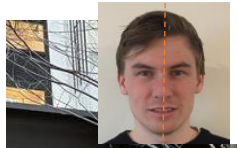
Working with ProteinFold

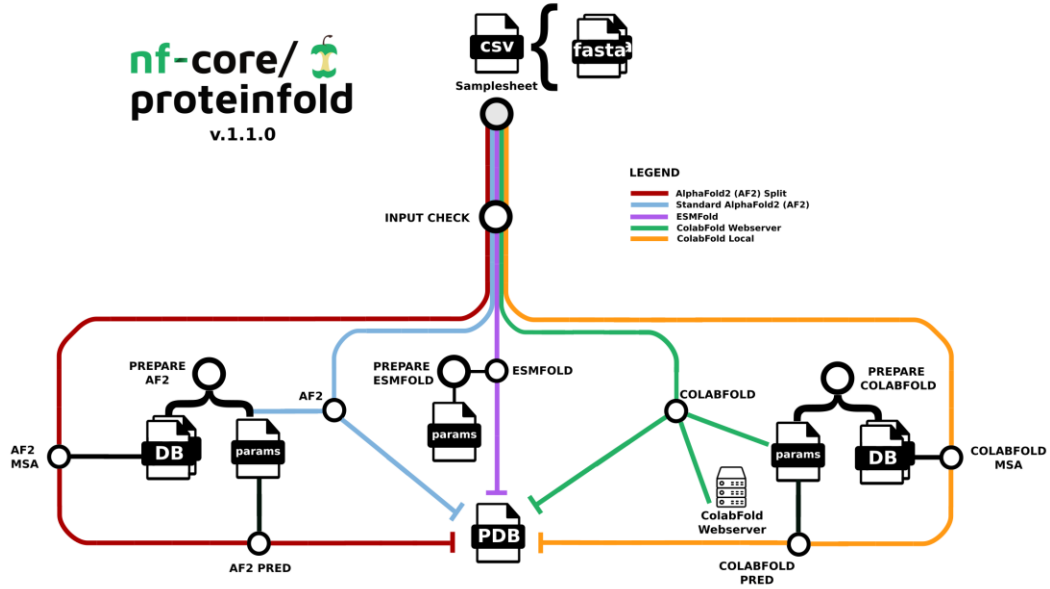
High performance protein structure prediction from a spreadsheet

Melbourne Bioinformatics Meetup - 2025 June 26

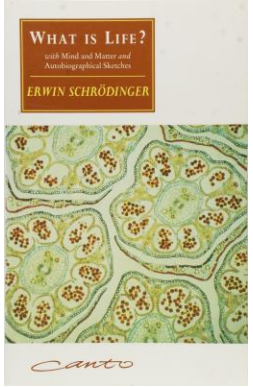
Dr Keiran Rowell



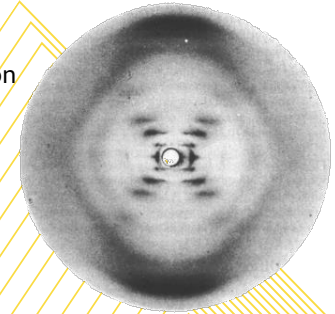





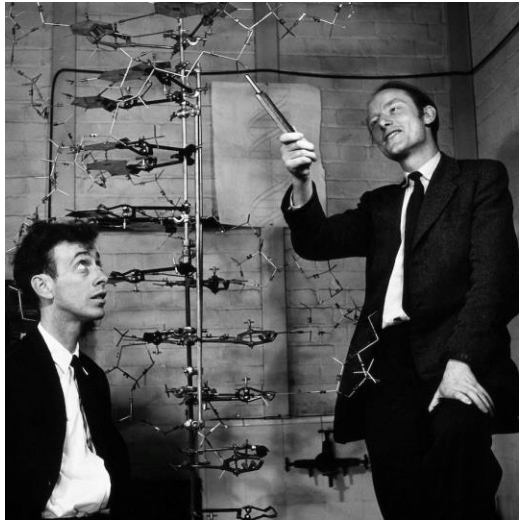
Models in Chemical Biology - function follows chemical form



Schrödinger speculates on the molecules of life **1944**

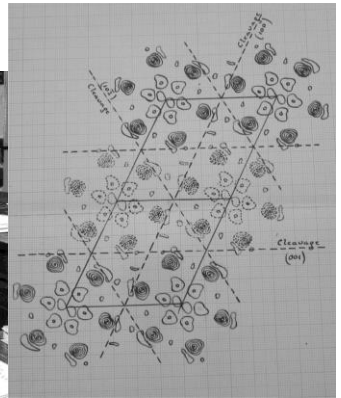


Wilkins & Fraklin's X-ray diffraction images of DNA **1953**



Watson & Crick's DNA double-helix model **1953**

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

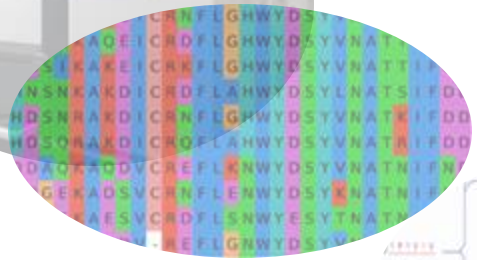


Kathleen Lonsdale resolves the structure of (hexamemthyl)benzene **1929**

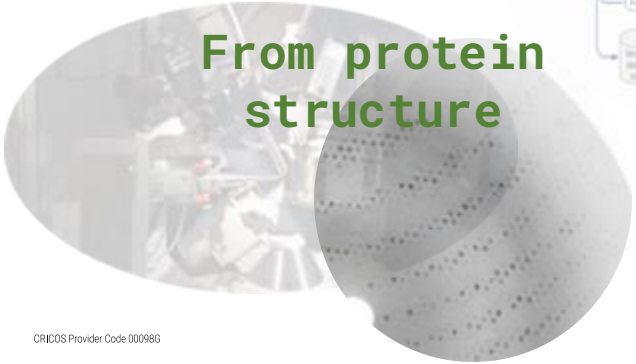
UNSW SYDNEY

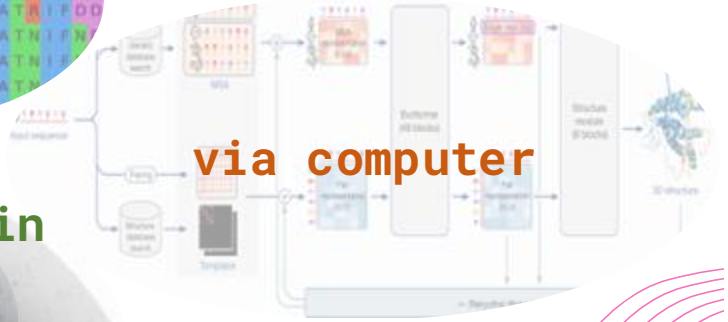
Deep Learning trained on experimental data

From genetic sequence



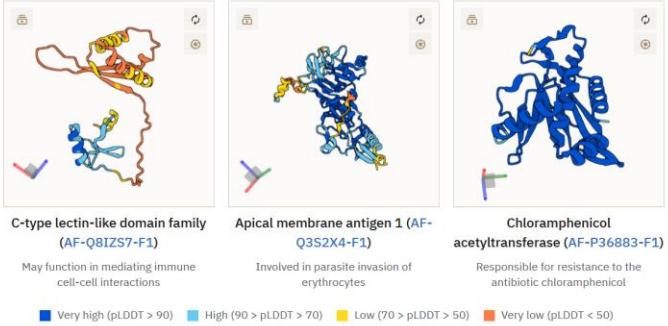
From protein structure





via computer

To modelled structure prediction

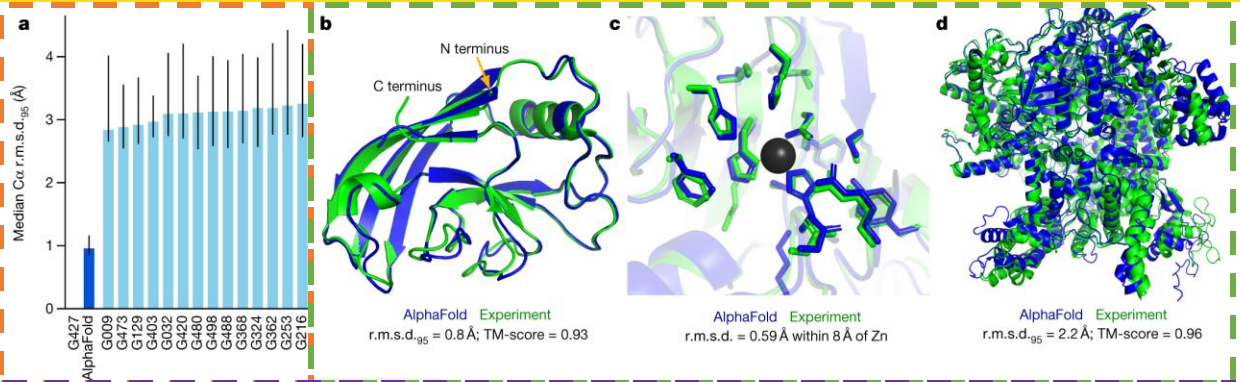


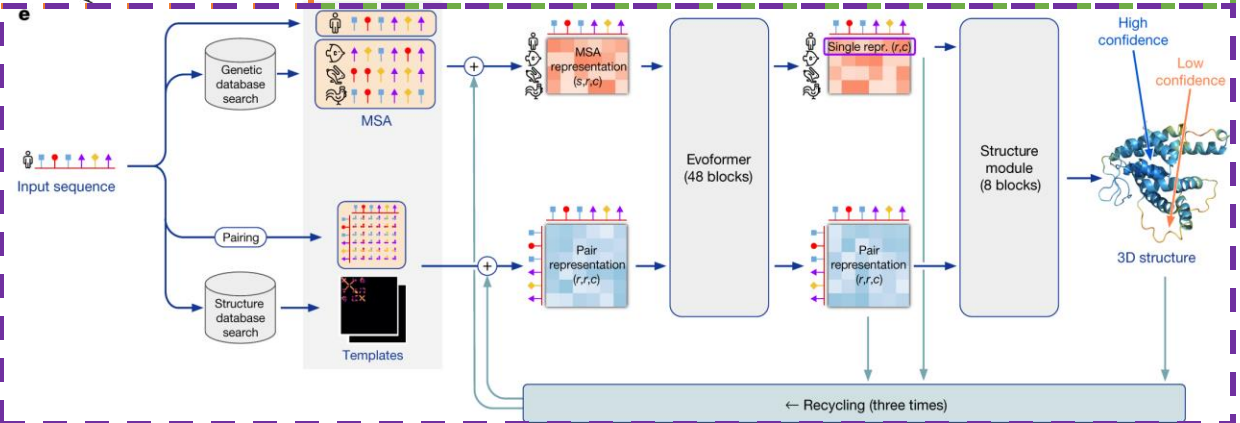
Source: https://www.ebi.ac.uk/training/online/courses/alphafold/inputs-and-outputs/evaluating-alphafolds-predicted-structures-using-confidence-scores/plddt-understanding-local-confidence/

AlphaFold2

A leap forward in computational prediction accuracy

Fig. 1: AlphaFold produces highly accurate structures – from *"Highly accurate protein structure prediction with AlphaFold"* - Nature, **596**, 583-89 (2021).





Improvements

- More than 2x as accurate as anyone else
- The global fold reliable when compared to very expensive experiments
- Set the algorithm architecture that is that standard today

Tutorial only on **protein folding** - **X**'Fold' programs

The largest Big Tech and academic groups in world have dedicated serious expertise and resources into developing these programs.

They are mostly free and open source (academic, non-commercial work).
If you develop the skills to run the code on local hardware.

Google DeepMind

https://github.com/google-deepmind/**alphafold**

https://github.com/google-deepmind/**alphafold3**

Facebook AI Research

https://github.com/facebookresearch/**esm**

ByteDance

https://github.com/bytedance/Protenix

https://github.com/aqlaboratory/**openfold**

https://github.com/sokrypton/**ColabFold**

MIT JAMEEL CLINIC

https://github.com/jwohlwend/**boltz**

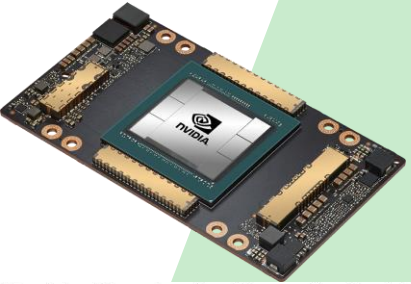
INSTITUTE FOR Protein Design
UNIVERSITY of WASHINGTON

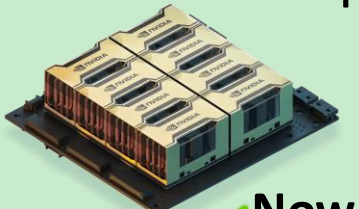
https://github.com/baker-laboratory/**RoseTTAFold-All-Atom**

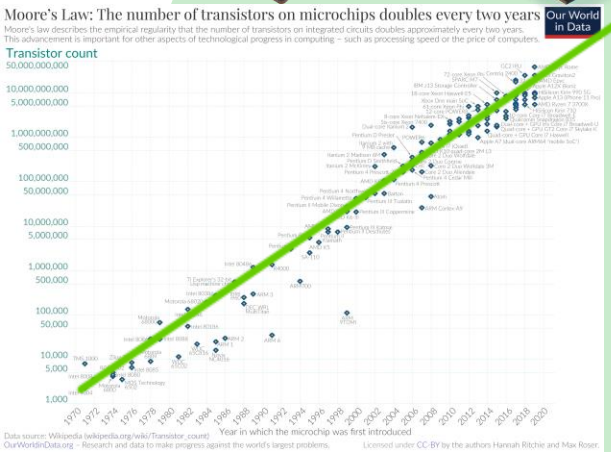
UNSW SYDNEY

Need a unified workflow method!

GPUs – Multiply things *really* fast

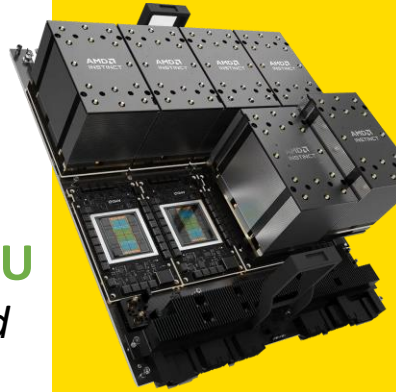






Now ~100 billion transistors in *each* GPU

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



NVIDIA - AMD

- Both design GPUs
- ½ of NVIDIA are software engineers

Sequence alignment is the rate limiting step!

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+DB retrieval.
  - 613 proteome 8 hrs vs 22 days. 4,622 protoeome 10 days vs ?? (2 years)

UNSW SYDNEY

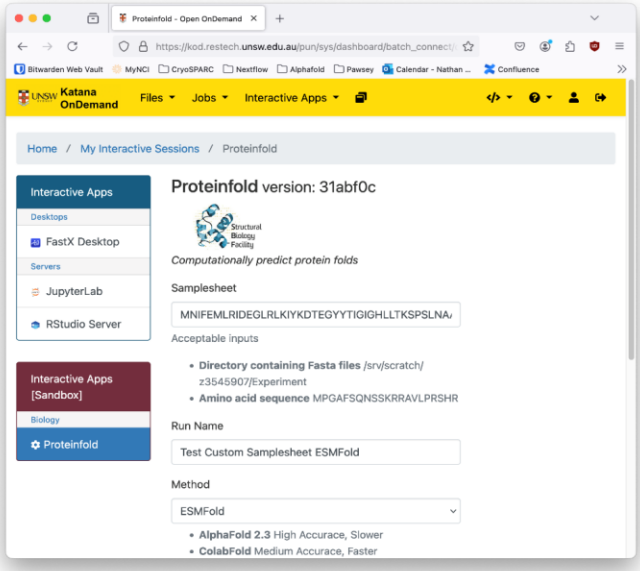
Where to turn? - Galaxy, Uni compute cluster, ProteinFold

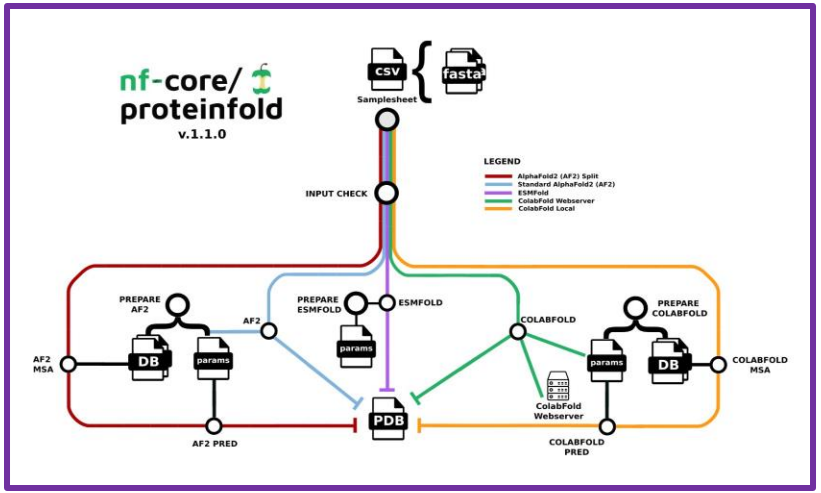
Australian BioCommons

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).

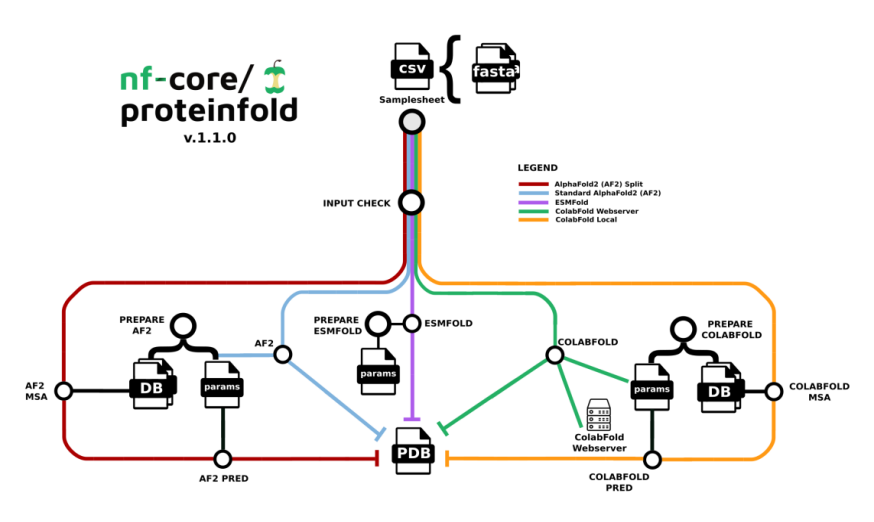




ABLeS: Pawsey AMD (1017) – NCI NVIDIA (za08)

UNSW SYDNEY

Live Demo – ProteinFold Terminal *vs* Web



Running the pipeline

The typical commands for running the pipeline on AlphaFold2, Colabfold and ESMFold modes are shown below.

AlphaFold2 regular can be run using this command:

```
nextflow run nf-core/proteinfold \
      --input samplesheet.csv \
      --outdir <OUTDIR> \
      --mode alphafold2 \
      --alphafold2_db <null (default) | DB_PATH> \
      --full_dbs <true/false> \
      --alphafold2_model_preset monomer \
      --use_gpu <true/false> \
      -profile <docker/singularity/.../institute>
```

To run the AlphaFold2 that splits the MSA calculation from the model inference, you can use the `--alphafold2_mode split_msa_prediction` parameter, as shown below:

Proteinfold

Computationally predict protein structures

Samplesheet

/srv/scratch/USER/fasta_files

Acceptable inputs

- **Directory containing Fasta file(s):** `/srv/scratch/z3141592/my_experiment`
- **Amino acid sequence:** `NLYIQWLKDGGPSSGRPPPS`

Warning! Please ensure your input data (e.g., FASTA file or run name) does not contain sensitive data. Katana is **NOT** suitable for sensitive or highly sensitive data. You should use the UNSW Data Classification scheme to classify your data and learn about managing your research data by visiting the Research Data Management Hub.

Run Name

test_run

Alphanumeric and "_" only

Method

Alphafold2

- **AlphaFold2.3** High Accuracy, Slower - Paper
- **ESMFold** Medium/Low Accuracy, Fastest (No Evolutionary Sequence Calculations) - Paper
- **RoseTTAFold-All-Atom** High Accuracy, Slower; optimised for atomic-level modeling - Paper

Mode

Monomer

- Only applies to AlphaFold2.3 and ESMFold
- Monomer_ptm for AlphaFold2.3 only

MSA Search Database

Full

- **Full** High Accuracy, Slower
- **Reduced** Optimised for speed

Facility Citation

doi.org/10.26190/4KQF-M552

Please cite the above DOI and include this following acknowledgement in any publication that uses this resource: "*The authors acknowledge use of facilities in the Structural Biology Facility within the Mark Wainwright Analytical Centre – UNSW, funded in part by the Australian Research Council Linkage Infrastructure, Equipment and Facilities Grant: ARC LIEF 190100165*"

Launch

* The Proteinfold session data for this session can be accessed under the data root directory.

What's new in ProteinFold v2? - adventures in dev

nf-core

Over the past year:

Programs:

- AlphaFold3 (BYO weights)
- Boltz (v2) – Unrestricted (MIT) AlphaFold3
- RosettaFold-All-Atom (Institute of Protein Design)
- HelixFold3 (Baidu) – AlphaFold3 re-implementation

Features:

- Quality Metrics – (MSA depth, pLDDT, PAE, iPTM)
- CPU vs GPU process labels – efficient computing
- NextFlow – understands on-prem or cloud scheduler
- Shared databases (3 TB) – who wants 5 UniRef30s?
- Containers – some repos already abandoned!
- MSA reuse (might be v2.1) – MSAs are slooow

Reports:

- HTML5 – interactive structure visualisation
- FoldSeek – auto fetch structural annotations
- MultiQC (might be v2.1) – Metrics for bulk folding
- Versioning – pipeline keeps track of container and database versions

UNSW SYDNEY

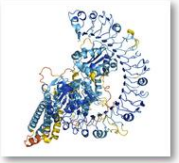
Self-learning – spend 3 hrs with the pros!

https://www.ebi.ac.uk/training/online/courses/alphafold/

ONLINE TUTORIAL

AlphaFold

A practical guide



Enter course

Mark as favourite

Time to complete:
3 hours

This course includes:

- Activities
- Quizzes
- Videos

Written by:
Paulyna Gabriela Magana Gomez
Oleg Kovalevskiy

Last reviewed:
December 2024

Proteins are essential components of life, predicting their 3D structure enables researchers to get an insight into its function and role. AlphaFold is an artificial intelligence (AI) system, developed by Google DeepMind, that predicts a protein's 3D structure based on its primary amino acid sequence. It regularly achieves accuracy competitive with experiment.

Course overview | Course contents | Getting started | Competencies

Feedback and help

Who is this course for?

This tutorial is aimed at researchers who are interested in using AlphaFold2 to predict protein structures and integrate these predictions into their projects. An undergraduate-level knowledge of protein structure and structural biology would be an advantage.

The content of this course provides an understanding of the fundamental concepts behind AlphaFold2, how users can run protein predictions and how AlphaFold2 has been used to enhance research.

Throughout the course there may be terms used you are unfamiliar with. If so, please review the Glossary of terms of help.

This training module on AlphaFold2 has been developed in collaboration with Google DeepMind.

Google DeepMind

Questions / Discussion

Facility: unsw.edu.au/research/facilities-and-infrastructure/find-a-facility/sbf
Pipeline: nf-co.re/proteinfold/
My site: keiran-rowell.github.io/