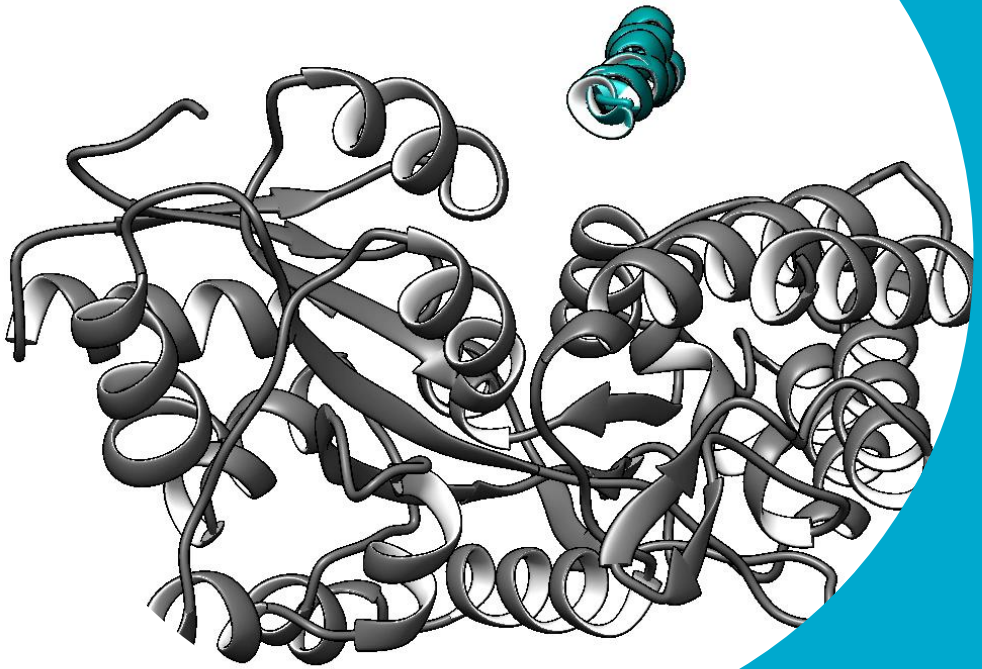
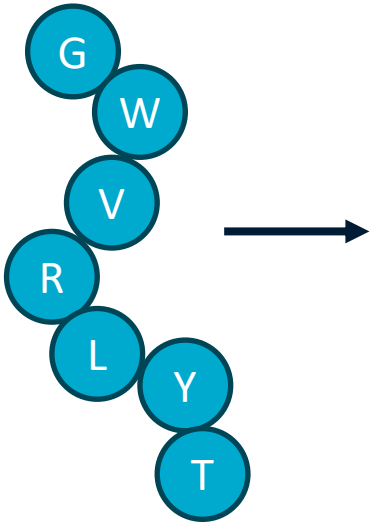
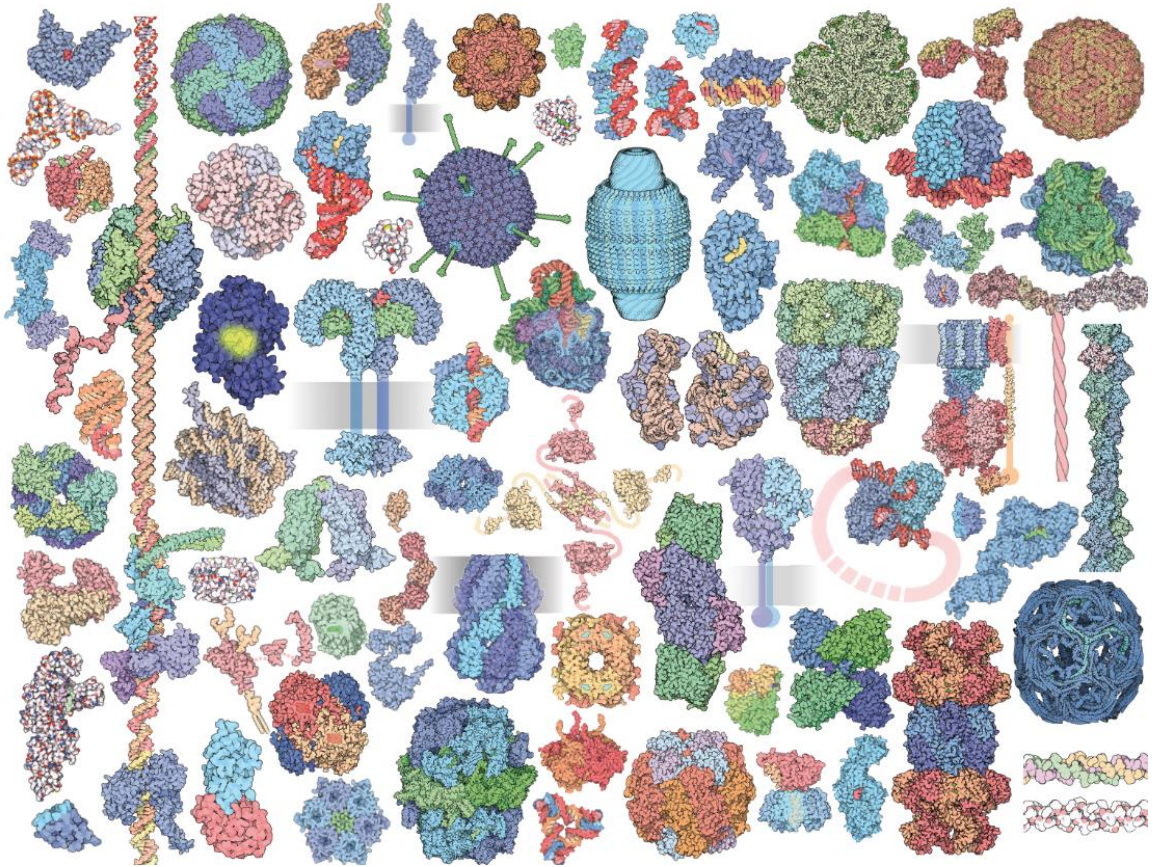
Fast-tracking biosensor design using high-throughput computational workflows



Helen Power| 29th Oct 24

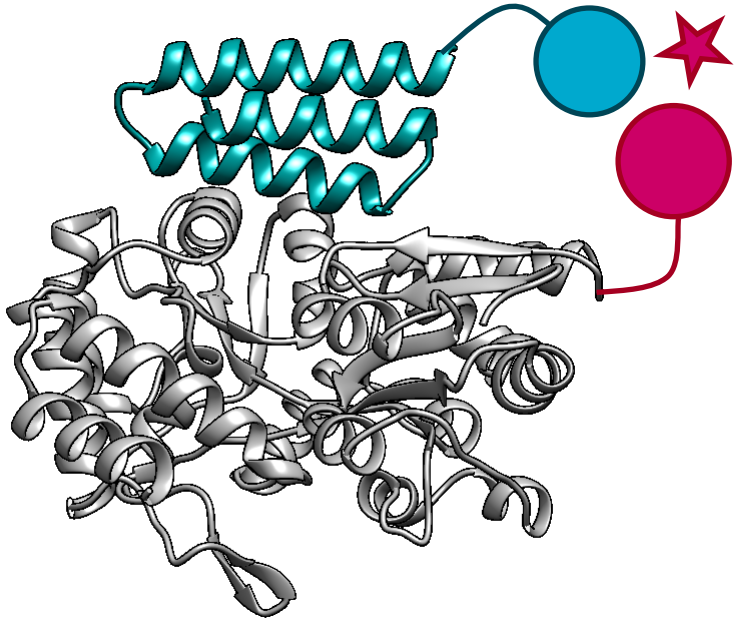




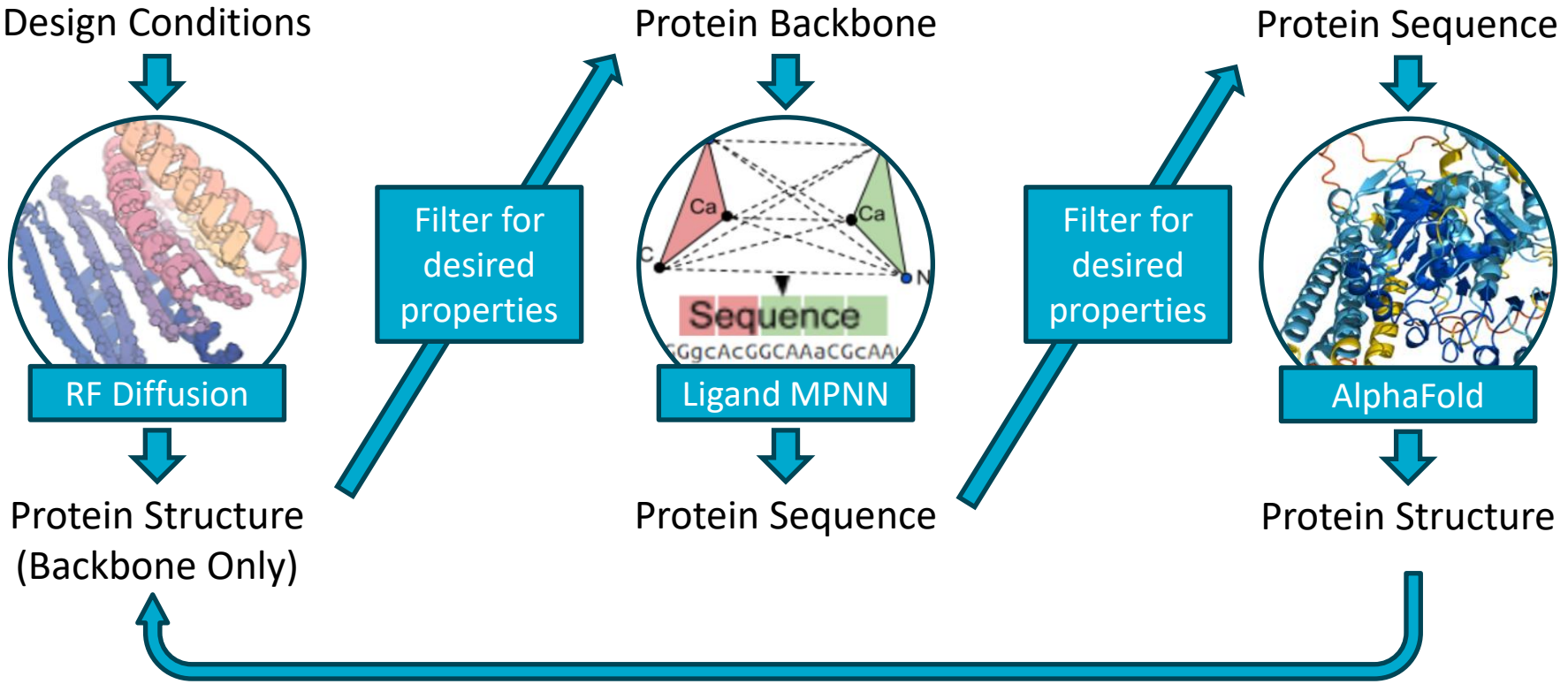
Acknowledgement: Illustration by David S. Goodsell, RCSB Protein Data Bank.

Protein Binders & Biosensors

- Mini-proteins that bind with high affinity and specificity to their target
- Perturb, inhibit, detect target proteins
- Biosensors for
  - Biomarkers of disease
  - Pathogens
  - Environmental contaminants
  - Research tools
  - …



Computational Binder Design Workflow



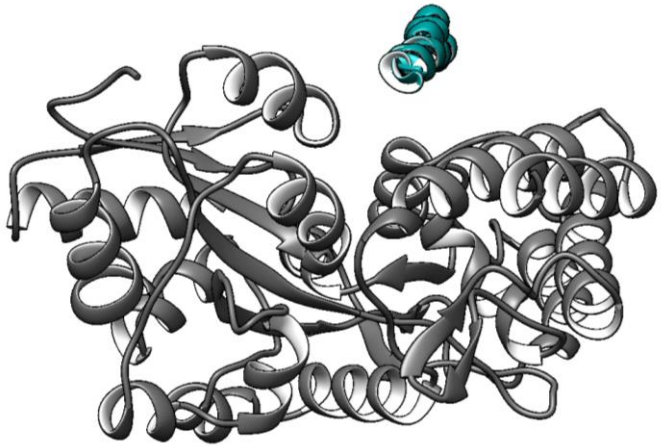
CSIRO HPC Systems

- **Petrichor**
  - CPU cluster
  - 404 x 64-core nodes with

  0.5, 1 or 4 TiB RAM
- **Virga**
  - GPU cluster
  - 111 Dual 36-core nodes with 512GB RAM and 4 NVIDIA H100 GPUs

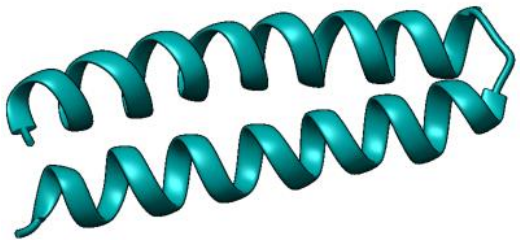
Structure Generation with **RFDiffusion**

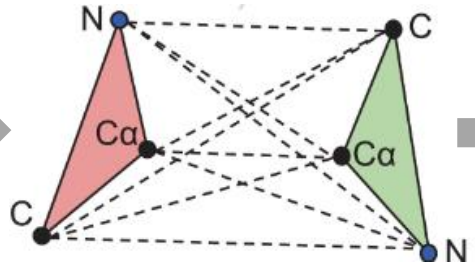
- Unconditional or conditional generation of protein structure
- Denoising diffusion probabilistic model
- **~1min** / design (H100 GPU)
- **5000** Binder Structures Generated



Sequence Prediction with **LigandMPNN**

- Protein Structure → Sequence
- Backbone coordinates processed to graph representation
- Amino acid identities iteratively decoded
- **<1s** per sequence
- 10 sequences / structure = **50,000** sequences

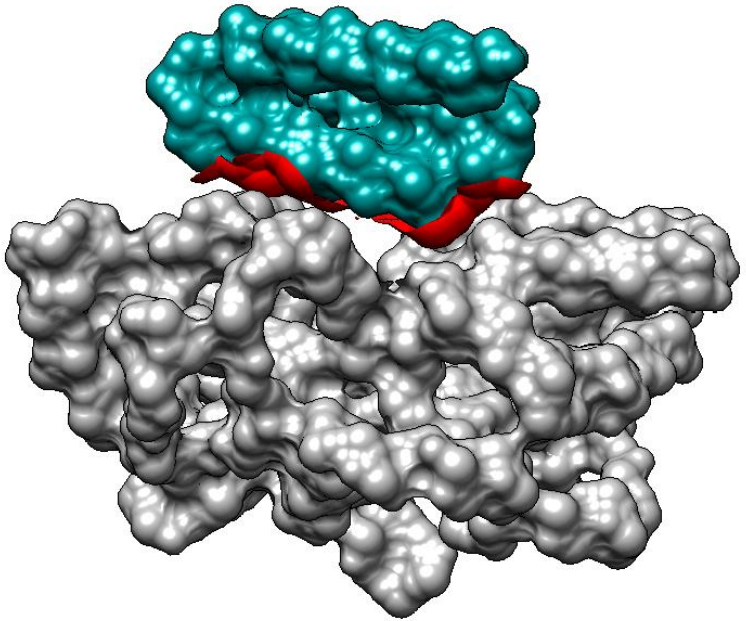




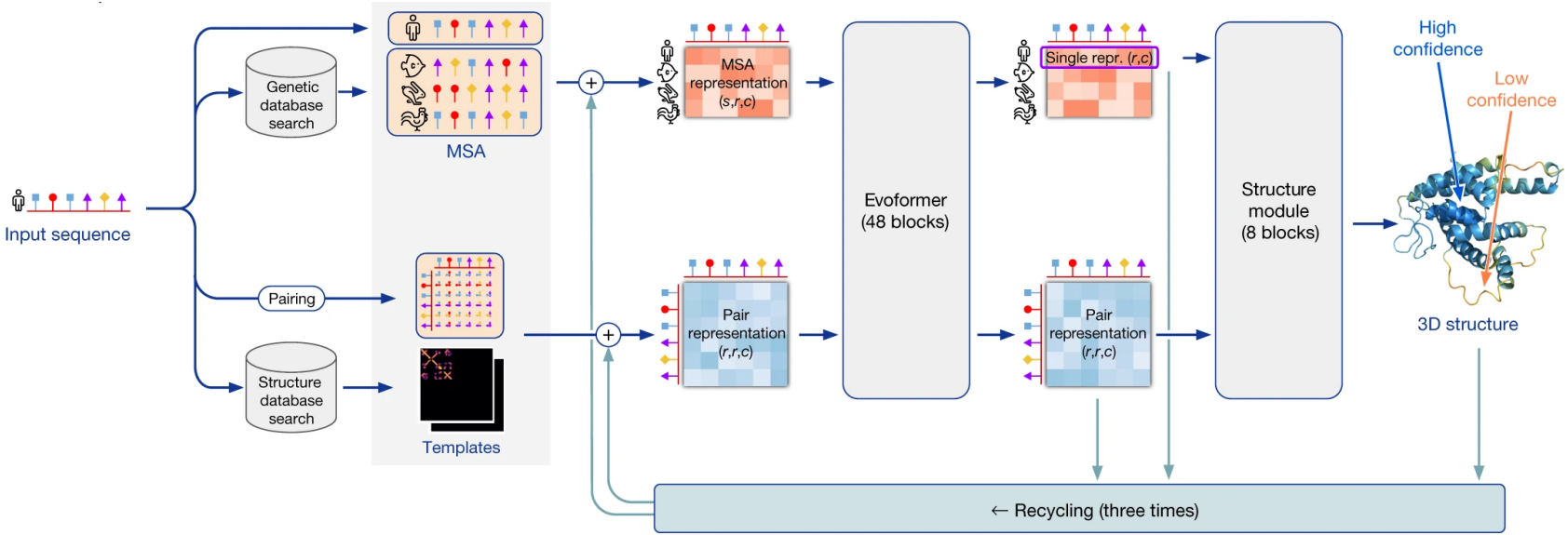
>EEKELAEELLRATVE
EAEYGGQALEAVAA
GDLATAEQLGEKAAE
LGKRAAEAGAKLKE

Analysing Structures with **Chimera** & **Rosetta**

- Physics-based tools
- Analysis of:
  - Contact surface area
  - Shape complementarity
  - Number of hydrogen bonds
  - Binding free energy (ddG)
  - Compactness



Structure Prediction with **AlphaFold2**



Acknowledgement: Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

AlphaFold2 Computational Challenges

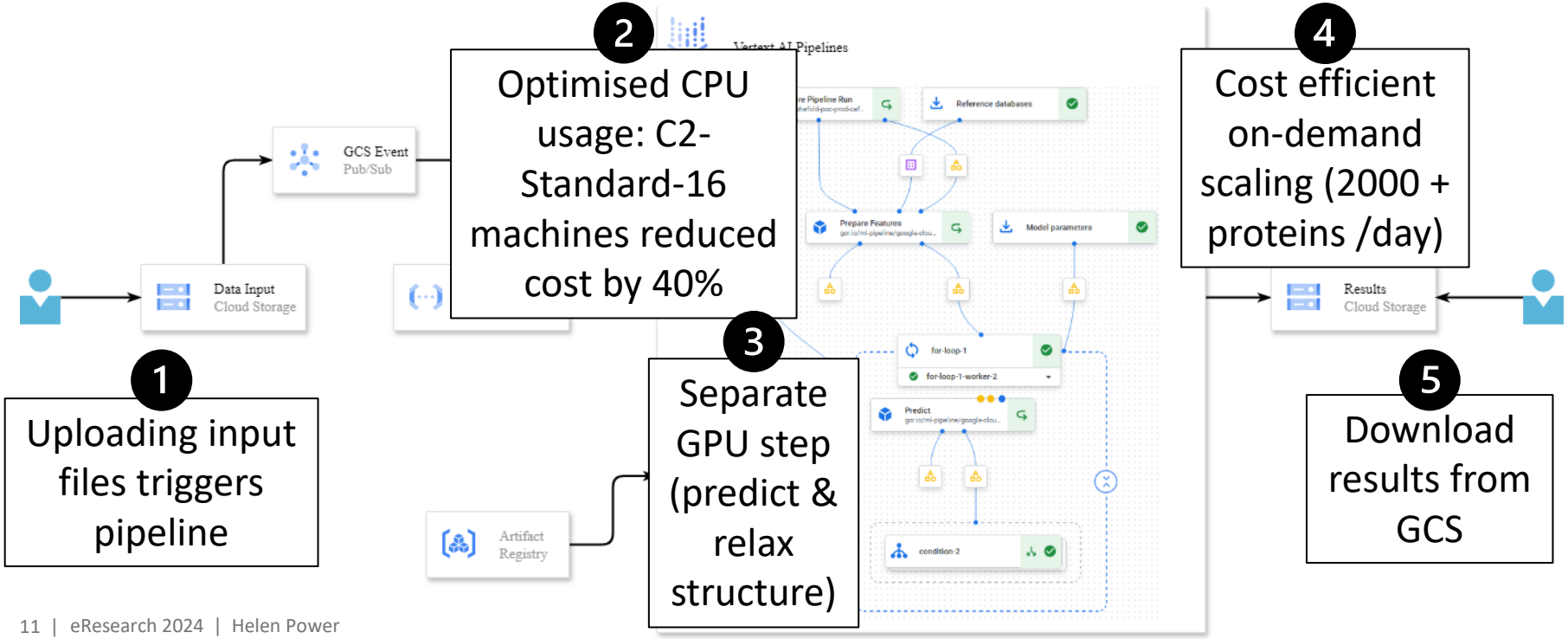
Long wall times	3hr for structures of length 400-450 amino acids
High Memory Requirements	80GB+ GPU memory for larger proteins
15,700 multimers to validate	15,700 x 3hr / H100 GPU = 1962.5 days!
Solution must be easy to implement, scalable, efficient and cost effective	

AlphaFold2 on Google's VertexAI

Joel Jasper

Cloud Platform Specialists

CSIRO scientists



1 Uploading input files triggers pipeline

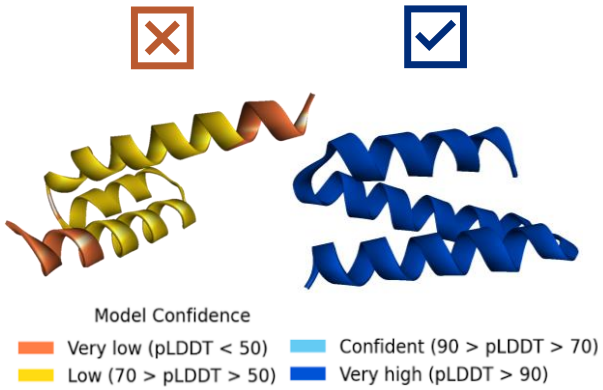
2 Optimised CPU usage: C2-Standard-16 machines reduced cost by 40%

3 Separate GPU step (predict & relax structure)

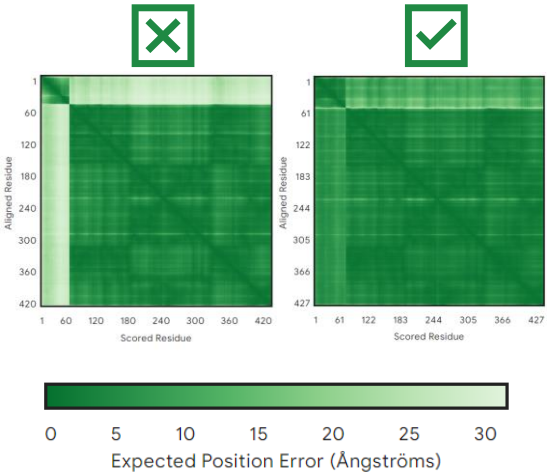
4 Cost efficient on-demand scaling (2000 + proteins /day)

5 Download results from GCS

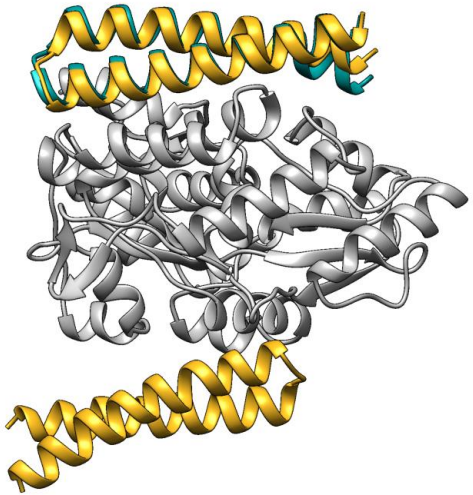
AlphaFold2 Results



✓High pLDDT



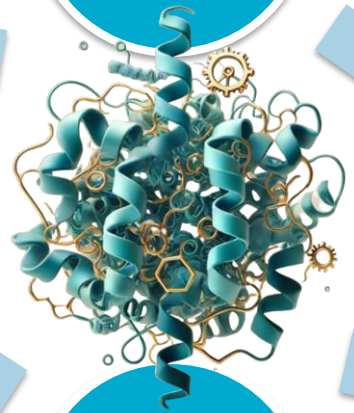
✓Low PAE



✓Low RMSD to RFDiffusion structure

Future Directions

Results will be used to inform future workflows.
E.g. which steps are unnecessary, filtering thresholds



Design ✓

Build ✓

Test

Learn

96 Binders Currently Being Validated Experimentally for Binding Affinity

Acknowledgments

Computing Infrastructure
Jasper Wang (Google)
Joel Ludbey
Ondrej Hlinka

ML/AI & In Silico Workflow
Andrew Warden
Marltan Wilson

Experimental Validation
Candice Jones
Tiffany Chau
Ullhas Kaarthi Moorthi