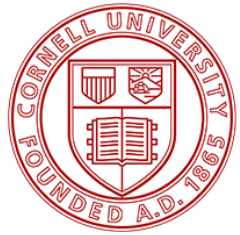
Optimizing AlphaFold for Accessible Computational Biology

Global Open OnDemand Conference 2025

Vinay Saji Mathew
PhD Student, IEOR
Laboratory for Intelligent Systems and Analytics, Penn State

March 28, 2025



PennState
Institute for Computational
and Data Sciences

Presentation Overview

1. Introduction
2. Objectives
3. Challenges & Solutions
4. Implementation
5. Impact
6. Future Work

Introduction

AlphaFold Achievement

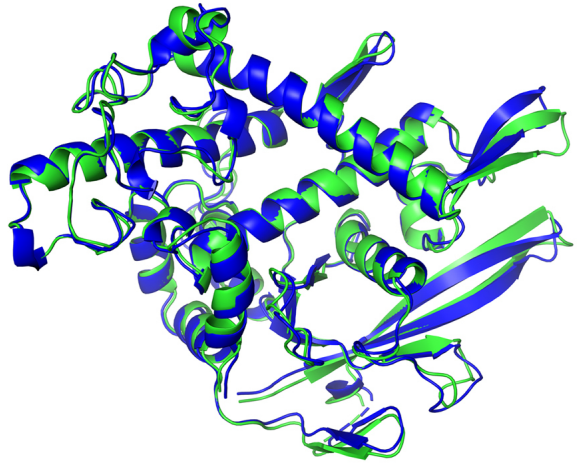
2024 Chemistry Nobel Prize Winner

Shared by John Jumper, Demis Hassabis, and David Baker

FASTA Sequence

```
>Protein
MKTIIALSYIFCLVFADYKDDDDK
FDKAKKLVFAATDGFYSVDVVK
...
```

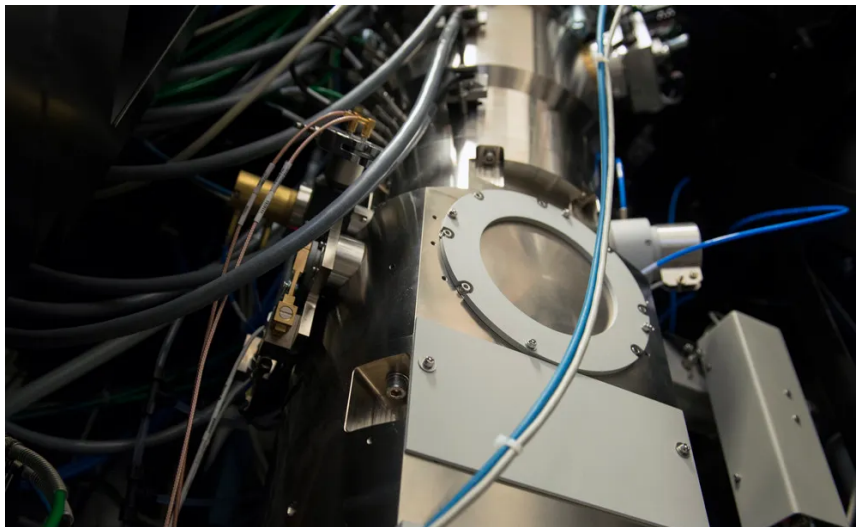
$\Rightarrow$



Biological Significance

Why Protein Structure Prediction Matters

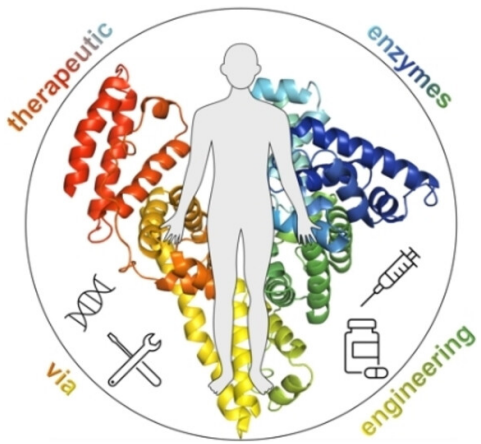
- Proteins are molecular machines governing biological processes
- AlphaFold achieves 92.5% experimental-level accuracy
- Billions of sequences produced annually
- Traditional methods (X-ray, cryo-EM) are resource-limited



Biological Significance

Several Research Applications

- Drug discovery through virtual screening
- Host-pathogen interaction modeling
- Enzyme and therapeutic protein engineering
- Evolutionary relationship analysis



Introduction

Current Challenges

- Computationally demanding
- GPU utilization inefficiencies
- **Accessibility limitations**

Objectives

1. Optimize computational workflow
2. Enhance accessibility
3. Remain open-source for deployment.

Challenges and Observations

Optimize computational workflow

- Complex deployment process (initially Docker-only, problematic for HPC systems)
- Current setup has 75% CPU-bound operations while GPU is idle
- Wasted GPU utilization
- Potential Multi Instance GPU Support (!)
- 5TB+ database requirements
  - Regular updates needed
  - **Redundant storage across users**
  - Cron job maintenance

Current HPC Deployment Issues

Current HPC Deployment Issues

- Resource wastage during CPU phases
- Complex setup requirements
- Duplicate database storage waste

Technical Implementation Details

Container Implementation

- Singularity/Apptainer based
- Modular design for CPU/GPU phases
- High-performance database access (VAST via NFS - ROAR)

System Requirements

- Minimum: Half an A100 GPU for vGPU (1/7th slice of a MIG A100)
- 5 TB database storage

Cross-Platform Validation

Benchmarked Clusters

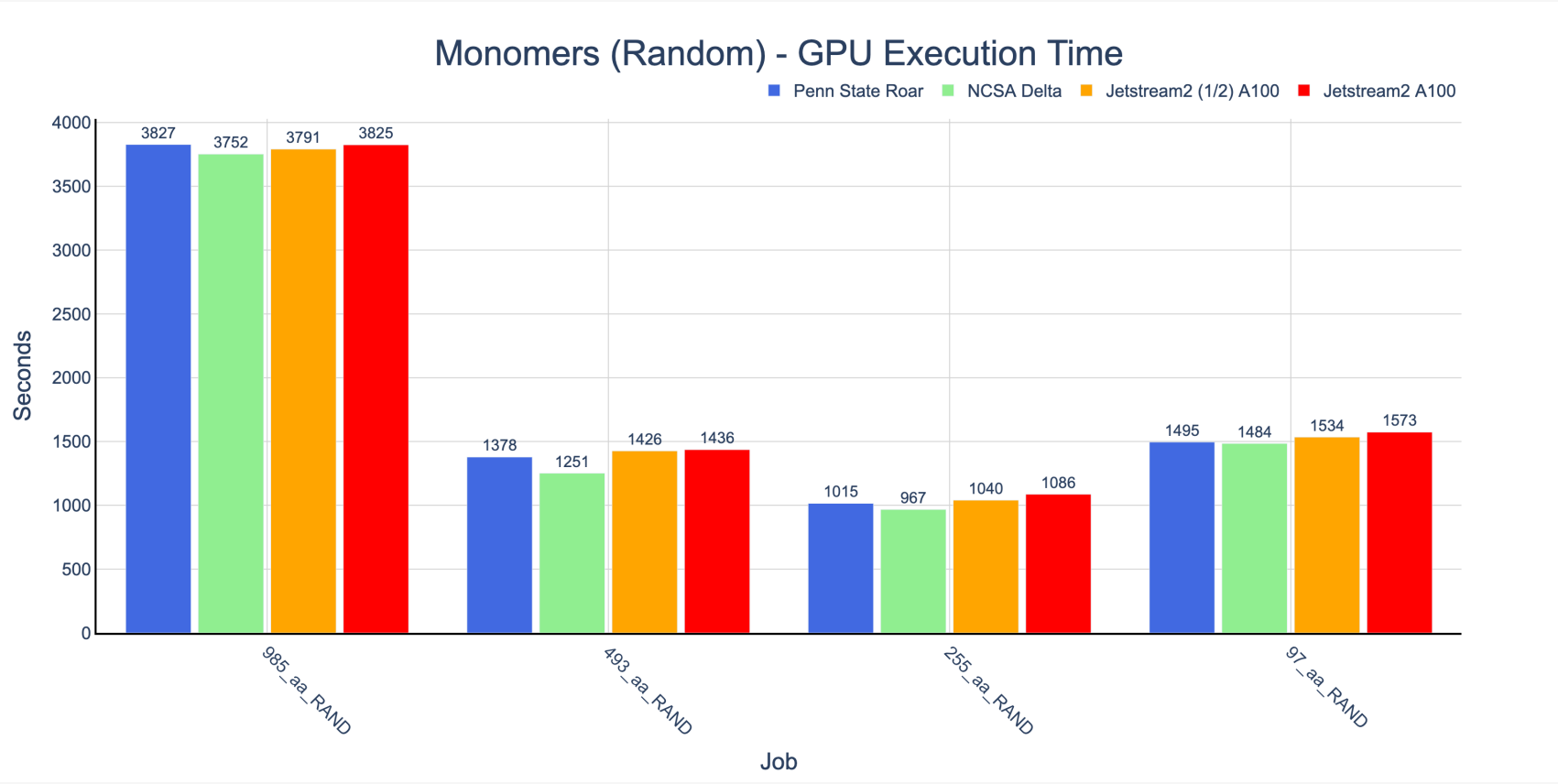
- NCSA Delta
- Jetstream2 by IU
- ROAR by Penn State

Performance Analysis - Monomers (CPU)

CPU Execution Time for Monomers

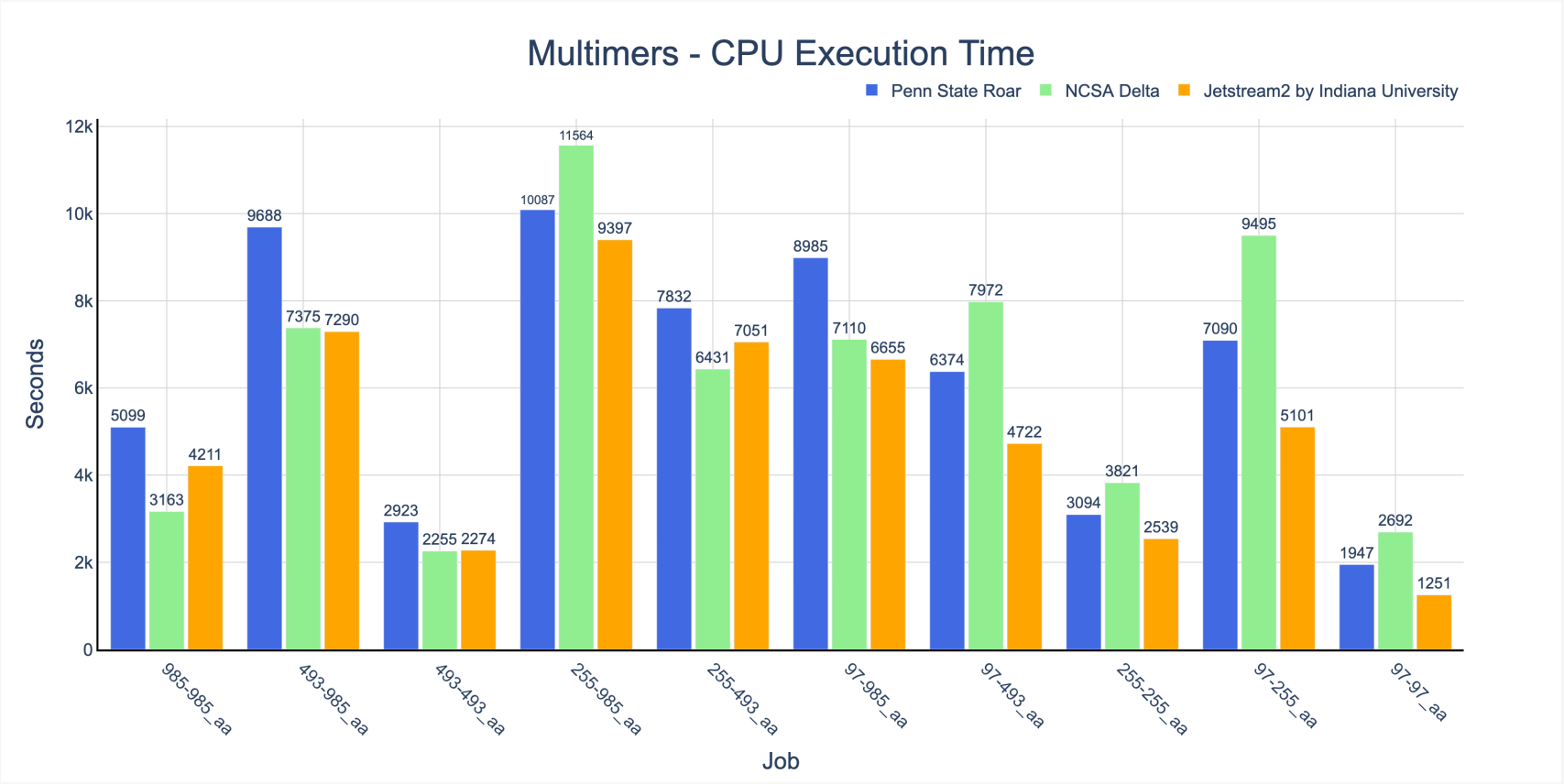
Performance Analysis - Monomers (GPU)

GPU Execution Time for Monomers



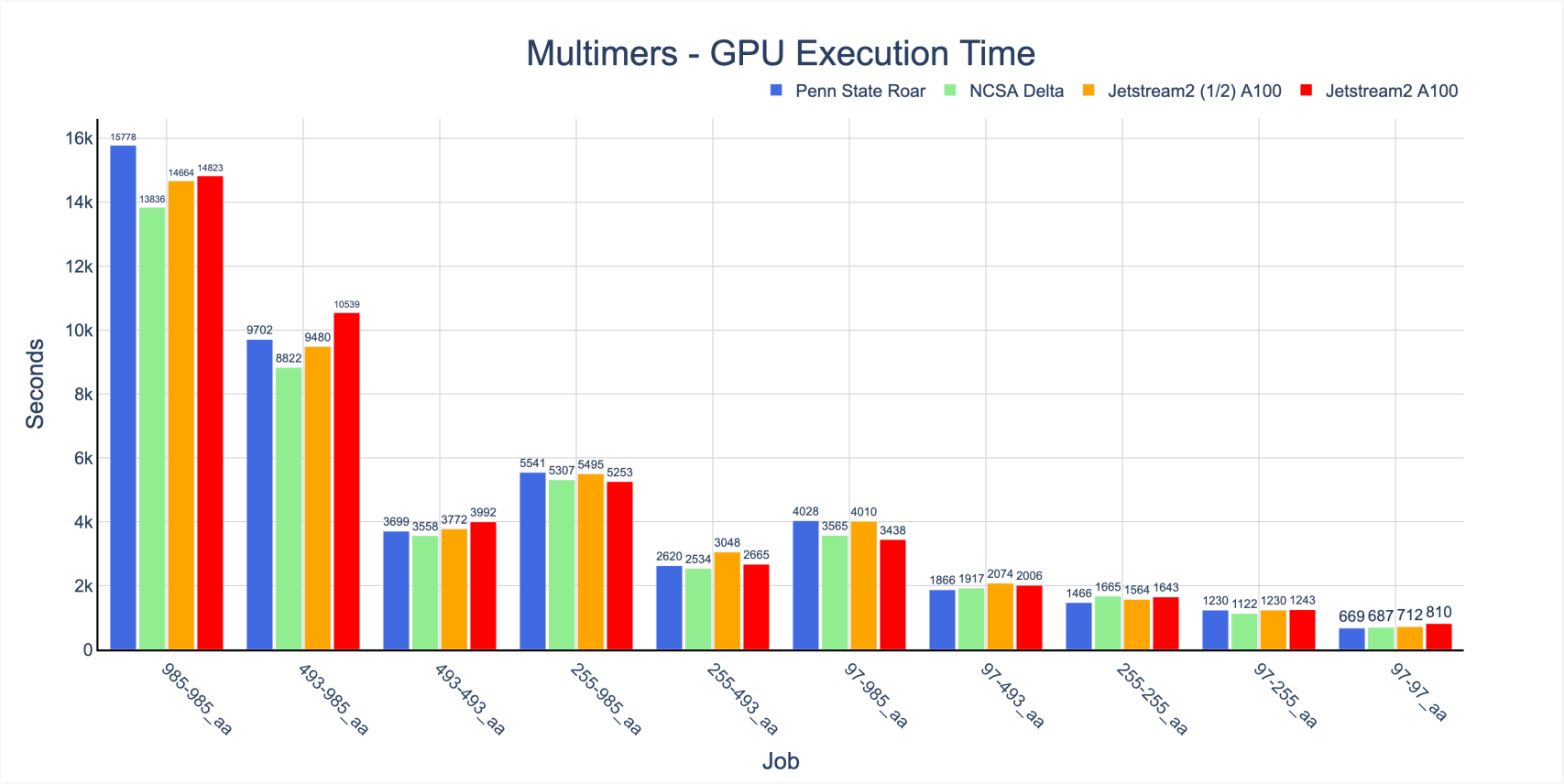
Performance Analysis - Multimers (CPU)

CPU Execution Time for Multimers



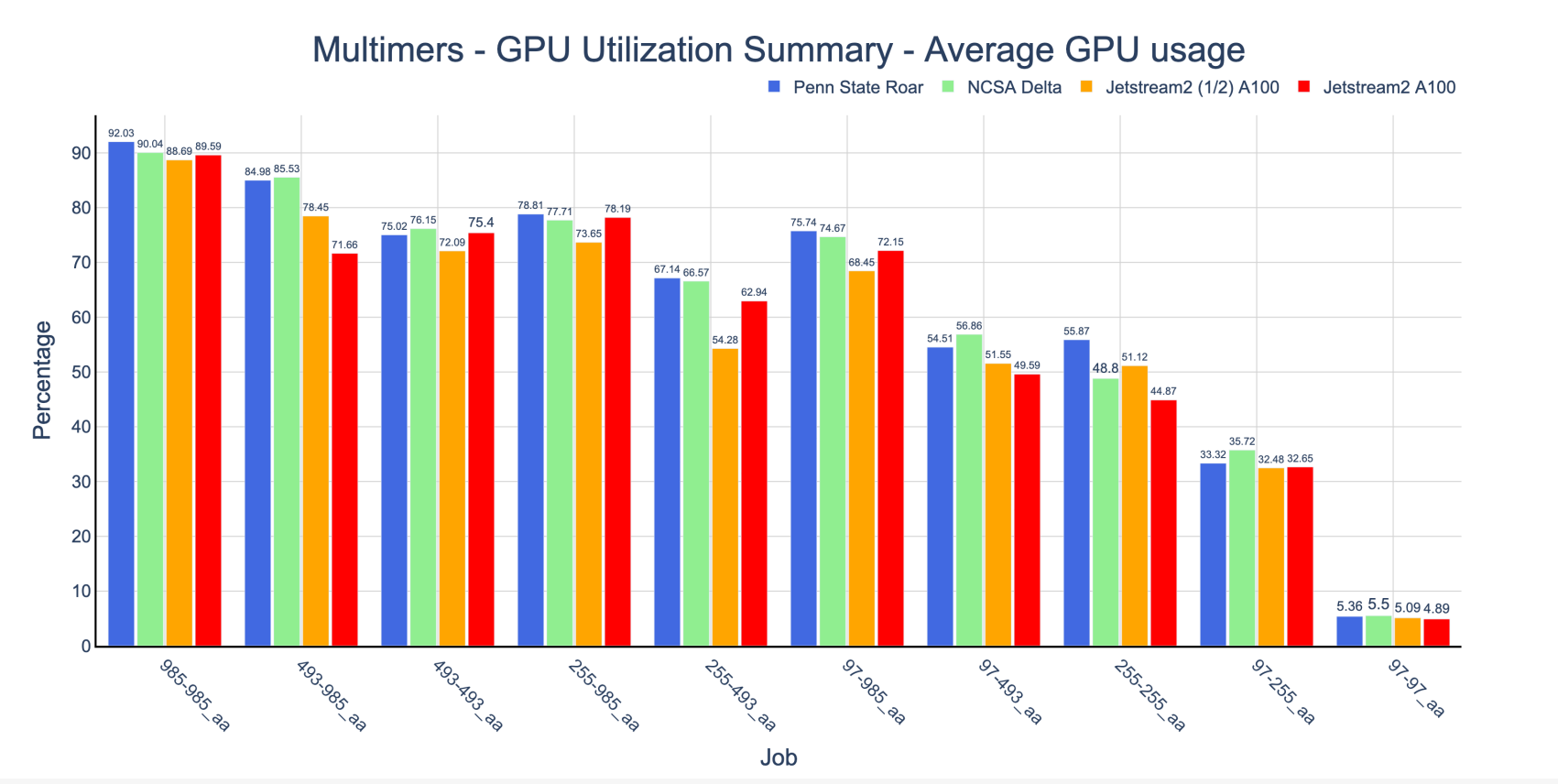
Performance Analysis - Multimers (GPU)

GPU Execution Time for Multimers



GPU Utilization Analysis

Multimer GPU Usage Summary



System-Specific Implementations

Cluster-Specific Observations

- **NCSA Delta & ROAR**
  - MIG-enabled GPU support
  - Parallel job submission
  - Queue-based workload distribution
  - Tested on MIG GPU partitions down to 1/7th of an A100. Performance degrades but completes on partitioned A100
- **Jetstream2**
  - Virtual GPU allocation
  - Anything less than half an A100 on vGPU setup crashes
  - Sequential job processing
  - Dedicated system assignment

Our Approach: Optimized OOD Implementation

Workflow Optimization

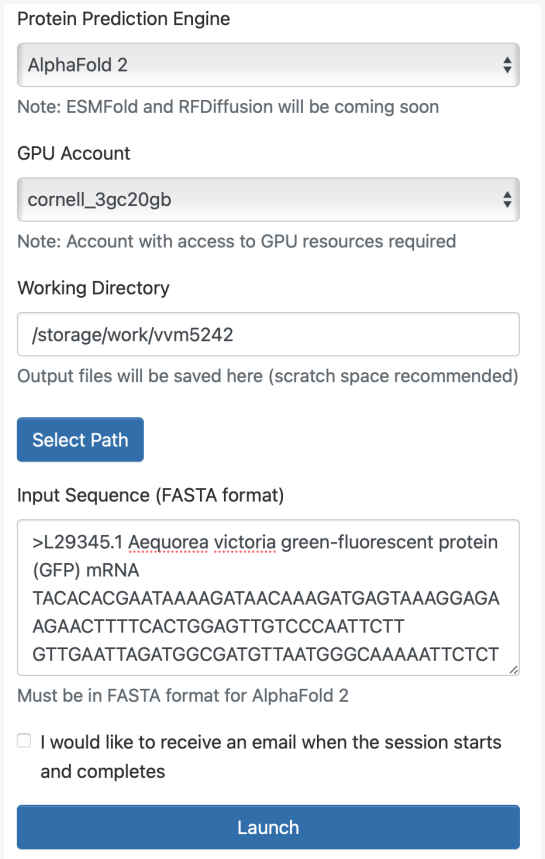
- CPU/GPU phase separation
- GPU allocation is dependent on successful completion and verification of output from MSA search.
- 75% reduction in GPU allocation

User Interface

- Easy Open On Demand interface
- Single input requirement
- Real time progress and Logs

AlphaFold2 User Interface

Open OnDemand Form



AlphaFold3 User Interface

Form Part 1

Protein Prediction

This app will generate a predicted structure for the input amino acid sequence using the selected engine.

Protein Prediction Engine

AlphaFold 3

Note: ESMFold and RFDiffusion will be coming soon

GPU Account

cornell_3gc20gb

Note: Account with access to GPU resources required

Working Directory

/storage/work/vvm5242

Output files will be saved here (scratch space recommended)

Select Path

Form Part 2

Input Sequence (JSON format)

```
{
  "name": "2PV7",
  "sequences": [
   {
    "protein": {
     "id": ["A", "B"],
     "sequence": "GMRESYAN"
    }
   }
  ],
  "modelSeeds": [1],
  "dialect": "alphafold3",
  "version": 1
 }
```

Must be in compatible JSON format (specifications are in the AlphaFold 3 documentation).

☑ I agree to Google's Terms of Service

Read the terms

Job Monitoring and Results

Progress During Run

Protein Prediction (33287050)

1 node | 1 core | Running

Host: >_ p-bc-5006

Delete

Created at: 2025-02-23 00:06:39 EST

Time Remaining: 23 hours and 56 minutes

Session ID: 106f12f0-435a-4466-b24e-541b472cc950

Problems with this session? Submit support ticket

Current Status

Job IDs: CPU = 33287054 | GPU = 33287055 View My Jobs

Run Directory: /storage/work/vvm5242/pp1740287199

Phase: AlphaFold 3 CPU Phase - MSA Generation

Step: Completed Jackhmmer search

25%

Progress:

- Detected AlphaFold 3 input format
- Getting protein MSAs
- Started Jackhmmer search
- Completed Jackhmmer search

CPU Phase Log

```
I0223 05:07:29.417441 22982725124096 folding_input.py:1044] Detected /root/af_inpu
```

After Completion

CPU Phase Log

```
I0213 21:42:51.090887 22878978326528 folding_input.py:1044] Detected /root/af_inpu
I0213 21:42:51.092551 22878978326528 pipeline.py:81] Getting protein MSAs for sequ
I0213 21:42:51.095661 22876702512704 jackhmmer.py:78] Query sequence: GMRESYANENQF
I0213 21:42:51.096306 22876696208960 jackhmmer.py:78] Query sequence: GMRESYANENQF
I0213 21:42:51.096395 22876700411456 jackhmmer.py:78] Query sequence: GMRESYANENQF
I0213 21:42:51.096657 22876698310208 jackhmmer.py:78] Query sequence: GMRESYANENQF
I0213 21:42:51.097247 22876702512704 subprocess_utils.py:68] Launching subprocess
I0213 21:42:51.097560 22876696208960 subprocess_utils.py:68] Launching subprocess
I0213 21:42:51.098170 22876700411456 subprocess_utils.py:68] Launching subprocess
I0213 21:42:51.098312 22876698310208 subprocess_utils.py:68] Launching subprocess
I0213 21:45:21.104391 22876698310208 subprocess_utils.py:97] Finished Jackhmmer in
I0213 21:51:36.808516 22876702512704 subprocess_utils.py:97] Finished Jackhmmer in
I0213 21:55:56.012930 22876696208960 subprocess_utils.py:97] Finished Jackhmmer in
I0213 22:00:35.508942 22876700411456 subprocess_utils.py:97] Finished Jackhmmer in
I0213 22:00:35.623793 22878978326528 pipeline.py:114] Getting protein MSAs took 16
```

GPU Phase Log

```
I0214 20:01:27.887957 23093135508608 xla_bridge.py:895] Unable to initialize backe
I0214 20:01:27.902343 23093135508608 xla_bridge.py:895] Unable to initialize backe
I0214 20:01:28.024057 23093135508608 folding_input.py:1044] Detected /root/af_outp
I0214 20:01:36.723238 23093135508608 pipeline.py:165] processing 2PV7, random_seed
I0214 20:01:36.760128 23093135508608 pipeline.py:258] Calculating bucket size for
I0214 20:01:36.760270 23093135508608 pipeline.py:264] Got bucket size 768 for inpu
Running AlphaFold 3. Please note that standard AlphaFold 3 model parameters are
only available under terms of use provided at
https://github.com/google-deepmind/alphafold3/blob/main/WEIGHTS_TERMS_OF_USE.md.
If you do not agree to these terms and are using AlphaFold 3 derived model
parameters, cancel execution of AlphaFold 3 inference with CTRL-C, and do not
use the model parameters.
Skipping running the data pipeline.
Found local devices: [CudaDevice(id=0)]
Building model from scratch...
Processing fold inputs.
```

Completion Information

Run Directory: /storage/work/vvm5242/pp1739482915

Structure Directory: /storage/work/vvm5242/pp1739482915/structure

Benefits and Outcomes

For HPC Centers

- Optimized resource utilization
- Ready for AlphaFold 2, 3 and Boltz
- Centralized compliance handling
- Available at: github.com/EpiGenomicsCode/ProteinStructure-OOD

For Researchers

- Simplified access to AlphaFold i.e., No coding requirements

Protein Design with RFDiffusion

Protein Design

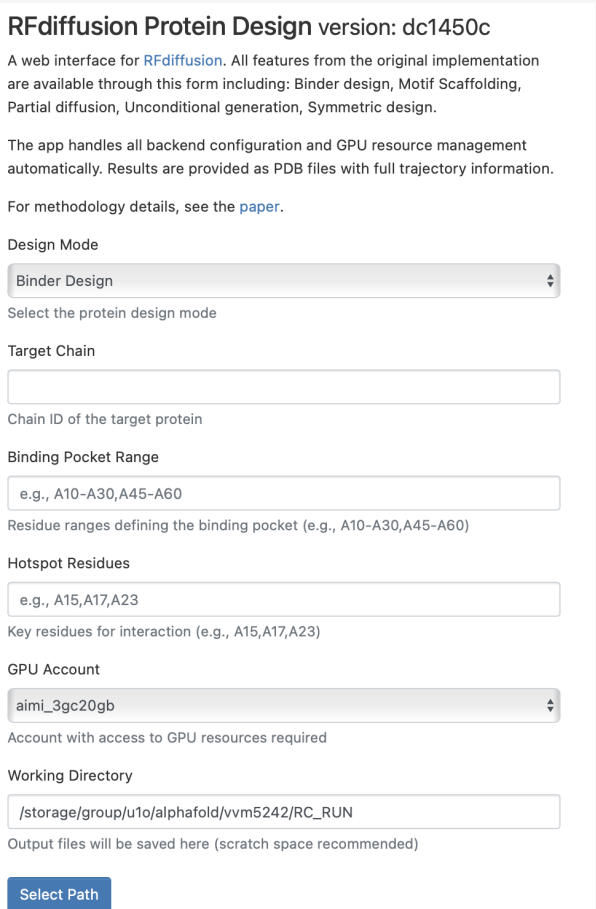
- Inverse problem to structure prediction
- Design proteins with specific functions
- Applications in drug development and enzyme engineering
- Complementary to AlphaFold's prediction capabilities

RFDiffusion OOD Implementation

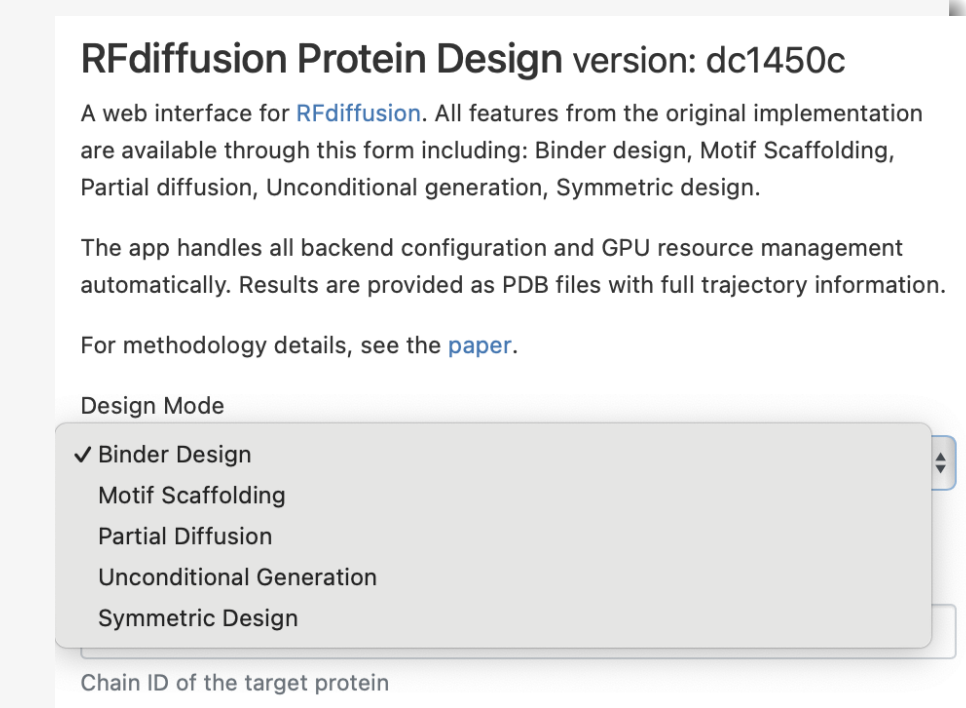
- Development app available at: github.com/EpiGenomicsCode/RFDiffusion-OOD
- Similar user-friendly interface

RFDiffusion Preview

RFDiffusion Interface

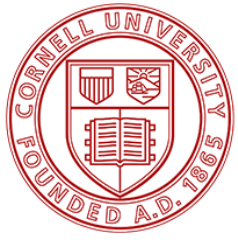


Design Modes



Acknowledgments

- Penn State Institute for Computational and Data Sciences (RRID: SCR_025154)
- Cornell BRC Epigenomics (RRID: SCR_021287)
- Penn State Center for Applications of Artificial Intelligence and Machine Learning to Industry Core Facility (RRID: SCR_022867)



PennState
Institute for Computational
and Data Sciences

Acknowledgments

- William Lai
- Gretta Kellogg
- Matt Hansen
- Chad Bahrmann
- Soundar Kumara

PennState
Institute for Computational
and Data Sciences