Automated Data Transfer for CryoEM

Matthew Cahn
System Administrator/Lecturer
Dept. of Molecular Biology / Research Computing
Princeton University
mcahn@princeton.edu

5/24/2023

Topics

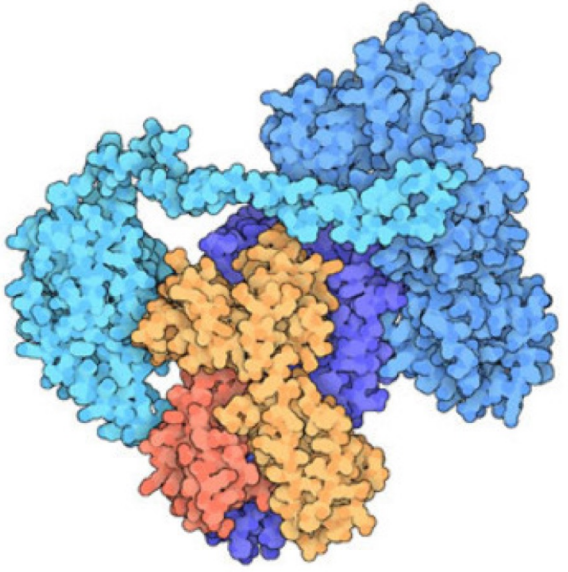
- Proteins
- CryoEM
- Data Transfer
  - Globus
  - AWS S3

All essential biological processes are carried out by proteins and protein complexes

PDB-101 April 2023 Molecule of the Month:

MHC I Peptide Loading Complex

Part of the immune system

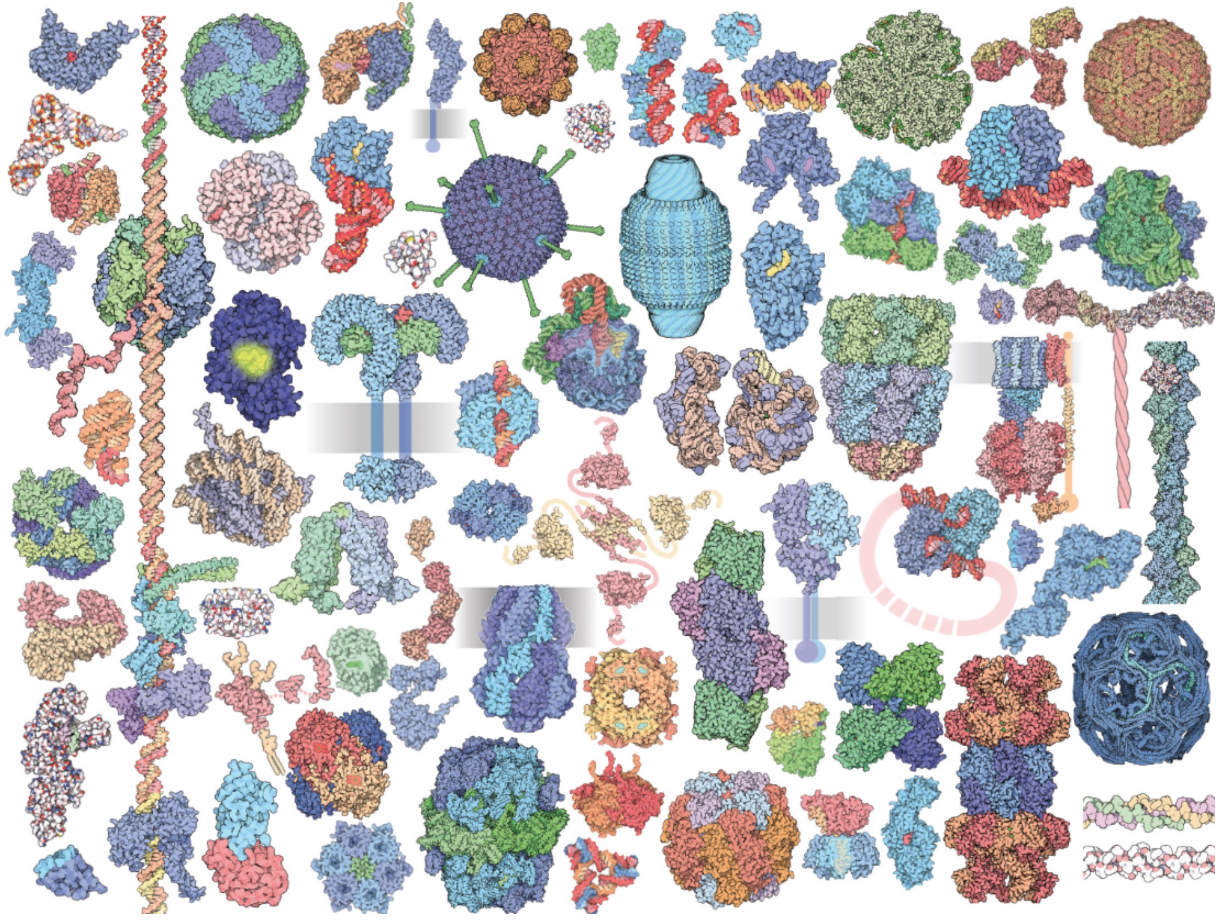


Source: Ellen Zhong,
Assistant Professor of Computer Science
Princeton University

Proteins

- Chains of Amino Acids
- Self-folding
- Structural Biology involves the determination of the structure of proteins and protein complexes

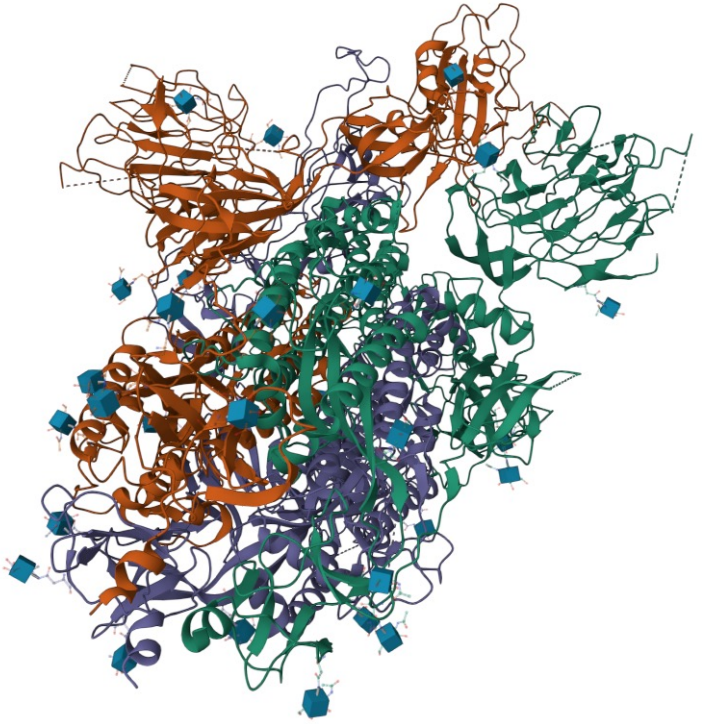
Huge variety of proteins



Goodsell et al. PLoS Biology 2015.

SARS-COV-2 Omicron Spike Trimer

- 1,285 amino acids
- 24,078 atoms
- ~6 nm in diameter
  - 1/12,000 of a human hair
  - 6.6 x 10 football fields



https://www.rcsb.org/structure/7WZ1
Deposited: 2022-02-16 Released: 2022-07-27
Deposition Author(s): Zhan, W.Q., Zhang, X., Chen, Z.G., Sun, L.
Funding Organization(s): Ministry of Science and Technology (MoST, China)

CryoEM

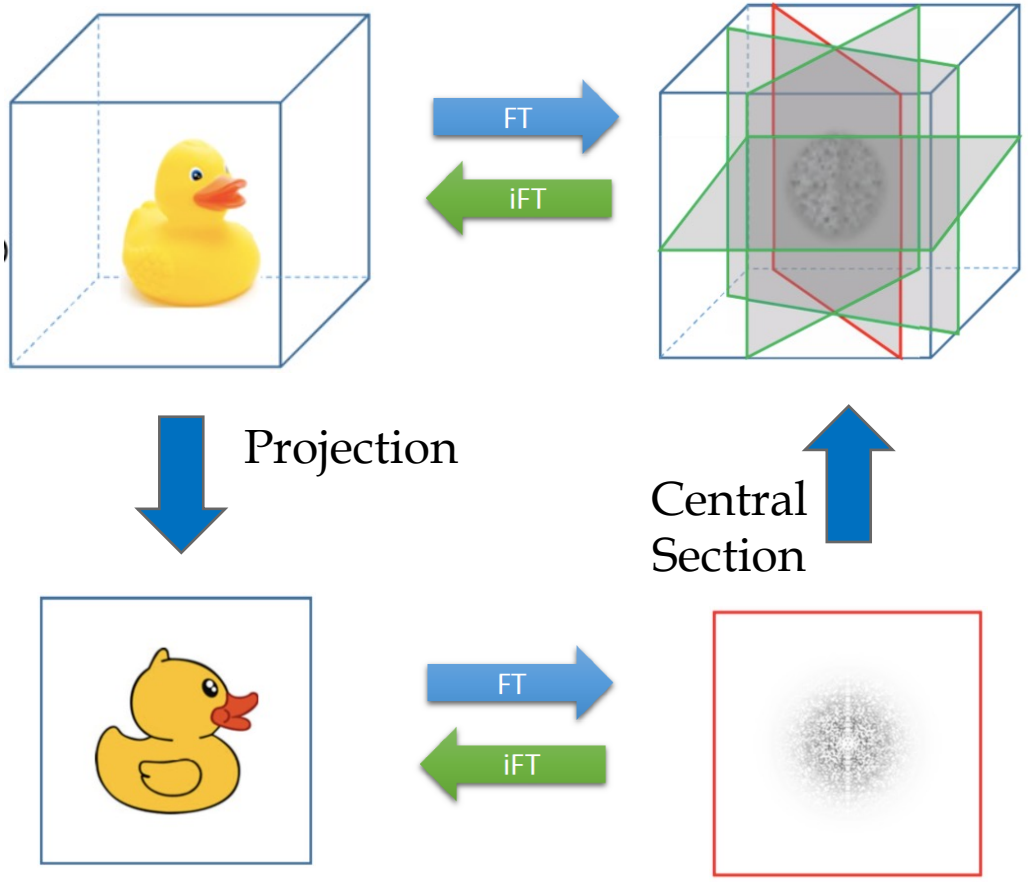
- Cryogenic electron microscopy
- Three-dimensional reconstruction of protein molecules
- 2017 Nobel Prize in Chemistry was awarded to Jacques Dubochet, Joachim Frank, and Richard Henderson for the invention of the cryoEM method.

The CryoEM Method

- Freeze sample rapidly in vitreous (non-crystalline) ice
- Collect thousands of movies in a specialized TEM (transmission electronic microscope)
- Construct a model of the protein from the movies

The Central Section Theorem

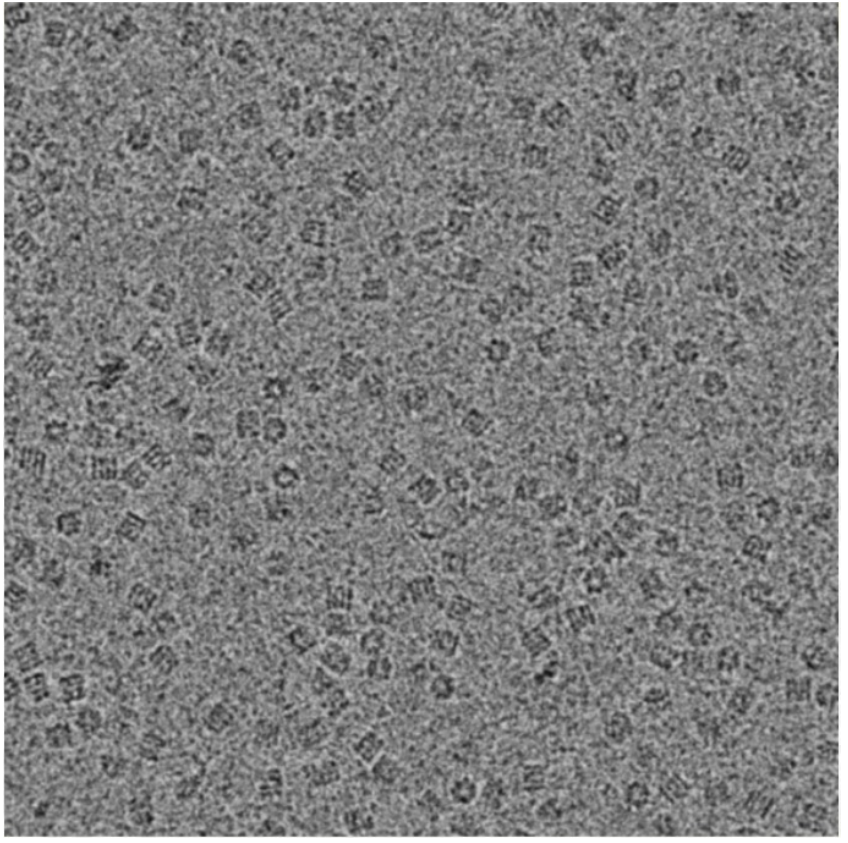
- The central section theorem states that the Fourier transform of a projection of a three-dimensional object is equivalent to a slice through the object's Fourier transform.



Source: Dr. Xiao Fan, Postdoctoral Research Associate, Princeton University

One (Motion-Corrected) Micrograph

Protein molecule in various orientations

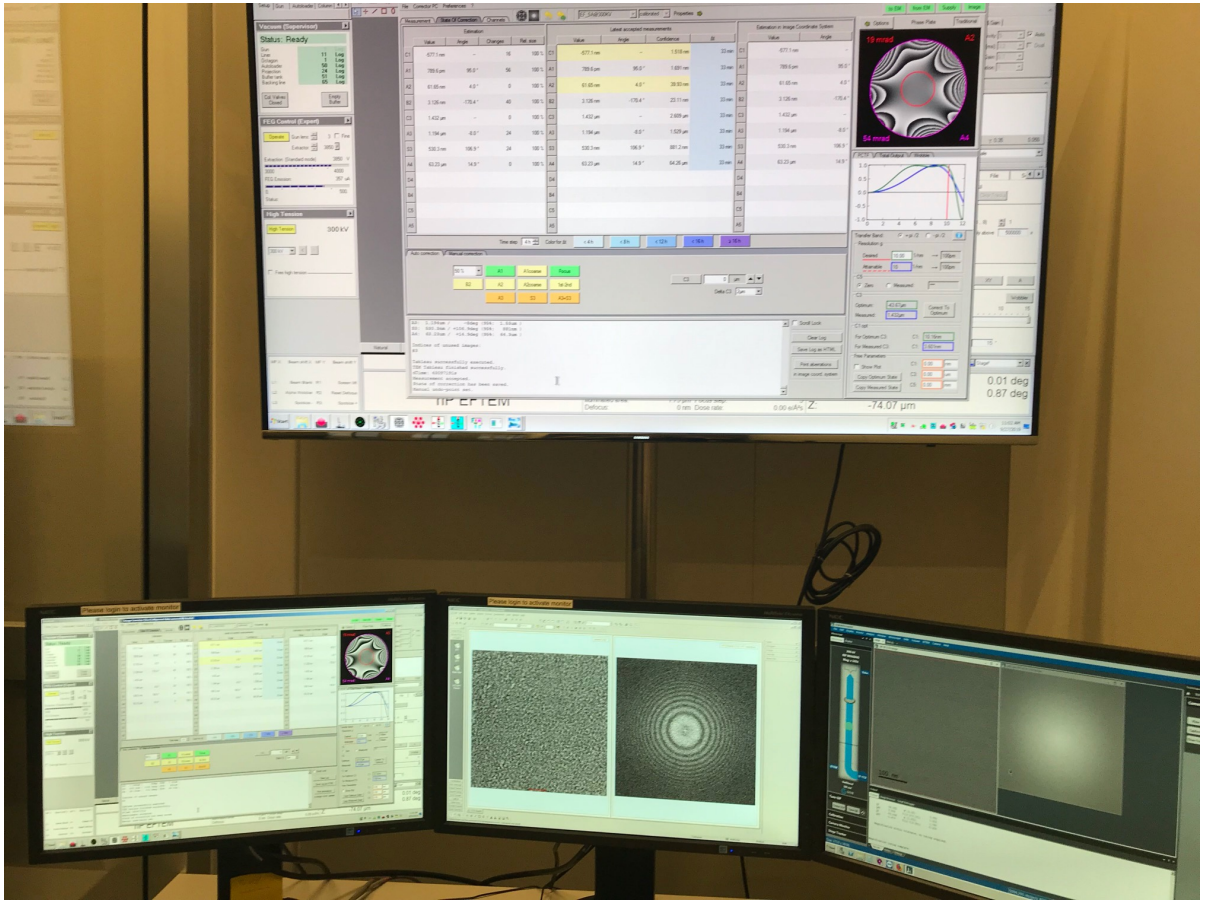


Source: Nanyang Technological University
https://www.embl-hamburg.de/biosaxs/courses/embo2017/slides/EMBO_cryoEM_2_Bhushan.pdf

ThermoFisher Krios Cryo-TEM Microscope

Power Supply and Data Acquisition Hardware

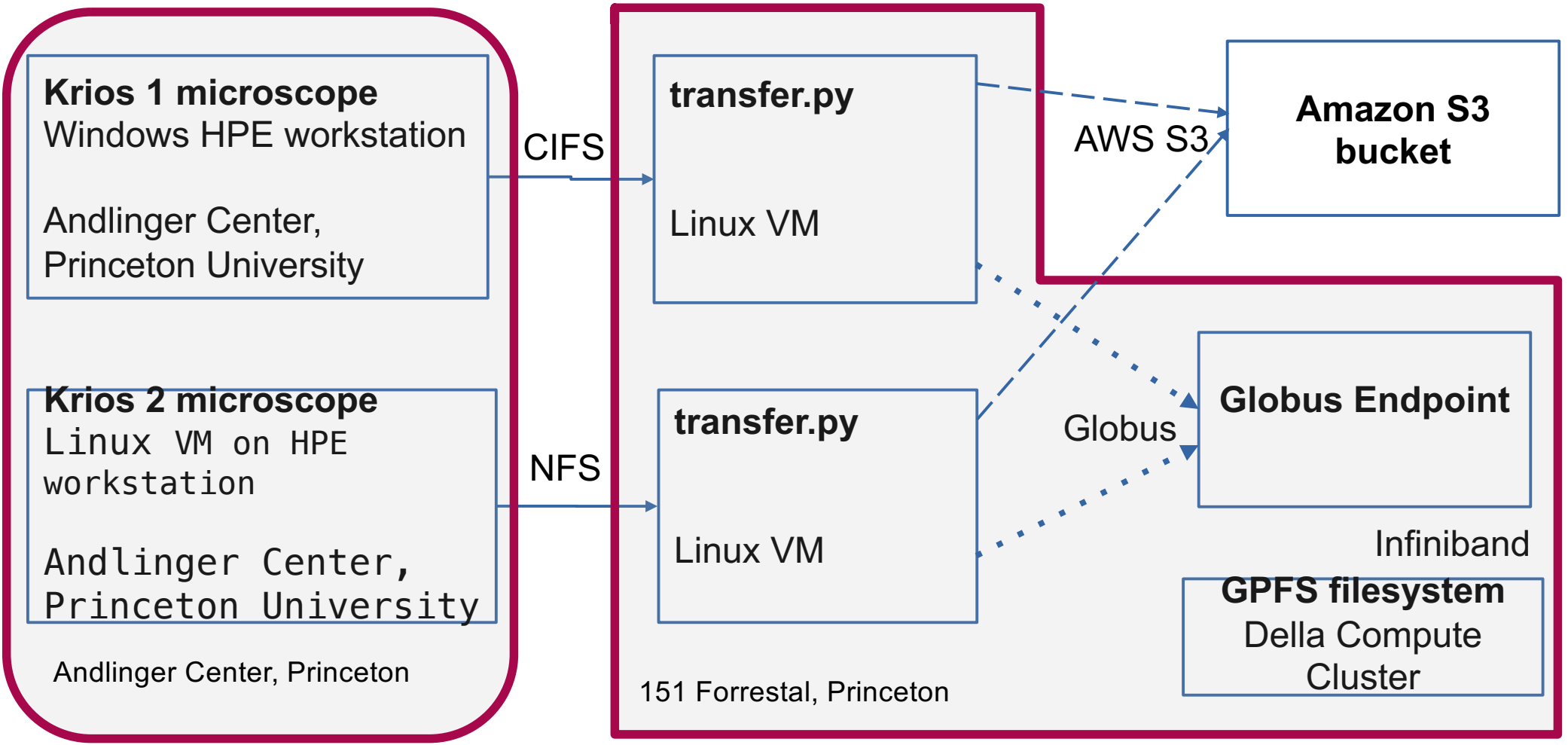
Krios User Interface



Data Volume

- ~2 TB per microscope (2) per day
- Plus output of processing
- ~ 2 PB since 2019
- Data must be transferred at least as fast as it is acquired.
  - Microscope ==> Compute cluster

Data Transfer



Transfer Program (1)

- Python
- Globus SDK
- AWS SDK (“boto3”)
- Runs on Linux VMs
- Monitored by cron

Transfer Program (2)

- Run forever
  - Find data to transfer
    - Directory name:
      - outgoing_<login name>_* – Globus transfer to compute cluster
      - outgoing_PHARMA_* – AWS S3 transfer to Amazon bucket
    - File age > 60 seconds
  - Transfer
  - Delete from microscope

Transfer Program (3)

- Globus transfer
  - Authorize
  - Submit transfer task
  - Wait for transfer task to complete
  - Delete transferred files from microscope
  - Clean up old empty directories

Transfer Program (4)

- AWS S3 transfer
  - Pass keys
  - Upload files

Transfer Program (5)

- Python logging module

```
2023-05-21-002148 INFO  Iteration:            44650
2023-05-21-002148 INFO  Checking:             for files to transfer in /krios2
2023-05-21-002148 INFO  Clock skew:           This machine's clock is  651.419 ahead
2023-05-21-002148 INFO  Priority file:        /krios2-offloaddata/priority.txt found
2023-05-21-002148 INFO  Found                 0 PHARMA files and 7 Princeton files
2023-05-21-002148 INFO  Will transfer         via Globus <filename>
...
2023-05-21-002148 INFO  Total size (MB):      5106.31 deadline: 1200 seconds
2023-05-21-002149 INFO  Request result:       The transfer has been accepted
2023-05-21-002209 INFO  Transfer result:      files: 7 files skipped: 0  files
transferred: 7
2023-05-21-002209 INFO  Effective Mb/sec.:    3059.14 Mb/sec.
2023-05-21-002209 INFO  Transferred:          <filename>
...
2023-05-21-002209 INFO  Rate Mb/sec.:         2042.52 Mb/sec.
2023-05-21-002209 INFO  Delete file:          <filename>
...
```

Processing

- RELION
  - Developed in the group of Sjors Scheres at the MRC Laboratory of Molecular Biology.
- CryoSPARC
  - Structura Biotechnology Inc.
  - Show CryoSPARC
- CryoDRGN
  - Heterogeneous reconstruction
  - Ellen D. Zhong , Tristan Bepler, Joseph H. Davis, Bonnie Berger
    - * Associate Professor, Dept. of Computer Science, Princeton University

Della High-Performance Computing Cluster

- 38 Nodes (computers) dedicated to cryoEM
- Four Nvidia A100 or V100 GPUs each
- 384 GB to 1 TB RAM

Questions