Folding Entire Proteomes

Predict a molecular inventory of your organism

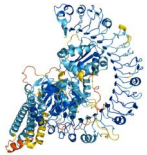
Proteomics bioinformatics Community Meeting - 2024 Dec 02

Dr Keiran Rowell

UNSW SYDNEY

UNSW Ramaciotti Centre for Genomics

Sequence genome → 4544 proteins

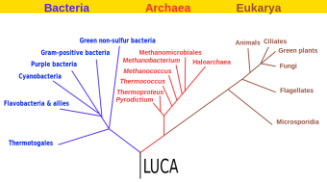


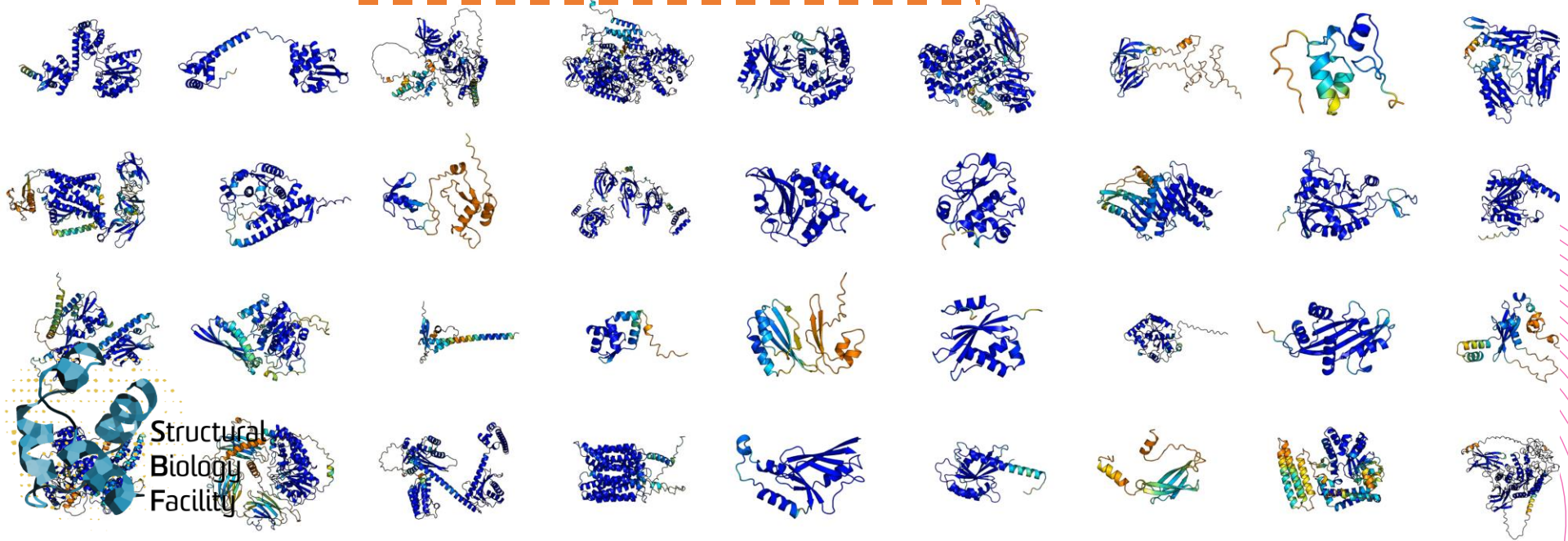
Google DeepMind

OpenFold — Democratizing AI for Biology

Evo-AI methods → 4544 structures

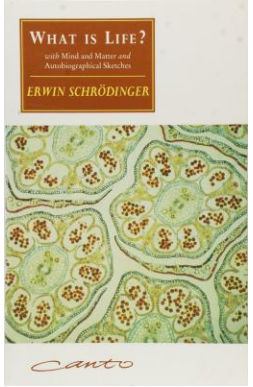
Struct search → AlphaFoldDB



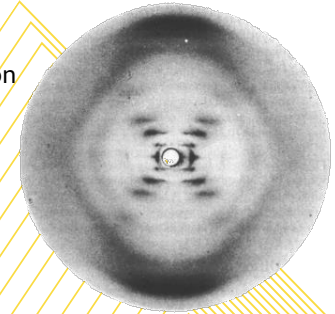


Structural Biology Facility

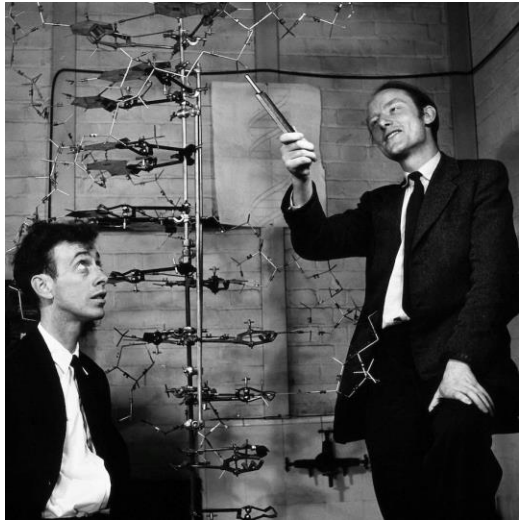
Models in Chemical Biology - function follows chemical form



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

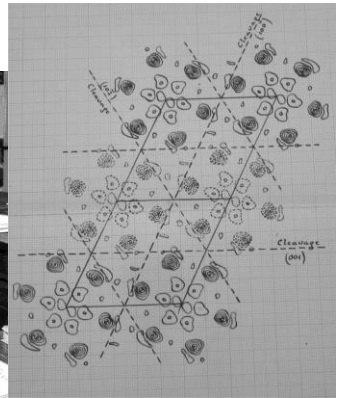


Wilkins & Fraklin's X-ray diffraction images of DNA **1953**



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

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

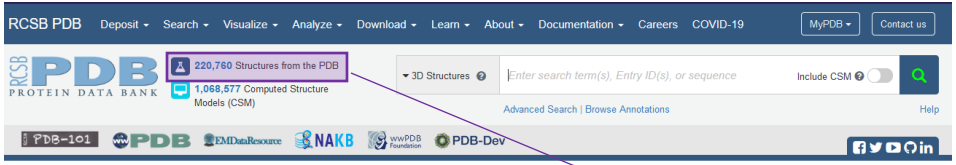


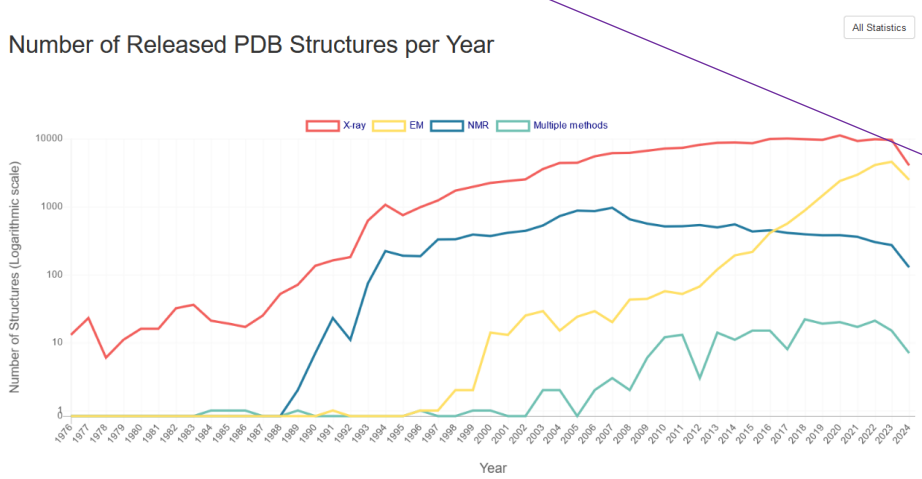
Kathleen Lonsdale resolves the structure of benzene **1924**

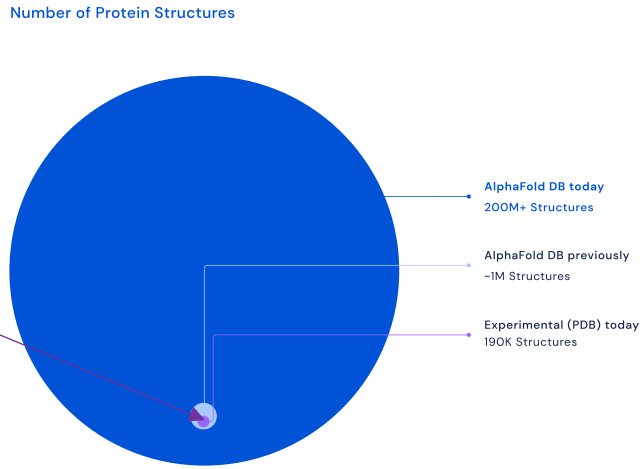
CRICOS Provider Code 00098G

UNSW SYDNEY

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







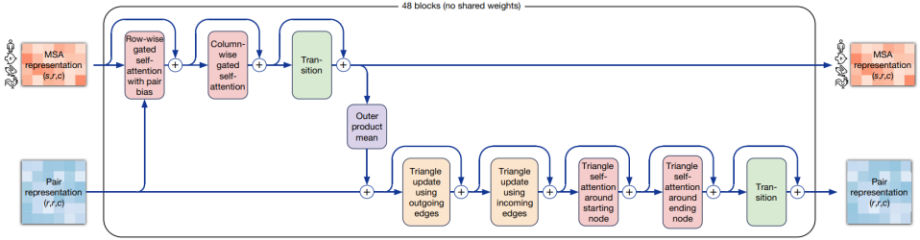
Compute power – Moore's law +
DNA Sequencing – super-Moore's law +
Deep Learning – context scaling laws
=
Comp.Struct.Bio. - **1000x** in 2 years

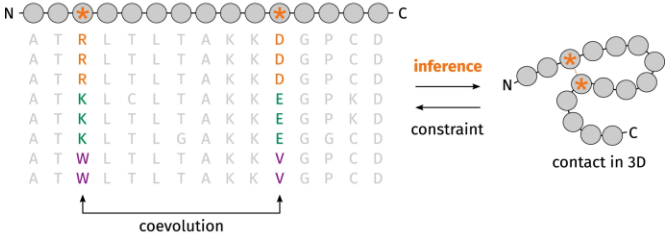
UNSW SYDNEY

CRICOS Provider Code 00098G

X'Fold' programs - co-evolution DeepLearning

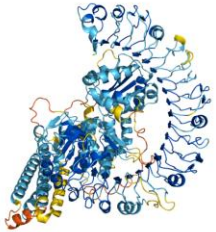
DL: No physics of folding – Transformers – 1D sequence -> 3D structure

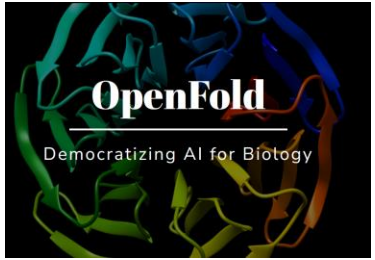




'Evo'formers (GPU+CPU+DBs)

(**Alpha**|Open|RoseTTA|Ab)**Fold**

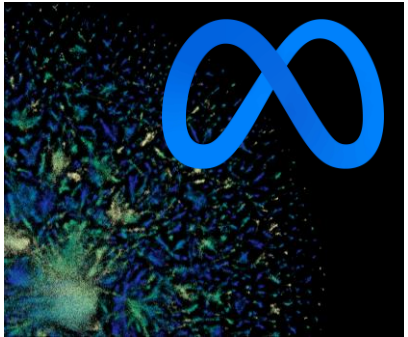




Google DeepMind

Protein-LMs (GPU)

ESMFold, ProtTrans



Struct Alignment (DBs)

Foldseek, Foldmason

UNSW SYDNEY

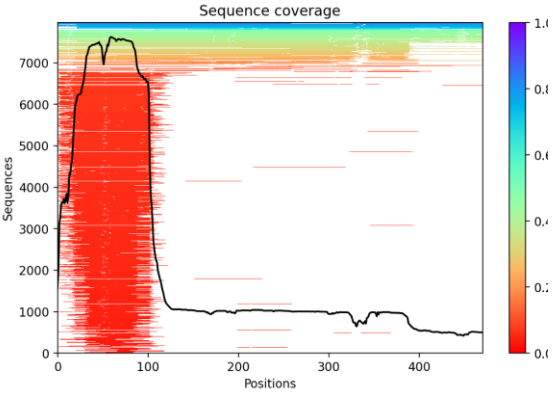
CRICOS Provider Code 00098G

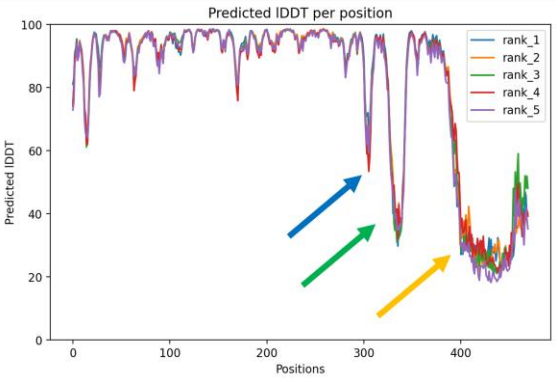
X'Fold' programs - No Physics

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

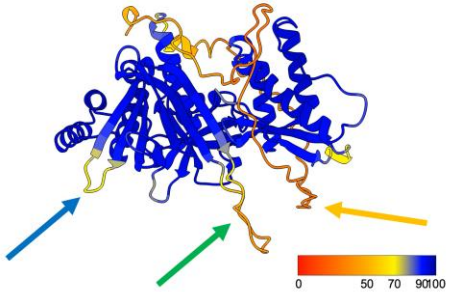
PDB ID: 7RDT





0000

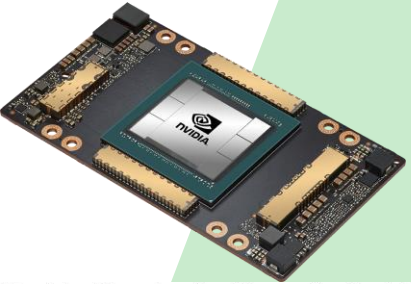
PDB ID: 7B3A

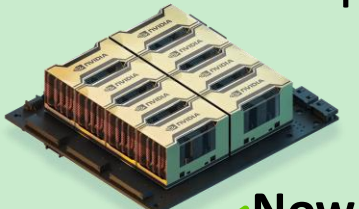


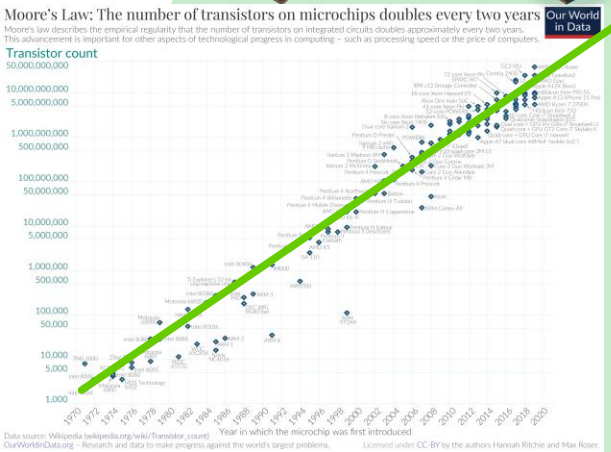
Confidence metrics: Dr Michael Healy, 'AlphaFold2 How-To Guide'

CRICOS Provider Code 00098G

GPUs – Multiply things *really* fast

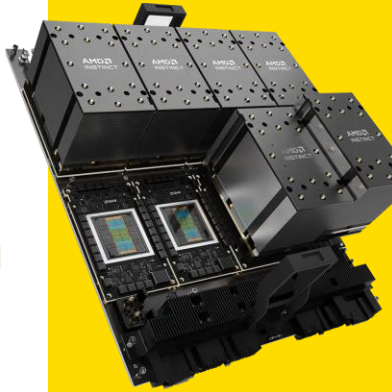






Now 100 billion transistors in *each* GPU

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

UNSW SYDNEY

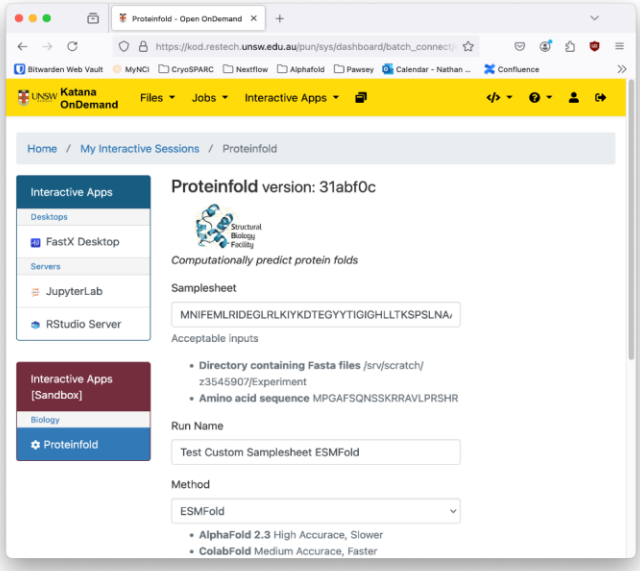
Where to turn? - Galaxy, Uni compute cluster, proteinfold

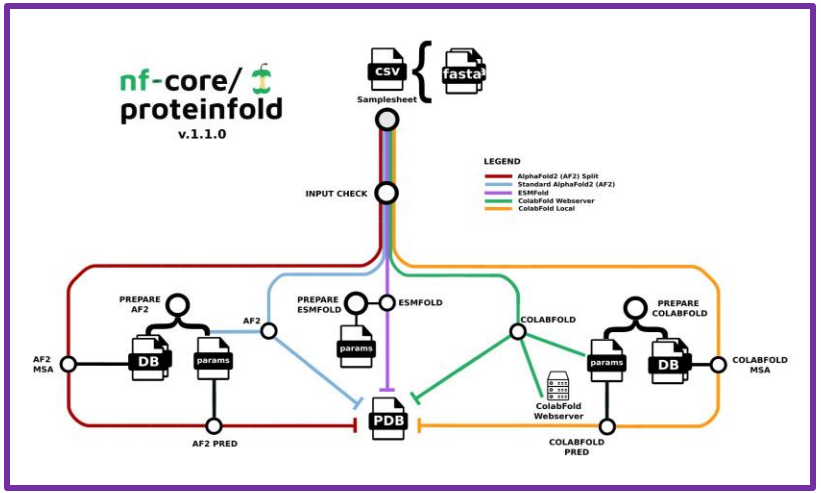
Australian BioCommons

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).





UNSW SYDNEY

CRICOS Provider Code 00098G

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 "KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQAT" https://api.esmatlas.com/foldSequence/v1/pdb/
```

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

```
z3374843@katana3:~ $ 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)
```

```
z3374843@katana3:~ $ 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
```

```
(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]
```

```
z3374843@katana3:~ $ 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
```

CRICOS Provider Code 00098G

Whole proteome?

github.com/keiran-rowell-unsw/Loki-ASV2_in_silico

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

ESMFold – 4378 prots, 10.5 hrs

File: esmfold_loki-ASV2_all_ORFs.log

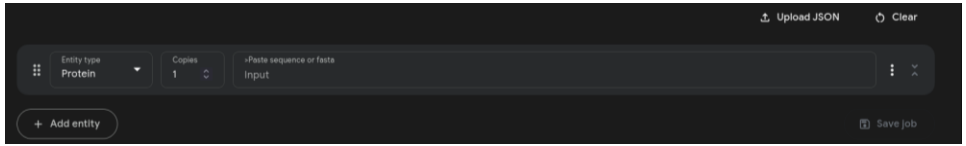
```
24/09/06 10:07:00 | INFO | root | Reading sequences from loki-ASV2_dfast_all_protein_ORFs.faa
24/09/06 10:07:00 | INFO | root | Loaded 4544 sequences from loki-ASV2_dfast_all_protein_ORFs.faa
24/09/06 10:07:00 | INFO | root | Loading model
24/09/06 10:07:50 | INFO | root | Starting Predictions
24/09/06 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.
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.
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.
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.
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.

24/09/06 20:36:11 | INFO | root | Failed (CUDA out of memory) on sequence LOKIASV2_24920 penicillin acylase family protein of length 884.
24/09/06 20:36:12 | INFO | root | Failed (CUDA out of memory) on sequence LOKIASV2_35940 hypothetical protein of length 884.
24/09/06 20:36:12 | INFO | root | Failed (CUDA out of memory) on sequence LOKIASV2_20420 hypothetical protein of length 885.
24/09/06 20:36:13 | INFO | root | Failed (CUDA out of memory) on sequence LOKIASV2_05820 DNA methyltransferase of length 887.
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.
24/09/06 20:36:14 | INFO | root | Failed (CUDA out of memory) on sequence LOKIASV2_43680 hypothetical protein of length 888.
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

```
python3 docker/run_docker.py \
    --fasta_paths=./fasta_dir/file.fasta \
    --max_template_date=2022-01-01 \
    --model_preset=monomer \
    --data_dir=/mnt/af2/ \
    #--use_precomputed_msas=true \
    --output_dir=/mnt/data/alphafold_output/ 2>&1 | tee -a alphafold2.log
```

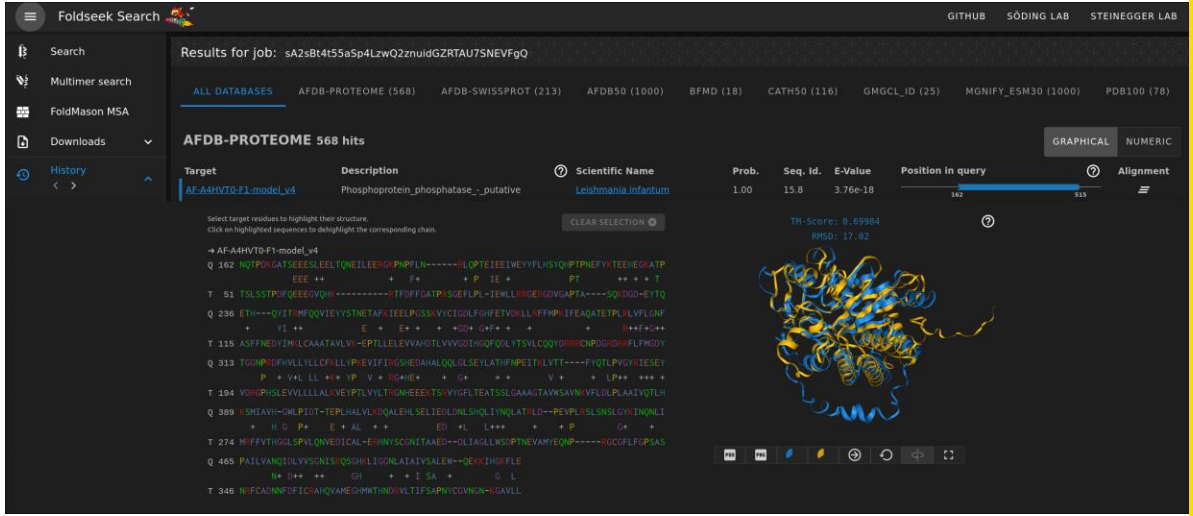
AlphaFold3 – 6 prots, pretty quick



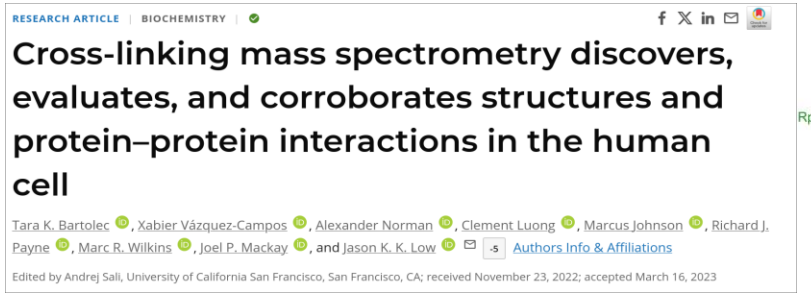
CRICOS Provider Code 00098G

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

search.foldseek.com



github.com/Rappsilber-Laboratory/AlphaLink2

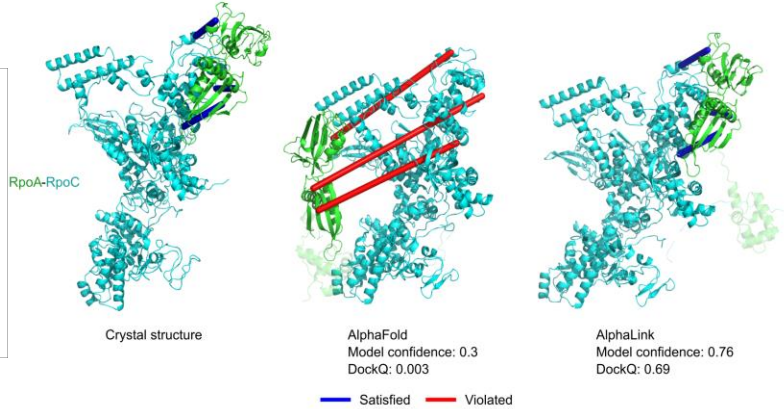


RESEARCH ARTICLE | BIOCHEMISTRY

Cross-linking mass spectrometry discovers, evaluates, and corroborates structures and protein–protein interactions in the human cell

Tara K. Bartolec, Xabier Vázquez-Campos, Alexander Norman, Clement Luong, Marcus Johnson, Richard J. Payne, Marc R. Wilkins, Joel P. Mackay, and Jason K. K. Low 5 Authors Info & Affiliations

Edited by Andrej Sali, University of California San Francisco, San Francisco, CA; received November 23, 2022; accepted March 16, 2023



UNSW SYDNEY

CRICOS Provider Code 00098G

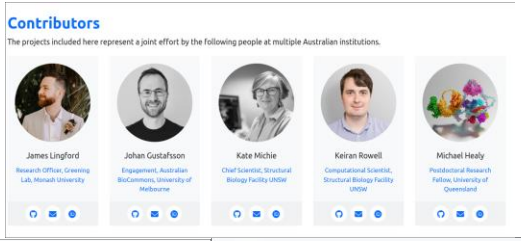
AlphaFold Practical Guide European Bioinformatics Institute Community

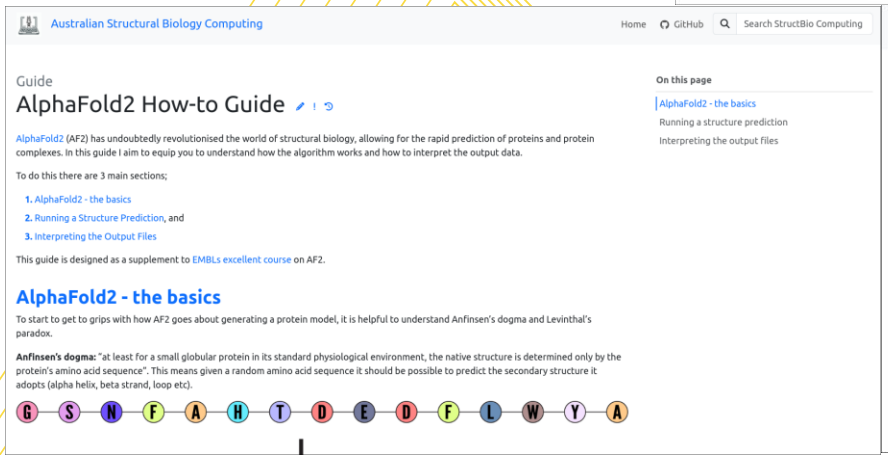
ebi.ac.uk/training/online/courses/alphafold

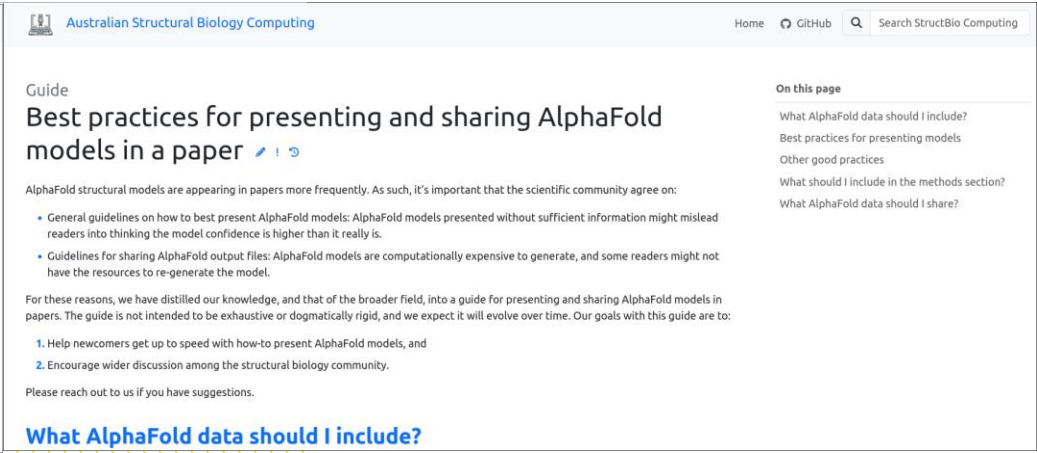
EMBL-EBI Google DeepMind

Australian Computational Structural Biology Community

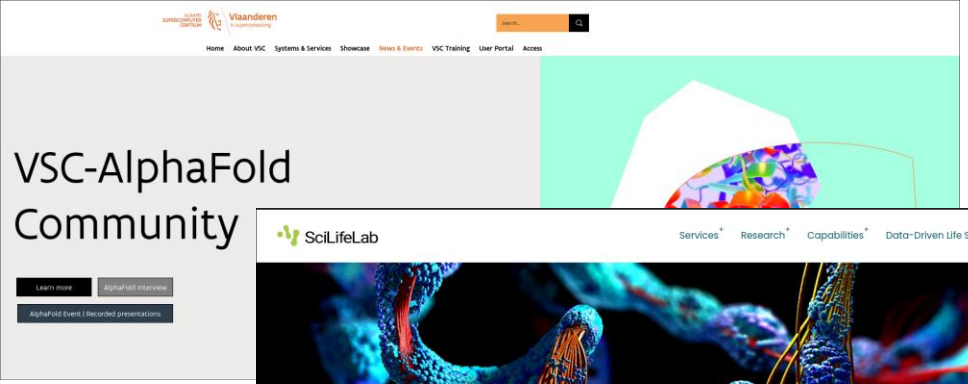
australian-structural-biology-computing.github.io/website/

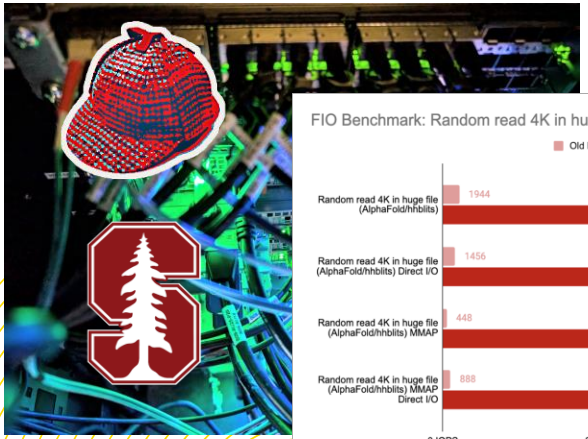


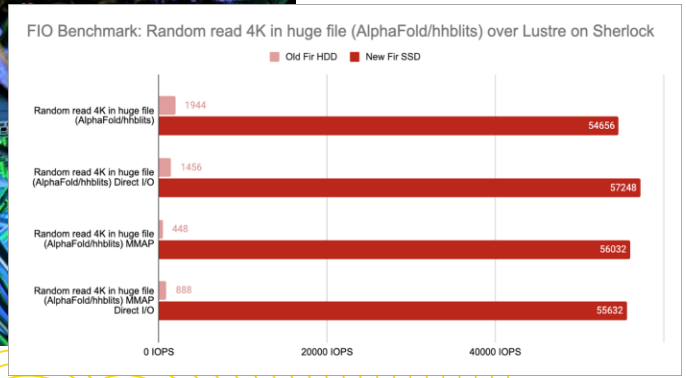


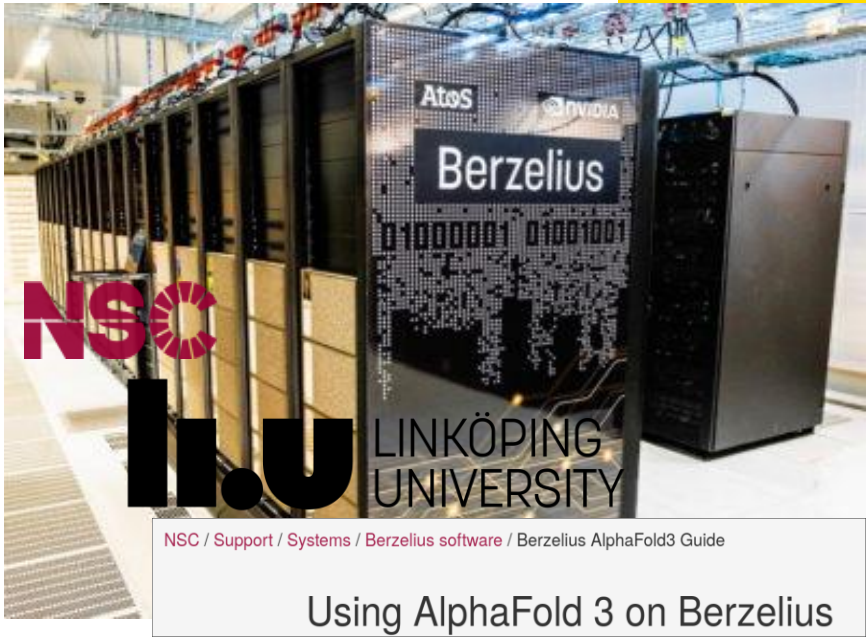


The world is doing this









Questions / Discussion