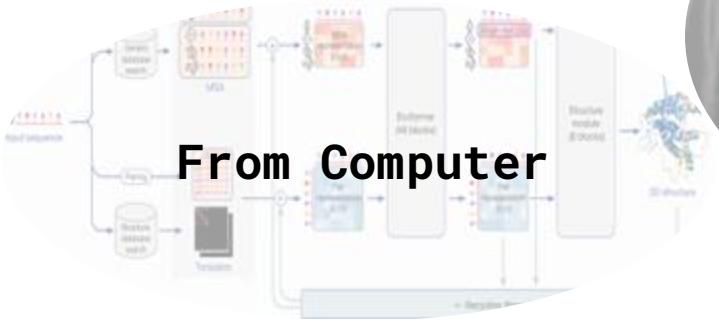
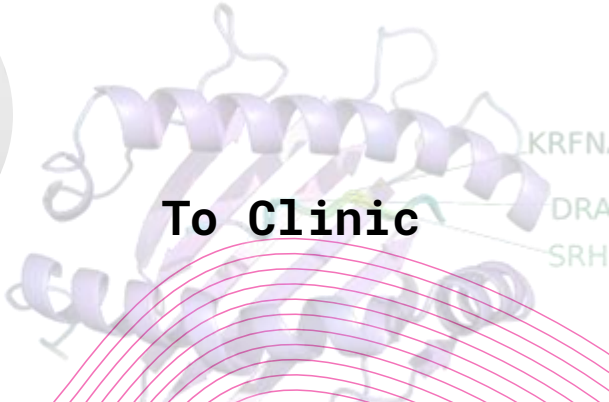
Computational revolution in Structural Biology

ResTech All Hands Meeting: 2024 July 13
Dr Keiran Rowell



From Computer

From Instrument



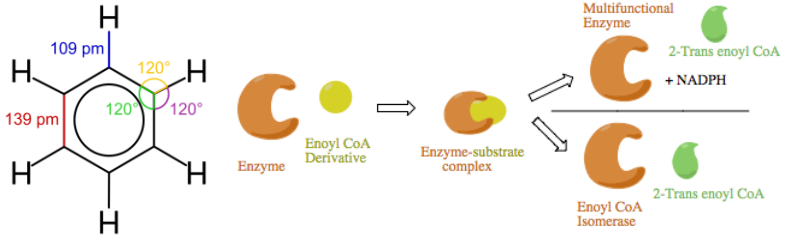
To Clinic

To Science

Structural Biology Facility

Structural Biology & Modelling

Molecules: 3D, represented by many models



'**Chemical Accuracy**' - models go beyond depictions, determine behaviour

Depends on application:

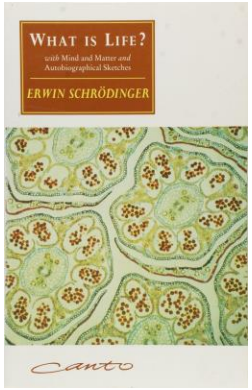
- Reaction energies: 1/100th of bond strength energy
- Protein structures: 0.1-0.3 nm *'atomic resolution'*
- Structural Evolution: 1-in-1-billion spatial alignments

Structures 'hand-made' -> routine in the last 3 years

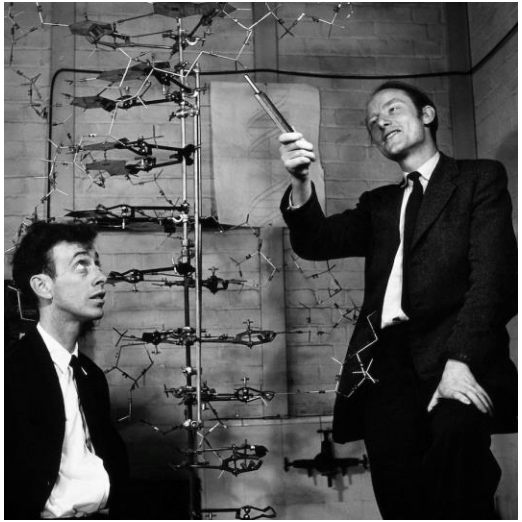
Comp. Struct. Bio. *was* academic, now determines science

UNSW SYDNEY

Models in Chemical Biology - function follows chemical form

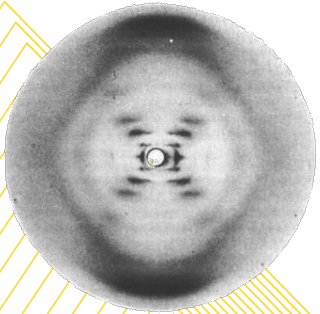


Schrödinger speculates on the molecules of life **1944**

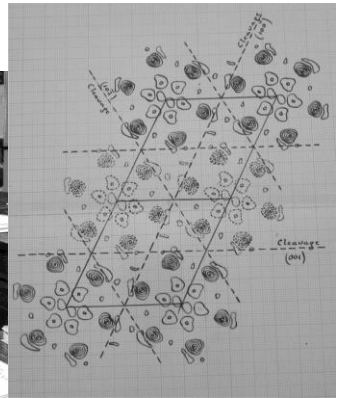


Watson & Crick's DNA double-helix model **1953**

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



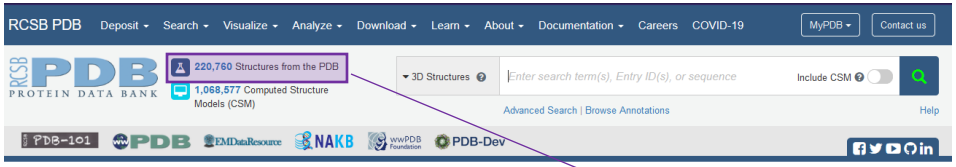
Wilkins & Fraklin's X-ray diffraction images of DNA **1953**

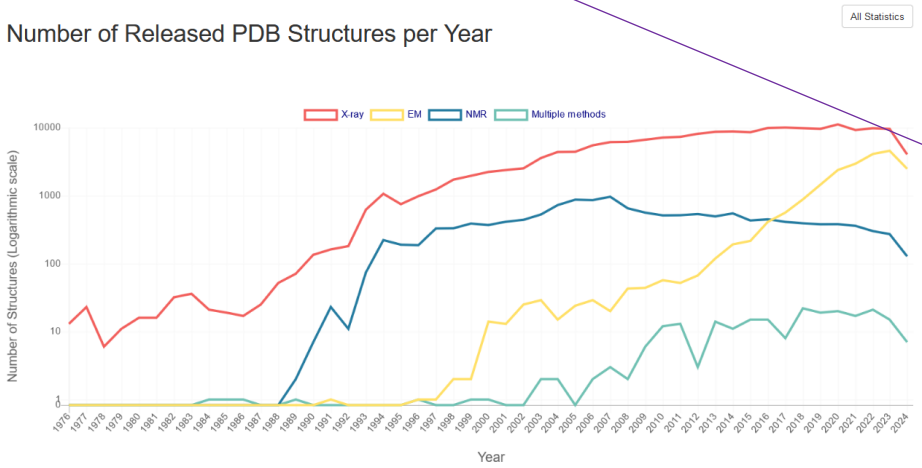


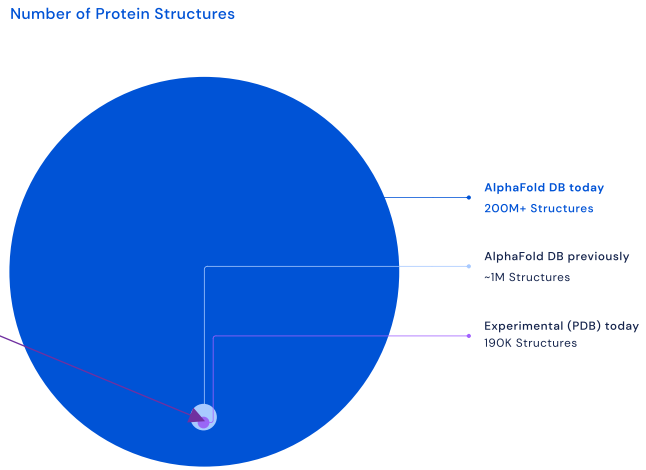
Kathleen Lonsdale resolves the structure of benzene **1924**

UNSW SYDNEY

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



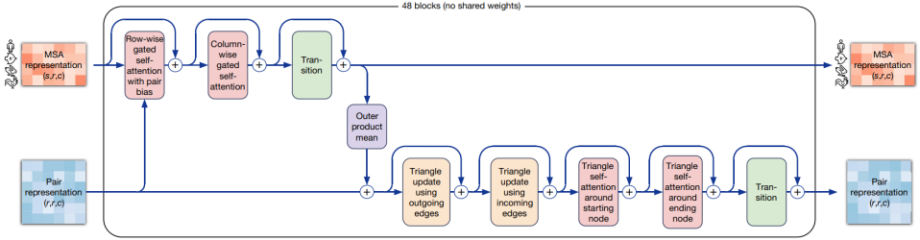


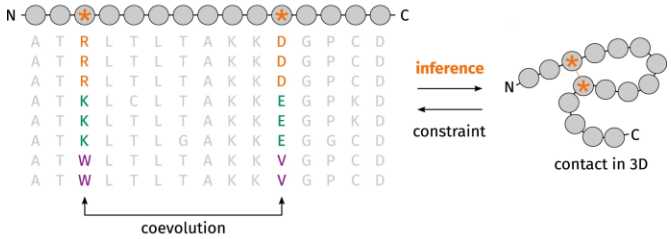


Compute power – Moore's law +
DNA Sequencing – super-Moore's law +
Deep Learning – context scaling laws
=
Comp.Struct.Bio. - **1000x** in 2 years

X'Fold' programs - co-evolution DeepLearning

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

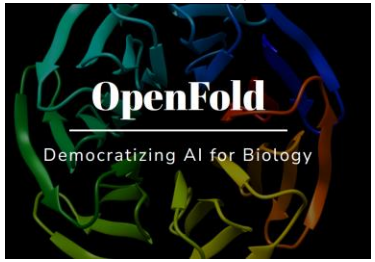


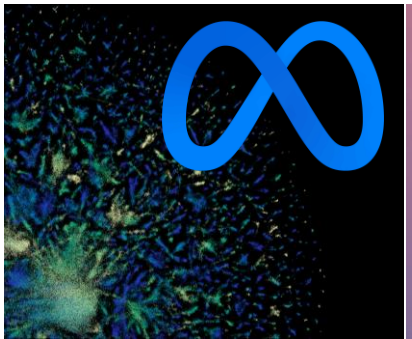


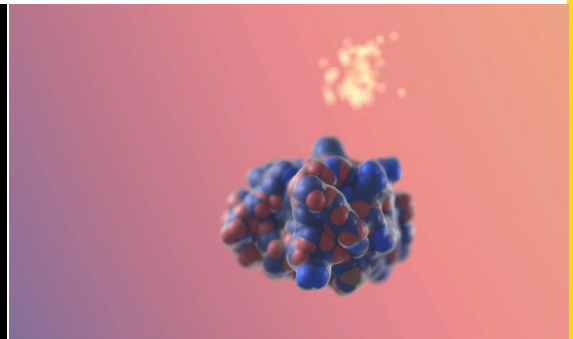
HPC & Hyperscalers: Batch, O(N^2) VRAM, O(N^3) time, fixed DBs, 1000s calcs, fast I/O

'Evo'formers (GPU+CPU+DBs) **Protein-LMs (GPU)** **Diffusion GenAI (GPU)**

(Alpha|Open|RoseTTA|Ab)Fold ESMFold, ProtTrans RFDiffusion, Chroma







Google DeepMind

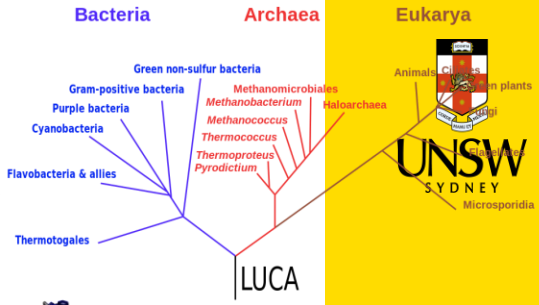
DL Biomolecular structures – Uses

Archaeal evolution – Michie/Burns Labs - ESMFold + FoldSeek

UNSW Ramaciotti Centre for Genomics

Sequence whole genome → 4622 proteins

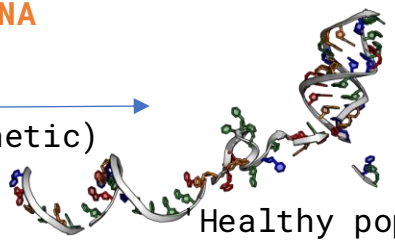
Structural → AFDB-search



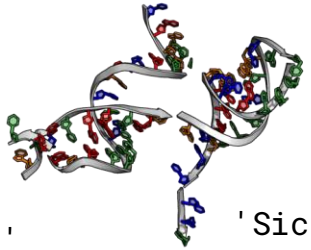
Medical genomics – Oates Lab – RFoldNA

gnomAD Genome Aggregation Database

160 variants → Neural disease (genetic)



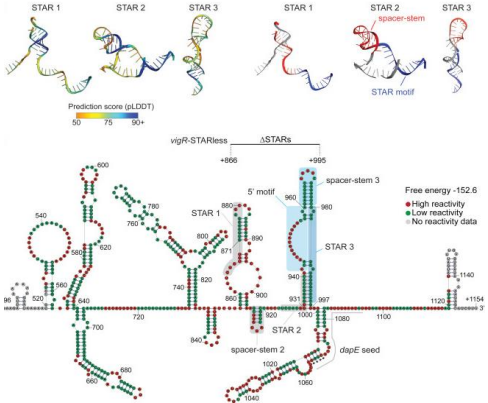
'Healthy population'



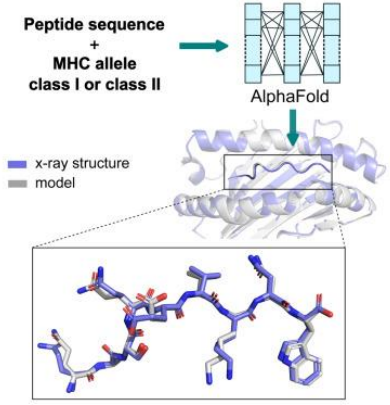
'Sick kid'

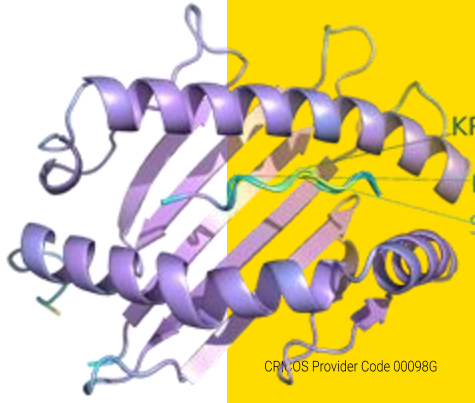
Antibiotic Resistance – Tree Lab – RFoldNA

'Golden Staph' vancomycin tolerance via RNA structures



Immune response – Tedla Lab –T(Cell)Fold





Workloads – Biomolecular Structures

(Alpha)Fold walltime:

- **Bulk fold a proteome** (A100-40)
  - Archeal virus – 613 proteins – AlphaFold 22 days - ESMFold 8 hrs
  - Lokiarcheon – 4622 proteins – AlphaFold ??(2yrs) - ESMFold 10 days

- **National use** (Galaxy Aus)
  - ~15k AlphaFold jobs. Median 10, mean 83. Top user 1000s
  - 55% of jobs are multiple proteins together - "multimers"

(Alpha)Fold VRAM use – courtesy of Daniel Cao @ HPE – thanks for benchmarking!

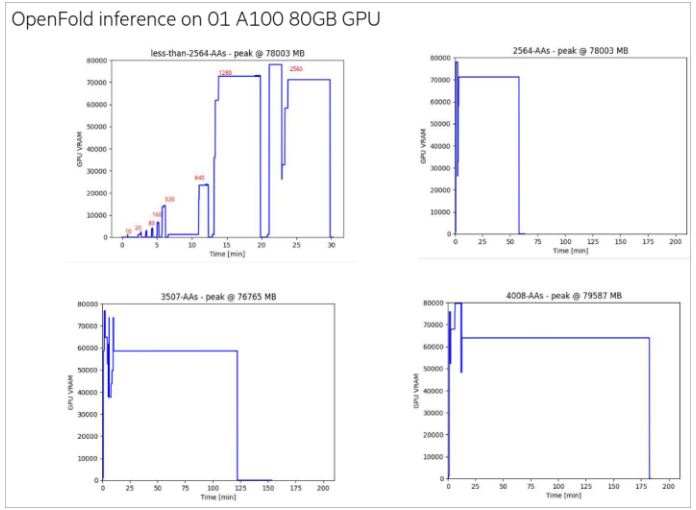
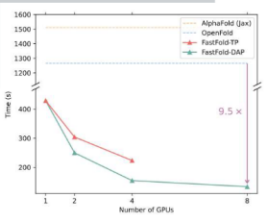


Table 1 – Inference latency for long sequence

Sequence Length	OpenFold (01 GPU)	FastFold (02 GPUs)	FastFold (04 GPUs)	FastFold (08 GPUs)
2564	~56 min	495.6 $s^{(64)}$	262.4 $s^{(64)}$	8 GPUs only offer marginal improvement, even when using FastFold DAP (Ref: FastFold paper below)
3013	~60 min	1391 $s^{(64)}$	419.5 $s^{(64)}$	
3507	~122 min	1426.17$^{(16)}$	OOM$^{(64)}$ -> 678.3$^{(16)}$	
4008	~180 min	2873.1 $s^{(16)}$	1494.6 $s^{(16)}$	
4516		OOM$^{(16)}$	OOM$^{(16)}$ -> OOM$^{(8)}$ -> OOM$^{(4)}$	
5005	OOM	OOM$^{(16)}$	OOM$^{(16)}$ -> OOM$^{(8)}$	

(64)/(16)/(8)/(4) Fixed chunk size of 64/16/8/4
To replicate the results, download the scripts from this GitHub repo



(c) Sequence Length = 2560

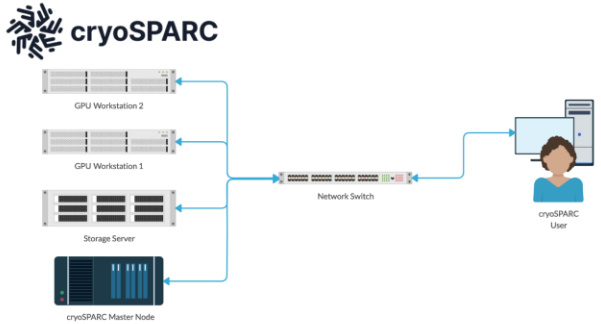
RFoldNA walltime: 160 variants in 24 hrs (A100-40)

RFDiffusion walltime: design ~100 scaffolds/hr, then AlphaFolded

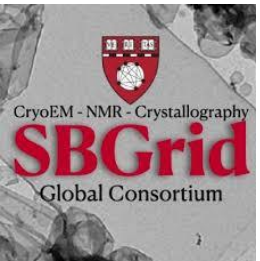
Cryo-Electron Density processing

Evolving ecosystem of EM-image software

The best (for ease of use)



- App-like user-friendly web interface, port forward
- Most popular suite, but not feature-complete
- Supports PBSPro integration for long image jobs



- 555 popular Structural Biology packages
- Version and dependency control (CUDA)
- UniMelb (Spartan) has SBGrid module
- source /programs/sbgrid.shrc

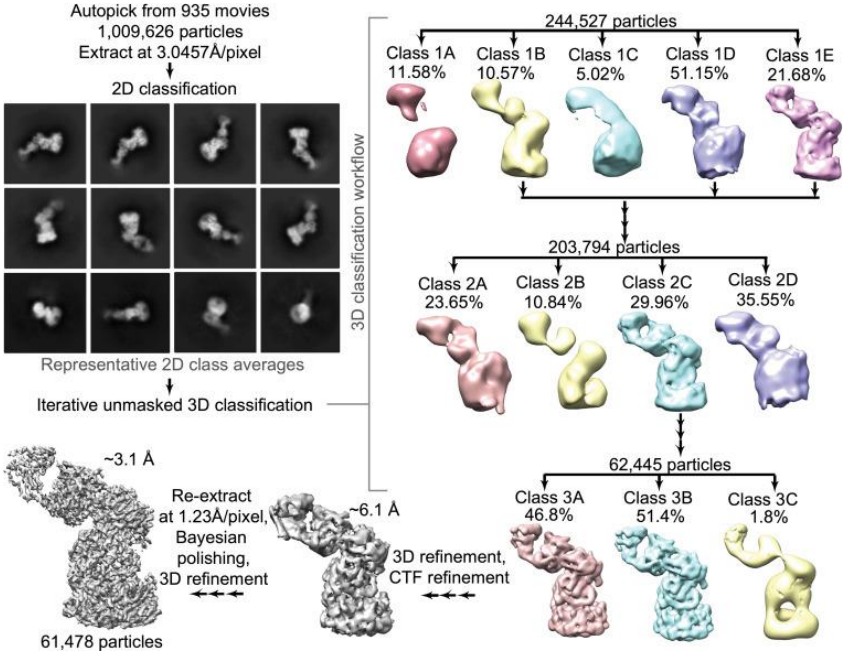
...and the rest

- Relion
- IMOD/Etomo
- EMAN2
- MotionCor2
- Gctf
- Scipion (Spanish CryoEM competitor to SBGrid)
- ISOLDE (interactive Molecular Dynamics refinement)

- Iterative refinement, not 'set-and-forget' batch
- Different graphics stacks (XCB/OpenGL, direct/indirect render)
- Some programs (e.g. IMOD) recommend against VNC remote view
- Monash 'MASSIVE' data visualisation HPC supports *some programs*

Workloads – CryoEM data transfer

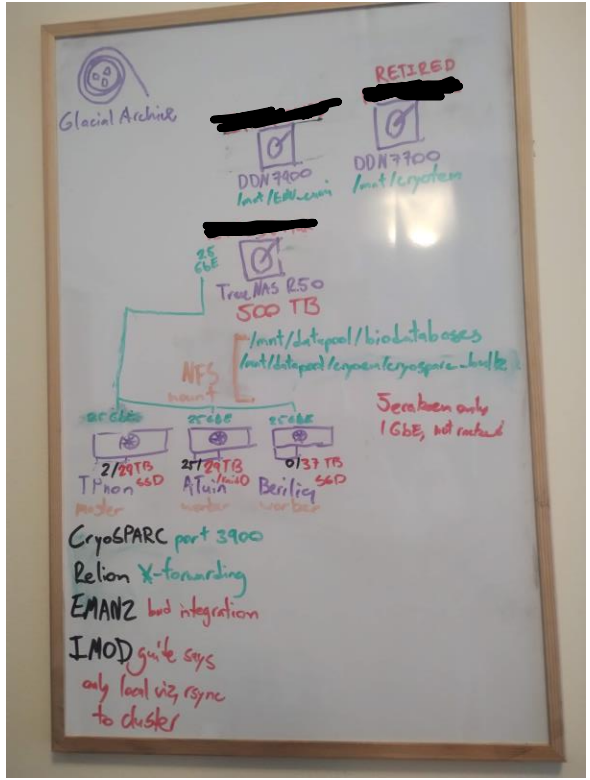
CryoEM structure determination – reprocess movies



Nguyen T.; Song J.; STAR Protocols 2(4):100852

TBs of movie files to transfer from EMU
(currently cifs share mounted on worker, want to change)

'On-the-fly' very useful (Daniel Luque)



Largely CryoSPARC portal
Remote desk (RustDesk) fallback

UNSW SYDNEY

Improved Workflows – HPC + Network

Data transfer – couple TBs/day/CryoEM scope -> GPU worker SSD cache

Next gen storage – I/O-bound, >4x AlphaFold sequence align – 7x CryoEM

GPU nodes for structure prediction – new Hopper-141GB nodes

GPU/CPU pipelines - BioCommons NextFlow proteinfold - NCI/Pawsey

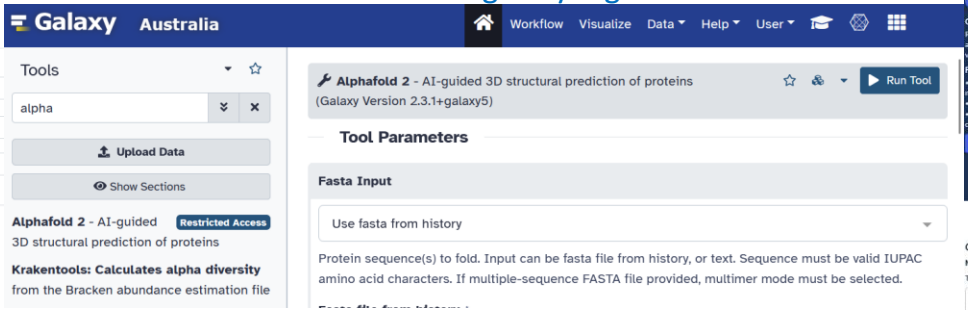
More efficient methods – MMSeq2 faster sequence alignment, ESMFold

Interface *wet-lab scientist friendly* - nice defaults, no CLI learning

UNSW SYDNEY

alphafold.restechlab.com

usegalaxy.org.au



neurosnap.ai

What's the Future for compute in Struct Bio?

UNSW SYDNEY

Protein-Interaction **screening:**

- CoFolding large #s of proteins and drug candidates

Workflow integration with sequence data**:**

- Ramaciotti RNASeq data in Katana -> workflow to RFoldNA 3D predictions

CryoEM <-> AlphaFold cross talk**:**

- Experimental density maps will help *de novo* structure prediction and vice versa

Molecular **Dynamics**:

- AlphaFold produces single static structures

GPU Quantum Chemistry of whole proteins**:**

- AlphaFold skips molecular physics. **QDX** (Barca group) has protein-scale QM calcs

UNSW SYDNEY

Questions / Discussion

How to make these tools accessible