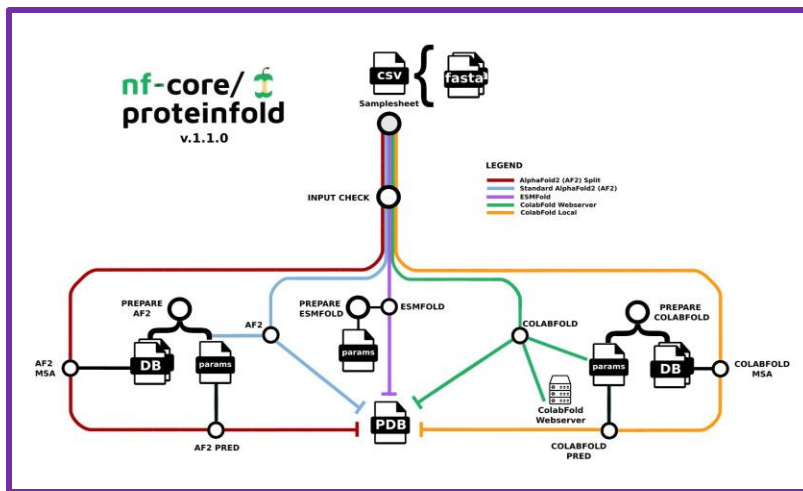


Australian AlphaFold Service

AlphaFold is an artificial intelligence (AI) system developed by [DeepMind](#) that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy that is competitive with experimental methods (see [Jumper et al. Nature 2021](#)).



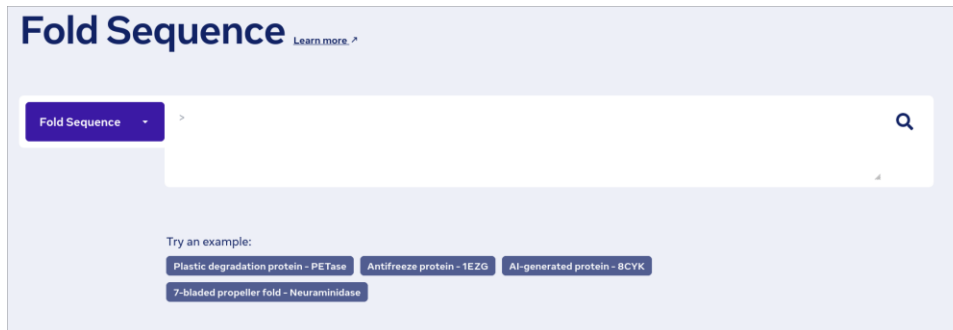
The screenshot shows the Proteinfold web interface. The top navigation bar includes 'Files', 'Jobs', and 'Interactive Apps'. The left sidebar lists 'Interactive Apps' with options like 'Desktop', 'FastX Desktop', 'Servers', 'JupyterLab', 'RStudio Server', 'Interactive Apps [Sandbox]', 'Biology', and 'Proteinfold'. The main content area displays 'Proteinfold version: 31abf0c' and 'Computationally predict protein folds'. A 'Samplesheet' section shows a sample input: 'MNIFEMLRIDGLRLKIYKDTEGYTTIGHLLTKSPSLNA'. Below this, 'Acceptable inputs' are listed: 'Directory containing Fasta files /srv/scratch/z3545907/Experiment' and 'Amino acid sequence MPGAFSQNSKRRVLRPSHR'. The 'Run Name' field is 'Test Custom Samplesheet ESMFold'. The 'Method' dropdown is set to 'ESMFold'. At the bottom, two options are listed: 'AlphaFold 2.3 High Accuracy, Slower' and 'ColabFold Medium Accuracy, Faster'.



How to run? - Install ESMFold, point at FASTA file

A) Use website (1-10 prots)

esmatlas.com/resources?action=fold



B) Use website on the command line (10-100 prots)

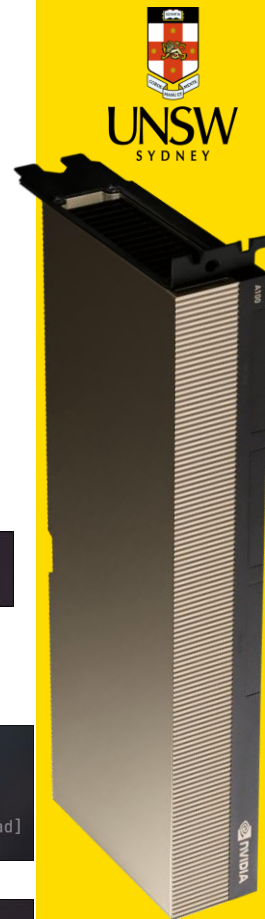
```
keiran@oleic in ~ via v3.11.9
❖ curl -X POST --data "KVFGRCELAAMKRHGLDNYRGYSLGNWVCAAKFESNFTQAT" https://api.esmatlas.com/foldSequence/v1/pdb/
```

C) Install on a GPU server (100-1,000s prots) [~10 minutes]

```
3374843@katana1:~$ pip install "fair-esm[esmfold]"
Defaulting to user installation because normal site-packages is not writeable
Collecting fair-esm[esmfold]
  Downloading fair_esm-2.0.0-py3-none-any.whl.metadata (37 kB)
3374843@katana1:~$ pip install 'dllogger @ git+https://github.com/NVIDIA/dllogger.git'
Defaulting to user installation because normal site-packages is not writeable
Collecting dllogger@ git+https://github.com/NVIDIA/dllogger.git
```

```
3374843@katana1:~$ pip install 'openfold @ git+https://github.com/aqlaboratory/openfold.git@4b41059694619831a7db195b7e0988fc4ff3a307'
Defaulting to user installation because normal site-packages is not writeable
Collecting openfold@ git+https://github.com/aqlaboratory/openfold.git@4b41059694619831a7db195b7e0988fc4ff3a307
```

```
(esmfold) keiran@zaphod:~$ esm-fold --help
usage: esm-fold [-h] -i FASTA -o PDB [-m MODEL_DIR]
               [--num-recycles NUM_RECYCLES]
               [--max-tokens-per-batch MAX_TOKENS_PER_BATCH]
               [--chunk-size CHUNK_SIZE] [--cpu-only] [--cpu-offload]
```



Whole proteome?

All seqs ESMFold (<900), AlphaFold2 (900-3,000), AlphaFold3 (3,000+)

ESMFold – 4378 prots, 10.5 hrs

File: esmfold_loki-ASV2_all_ORFs.log									
1	24/09/06 10:07:00	INFO	root	Reading sequences from loki-ASV2_dfast_all_protein_ORFs.faa					
2	24/09/06 10:07:00	INFO	root	Loaded 4544 sequences from loki-ASV2_dfast_all_protein_ORFs.faa					
3	24/09/06 10:07:00	INFO	root	Loading model					
4	24/09/06 10:07:50	INFO	root	Starting Predictions					
5	24/09/06 10:07:54	INFO	root	Predicted structure for LOKIASV2_25920 hypothetical protein with length 29, pLDDT 64.7, pTM 0.202 in 0.1s (amortized, batch size 29). 1 / 4544 completed.					
6	24/09/06 10:07:54	INFO	root	Predicted structure for LOKIASV2_31160 hypothetical protein with length 29, pLDDT 61.6, pTM 0.110 in 0.1s (amortized, batch size 29). 2 / 4544 completed.					
7	24/09/06 10:07:54	INFO	root	Predicted structure for LOKIASV2_33320 hypothetical protein with length 29, pLDDT 80.2, pTM 0.453 in 0.1s (amortized, batch size 29). 3 / 4544 completed.					
8	24/09/06 10:07:54	INFO	root	Predicted structure for LOKIASV2_15650 hypothetical protein with length 30, pLDDT 74.9, pTM 0.352 in 0.1s (amortized, batch size 29). 4 / 4544 completed.					
9	24/09/06 10:07:54	INFO	root	Predicted structure for LOKIASV2_32700 hypothetical protein with length 31, pLDDT 78.0, pTM 0.467 in 0.1s (amortized, batch size 29). 5 / 4544 completed.					
4379	24/09/06 20:36:11	INFO	root	Failed (CUDA out of memory) on sequence LOKIASV2_24920 penicillin acylase family protein of length 884.					
4380	24/09/06 20:36:12	INFO	root	Failed (CUDA out of memory) on sequence LOKIASV2_35940 hypothetical protein of length 884.					
4381	24/09/06 20:36:12	INFO	root	Failed (CUDA out of memory) on sequence LOKIASV2_20420 hypothetical protein of length 885.					
4382	24/09/06 20:36:13	INFO	root	Failed (CUDA out of memory) on sequence LOKIASV2_05820 DNA methyltransferase of length 887.					
4383	24/09/06 20:36:14	INFO	root	Failed (CUDA out of memory) on sequence LOKIASV2_29000 glycoside hydrolase family 31 protein of length 888.					
4384	24/09/06 20:36:14	INFO	root	Failed (CUDA out of memory) on sequence LOKIASV2_43680 hypothetical protein of length 888.					
4385	24/09/06 20:36:14	INFO	root	Failed (CUDA out of memory) on sequence LOKIASV2_45440 hypothetical protein of length 888.					

AlphaFold2.3 – 160 prots, 10.5 hrs

File: run_AF2_TEMPLATE.sh									
1	python3 docker/run_docker.py \								
2	--fasta_paths=./fasta_dir/file.fasta \								
3	--max_template_date=2022-01-01 \								
4	--model_preset=monomer \								
5	--data_dir=/mnt/af2/ \								
6	#--use_precomputed_msas=true \								
7	--output_dir=/mnt/data/alphafold_output/ 2>&1 tee -a alphafold2.log								

AlphaFold3 – 6 prots, pretty quick

Upload JSON

Clear

Entity type

Protein

Copies

1

Paste sequence or fasta

Input

+

Add entity

Save job

What to do? - Pull up **Homologues**, compare w/ **MS**

search.foldseek.com



Foldseek Search

Results for job: sA2s8t4t55aSp4LzwQ2znuidGZRTAU75NEVFgQ

ALL DATABASES AFD8-PROTEOME (568) AFD8-SWISSPROT (213) AFD850 (1000) BFMD (18) CATH50 (116) GMGCL_ID (25) MGNIFY_ESM30 (1000) PDB100 (78)

AFD8-PROTEOME 568 hits

Target	Description	Scientific Name	Prob.	Seq. Id.	E-Value	Position in query	Alignment
AF-A4HYD-F1-model_v4	Phosphoprotein_phosphatase_ putative	Leishmania infantum	1.00	15.8	3.76e-18	162 515	

TH-Score: 0.69984 RMSD: 17.82

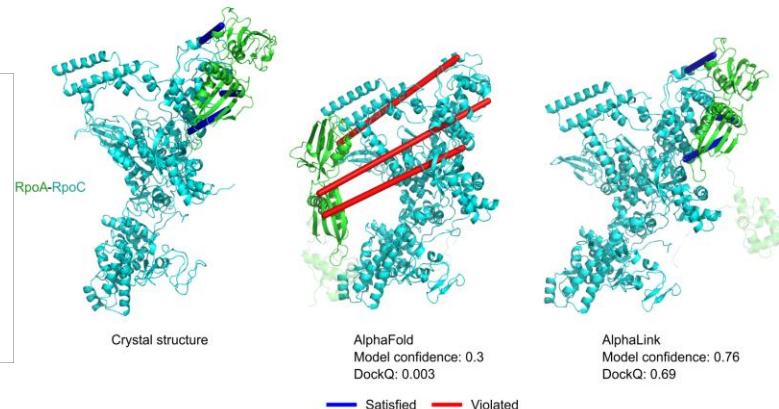
Crystal structure

AlphaFold Model confidence: 0.3 DockQ: 0.003

AlphaLink Model confidence: 0.76 DockQ: 0.69



github.com/Rappsilber-Laboratory/AlphaLink2



ebi.ac.uk/training/online/courses/alphafold

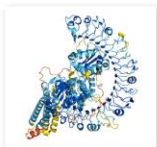
Australian Computational Structural Biology Community

australian-structural-biology-computing.github.io/website/

ONLINE TUTORIAL

AlphaFold

A practical guide



Enter course

♥ Mark as favourite

Time to complete:
3 hours**Contributors**

The projects included here represent a joint effort by the following people at multiple Australian institutions.



James Lingford

Research Officer, Greening
Lab, Monash University

Johan Gustafsson

Engagement, Australian
BioCommons, University of
Melbourne

Kate Michie

Chief Scientist, Structural
Biology Facility (SHF)

Keiran Rowell

Computational Scientist,
Structural Biology Facility
Unit

Michael Healy

Principal Research Fellow,
University of
Queensland

Australian Structural Biology Computing

Home GitHub Search StructBio Computing

Guide

AlphaFold2 How-to Guide

AlphaFold2 (AF2) has undoubtedly revolutionised the world of structural biology, allowing for the rapid prediction of proteins and protein complexes. In this guide I aim to equip you to understand how the algorithm works and how to interpret the output data.

To do this there are 3 main sections;

1. AlphaFold2 - the basics
2. Running a Structure Prediction, and
3. Interpreting the Output Files

This guide is designed as a supplement to EMBL's excellent course on AF2.

AlphaFold2 - the basics

To start to get to grips with how AF2 goes about generating a protein model, it is helpful to understand Anfinsen's dogma and Levinthal's paradox.

Anfinsen's dogma: "at least for a small globular protein in its standard physiological environment, the native structure is determined only by the protein's amino acid sequence". This means given a random amino acid sequence it should be possible to predict the secondary structure it adopts (alpha helix, beta strand, loop etc).



On this page

[AlphaFold2 - the basics](#)

Running a structure prediction
Interpreting the output files

Australian Structural Biology Computing

Home GitHub Search StructBio Computing

Guide

Best practices for presenting and sharing AlphaFold models in a paper

AlphaFold structural models are appearing in papers more frequently. As such, it's important that the scientific community agree on:

- General guidelines on how to best present AlphaFold models: AlphaFold models presented without sufficient information might mislead readers into thinking the model confidence is higher than it really is.
- Guidelines for sharing AlphaFold output files: AlphaFold models are computationally expensive to generate, and some readers might not have the resources to re-generate the model.

For these reasons, we have distilled our knowledge, and that of the broader field, into a guide for presenting and sharing AlphaFold models in papers. The guide is not intended to be exhaustive or dogmatically rigid, and we expect it will evolve over time. Our goals with this guide are to:

1. Help newcomers get up to speed with how-to present AlphaFold models, and
2. Encourage wider discussion among the structural biology community.

Please reach out to us if you have suggestions.

What AlphaFold data should I include?

On this page

What AlphaFold data should I include?

Best practices for presenting models

Other good practices

What should I include in the methods section?

What AlphaFold data should I share?

The world is doing this

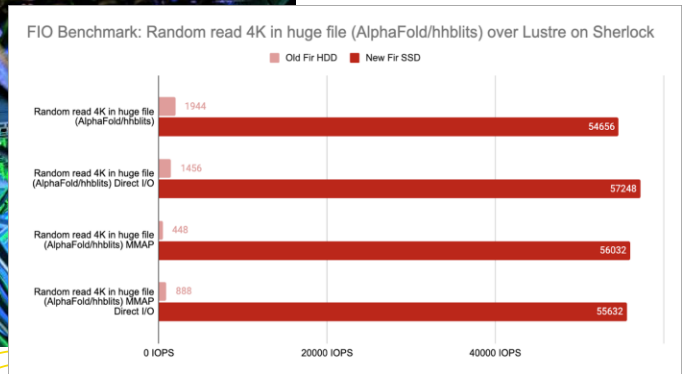
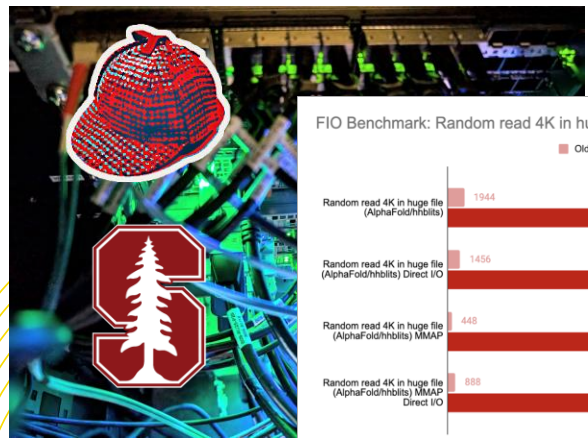


[Services](#)
[Research](#)
[Capabilities](#)
[Data-Driven Life Science](#)
[Data](#)
[Calendar](#)
[Training](#)
[News](#)
[About us](#)
[Contact](#)

SciLifeLab / News / AlphaFold 2 and SciLifeLab: advancing structural biology beyond protein folding

AUGUST 26, 2021

AlphaFold 2 and SciLifeLab: advancing structural biology beyond protein folding



NSC

li.u

LINKÖPING UNIVERSITY

NSC / Support / Systems / Berzelius software / Berzelius AlphaFold3 Guide

Using AlphaFold 3 on Berzelius

Questions / Discussion