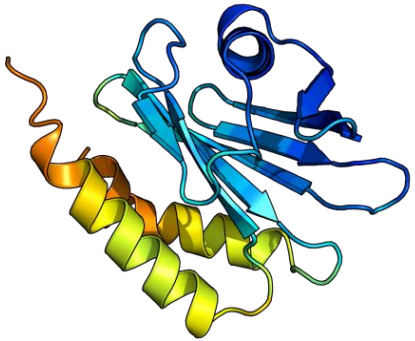
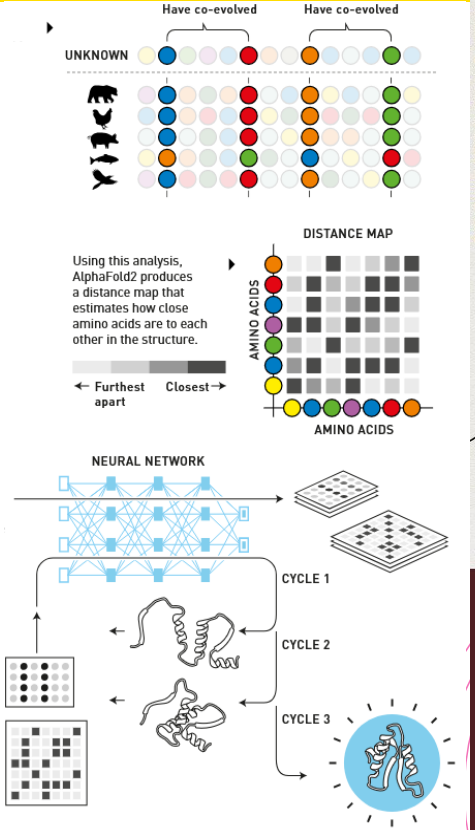
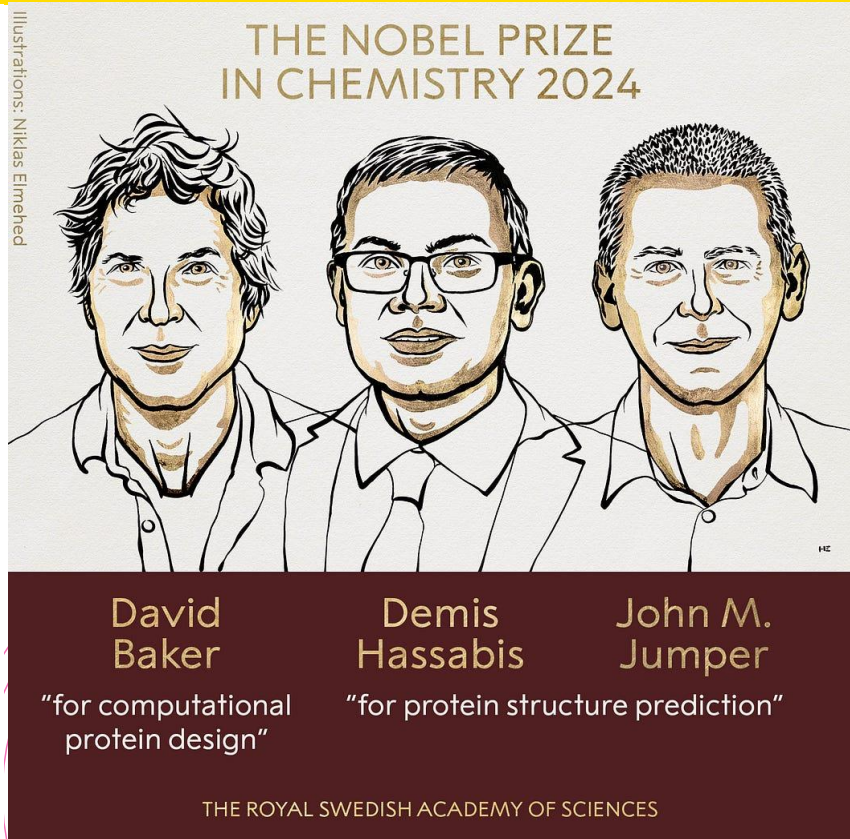
Protein Structure Prediction

Source: https://www.nobelprize.org/prizes/chemistry/2024/press-release/







Structural Biology Facility

Tutorial structure

Q&A (a few mins) between each section (yellow slides)

How did we get here? – 30 mins

- A quick history of protein structure folding – how did we get to **AlphaFold**?
- Looking under the hood – roughly how does AlphaFold work
- An explosion of code choice – awareness of 'right tool' for the job

Biotech applications, and access – 30 mins

- Folding all proteins in an organism
- Resolving previously unsolved crystal structures
- Building previously unseen giant assemblies
- Access methods - licence restrictions on code

Break! – 10 mins

Do it yourself – 1 hr

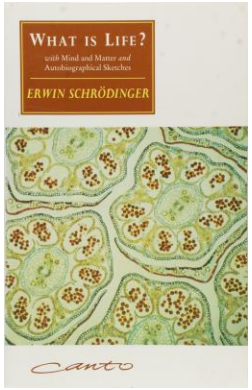
- Running **AlphaFold2** in a Google Colab notebook

Live demo – 30 mins

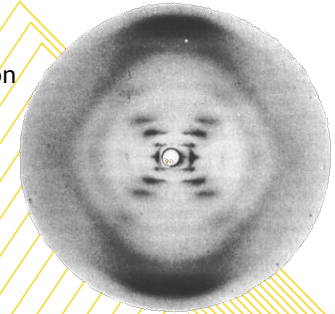
- ProteinFold – **X**'Fold' on a high performance computer terminal
- On-Demand ProteinFold – automated webform through UNSW's Katana OnDemand

A brief history..

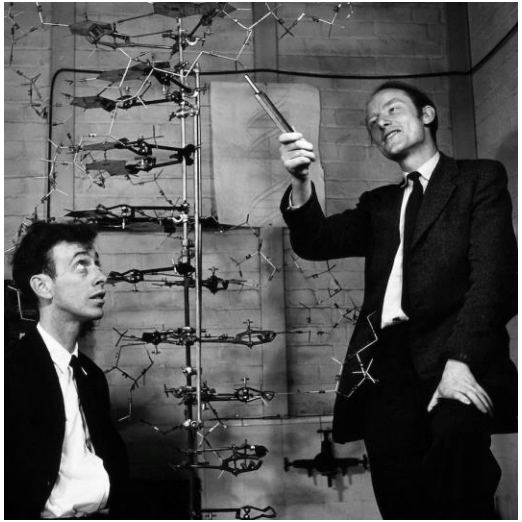
Models in Chemical Biology - function follows chemical form



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



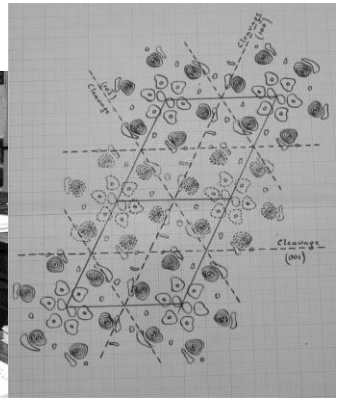
Wilkins & Fraklin's X-ray diffraction images of DNA **1953**



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

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

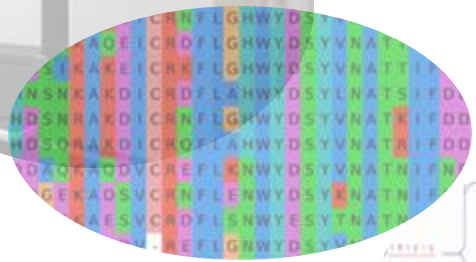
Kathleen Lonsdale resolves the structure of benzene **1924**



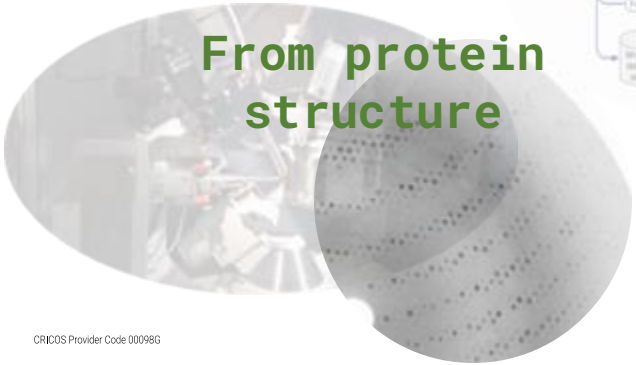
UNSW SYDNEY

Deep Learning trained on experimental data

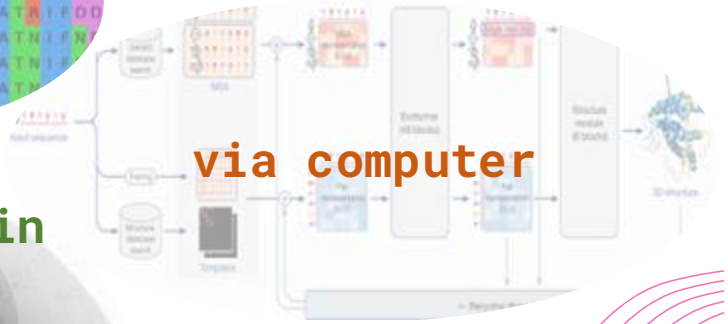
From genetic sequence



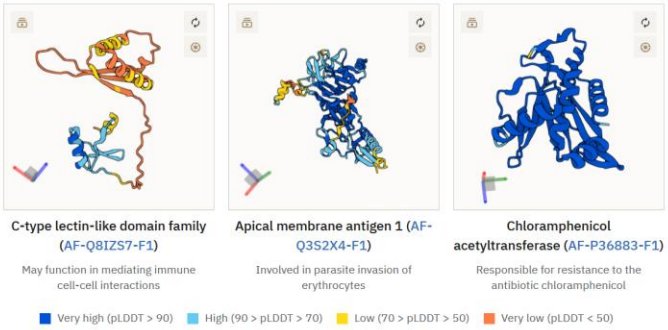
From protein structure



via computer

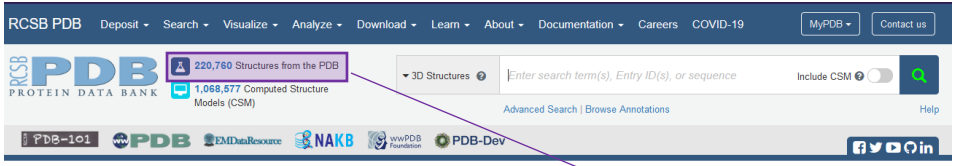


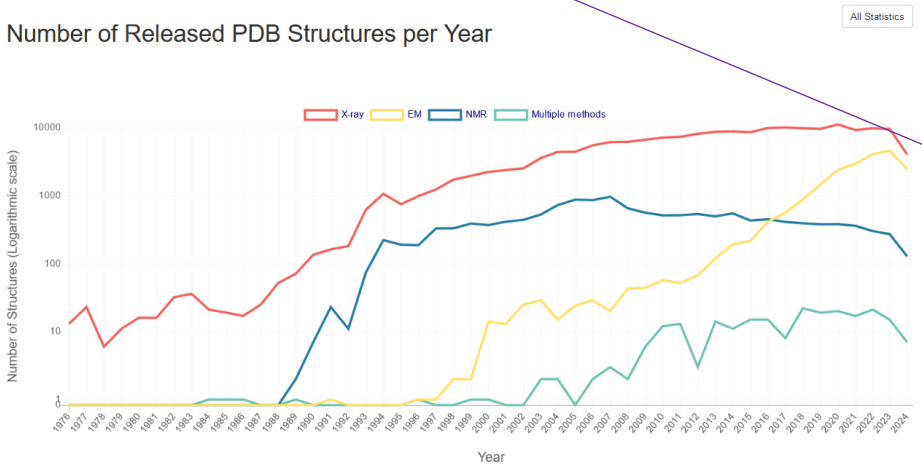
To modelled structure prediction

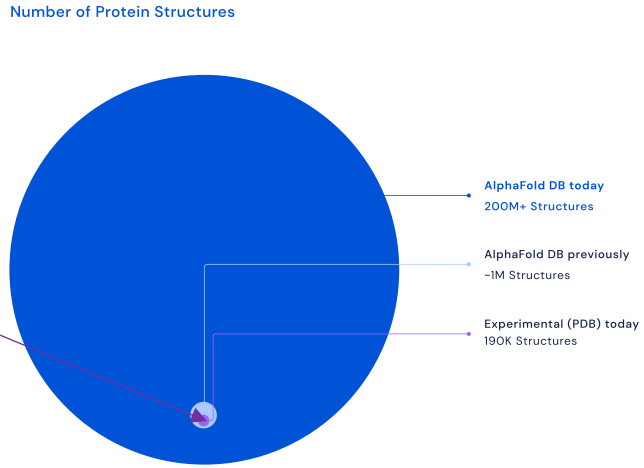


Source: https://www.ebi.ac.uk/training/online/courses/alphafold/inputs-and-outputs/evaluating-alphafolds-predicted-structures-using-confidence-scores/plddt-understanding-local-confidence/

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







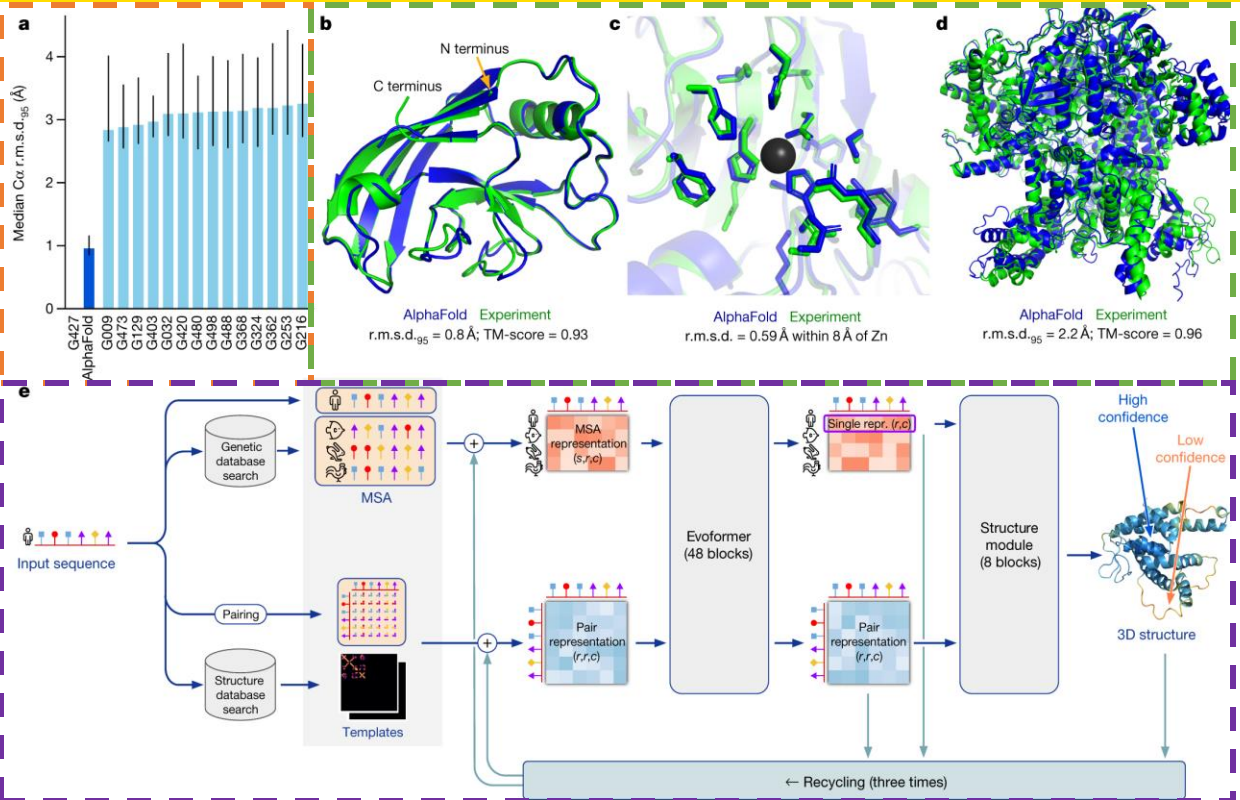
Protein Data Bank (**PDB**)

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

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

Under the hood..

X'Fold' programs – the way they fold is very strange

No (explicit) Physics or Chemistry

Large Language Models (*e.g.* GPT) learns context from pure text tokens

EvoFormers learns 1D -> 2D -> 3D from sequence and solved structures in Protein Data Bank (**PDB**)

PDB ID: 7RDT

PDB ID: 7B3A

UNSW SYDNEY

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

AlphaFold2 architecture

ingenious algorithm design made this breakthrough

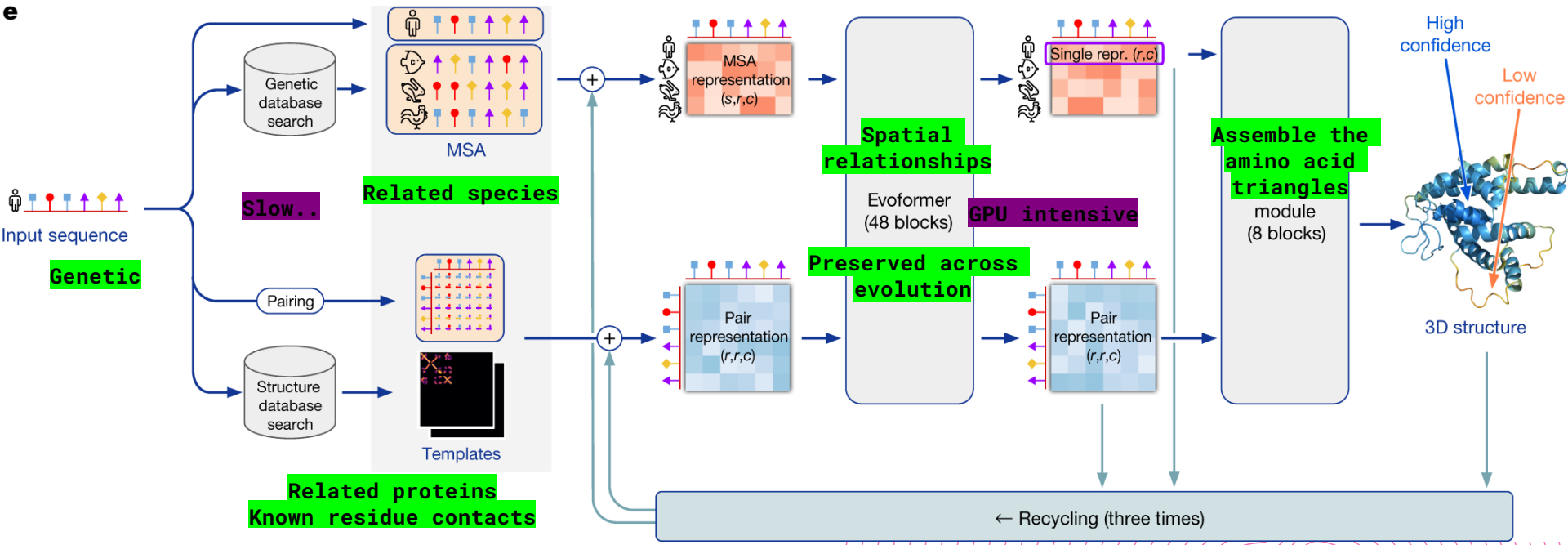
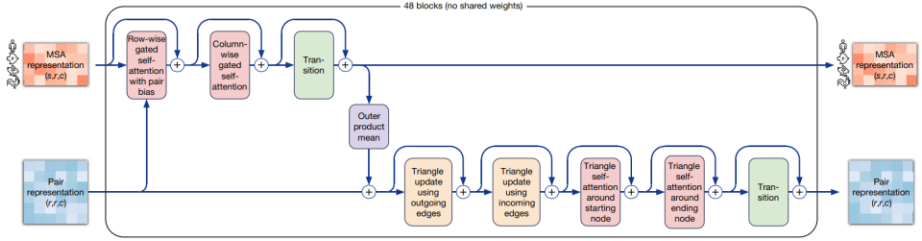


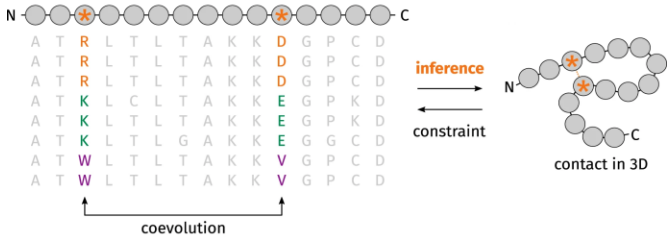
Fig. 1: AlphaFold produces highly accurate structures – from *"Highly accurate protein structure prediction with AlphaFold"* - Nature, **596**, 583-89 (2021).

Code choice..

X'Fold' programs - co-evolution DeepLearning

DeepLearning: No folding physics – Transformers – 1D sequence -> 3D structure

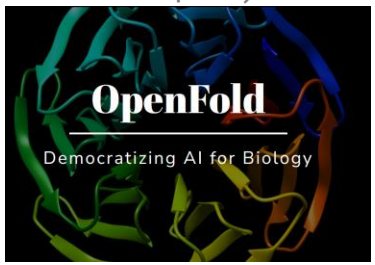




High Performance Computing: Not for domestic computers. Use your uni's HPC cluster.

'Evo'formers

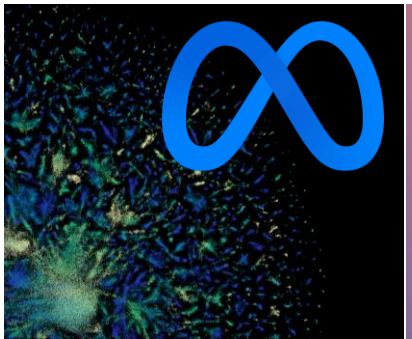
(Alpha|Open|RoseTTA|Ab)Fold



Google DeepMind

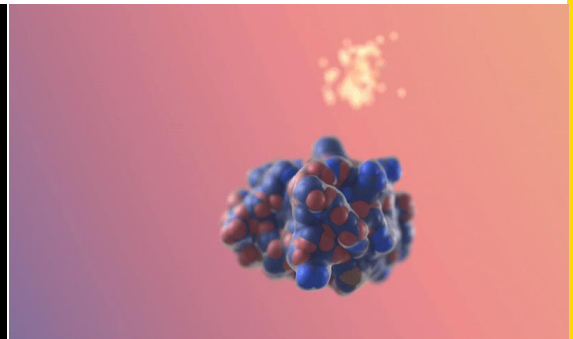
Protein-LLMs

ESMFold, ProtTrans



GenAI - Diffusion

RFDiffusion, Chroma



UNSW SYDNEY

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

Let's do Biology!

Focus on **easiest AlphaFold2** method

X'Fold' programs – each has its place

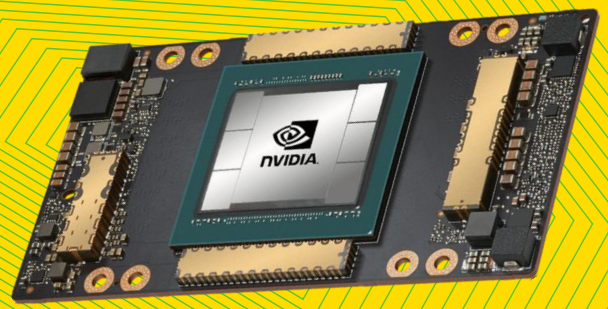
Computing is all about **trade-offs**:

speed ⚔️ accuracy ⚔️ cost

Algorithms are also about **hand-offs**:

💾 Storage 🤝 📦 memory 🤝 Processors 🤝 ⏳ run time

COMPUTERS ARE FAST



$10,000s a pop in the AI arms race

10s/100s/1,000s structures/day

Programming is all about **re-use**:

Code is crystallized labour that can be run x***billions***/second

Researchers adapt & script the last-mile for their usage

The computing stack is also about **access**:

Programmers 🤝 pipetters | customisability ⚔️ ease-of-use

Data processing ⚔️ restricted graphical interfaces

Free & Open Source licences ⚔️ recoup development cost

Let's do Biology!

Focus on **easiest-to-use** **AlphaFold2** method

UNSW SYDNEY

BioTech applications..

Folding Entire Proteomes

Predict a molecular inventory of your organism

(my predictions from bioinformatician's closed genome)

A mix of AlphaFold2, ESMFold and AlphaFold3

UNSW SYDNEY

UNSW Ramaciotti Centre for Genomics

Sequence genome

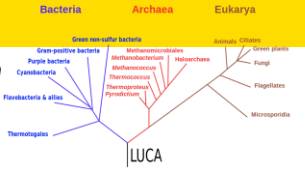
4544 proteins

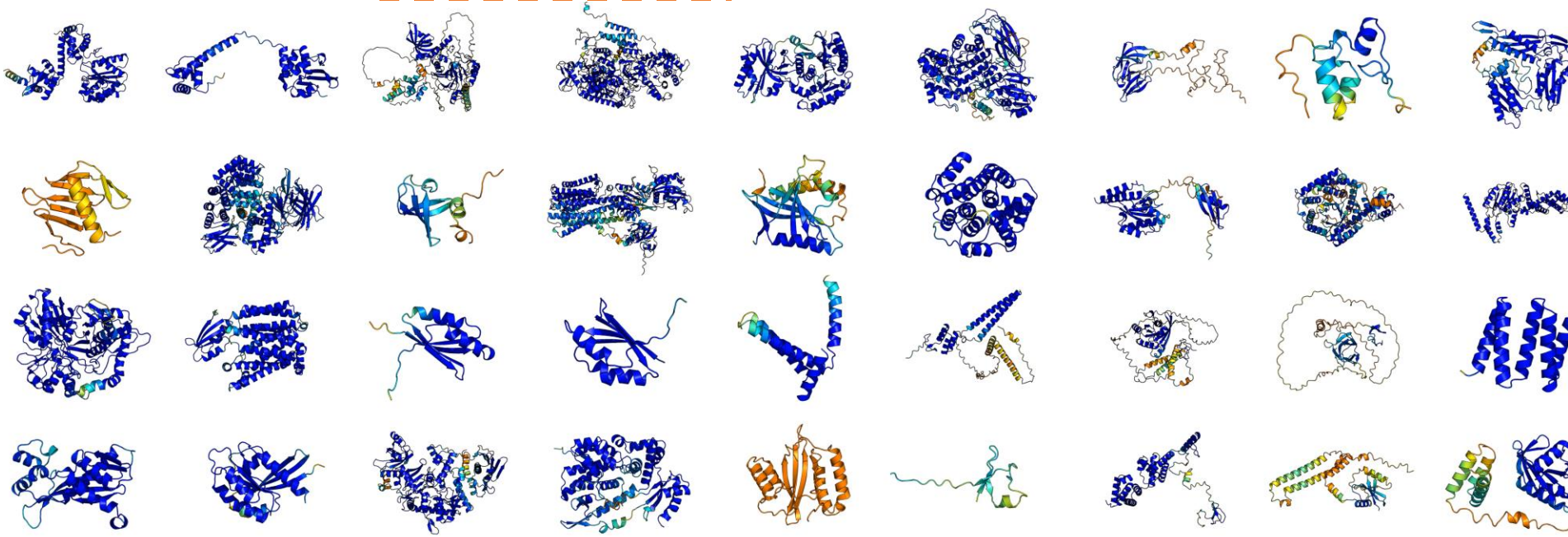
Evo-AI methods

4544 structures

Struct search

AlphaFoldDB





Clarifying Experiments

Using computer models on unsolved structure problems

STRUCTURAL BIOLOGY
ISSN: 2059-7983

Volume 78 | **Part 1** | January 2022 | Pages 1-13
https://doi.org/10.1107/S2059798321012122
OPEN ACCESS
Cited by 121
Part of a special issue
Part of a special issue

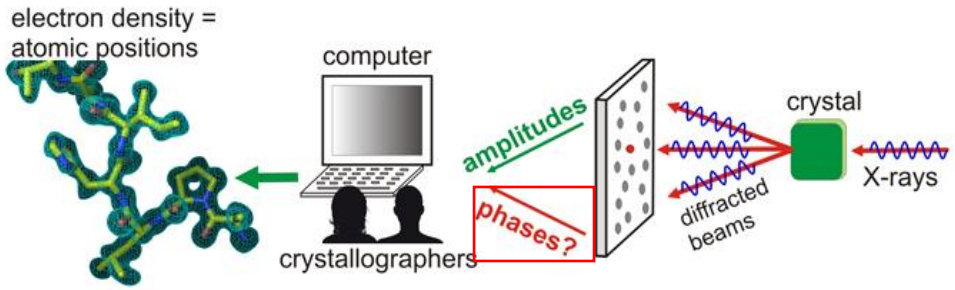
Implications of *AlphaFold2* for crystallographic phasing by molecular replacement

Airlie J. McCoy,[a]* Massimo D. Sammito[a] and Randy J. Read[a]

[a]Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge CB2 0XY, United Kingdom
*Correspondence e-mail: ajm201@cam.ac.uk

Edited by D. J. Rigden, University of Liverpool, United Kingdom (Received 18 May 2021; accepted 13 November 2021)

The *AlphaFold2* results in the 14th edition of Critical Assessment of Structure Prediction (CASP14) showed that accurate (low root-mean-square deviation) *in silico* models of protein structure domains are on the horizon, whether or not the protein is related to known structures through high-coverage sequence similarity. As highly accurate models become available, generated by harnessing the power of correlated mutations and deep learning, one of the aspects of structural biology to be impacted will be methods of phasing in crystallography. Here, the data from CASP14 are used to explore the prospects for changes in phasing methods, and in particular to explore the prospects for molecular-replacement phasing using *in silico* models.



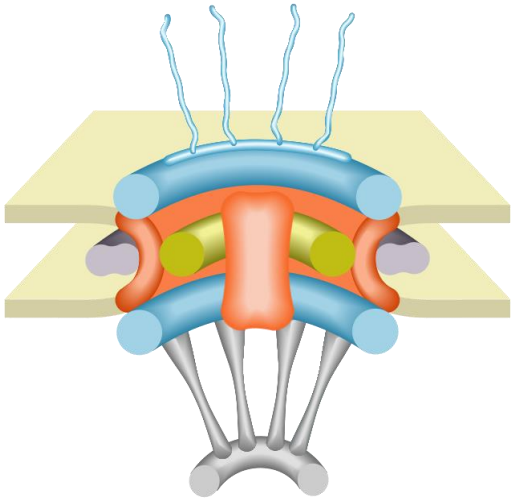
"Impossible" X-ray microscope

Source: LibreTexts Chemistry
1.7: Structural Resolution
(accessed March 2025)

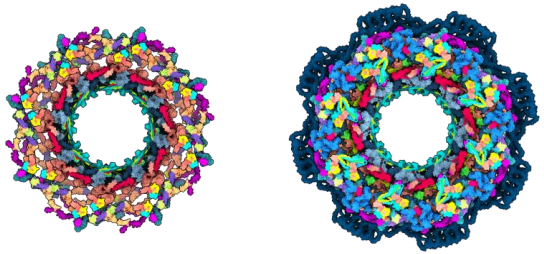
Assembly Building from AlphaFold + Electron Microscopy

Building the **"monstorous maw"**

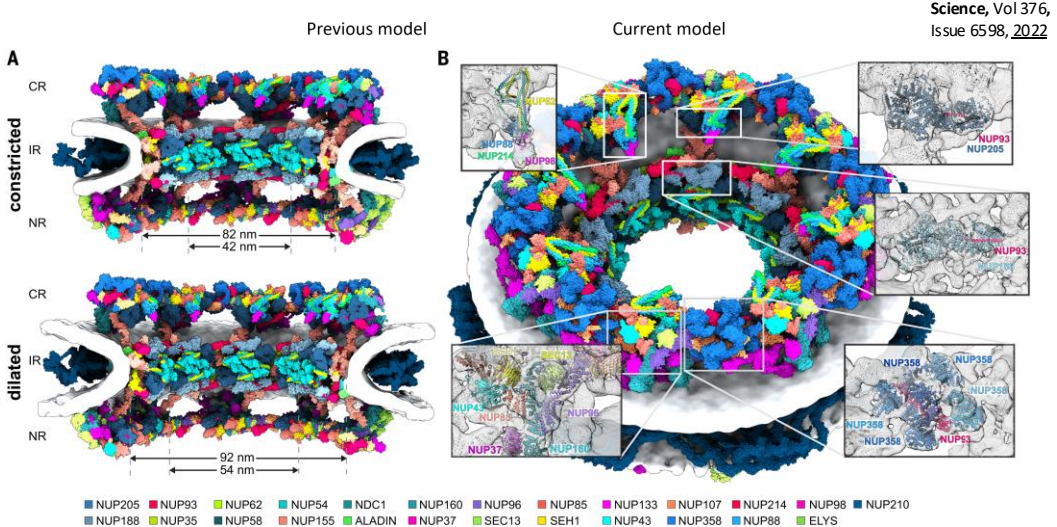
Nuclear Pore Complex



Transport of large molecules to cell



Source: https://www.embl.org/news/science/puzzling-out-the-structure-of-a-molecular-giant/



Science, Vol 376, Issue 6598, 2022

Accessing software..

and licence restrictions…

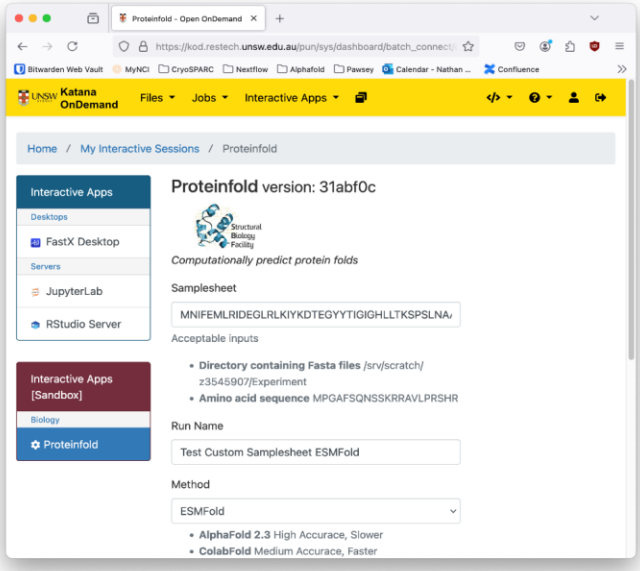
Where to run? Galaxy, Uni compute cluster, proteinfold

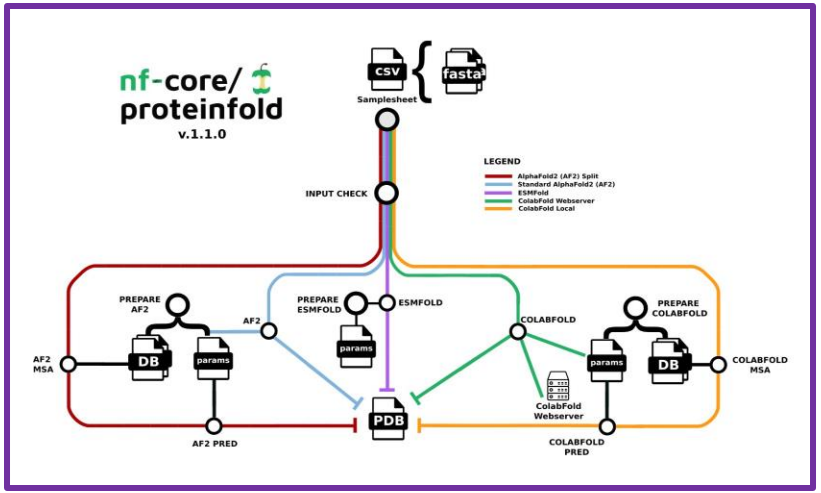
Australian BioCommons

ABOUT ACTIVITIES SERVICES TRAINING & EVENTS DOMAINS NEWS CONTACT HELP

Australian AlphaFold Service

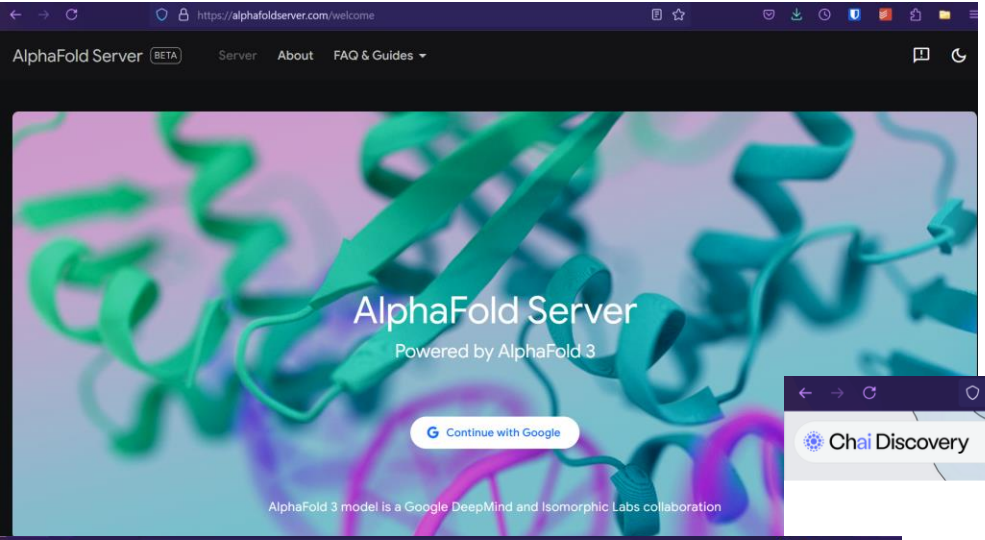
AlphaFold is an artificial intelligence (AI) system developed by DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy that is competitive with experimental methods (see Jumper *et al. Nature* 2021).

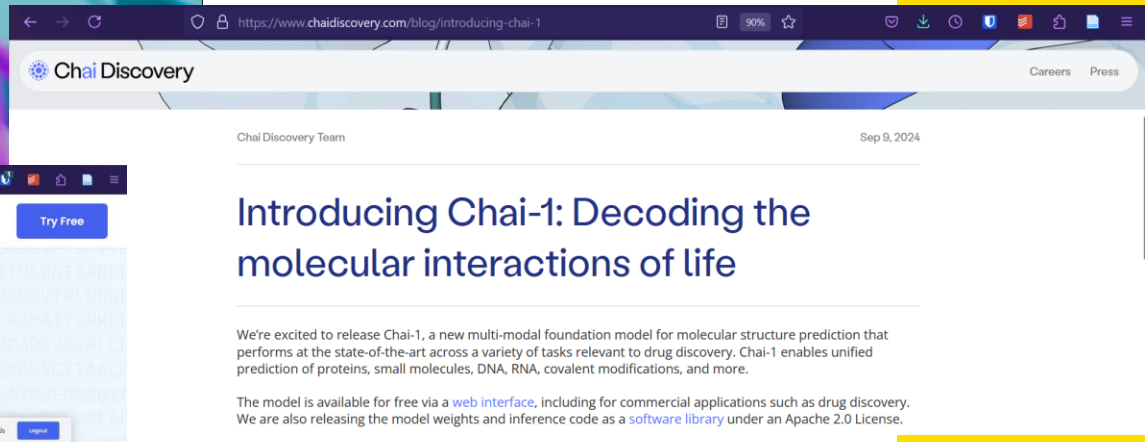


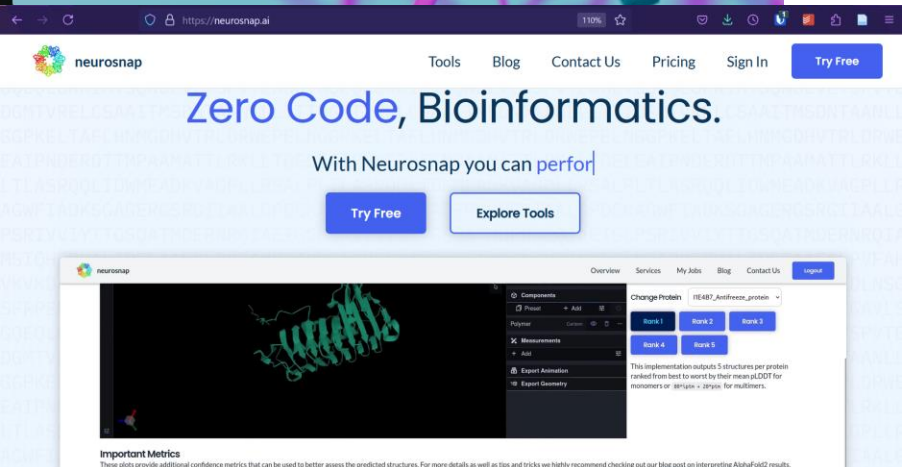


UNSW SYDNEY

Dedicated web forms - portals built by companies

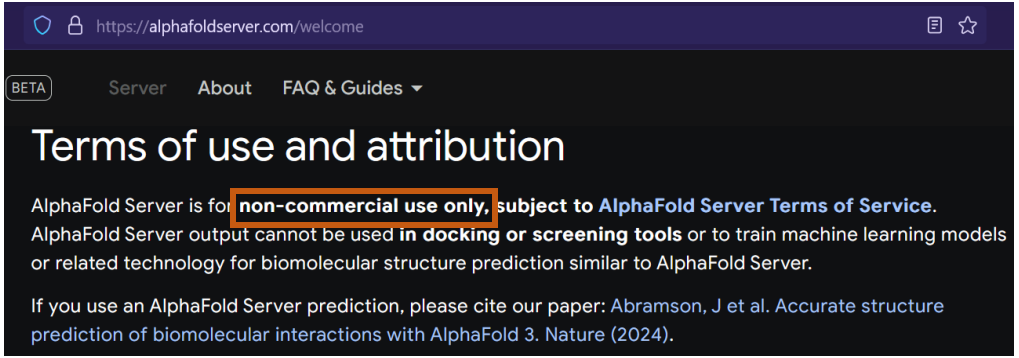


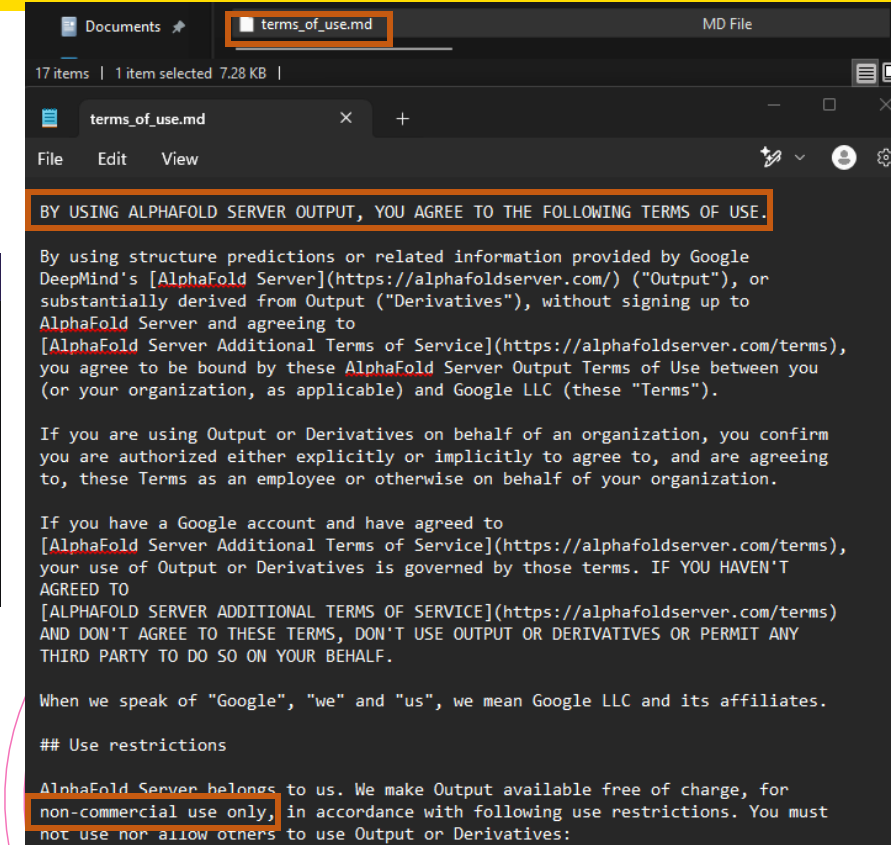




Licensing restrictions

Think before you input!

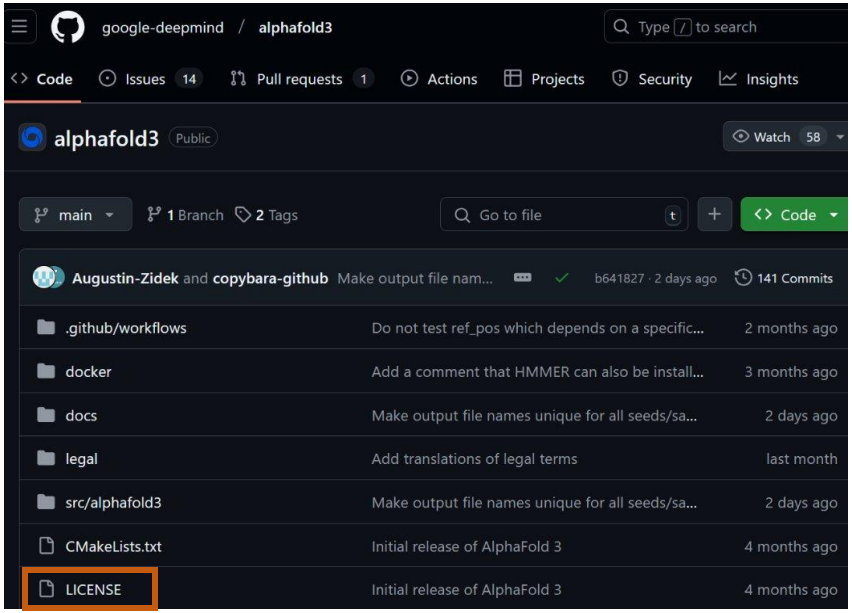




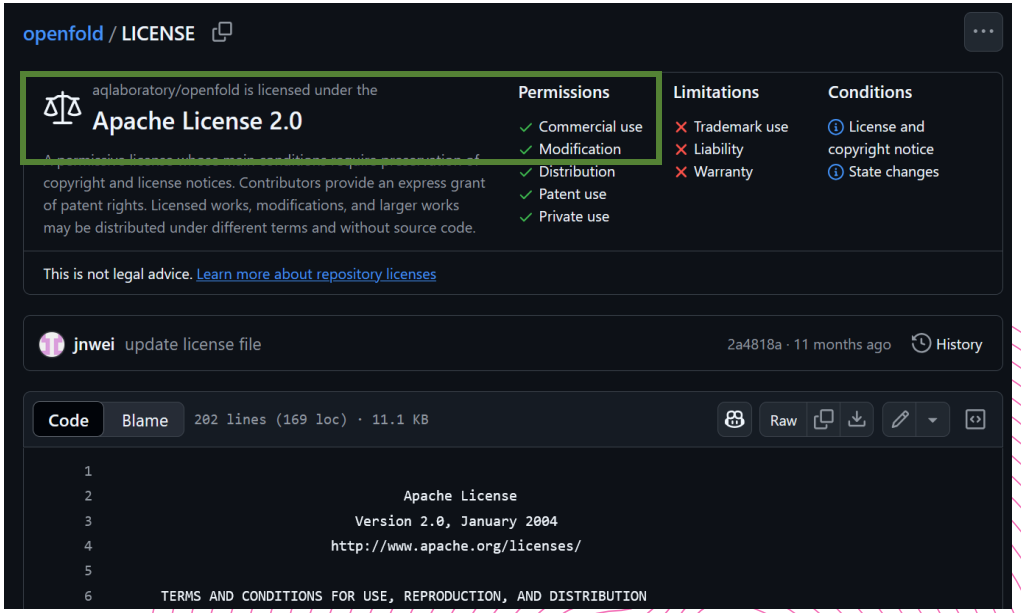
Licensing restrictions

find the LICENSE file summary

Restricted



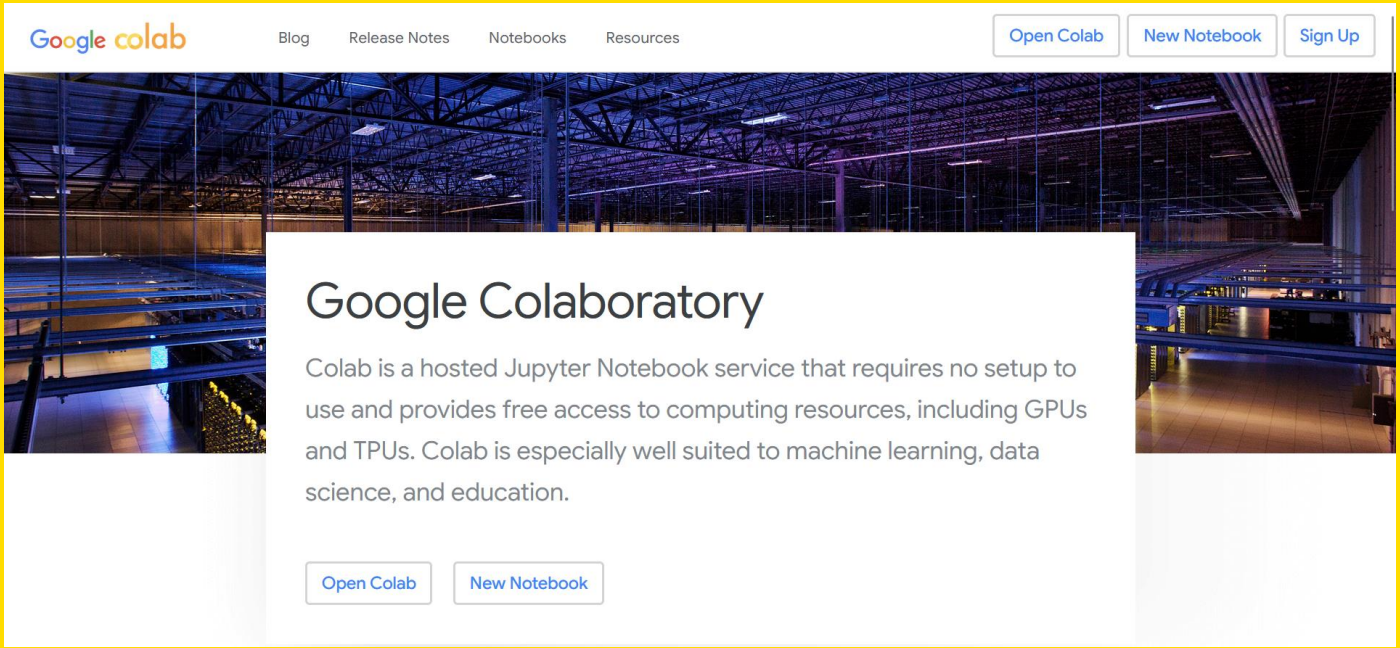
Unrestricted



Break!

Get your laptops ready to run your own structure prediction :)

ColabFold – running AlphaFold2 on Colab



https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

ColabFold – running AlphaFold2 on Colab

https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

Architecture compared to `input_options` (mostly default)

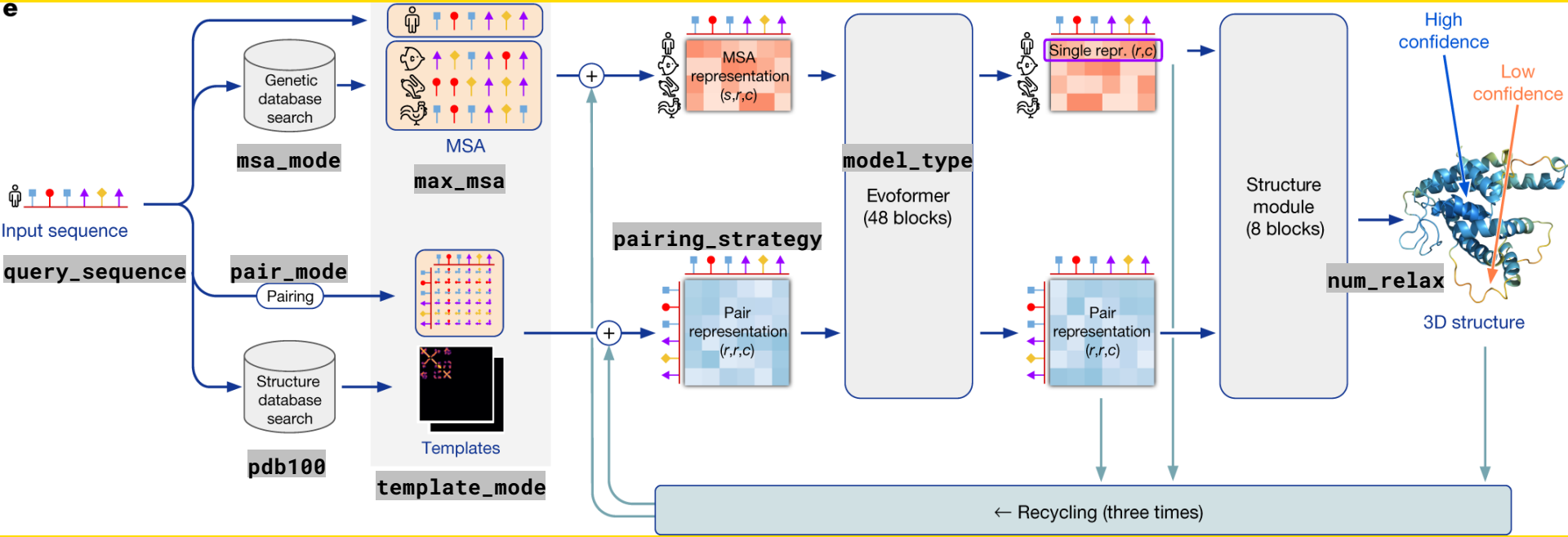
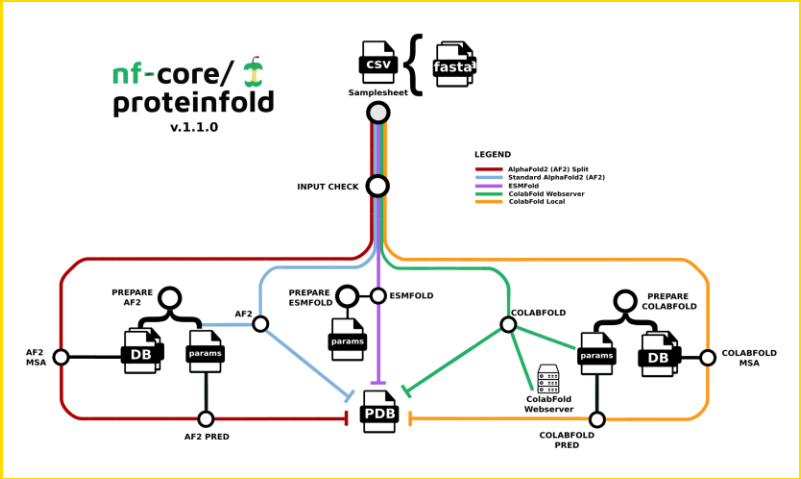


Fig. 1: AlphaFold produces highly accurate structures – from *"Highly accurate protein structure prediction with AlphaFold"* - Nature, **596**, 583-89 (2021).

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 version: 0.1.6a-1-g77aceb9

Structural Biology Facility

Computationally predict protein structures

Samplesheet

/srv/scratch/USER/test_run_files

Acceptable inputs

- **Directory containing Fasta files:** `/srv/scratch/z3141592/my_experiment`
- **Amino acid sequence:** `NLYIQWLKDGGPSSGRPPPS`

Run Name

test_run

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
- **HelixFold3** High Accuracy, Slower (Restricted AlphaFold3 method) - Pre-print
  WARNING: Non-commercial use only!

Mode

Monomer

Email Address

k.rowell@unsw.edu.au

Launch

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

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