Distributed prediction of protein structures with alphafold

Glen Hocky

ParslFest 2021

Department of Chemistry, NYU

(in collaboration with Parallel Works, Inc)

Acknowledgements

Hocky Research Group, Fall 2021

Funding and Support

New York University
Faculty of Arts and Sciences

Department of Energy
SBIR DE-SC0019695

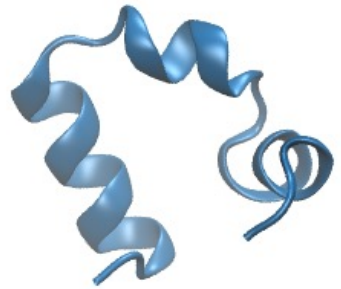
Collaborators

Parallel Works, Inc
Shenglong Wang, NYU HPC

BIG PICTURE

Protein folding problem

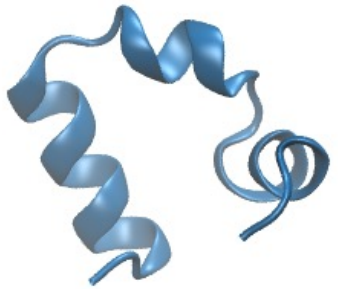
?



Structure prediction problem

DAYAQWLKDGGPSSGRPPPS

?



PAST WORK

Protein folding with Swift (2007-2009), on BlueGene/P

Towards petascale *ab initio* protein folding through parallel scripting

Glen Hocky[1], Michael Wilde[2,3*], Joe DeBartolo[4,5], Mihael Hategan[2], Ian Foster[2,3,6], Tobin R. Sosnick[2,4,5*], Karl F. Freed[1,2,7*]
[1]Department of Chemistry, University of Chicago
[2]Computation Institute, University of Chicago & Argonne National Laboratory, USA
[3]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne IL, USA
[4]Department of Biochemistry and Molecular Biology, University of Chicago
[5]Institute for Biophysical Dynamics, University of Chicago
[6]Department of Computer Science, University of Chicago, IL, USA
[7]James Franck Institute, University of Chicago
*wilde@mcs.anl.gov, trsosnic@uchicago.edu, freed@uchicago.edu

Abstract

Petascale computers allow scientists and engineers not only to address old problems better, but also to consider new methods and new problems. We report here on work that both applies new methods and tackles new problems in the area of structural biology. The project combines an efficient protein structure prediction algorithm implemented in the Open Protein System (OOPS) system with the Swift parallel scripting system to enable the rapid and flexible composition of OOPS components into parallel program

As OOPS becomes more accurate and efficient, a number of related computational challenges emerge in our desire to tackle proteins of increasing size because current prediction methods have limited accuracy even for proteins on the order of 100 residues when homology-based information is minimal. To predict the structures of larger and multi-domain proteins, statistical sampling becomes a limiting factor, and thus we require significantly more computing resources.

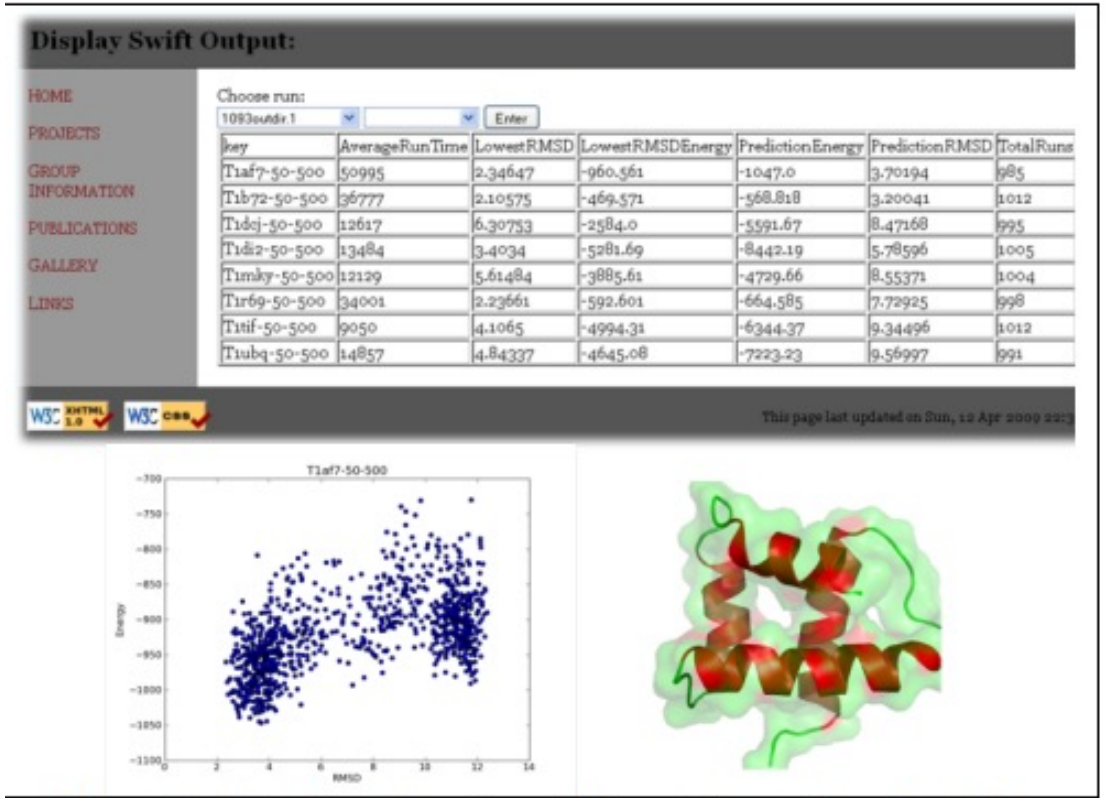


Figure 6 – Results of running eight proteins on 2 racks (8192 CPUS) on Argonne's BG/P, Intrepid. Below are results from this investigation for T1af7. On the left is a scatter plot showing the correlation between our statistical energy potential and accuracy of the protein structures for the 985 simulations that ran to completion. On the right is an image showing the lowest RMSD structure. This table, plot and image were all automatically generated by our scripting mechanism, and the table is presented by a simple CGI script at our web site [2].

Parallel scripting for applications at the petascale and beyond

Authors: Michael Wilde, Ian Foster, Kamil Iskra, Pete Beckman, Zhao Zhang, Allan Espinosa, Mihael Hategan, Ben Clifford, Ioan Raicu

Publication date: 2009/11/13

ALPHAFOLD ADVANCE

NEWS | 30 November 2020

‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures

Google’s deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

AlphaFold Is The Most Important Achievement In AI—Ever

Rob Toews Contributor

AI

I write about the big picture of artificial intelligence.

Follow

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]✉

RESEARCH

RESEARCH ARTICLE

PROTEIN FOLDING

Accurate prediction of protein structures and interactions using a three-track neural network

Minkyung Baek[1,2], Frank DiMaio[1,2], Ivan Anishchenko[1,2], Justas Dauparas[1,2], Sergey Ovchinnikov[3,4], Gyu Rie Lee[1,2], Jue Wang[1,2], Qian Cong[5,6], Lisa N. Kinch[7], R. Dustin Schaeffer[6], Claudia Millán[8], Hahnbeom Park[1,2], Carson Adams[1,2], Caleb R. Glassman[9,10,11], Andy DeGiovanni[12], Jose H. Pereira[12], Andria V. Rodrigues[12], Alberdina A. van Dijk[13], Ana C. Ebrecht[13], Diederik J. Opperman[14], Theo Sagmeister[15], Christoph Buhlheller[15,16], Tea Pavkov-Keller[15,17], Manoj K. Rathinaswamy[18], Udit Dalwadi[19], Calvin K. Yip[19], John E. Burke[18], K. Christopher Garcia[9,10,11,20], Nick V. Grishin[6,7,21], Paul D. Adams[12,22], Randy J. Read[8], David Baker[1,2,23]*

ALPHAFOLD ADVANCE

NEWS | 30 November 2020

'It will change ever
DeepMind's AI ma
gigantic leap in so
protein structures

Google's deep-learning program for determini
stands to transform biology, say scientists.

AlphaFold Is Th
Important
Achievement In
Ever

Rob Toews Contributor

AI

*I write about the big picture of artific
intelligence.*

deepmind / alphafold Public

Wat

Code | Issues 60 | Pull requests 15 | Actions | Projects | Security | Insights

main | 1 branch | 2 tags | Go to file | Add file | Code

Augustin-Zidek and **Copybara-Service** Accept any ordering given by ListDir i... ... 1d43aaf on Sep 10 | 25 commits

alphafold	Skip obsolete PDB templates that don't have a replacement.	2 months ago
docker	Fix a few typos.	3 months ago
imgs	Initial release of AlphaFold.	3 months ago
notebooks	Fix TensorFlow versions in AlphaFold Colab notebook.	2 months ago
scripts	Remove a redundant space.	3 months ago
.dockerignore	Collapse hh-suite install steps into single layer.	3 months ago
CONTRIBUTING.md	Initial release of AlphaFold.	3 months ago
LICENSE	Initial release of AlphaFold.	3 months ago
README.md	Update the bibtex citation with the issue number and pages	2 months ago
requirements.txt	Switch to Tensorflow CPU-only. GPU not needed for data pipeline.	3 months ago
run_alphafold.py	Use pLDDT in the B-factor column of the output PDBs.	2 months ago
run_alphafold_test.py	Accept any ordering given by ListDir in the assert.	2 months ago
setup.py	Use tensorflow-cpu in setup.py as well.	2 months ago

Article

ructure prediction

Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4],
asuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4],
[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4],
[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4],
Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1],
r[2,3], Michalina Pacholska[1], Tamas Berghammer[1],
r[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1],
[1,4]✉

ructures and
neural network

s Dauparas[1,2], Sergey Ovchinnikov[3,4],
ustin Schaeffer[6], Claudia Millán[8],
Andy DeGiovanni[12], Jose H. Pereira[12],
[13], Diederik J. Opperman[14],
Theo Sagmeister[15], Christoph Buhlheller[15,16], Tea Pavkov-Keller[15,17], Manoj K. Rathinaswamy[18],
Udit Dalwadi[19], Calvin K. Yip[19], John E. Burke[18], K. Christopher Garcia[9,10,11,20], Nick V. Grishin[6,7,21],
Paul D. Adams[12,22], Randy J. Read[8], David Baker[1,2,23]*

WHY ARE WE EXCITED?

Before 2021:

Protein

In PDB?

yes

Do simulation

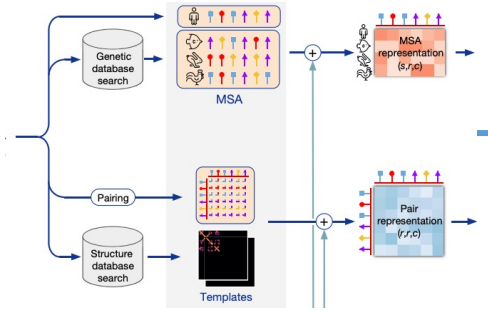
no

Project over!

Mid 2021+:

Protein

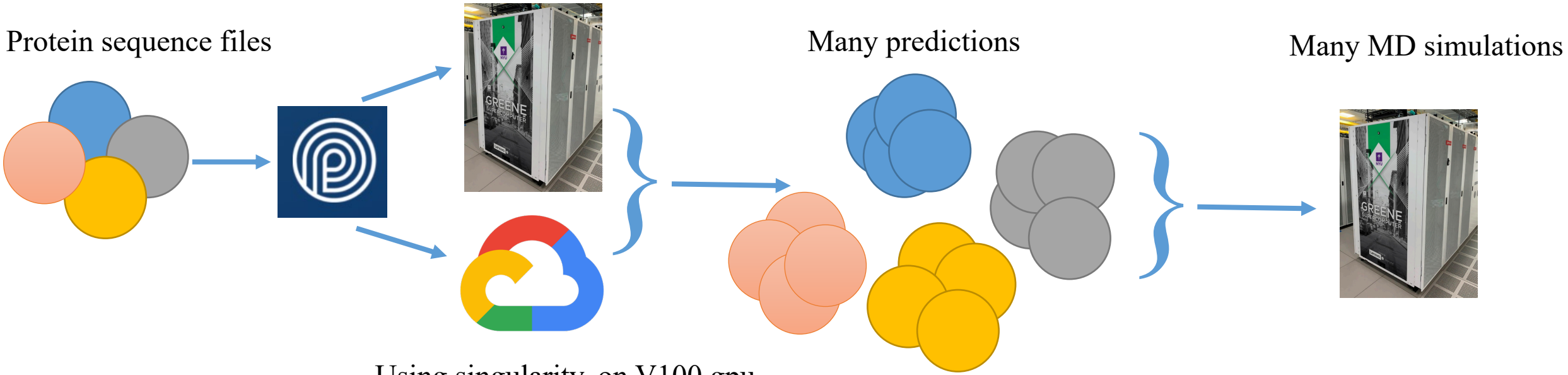
Predict structure



CHALLENGES

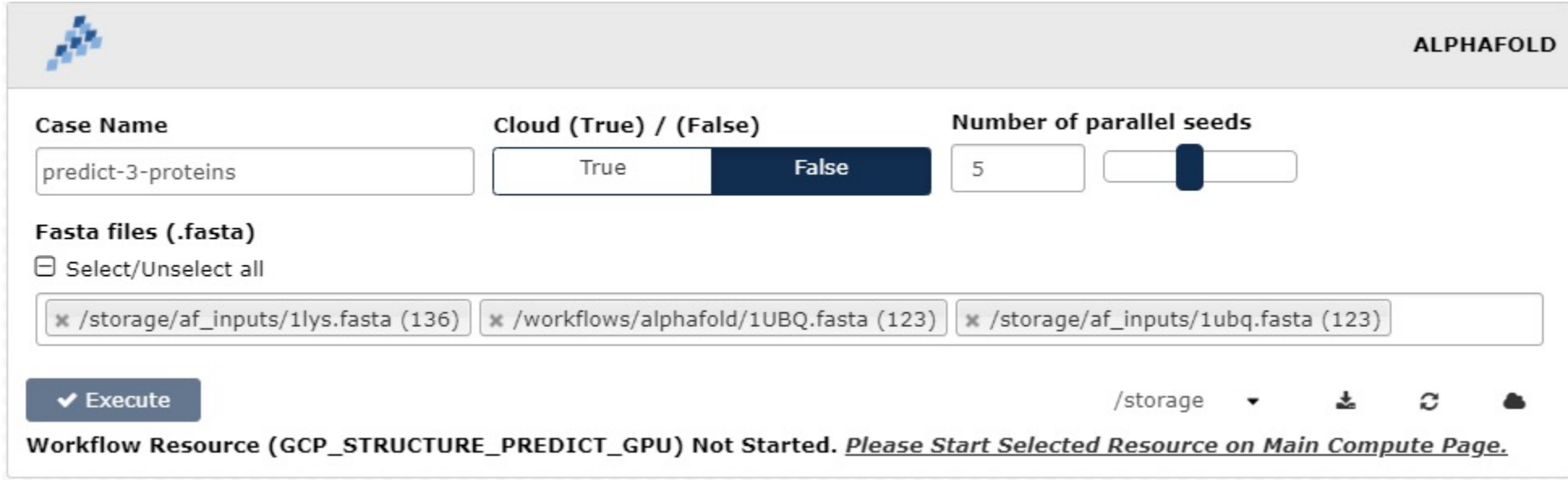
- Alphafold and similar software are computationally expensive
- Hard to install and require a lot of data
- Not high-throughput “out of the box”

Parallel workflow w/ parsl:



Using singularity, on V100 gpu
Currently not full AF

PW In and Out




```
95  send_files = [Path(f) for f in glob.glob("*.sh")+glob.glob("*.py")+glob.glob("*.tcl")]
96
97  for run_file_name in run_file_names:
98
99      run_file = Path(run_file_name)
100     for i in range(1,n_seeds+1):
101         r = run_alphafold(
102             runscript=runscript,
103             random_seed=i,
104             inputs = [run_file]+send_files,
105             outputs=[out_dir,Path("af.stdout"),Path("af.stderr")])
106         runs.append(r)
107
108 print("Running",len(runs),"alphafold executions...")
109 [r.result() for r in runs]
110
111 gen_table(pwargs.outcsv,pwargs.outhtml)
```

PW In and Out

Parametric Variables

- label
- model
- seed
- is_relaxed
- rg
- plddt

Cases

label: 1yrf … 1lys (1yrf, 1ubq, 1lys)

model: 1 … 5 (1, 2, 3, 4, 5)

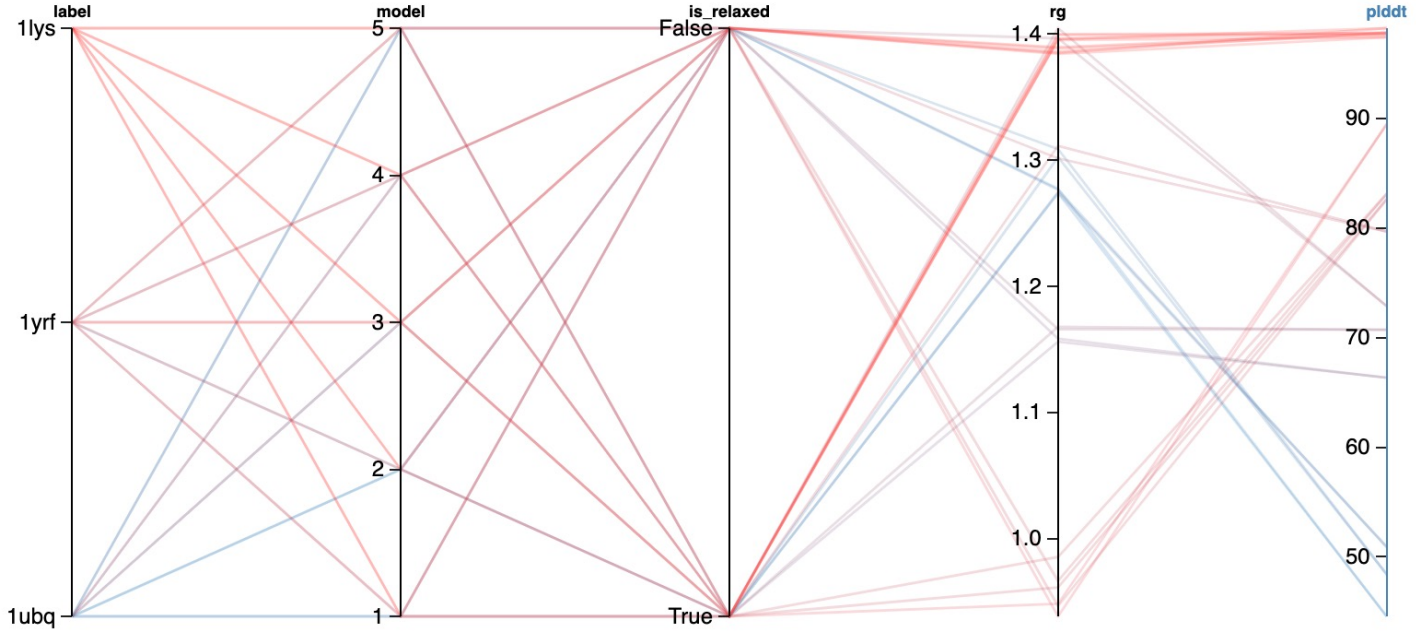
seed: 1

is_relaxed: True … False (True, False)

Reset Selection | Exclude Selection | Zoom to Selection | Save Selection to File

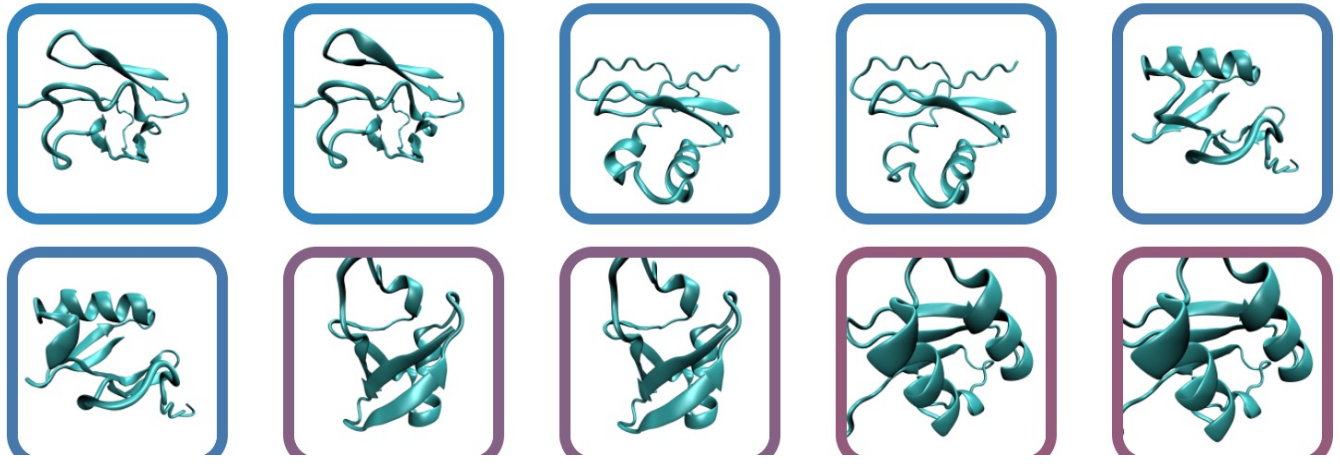
Thumbnail Size: L M S

Open in New Window



Sort by: plddt

Metric: model



FINAL THOUGHTS

- Use of containers and platform specific bash scripts make parallel script more platform agnostic (but is there a place for app definitions per site?)
- Use of PW platform enables use of alphafold and visualization of results for total novice, using local or cloud resources (but see ongoing work on google colab notebooks for non-high-throughput cases)
- Farming out predictions may be good case for funcX, but composing with other functions like generating MD inputs or viz files still good case for parsl