Applications of Structural Biology AI tools @ ASGC

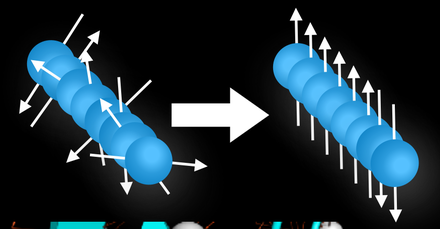
ASGC Oct. 30, 2024

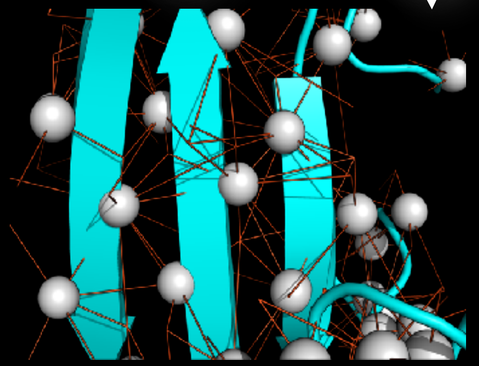
Kuen-Phon Wu | 吳昆峯

Institute of Biological Chemistry, Academia Sinica | 中央研究院生物化學所

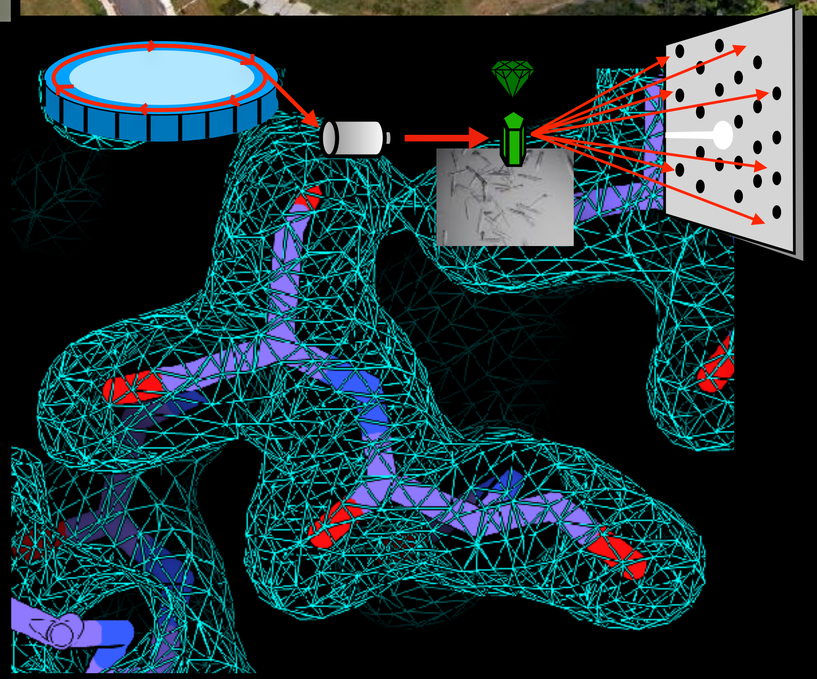
The major tools in structural biology

NMR

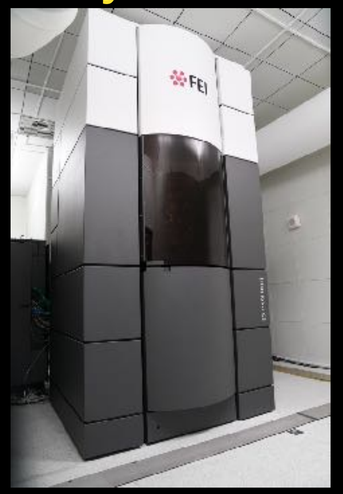


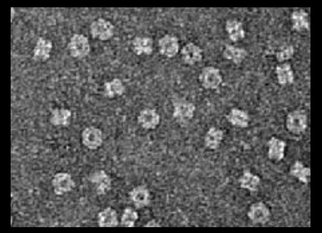


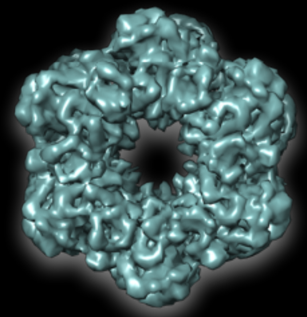
X-ray Crystallography



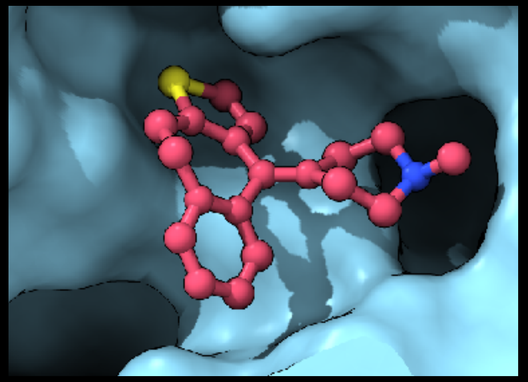
CryoEM



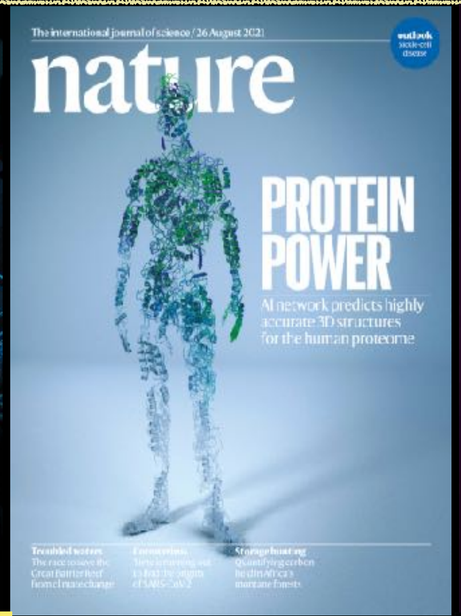


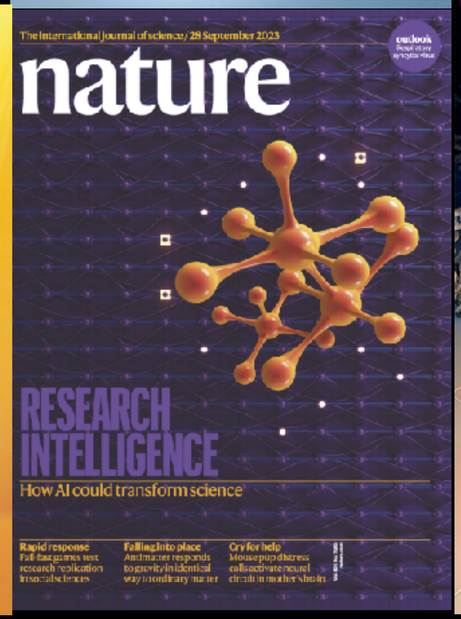


Computation



AI in Structural Biology is everywhere



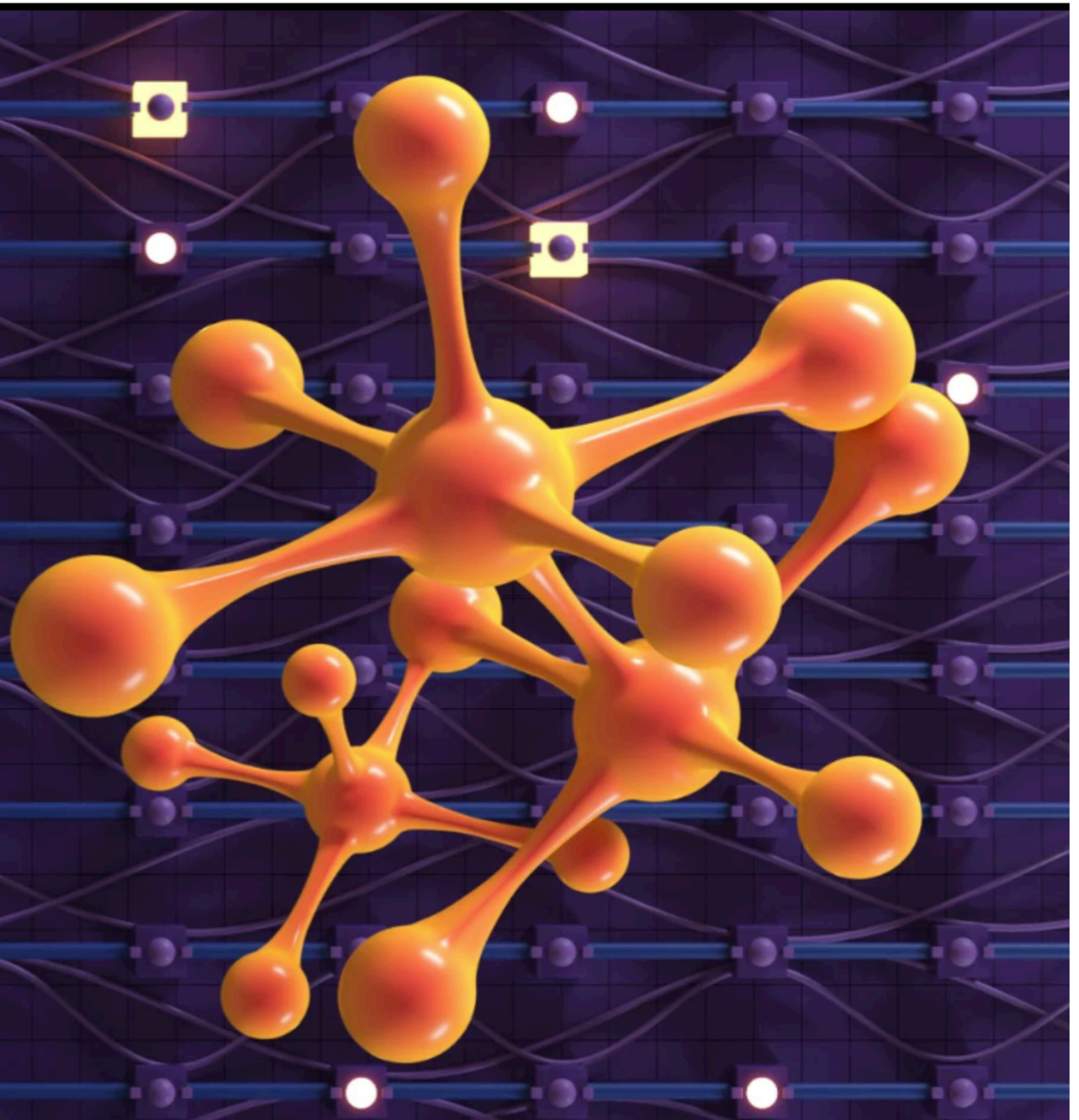


nature

EDITORIAL | 27 September 2023

AI will transform science — now researchers must tame it

Science and the new age of AI



Updated 10 October 2023

Credit: Carlo Cadenas

The Nobel Prize in Chemistry 2024

The Nobel Prize in Chemistry 2024 was divided, one half awarded to David Baker "for computational protein design", the other half jointly to Demis Hassabis and John M. Jumper "for protein structure prediction"

David Baker (1/2)

Demis Hassabis (1/4)

John M. Jumper (1/4)

Applications of AI in Structural Biology

Structure Prediction

Protein Design

Small to Large (>3000aa) protein

Monomer to Oligomer

Low-order to high-order complex

Docking

Protein-Ligand

Protein-Protein

Data Analysis

NMR Spectral Assignment

Automation in crystallography

CryoEM particle picking/model building

A new automation solution

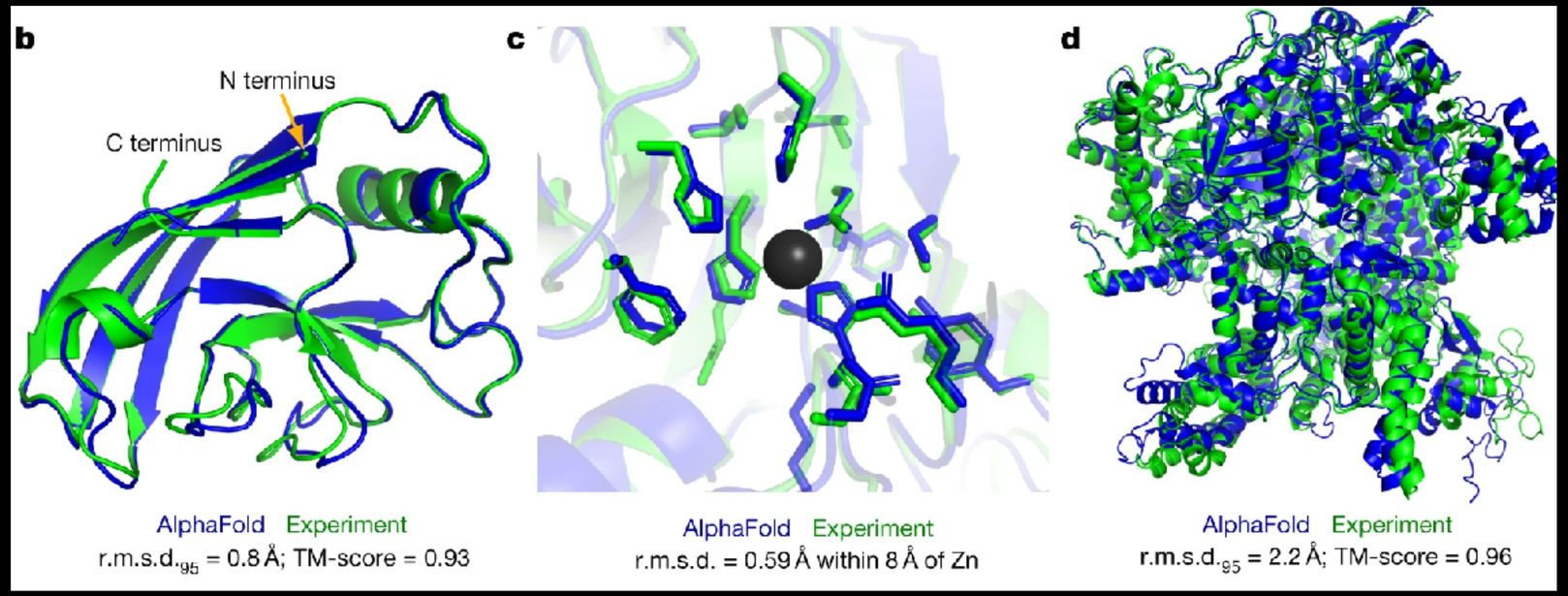
AlphaFold provides a new aspect of structure determination

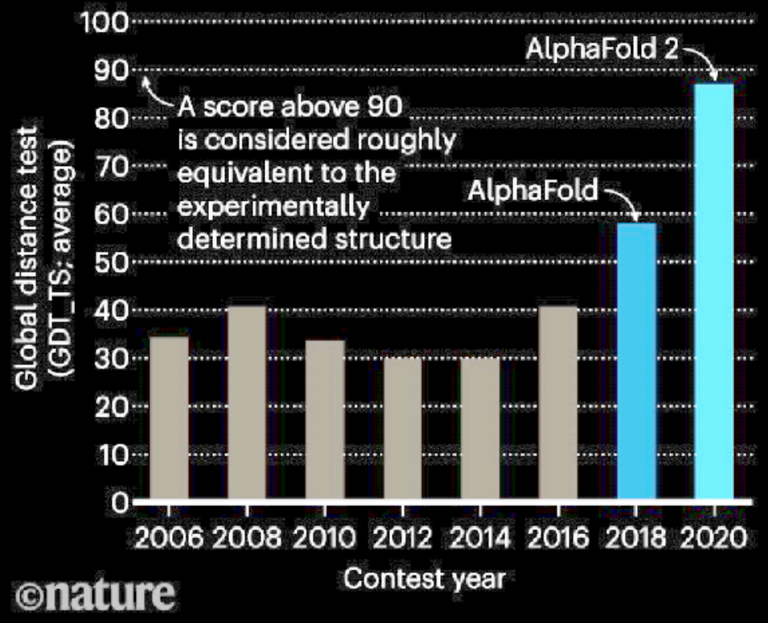
Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4]✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉





AlphaFold 3.0 was released in 2024

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Published: 08 May 2024

Accurate structure prediction of biomolecular interactions with AlphaFold 3

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, ... John M. Jumper + Show authors

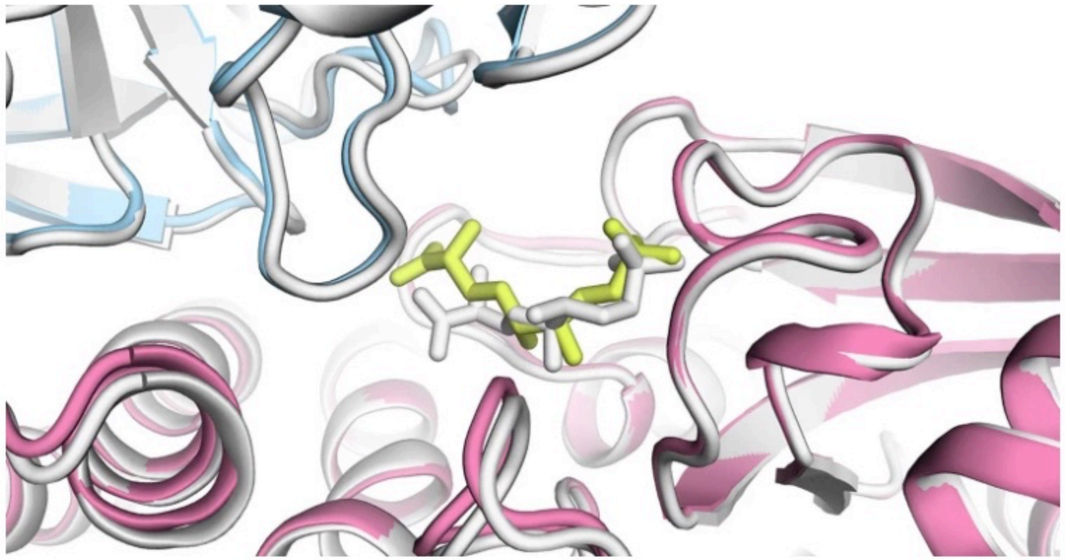
Nature (2024) | Cite this article

NEWS | 08 May 2024

Major AlphaFold upgrade offers boost for drug discovery

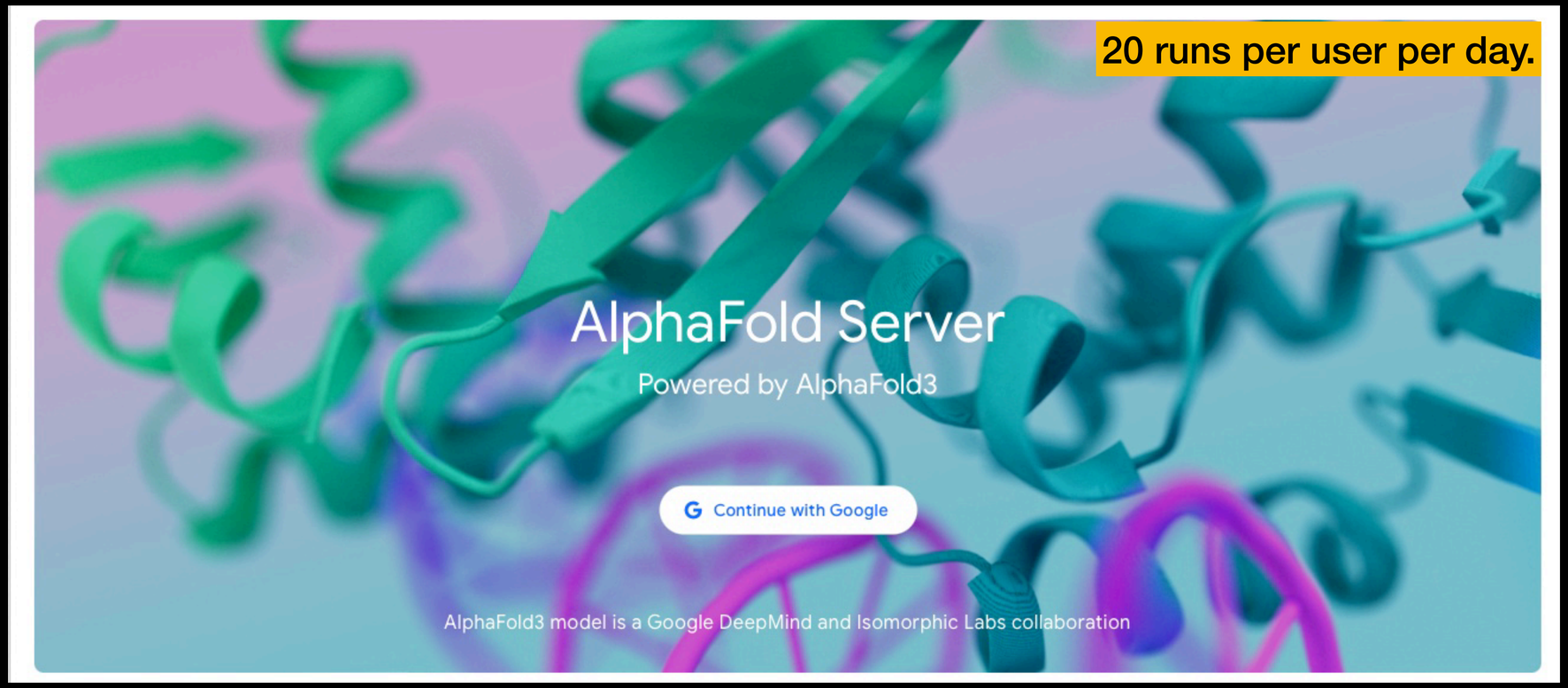
Latest version of the AI models how proteins interact with other molecules – but DeepMind restricts access to the tool.

By Ewen Callaway



An AlphaFold3 model of a bacterial enzyme bound to a chemical. Credit: Isomorphic Labs

AlphaFold 3.0 is a web-based service



From DeepMind:
Introducing AlphaFold 3, a new AI model developed by Google DeepMind and Isomorphic Labs. By accurately predicting the structure of proteins, DNA, RNA, ligands and more, and how they interact, we hope it will transform our understanding of the biological world and drug discovery.

ColabFold: AF2 core but faster than AF2

RTX3090 24GB V100 24GB A100 24GB A100 24GB

nature methods

BRIEF COMMUNICATION

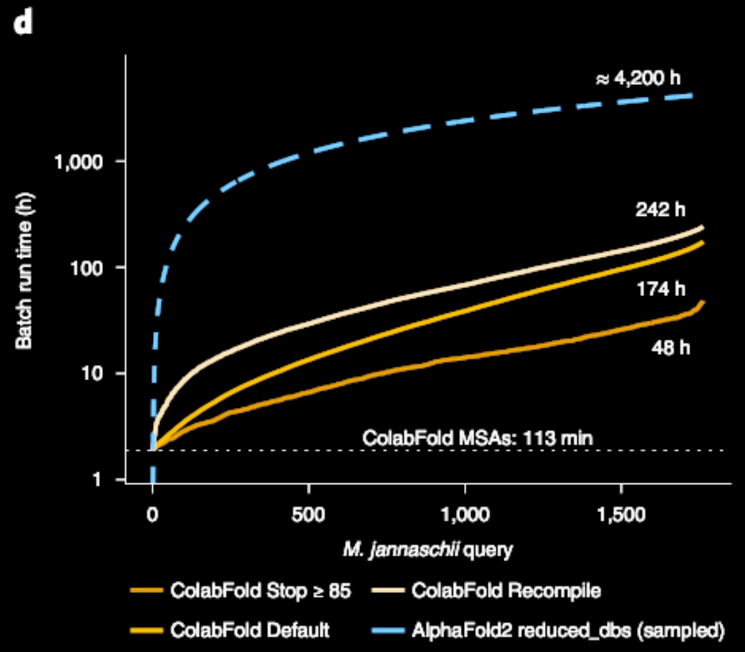
https://doi.org/10.1038/s41592-022-01488-1

Check for updates

OPEN

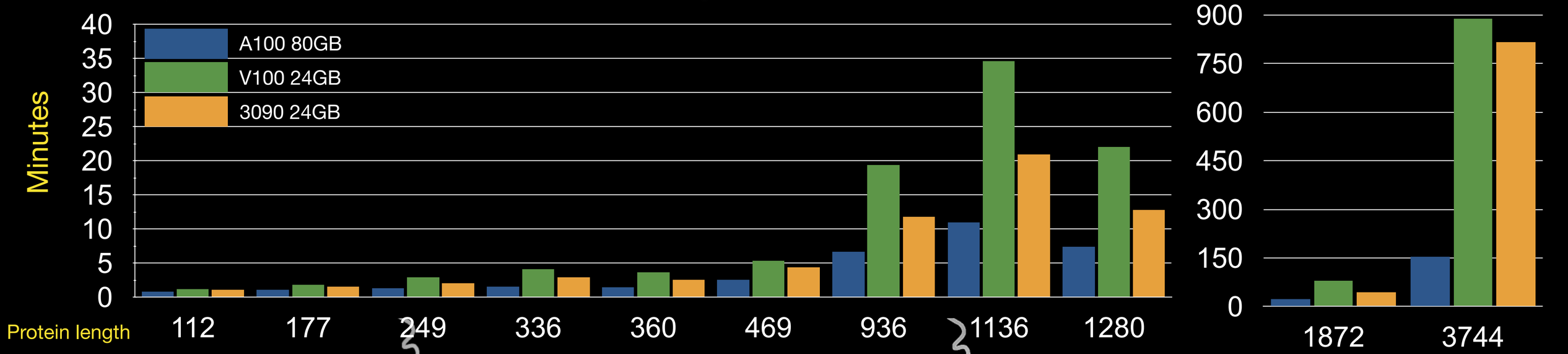
ColabFold: making protein folding accessible to all

Milot Mirdita[1,10] ✉, Konstantin Schütze[2], Yoshitaka Moriwaki[3,4], Lim Heo[5], Sergey Ovchinnikov[6,7,10] ✉ and Martin Steinegger[2,8,9,10] ✉

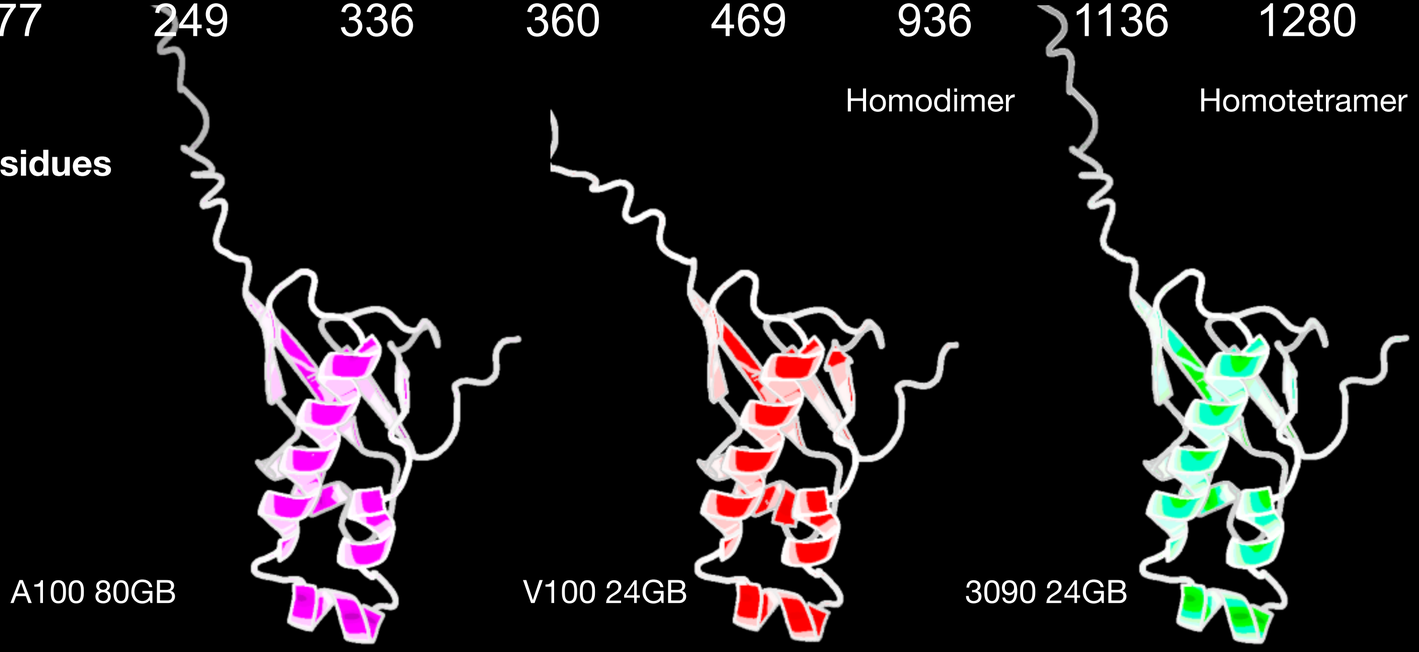


ColabFold benchmarks @ ASGC

https://kpwulab.com/2023/11/25/alphafold-benchmarks-on-asgc/

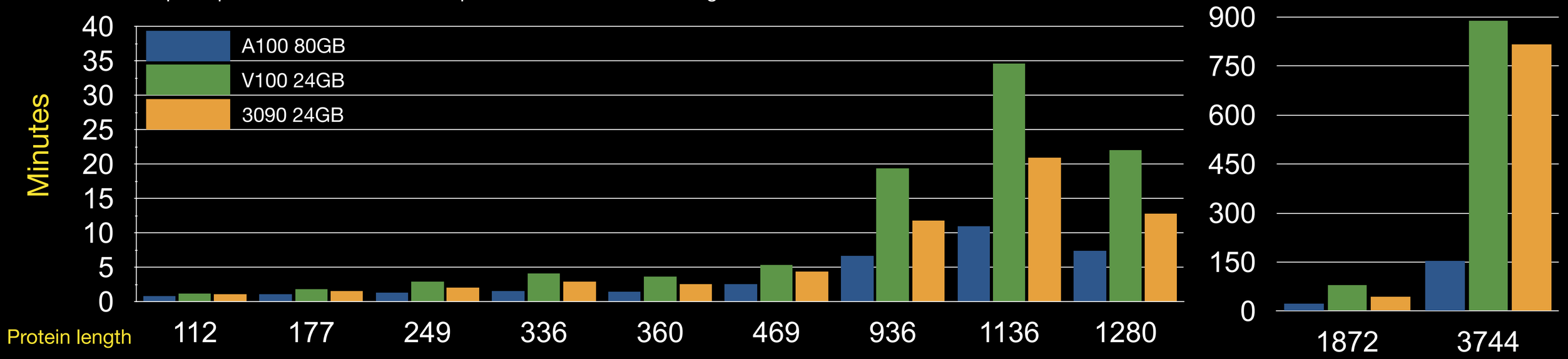


112 residues

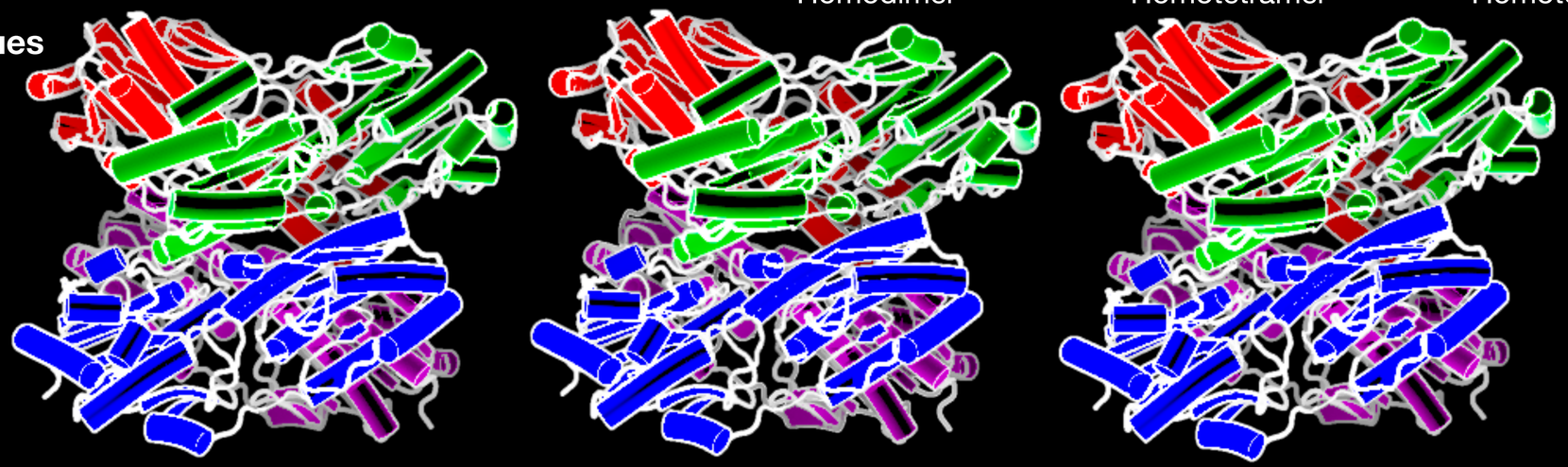


ColabFold benchmarks @ ASGC

https://kpwulab.com/2023/11/25/alphafold-benchmarks-on-asgc/



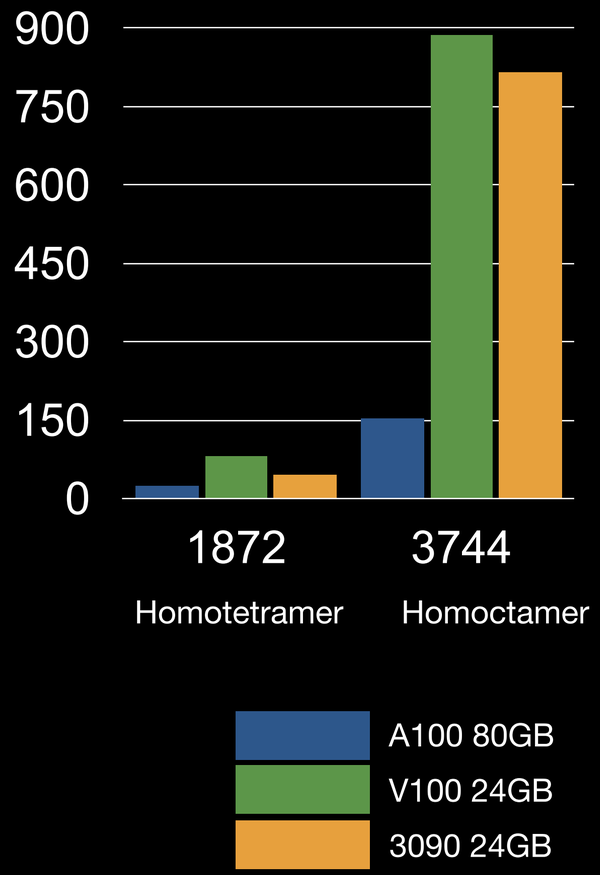
1280 residues

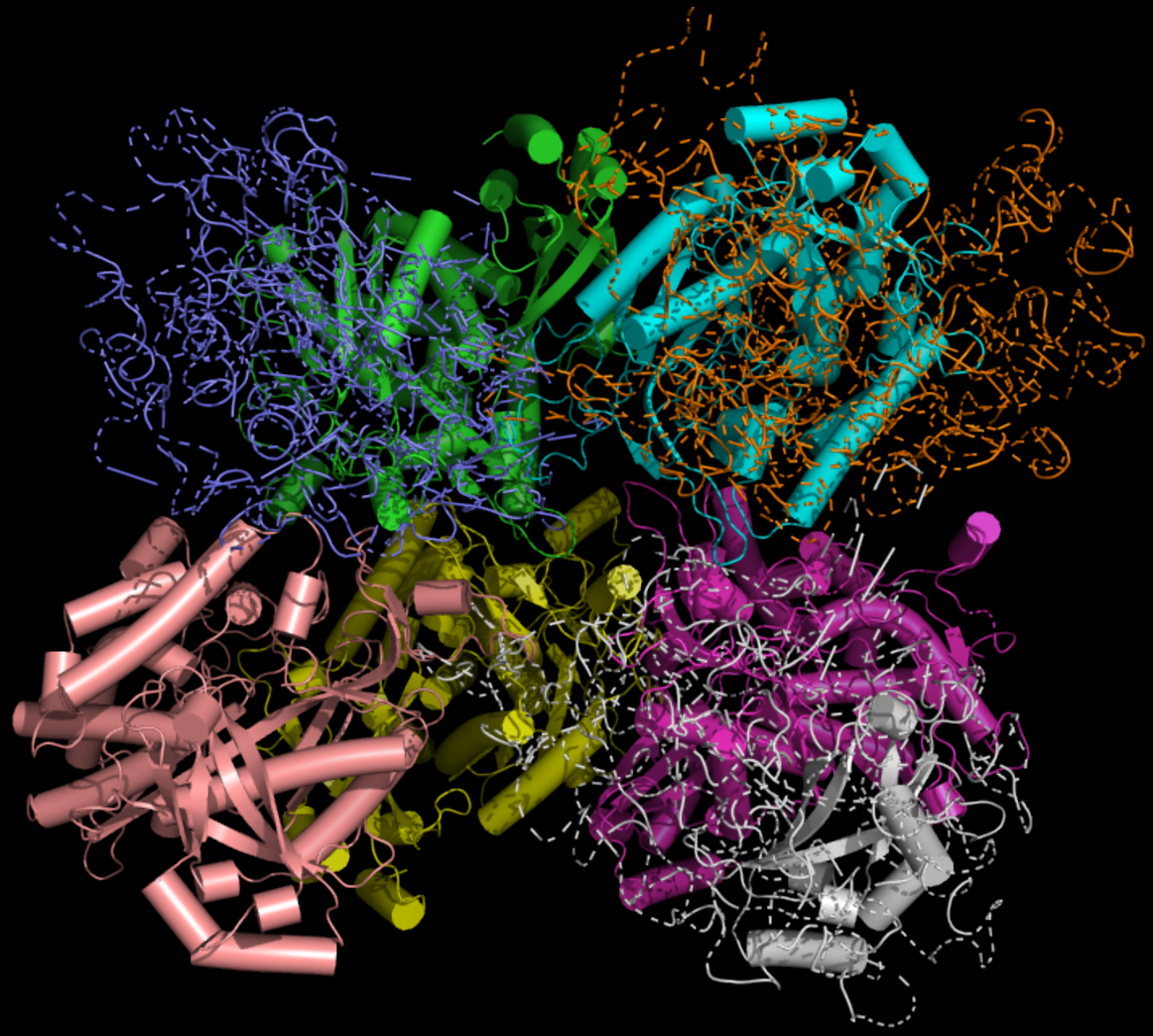


A100 80GB V100 24GB 3090 24GB

ColabFold benchmarks @ ASGC

24GB memory is not sufficient to handle >2500 residues in our experiences



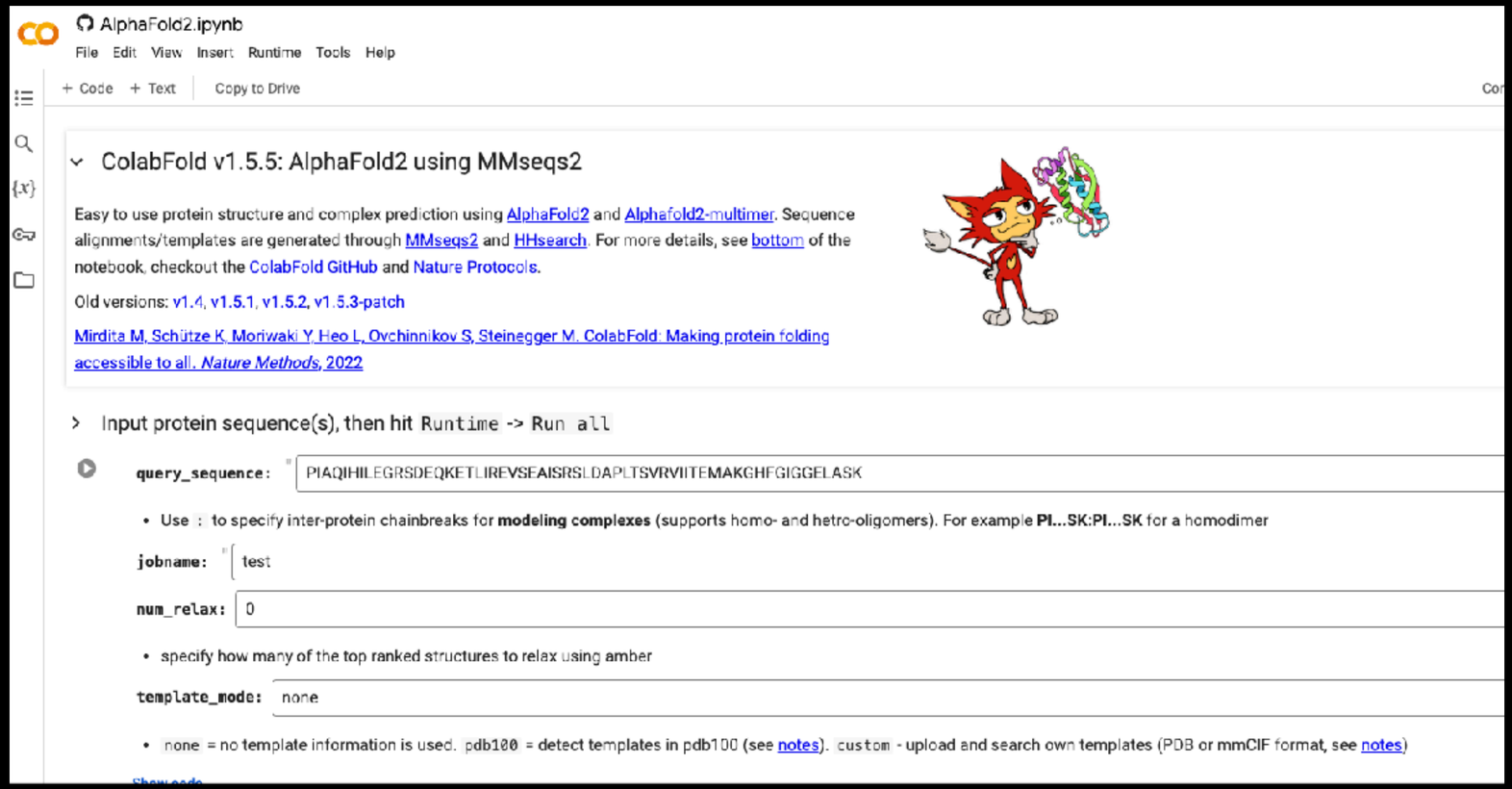


3090 24GB

Run ColabFold @ DiCOS

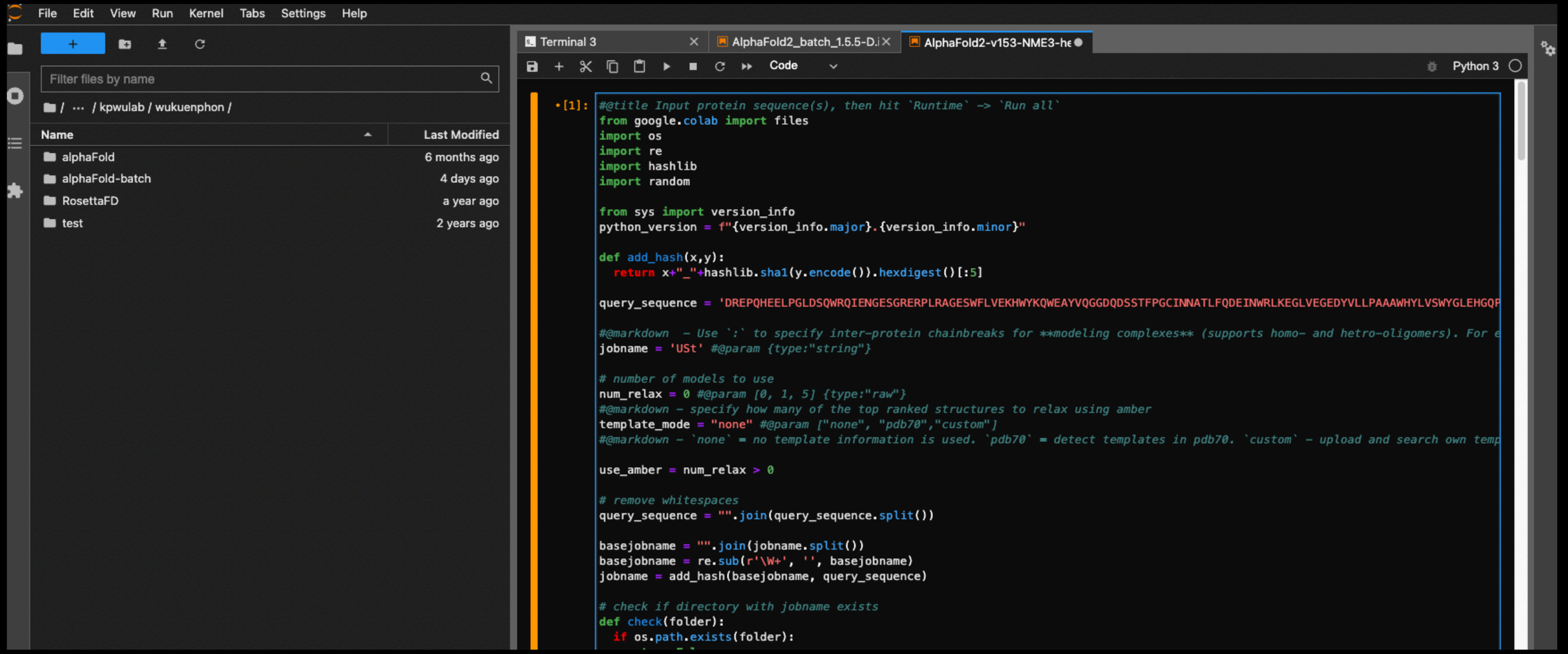
- ColabFold provides a Jupyter script to set up input parameters.
- Easy for new comers, no need to know much about Linux and HPC settings

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



Run ColabFold @ DiCOS

- ColabFold provides a Jupyter script to set up input parameters.
- Easy for new comers, no need to know much about Linux and HPC settings



Terminal 3 | AlphaFold2_batch_1.5.5-D.i

Code | Python 3

#ColabFold v1.5.5: AlphaFold2 w/ MMseqs2 BATCH

Easy to use AlphaFold2 protein structure (Jumper et al. 2021) and complex (Evans et al. 2021) prediction using multiple sequence alignments generated through MMseqs2. For details, refer to our manuscript:

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022

Usage

`input_dir` directory with only fasta files or MSAs stored in Google Drive. MSAs need to be A3M formatted and have an `.a3m` extention. For MSAs MMseqs2 will not be called.

`result_dir` results will be written to the result directory in Google Drive

Old versions: v1.4, v1.5.1, v1.5.2, v1.5.3-patch

For more details, see bottom of the notebook and checkout the ColabFold GitHub.

News

- 2023/07/31: The ColabFold MSA server is back to normal. It was using older DB (UniRef30 2202/PDB70 220313) from 27th ~8:30 AM CEST to 31st ~11:10 AM CEST.
- 2023/06/12: New databases! UniRef30 updated to 2023_02 and PDB to 230517. We now use PDB100 instead of PDB70 (see notes in the **main** notebook).
- 2023/06/12: We introduced a new default pairing strategy: Previously, for multimer predictions with more than 2 chains, we only pair if all sequences taxonomically match ("complete" pairing). The new default "greedy" strategy pairs any taxonomically matching subsets.

[1]:
```
#@title Mount google drive
from sys import version_info
python_version = f"{version_info.major}.{version_info.minor}"
```

[2]:
```
#@title Input protein sequence, then hit `Runtime` -> `Run all`

input_dir = 'B/seq' #@param {type:"string"}
result_dir = 'B/results' #@param {type:"string"}

# number of models to use
#@markdown ---
#@markdown ### Advanced settings
msa_mode = "MMseqs2 (UniRef+Environmental)" #@param ["MMseqs2 (UniRef+Environmental)", "MMseqs2 (UniRef only)","single_sequence",
num_models = 5 #@param [1,2,3,4,5] {type:"raw"}
num_recycles = 3 #@param [1,3,6,12,24,48] {type:"raw"}
stop_at_score = 100 #@param {type:"string"}
#@markdown - early stop computing models once score > threshold (avg. plddt for "structures" and ptmscore for "complexes")
use_custom_msa = False
num_relax = 1 #@param [0, 1, 5] {type:"raw"}
use_amber = num_relax > 0
relax_max_iterations = 200 #@param [0,200,2000] {type:"raw"}
use_templates = False #@param {type:"boolean"}
do_not_overwrite_results = True #@param {type:"boolean"}
zip_results = False #@param {type:"boolean"}
```

Batch mode runs more

- ColabFold provides a Jupyter script to set up input parameters.
- Easy for new comers, no need to know much about Linux and HPC settings

- ColabFold Batch (now v1.5.5) can run multiple sequences (one by one) in one click!

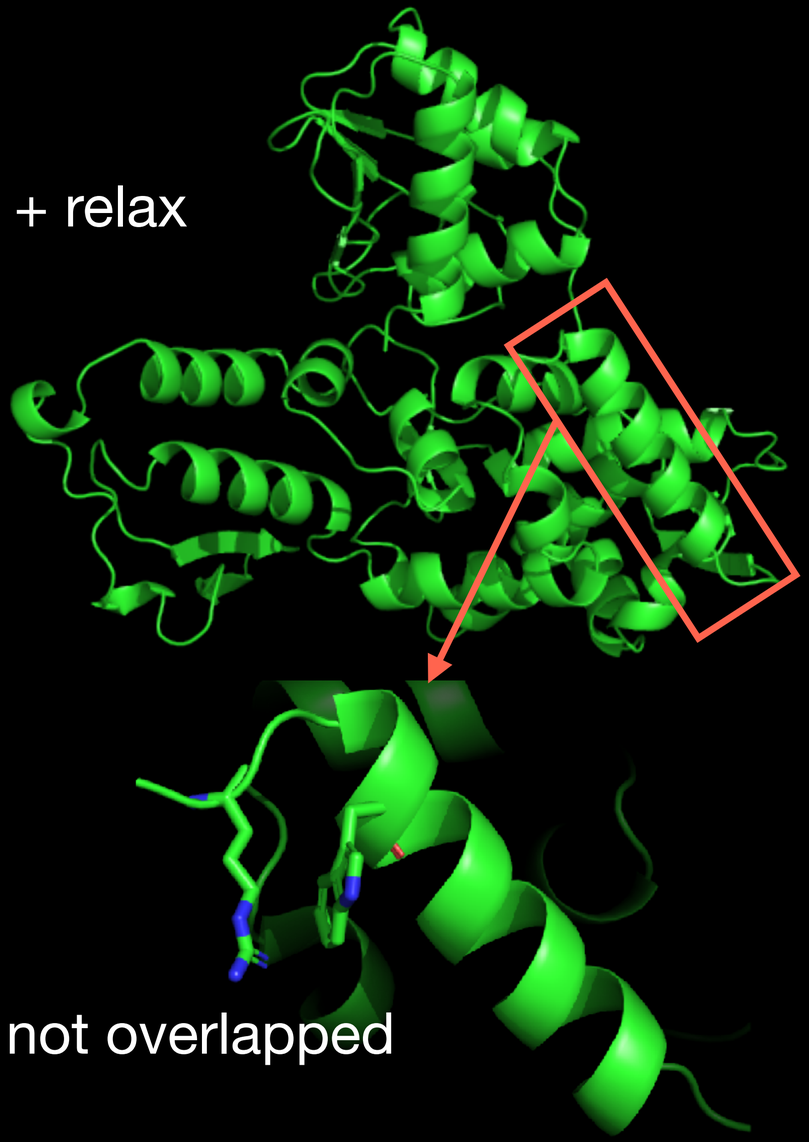
- 100 sequences (~500 residues) took about :
  - 10-12 hours without “relax”
  - 20-24 hours with “relax”

Run ColabFold @ DiCOS with “relax”

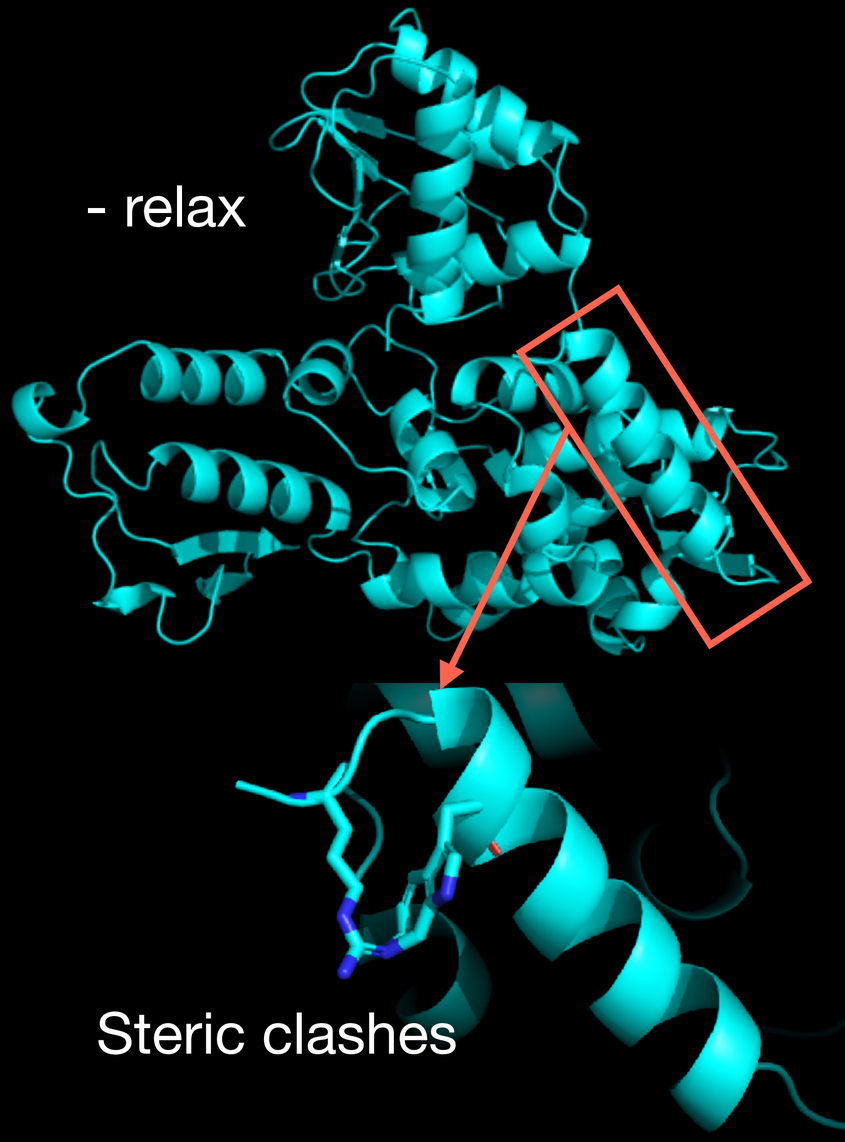
```
#@title Input protein sequence, then hit `Runtime` -> `Run all`

input_dir = 'B/seq' #@param {type:"string"}
result_dir = 'B/results' #@param {type:"string"}

# number of models to use
#@markdown ---
#@markdown ### Advanced settings
msa_mode = "MMseqs2 (UniRef+Environmental)" #@param ["MMseqs2 (
num_models = 5 #@param [1,2,3,4,5] {type:"raw"}
num_recycles = 3 #@param [1,3,6,12,24,48] {type:"raw"}
stop_at_score = 100 #@param {type:"string"}
#@markdown - early stop computing models once score > threshold
use_custom_msa = False
num_relax = 1 #@param [0, 1, 5] {type:"raw"}
use_amber = num_relax > 0
relax_max_iterations = 200 #@param [0,200,2000] {type:"raw"}
use_templates = False #@param {type:"boolean"}
do_not_overwrite_results = True #@param {type:"boolean"}
zip_results = False #@param {type:"boolean"}
```

- ColabFold and AlphaFold both calculate “5” models and rank the models with multiple metrics.
  - Numbered as 1, 2, 3, 4, 5
  - Model 1 is the best ranked structure (more reliable).
- It is known that AlphaFold2 (not 3) doesn’t place the sidechain well. It often create steric clashes. Hence low reliability for drug discovery.
- ColabFold integrates “Amber” to run “relax” process — using simple *relax function* in molecular dynamic simulation to place the sidechain in a more appropriate space. —> increase quality of predicted model.
- The ‘relax” calculation takes another 5-10 minutes (depending on protein size).

Amber relax improves the quality of predicted structure

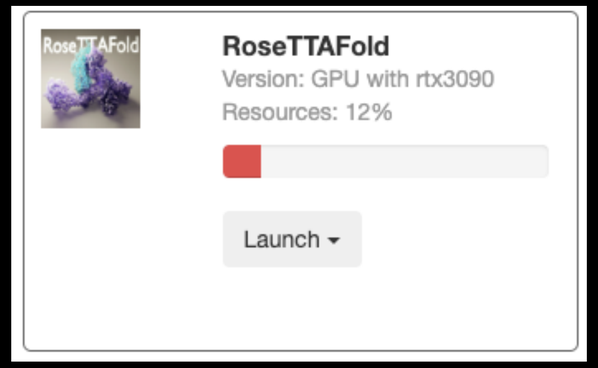


Lys-Trp are not overlapped

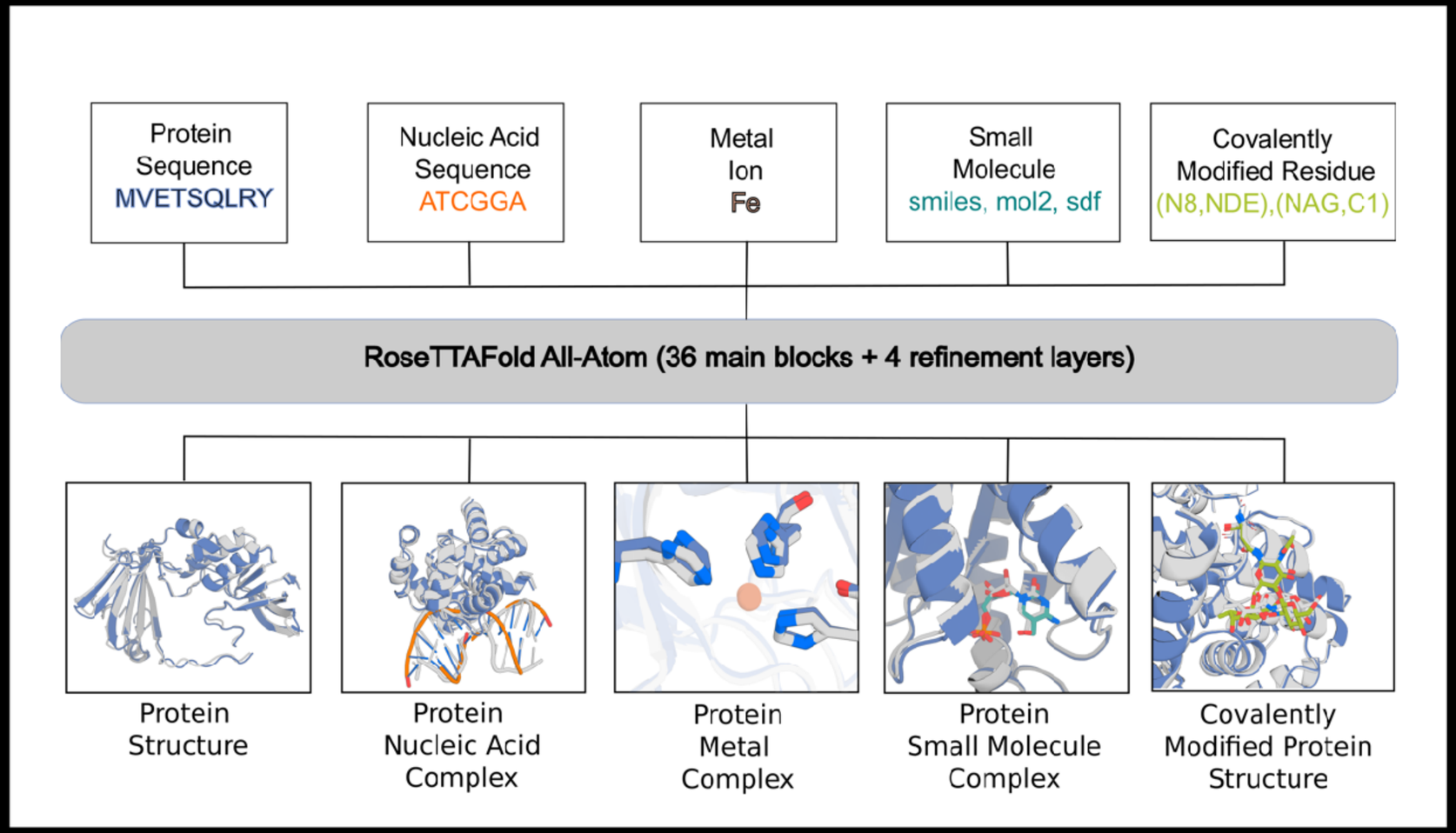


Steric clashes

RosettaFold and its succedent

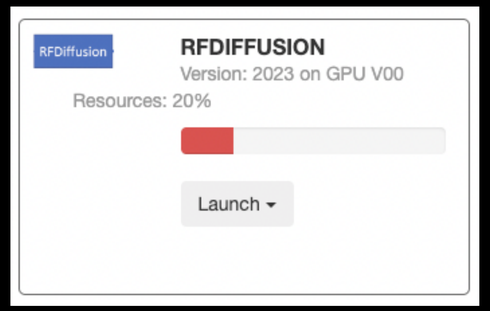


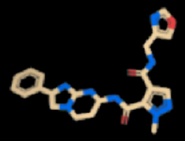
- Rosetta (original)
- RoseTTAFold (Deep learning)
- RoseTTAFold All-Atom
- RFDiffusion All-Atom

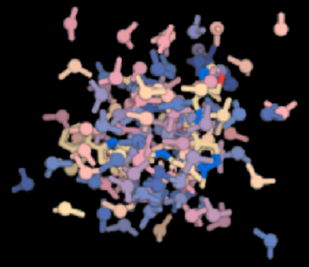


https://github.com/baker-laboratory

Generated AI and structural biology





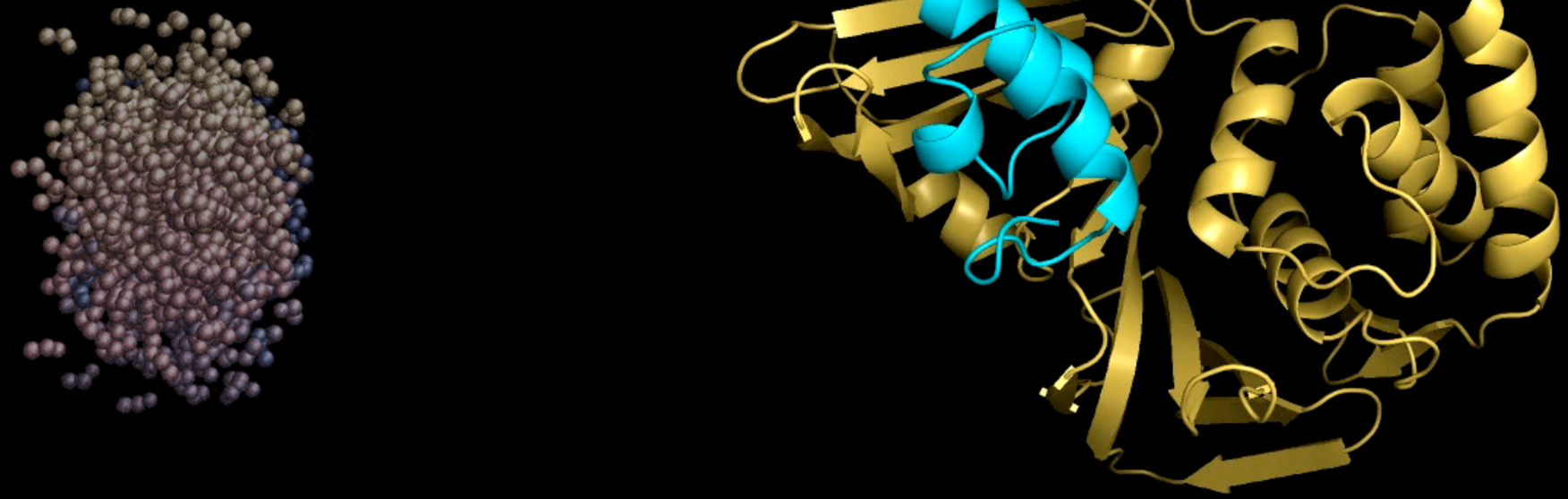


- RFDiffusion (RFD)
- RFDiffusion All-Atom (RFDAA)

https://github.com/baker-laboratory/rf_diffusion_all_atom

Described at Baker Lab:
“Inspired by AI image generators, RFdiffusion can be used to create novel protein structures in seconds. It sculpts clouds of disconnected atoms into novel protein backbones, yielding monomers, oligomers, binders, and more. Initially trained only on amino acids, RFdiffusion All-Atom now builds molecules using all of life’s building blocks, including DNA, RNA, ions, and small molecules.”

Generated AI and structural biology



https://github.com/baker-laboratory/rf_diffusion_all_atom

BindCrarft—An integrated pipepline for protein design

BindCraft: one-shot design of functional protein binders

Martin Pacesa[1]‡, Lennart Nickel[1]‡, Joseph Schmidt[1], Ekaterina Pyatova[1], Christian Schellhaas[1], Lucas Kissling[2], Ana Alcaraz-Serna[3], Yehlin Cho[4], Kourosh H. Ghamary[5], Laura Vinué[5], Brahm J. Yachnin[5], Andrew M. Wollacott[5], Stephen Buckley[1], Sandrine Georgeon[1], Casper A. Goverde[1], Georgios N. Hatzopoulos[6], Pierre Gönczy[6], Yannick D. Muller[3], Gerald Schwank[2], Sergey Ovchinnikov[4]*, Bruno E. Correia[1]*

[1] Laboratory of Protein Design and Immunoengineering, École Polytechnique Fédérale de Lausanne and Swiss Institute of Bioinformatics; Lausanne, Switzerland

[2] Institute of Pharmacology and Toxicology, University of Zurich, Zurich, Switzerland

[3] Division of Immunology and Allergy, Lausanne University Hospital and University of Lausanne; Lausanne, Switzerland

[4] Massachusetts Institute of Technology; Cambridge, MA, USA

[5] Visterra Inc., Waltham, MA, USA

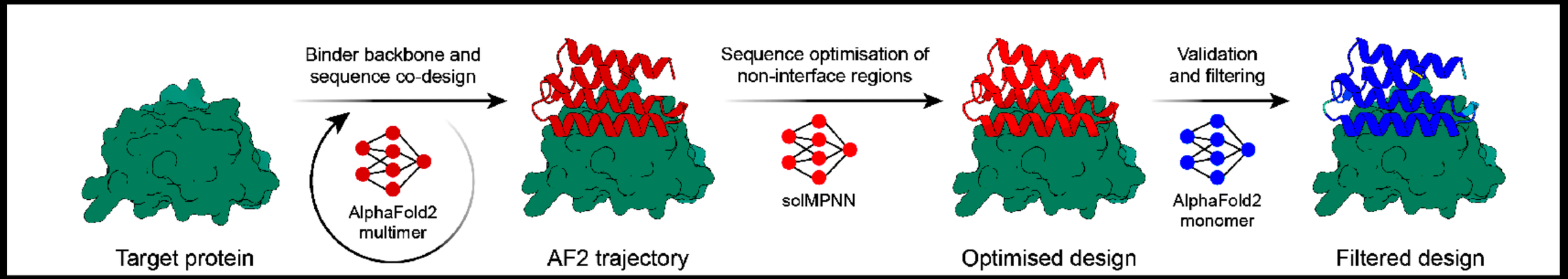
[6] Swiss Institute for Experimental Cancer Research (ISREC), School of Life Sciences, Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland.

CSH Cold Spring Harbor Laboratory

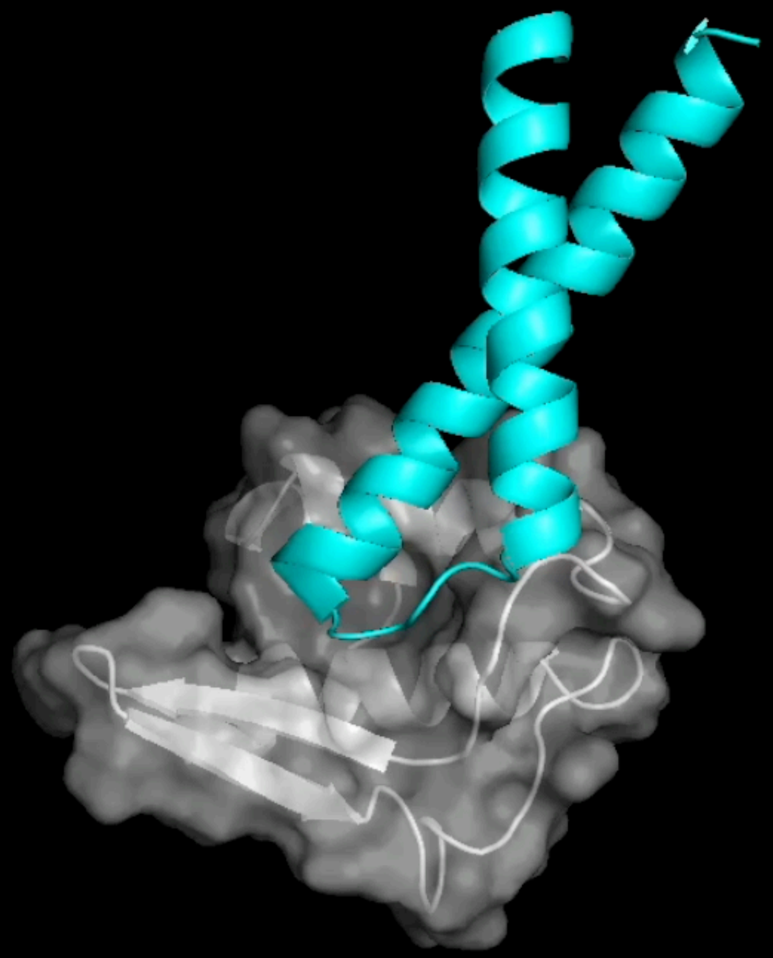
bioRχiv THE PREPRINT SERVER FOR BIOLOGY

https://www.biorxiv.org/content/10.1101/2024.09.30.615802v1
Posted online October 1, 2024

https://github.com/martinpacesa/BindCraft



BindCrarft—An integrated pipepline for protein design

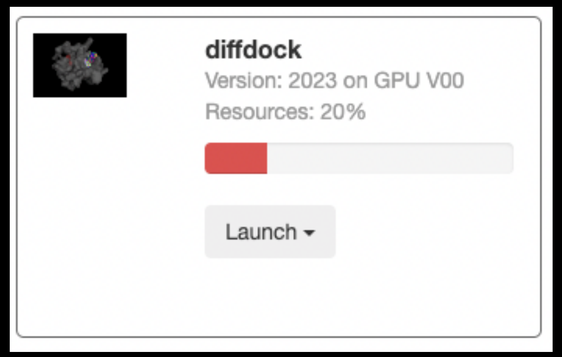


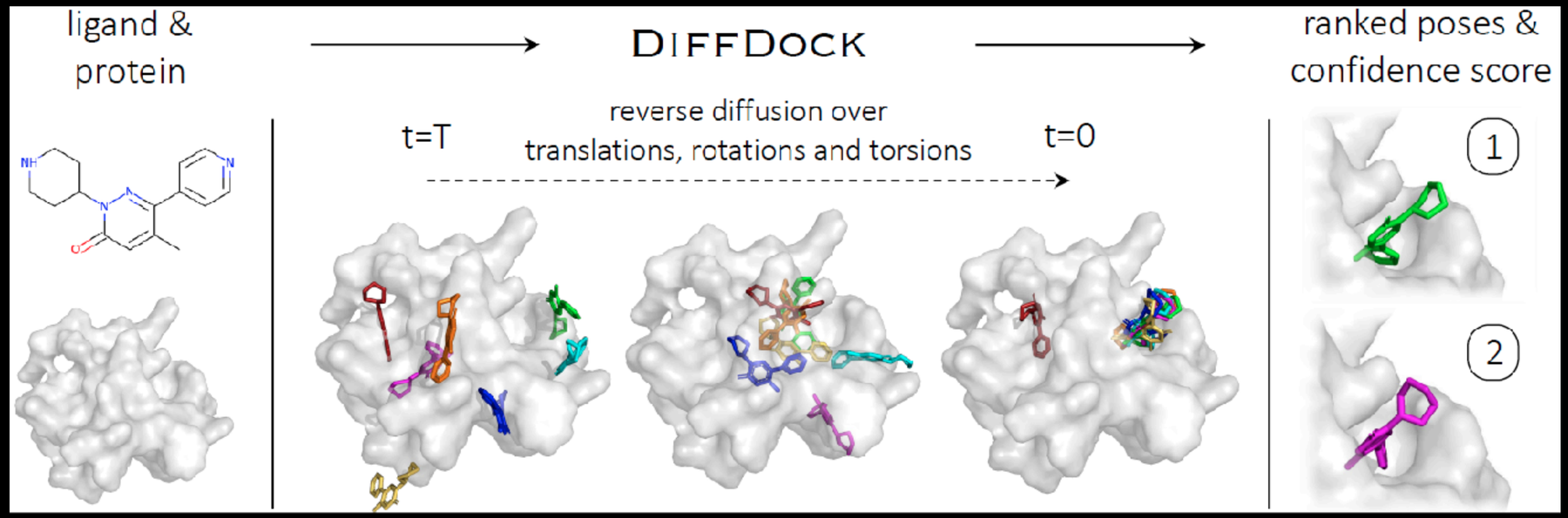
In our own test

- Generate 100 “acceptable” model
- Nvidia Rtx 4070 (16GB): 19-25 hours
- Nvidia Rtx 3090 (24GB): 20-35 hours
- Nvidia Rtx 2080 (12GB): 92-120 hours

- A100 (40GB or higher) is suggested by the author.

A diffusion model to protein-ligand docking



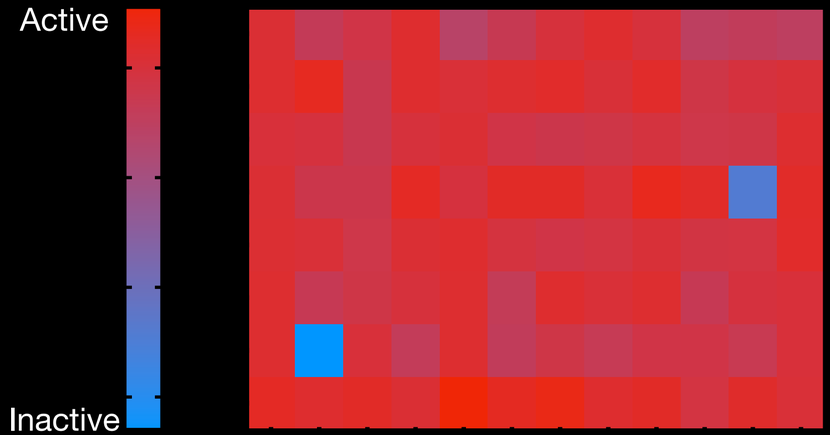


https://github.com/gcorso/DiffDock

~4000 candidates

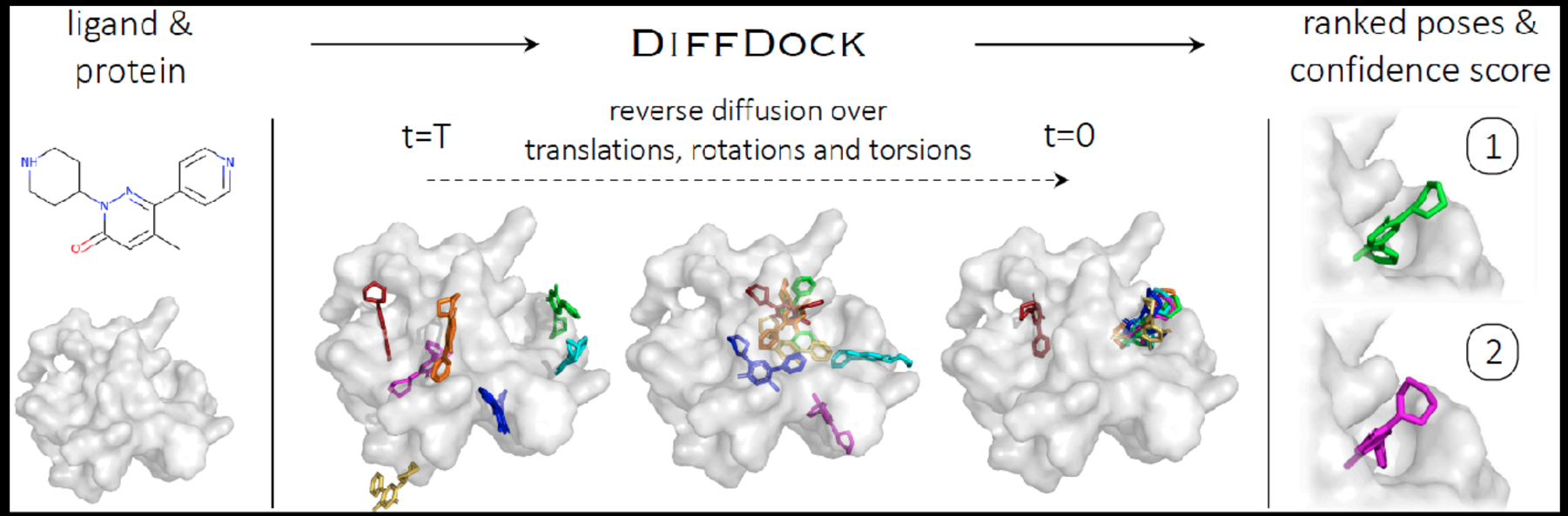
96 for in vitro assays

Two hits: IC_{50} at 1~10 μM



Unpublished data

A diffusion model to protein-ligand docking

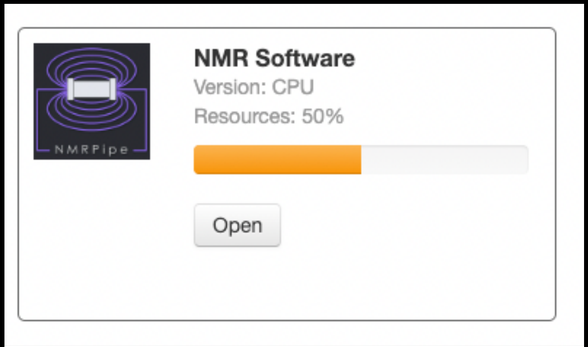


https://github.com/gcorso/DiffDock

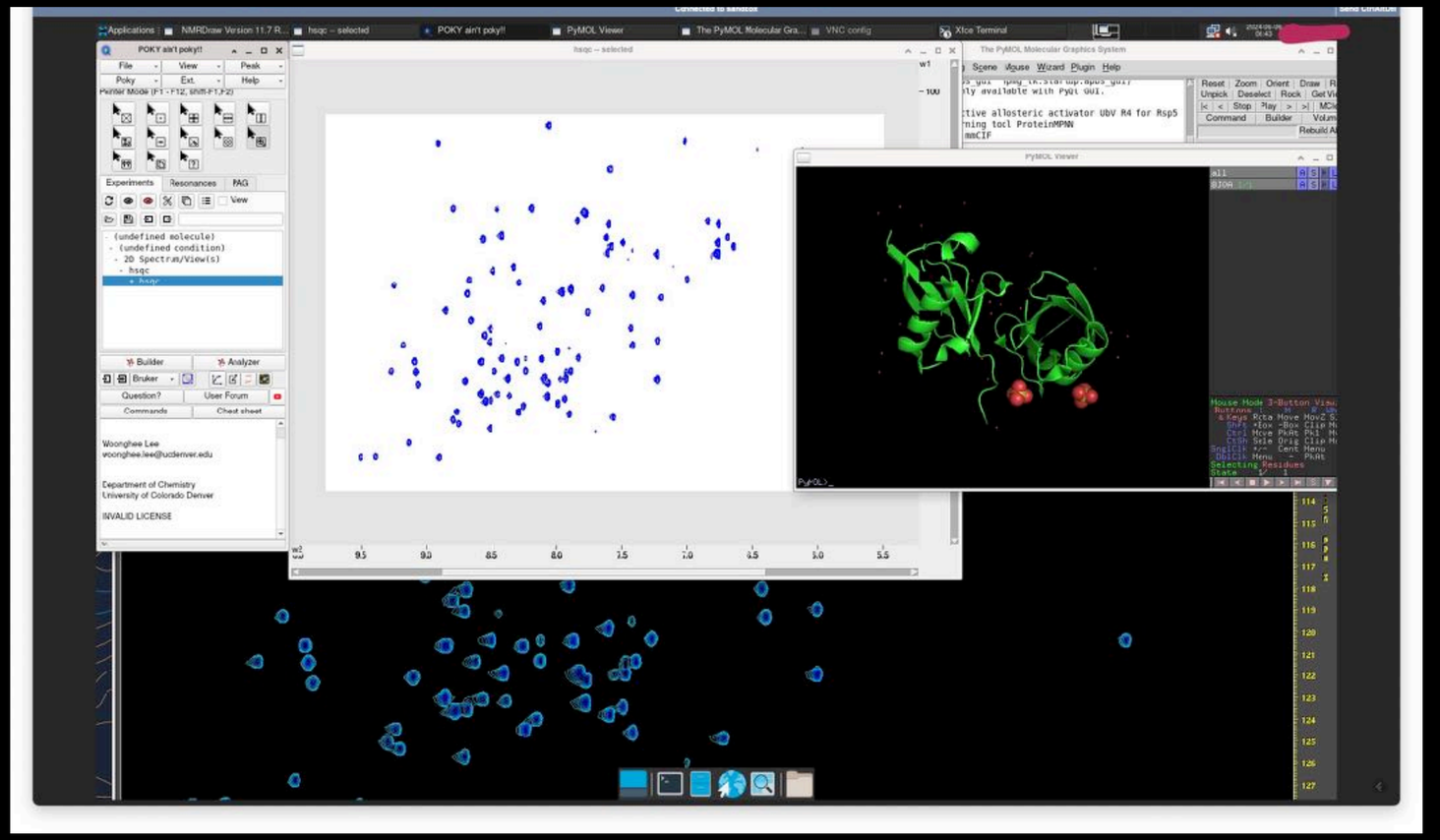
For a job docking ~4000 compounds, it took:

- 3-4 weeks by 24-core, 48-thread CPUs (Intel Xeon Silver 4116 2.1 GHz)
- 1-2 days by ONE Nvidia Rtx 4090 24GB GPU card (no parallel setting)

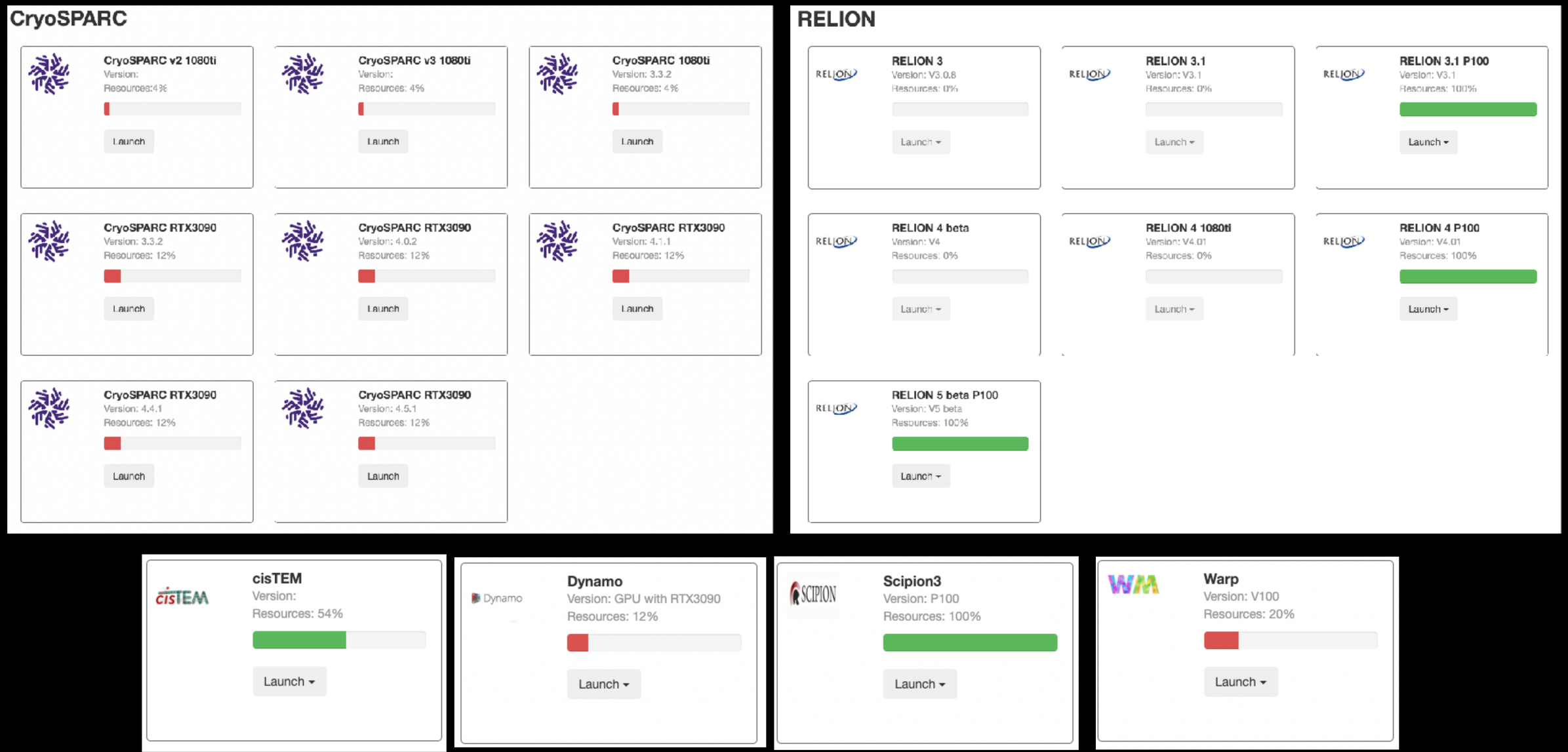
Everything in the cloud —NMR application



- DiCOS system has a NMR app
- Please go to https://dicos.grid.sinica.edu.tw/resources to apply an account (not limited to AS users)
- So far, NMRpipe, masterHI, Poky, NMRFam-Sparky, Xplor-NIH are installed (need to debug a bit more)



Everything in the cloud —cryoEM application



Suite for AI-based protein-related software

Structure Prediction	Protein Design	Ligands	Data Analysis
AlphaFold-Full DB	RFDiffusion	DiffDock	CryoSPARC
AlphaFold/ColabFold	RFD-all atom	LigandMPNN	RELION
RosettaFold	RosettaFold-PPI	SBDD	Warp
RosettaFold-NA	ProteinMPNN		cisTEM
	BindCraft		NMR-app bundle
			CryoDRGN

Thanks ASGC for the stable and affordable resource

ASGC研究型服務收費 (ASGC Pricing)

本表各系統每二十四小時使用量之單價(以 CPUCore-Day 或 GPUBoard-Day 計算)。

(Last Updated: Jun. 2024. Pricing model will be reviewed and revised annually.)

Computing Pricing 計算服務			
Model	**Computing Usage * NTD$**	**國內非學術單位使用者**	**國外學術單位使用者**
CPU-FDR5	CPUCore-Day * NTD$0.24		
CPU-EDR1	CPUCore-Day * NTD$1.2		
CPU-HDR1	CPUCore-Day * NTD$0		
GPU-1080Ti	GPUBoard-Day * NTD$1	加計 50%	加計 50%
GPU-P100	GPUBoard-Day * NTD$8		
GPU-RTX3090	GPUBoard-Day * NTD$40		
GPU-V100	GPUBoard-Day * NTD$35		
GPU-A100	GPUBoard-Day * NTD$120		
Storage Pricing (TB) 儲存與擷取服務			
Storage	**Unit Cost(NTD$)**	**國內非學術單位使用者**	**國外學術單位使用者**
YEAR	NTD$1000 * per TB-year	加計 50%	加計 50%
DAY	NTD$3 * per TB-day		
Data Transmission 資料傳輸			
目前未納入計費			
Customization Service 進階服務			
依據所需人時計算。額外需開發之軟體、系統或使用介面等，將另按工時計費(每 168 man-hr 為 NTS 100,000)