nextPYP

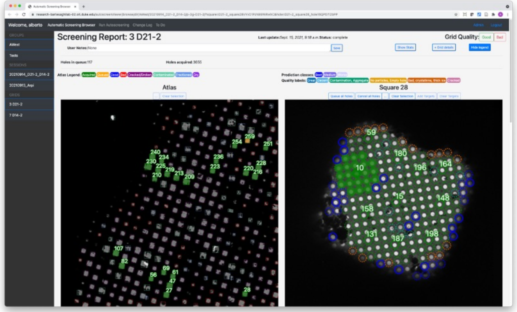
Cryo-ET data processing in nextPYP

Bartesaghi Lab

Duke
UNIVERSITY

Tools for cryo-EM/ET data collection and analysis

SmartScope: automated screening



Screening Report: 3 D21-2

Last update: Sept. 15, 2021, 9:18 a.m. Status: complete

Grid Quality:

User Notes: None

Holes in queue: 117

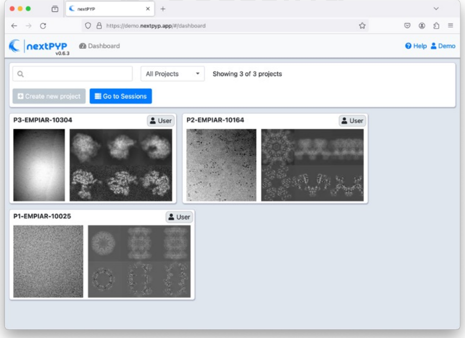
Holes acquired: 3655

Atlas Legend: Acquired Queued Good Bad Cracked/Broken Contaminated Fractioned Dry

Prediction classes: Best Medium Worst

Quality labels: Great Decent Contamination, Aggregate No particles, Empty hole Bad, crystalline, thick ice Cracked

nextPYP: cryo-EM/ET data processing



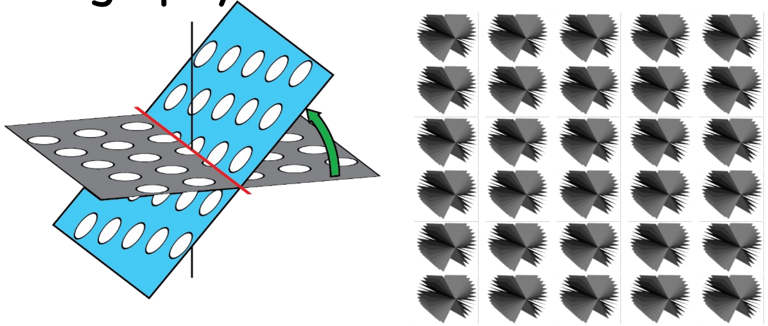
Showing 3 of 3 projects

P3-EMPIAR-10304

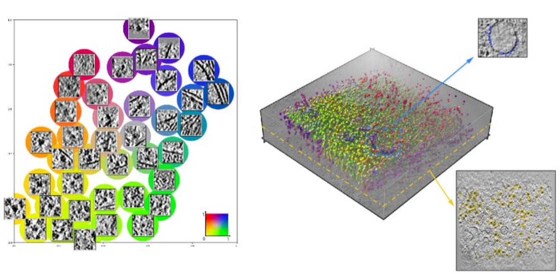
P2-EMPIAR-10164

P1-EMPIAR-10025

BISECT: Beam-image shift tomography



MiLoPYP: cellular pattern mining

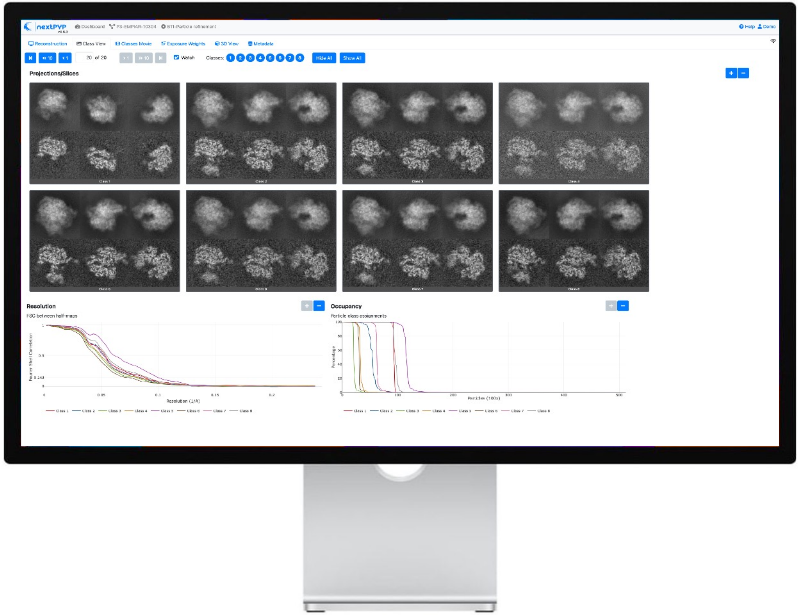


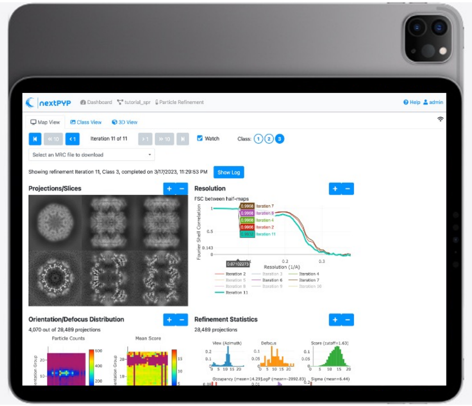
nextPYP

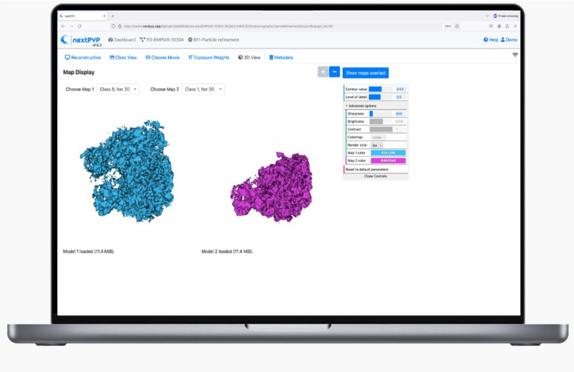
A scalable, end-to-end platform for single-particle cryo-EM/ET image analysis

- Portable, fully-featured web-based GUI and CLI
- End-to-end single-particle cryo-EM/ET pipelines
- Small storage footprint
- Distributed processing for scalability to large datasets
- Free and open source:
    - github.com/nextpyp
    - BSD 3-Clause license

Portability, web-based platform





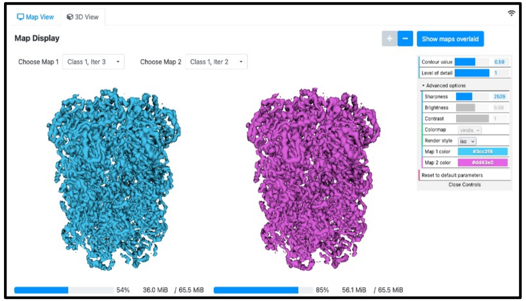


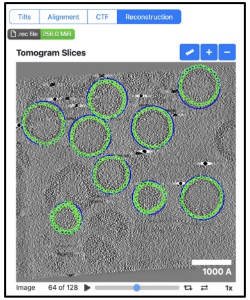
Integrated data visualization tools

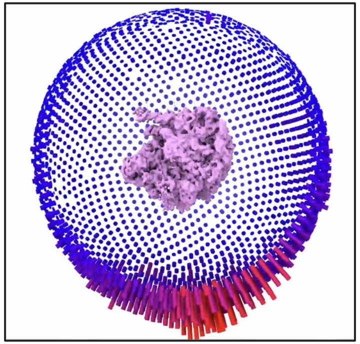
Showing refinement Iteration 11, Class 3, completed on 3/17/2023, 11:29:53 PM

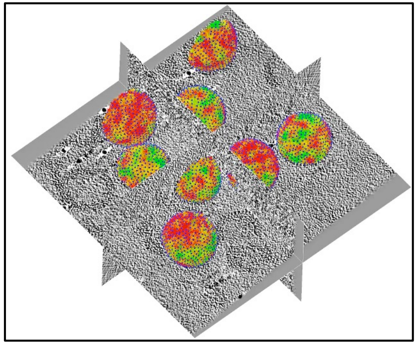
4,070 out of 28,489 projections

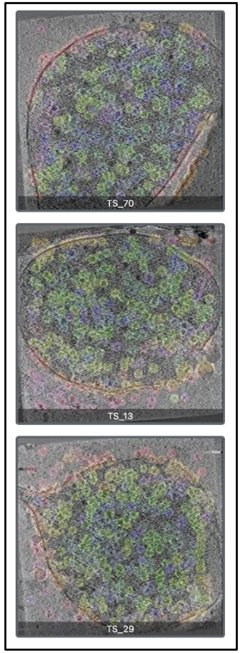
28,489 projections





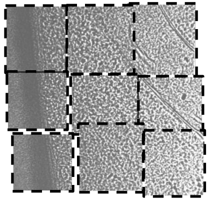




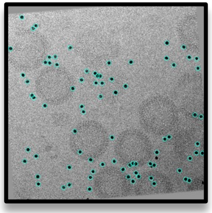


From raw tilt-series to 3D structure

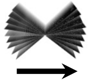
1. Frame/tilt-series alignment



Patch-based

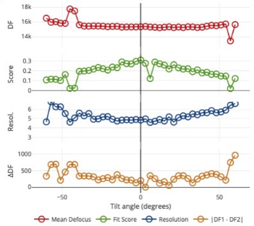


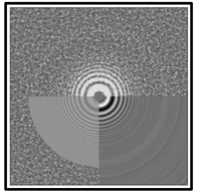
Fiducial-based

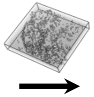


Aligned tilt-series

2. CTF estimation and reconstruction

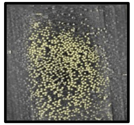




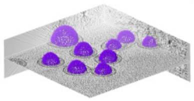


Binned tomogram

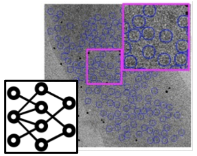
3. Particle picking



Size-based



Geometry-based



Neural network-based

4. Constrained refinement

Ab-initio

↓

Particle and tilt-geometry refinement

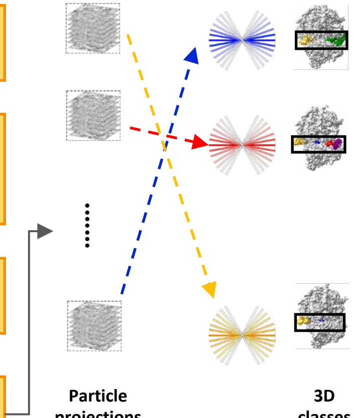
↓

Region-based refinement

↓

Particle-based CTF refinement

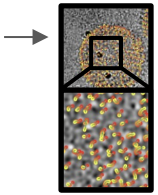
5. Constrained classification



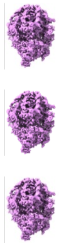
Particle projections

3D classes

6. Movie-frame refinement

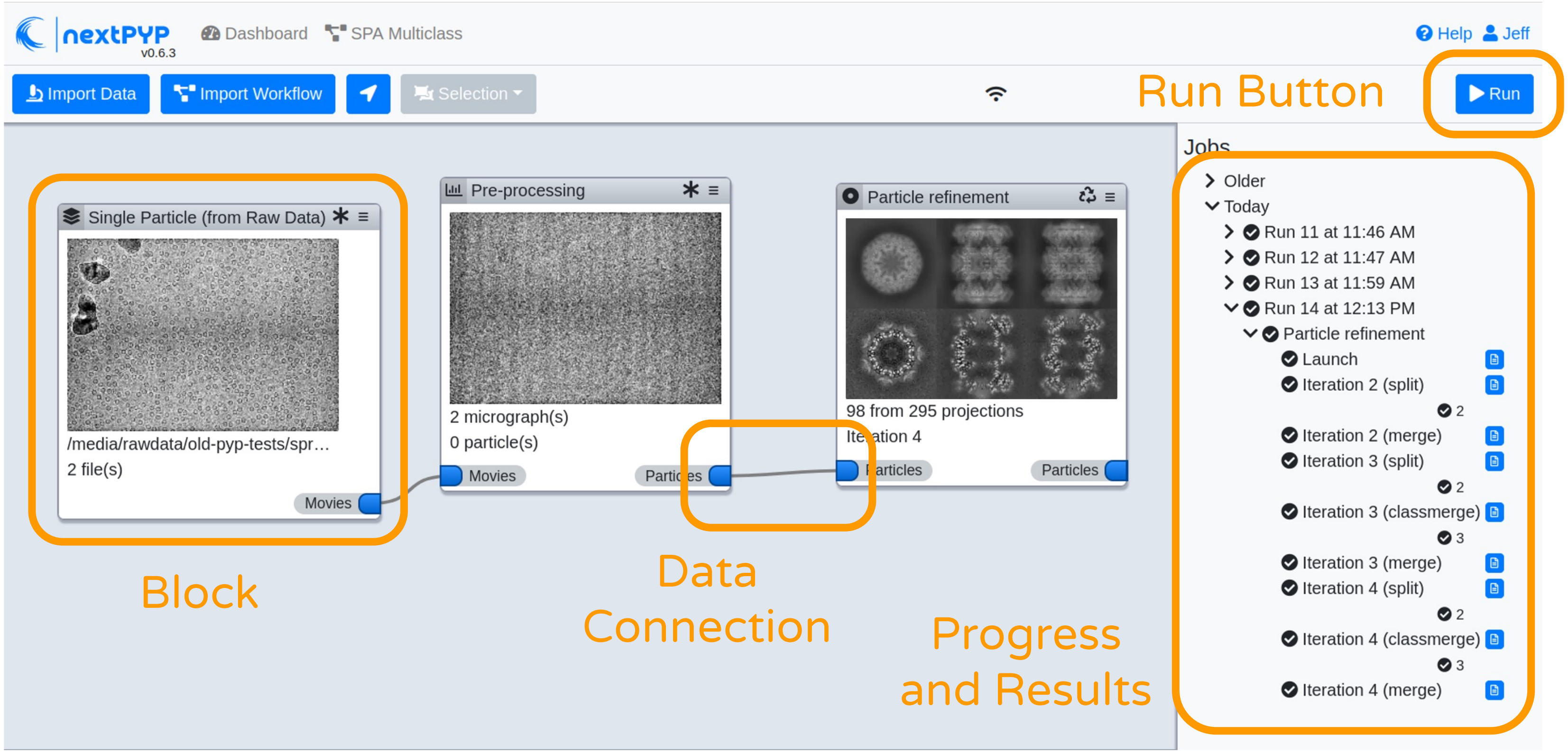


Particle trajectories



Final maps

Intuitive block-based user interface



/media/rawdata/old-pyp-tests/spr…
2 file(s)

2 micrograph(s)
0 particle(s)

98 from 295 projections
Iteration 4

- Older
- Today
  - Run 11 at 11:46 AM
  - Run 12 at 11:47 AM
  - Run 13 at 11:59 AM
  - Run 14 at 12:13 PM
    - Particle refinement
      - Launch
      - Iteration 2 (split)
        - 2
      - Iteration 2 (merge)
      - Iteration 3 (split)
        - 2
      - Iteration 3 (classmerge)
        - 3
      - Iteration 3 (merge)
      - Iteration 4 (split)
        - 2
      - Iteration 4 (classmerge)
        - 3
      - Iteration 4 (merge)

Run Button

Block

Data Connection

Progress and Results

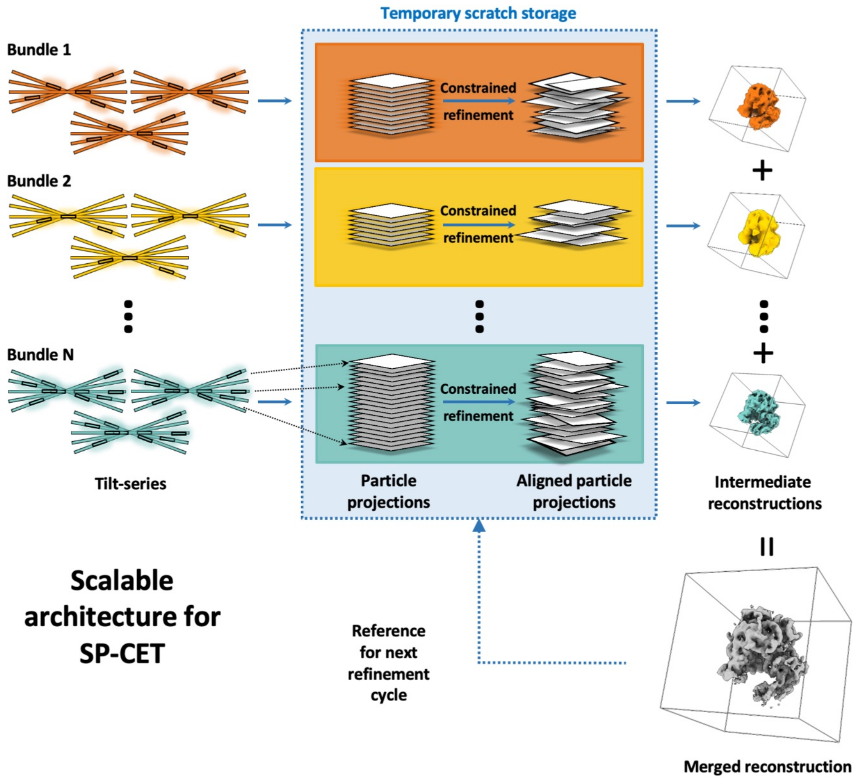
Scalability

Storage efficiency

- Heavily binned tomograms
- No sub-volumes
- No particle-stacks

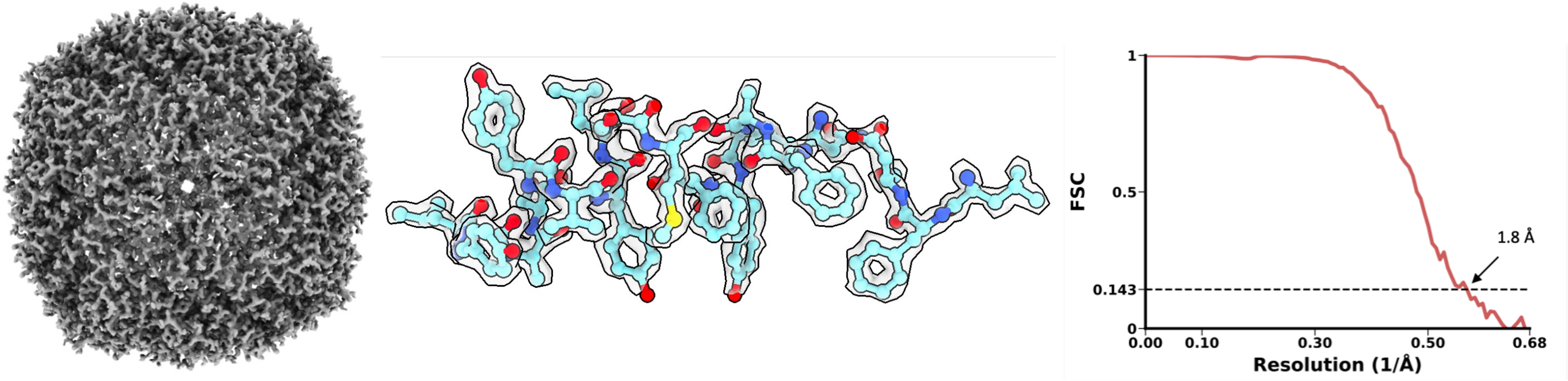
Scalability to large datasets

- High-core CPU architectures
- Fast local I/O (SSDs)



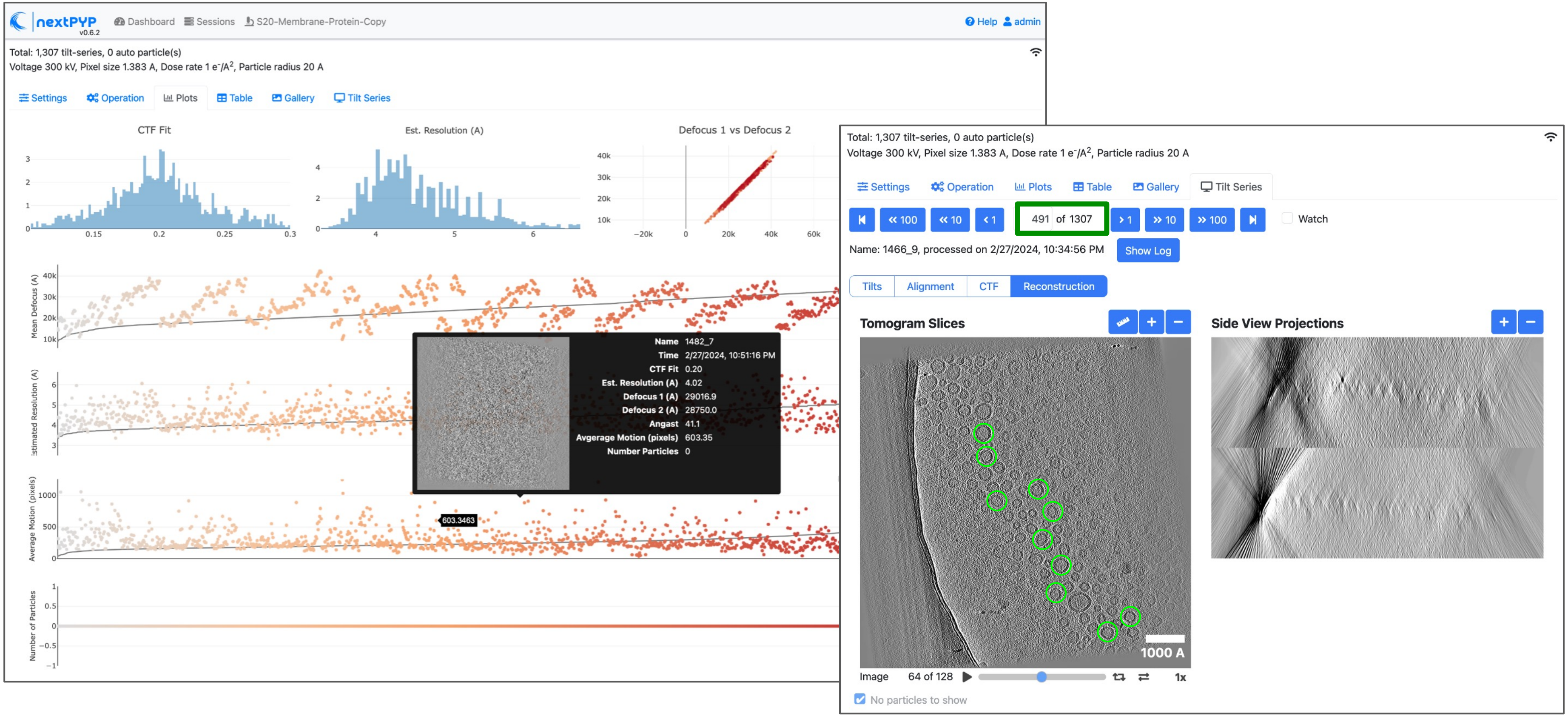
Scalable architecture for SP-CET

Validation using tilt-series of apoferritin



1.8 Å resolution map (32k particles, 100 tilt series, EMPIAR-11273)

On-the-fly pre-processing



Total: 1,307 tilt-series, 0 auto particle(s)
Voltage 300 kV, Pixel size 1.383 A, Dose rate 1 e^-/A^2, Particle radius 20 A

Name 1482_7
Time 2/27/2024, 10:51:16 PM
CTF Fit 0.20
Est. Resolution (A) 4.02
Defocus 1 (A) 29016.9
Defocus 2 (A) 28750.0
Angast 41.1
Avgerage Motion (pixels) 603.35
Number Particles 0

Total: 1,307 tilt-series, 0 auto particle(s)
Voltage 300 kV, Pixel size 1.383 A, Dose rate 1 e^-/A^2, Particle radius 20 A

491 of 1307

Name: 1466_9, processed on 2/27/2024, 10:34:56 PM

Image 64 of 128

☑ No particles to show

Other notable features

- Easy to install and configure
- Standalone or cluster mode (CPU or GPU)
- Provides easy access to a range of external tools:
  - cryoCARE, IsoNet, tomoDRGN, Topaz, AreTomo2, MotionCor3, membrain-seg, etc.
- Interoperable with other tomography packages

Version history

Oct. 23’	…	Apr. 24’	Apr. 25’
v0.5.0	…	v0.6.5	v0.7.0

What’s new in 0.7.0

- New dedicated blocks for particle picking
- New blocks for ML-based operations (denosing, picking, heterogeneity)
- 2x faster refinement, 50% reduction in storage
- Easier to install, maintain, and configure (SLURM templates, OS users)
- Better compatibility with external packages
- Updated documentation and many other bug fixes and UI improvements

nextPYP

Introduction to practical

Bartesaghi Lab

Duke
UNIVERSITY

Acknowledgments

Duke
OIT

Research sponsors

National Institute
of General Medical
Sciences

CHAN
ZUCKERBERG
INITIATIVE

NIEHS

NIH
Office of AIDS Research

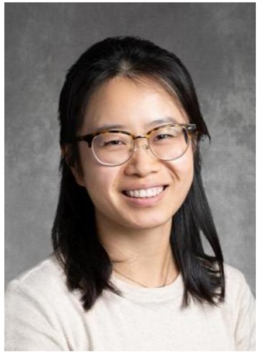
DUKE
CENTER FOR HIV
STRUCTURAL BIOLOGY

R01-GM141223, CZI-2021-234602, OAR-NIEHS-P00258, U54-AI170752

The nextPYP team


Jeff Martin


Abby Watson


Laura He

Former contributors: Y. Zhou, H. Liu, Q. Huang, W. Jin, X. Du, J. Piland, and J. Mandel.

Setup

Web-access to instances running in the Bartesaghi Lab

- Compute resources

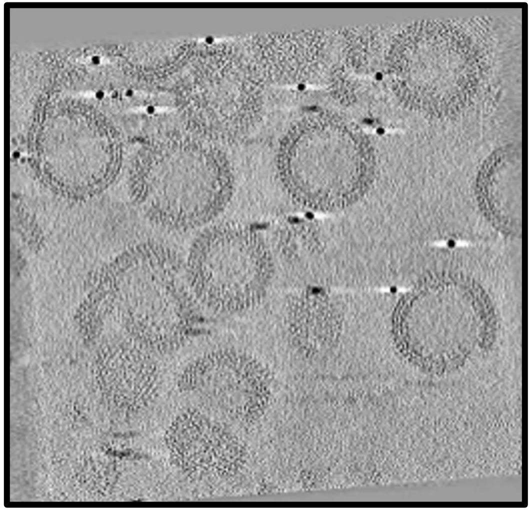
CPU cores	RAM (GB/core)	SSDs (GB/core)
372	4	45

- We won’t run any operations requiring GPUs
- We will use pre-calculated initial alignments (time constraints)
- Access will remain active until Sunday 4/6 9PM EST

Datasets

Dataset 1

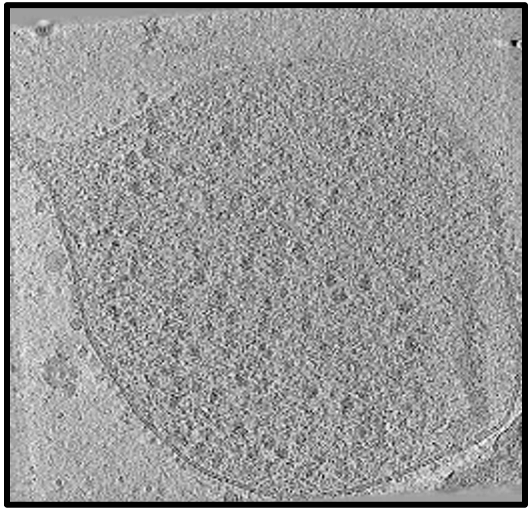
Immature HIV-1 Gag



EMPIAR-10164

Dataset 2

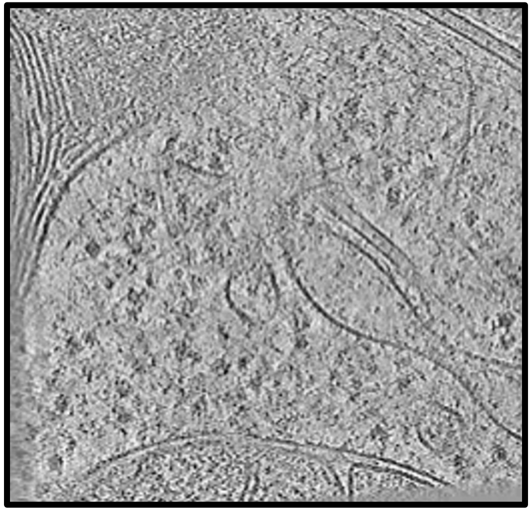
Whole bacterial cells



EMPIAR-10499

Dataset 3

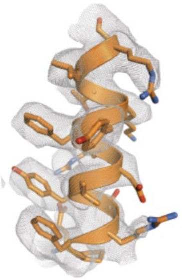
FIB-milled cells



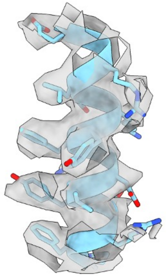
EMPIAR-10987

Goal of practical: structure of HIV-1 CA-SP1

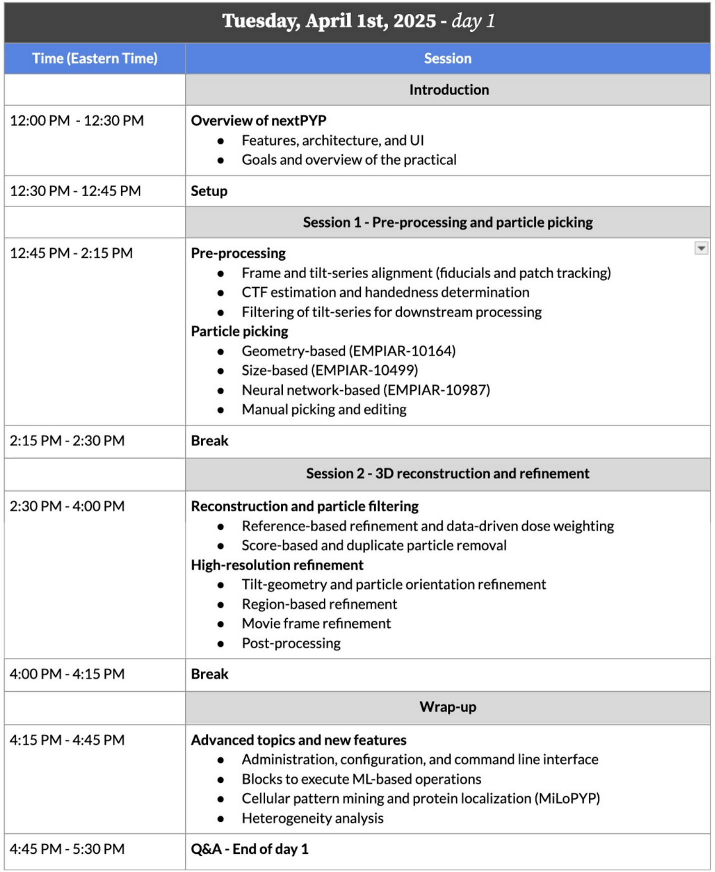
EMPIAR-10164	Schur et al., Science, 2016	nextPYP practical
Tilt-series	43	3
Particles	64,340	5,000
Resolution	3.9 Å	~3.9 Å



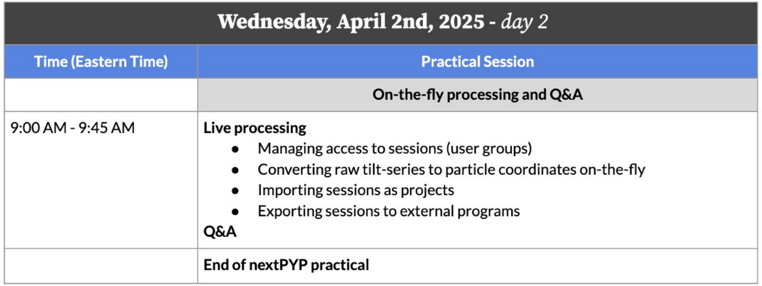
EMD-4015



Agenda



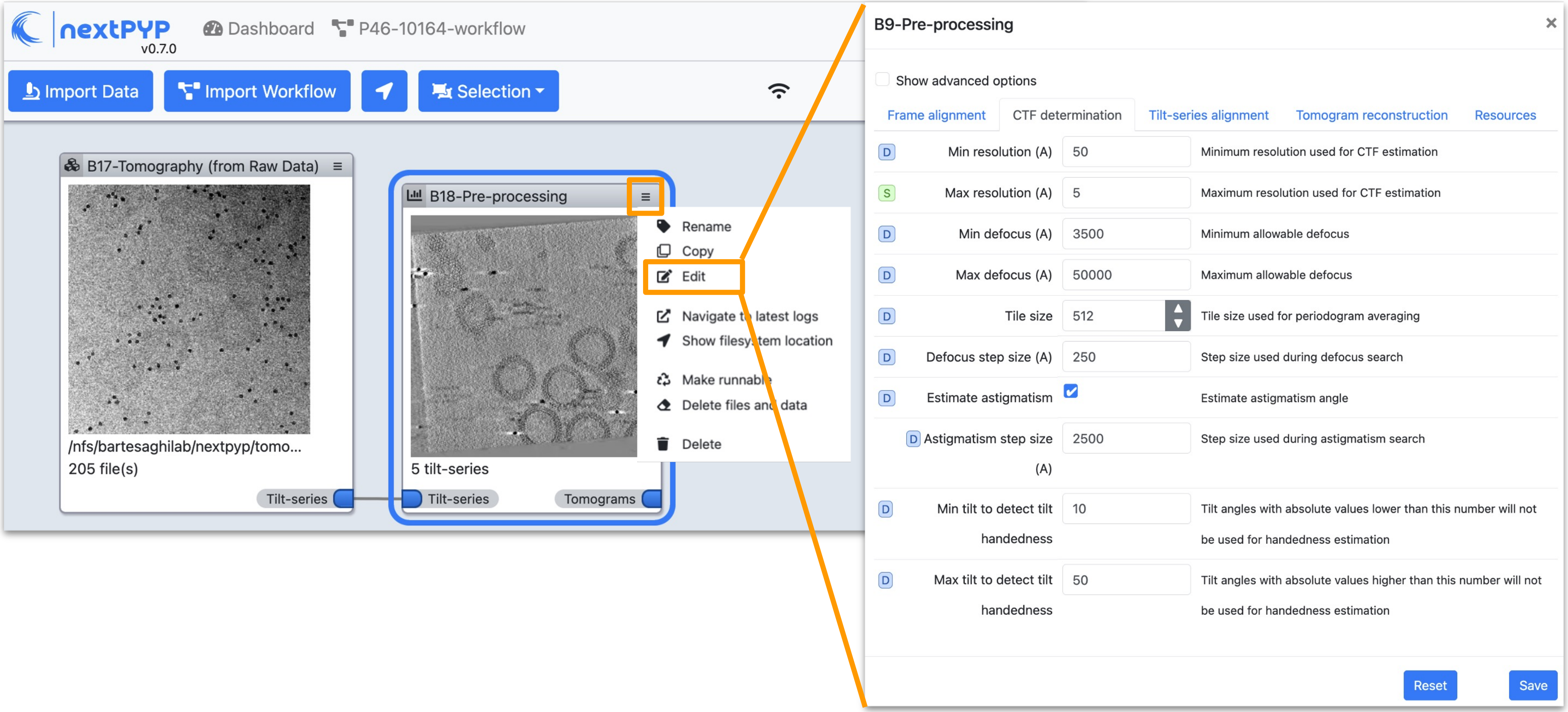
Tuesday, April 1st, 2025 - *day 1*	
Time (Eastern Time)	**Session**
	Introduction
12:00 PM - 12:30 PM	**Overview of nextPYP** • Features, architecture, and UI • Goals and overview of the practical
12:30 PM - 12:45 PM	**Setup**
	Session 1 - Pre-processing and particle picking
12:45 PM - 2:15 PM	**Pre-processing** • Frame and tilt-series alignment (fiducials and patch tracking) • CTF estimation and handedness determination • Filtering of tilt-series for downstream processing **Particle picking** • Geometry-based (EMPIAR-10164) • Size-based (EMPIAR-10499) • Neural network-based (EMPIAR-10987) • Manual picking and editing
2:15 PM - 2:30 PM	**Break**
	Session 2 - 3D reconstruction and refinement
2:30 PM - 4:00 PM	**Reconstruction and particle filtering** • Reference-based refinement and data-driven dose weighting • Score-based and duplicate particle removal **High-resolution refinement** • Tilt-geometry and particle orientation refinement • Region-based refinement • Movie frame refinement • Post-processing
4:00 PM - 4:15 PM	**Break**
	Wrap-up
4:15 PM - 4:45 PM	**Advanced topics and new features** • Administration, configuration, and command line interface • Blocks to execute ML-based operations • Cellular pattern mining and protein localization (MiLoPYP) • Heterogeneity analysis
4:45 PM - 5:30 PM	**Q&A - End of day 1**



Wednesday, April 2nd, 2025 - *day 2*	
Time (Eastern Time)	**Practical Session**
	On-the-fly processing and Q&A
9:00 AM - 9:45 AM	**Live processing** • Managing access to sessions (user groups) • Converting raw tilt-series to particle coordinates on-the-fly • Importing sessions as projects • Exporting sessions to external programs **Q&A**
	End of nextPYP practical

- **Movie-frame alignment**
  - Unblur
  - Motioncor3
- **Tilt-series alignment**
  - IMOD (fiducials, patch tracking)
  - AreTomo2
- **CTF determination**
  - Per-tilt estimation
  - Handedness determination
- **Tomogram reconstruction**
  - WBP, fakeSIRT (IMOD)
  - WBP/SART (AreTomo2)
  - Erase gold fiducials

Session 1 Pre-processing

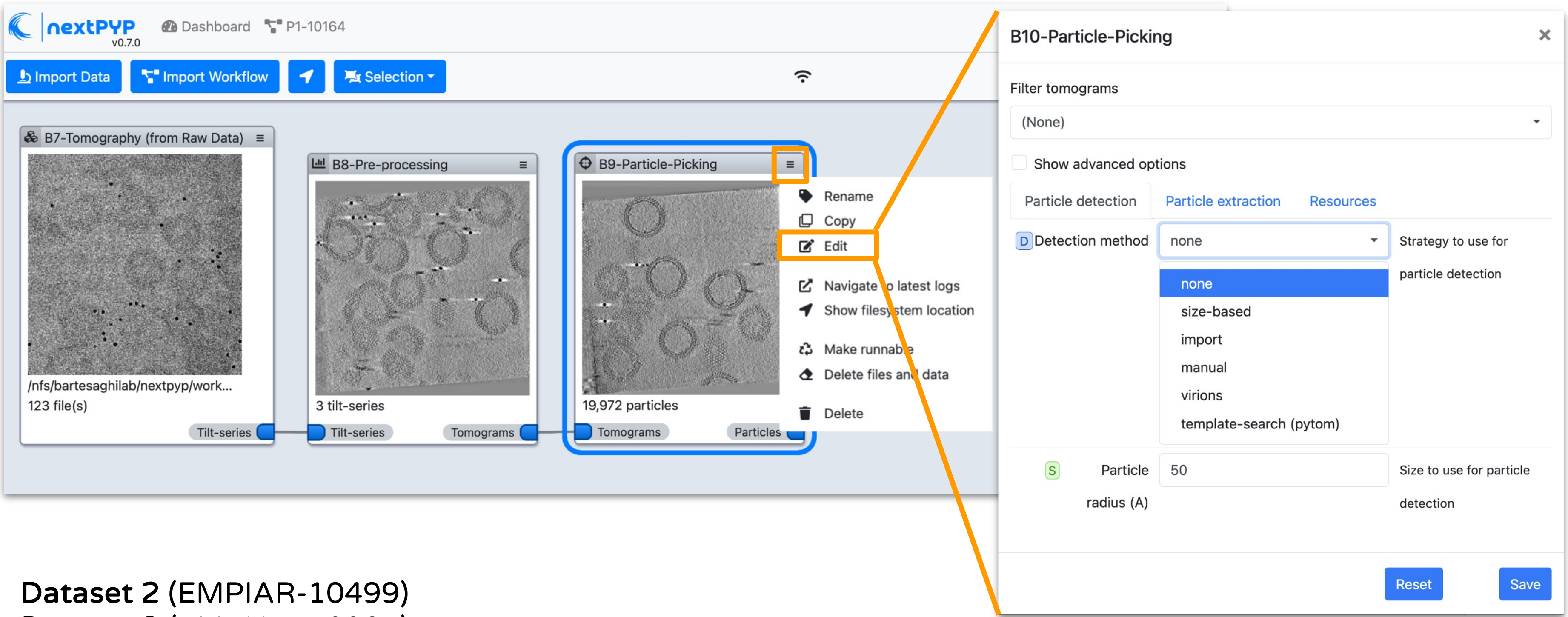


☐ Show advanced options

D	Min resolution (A)	50	Minimum resolution used for CTF estimation
S	Max resolution (A)	5	Maximum resolution used for CTF estimation
D	Min defocus (A)	3500	Minimum allowable defocus
D	Max defocus (A)	50000	Maximum allowable defocus
D	Tile size	512	Tile size used for periodogram averaging
D	Defocus step size (A)	250	Step size used during defocus search
D	Estimate astigmatism	☑	Estimate astigmatism angle
D	Astigmatism step size (A)	2500	Step size used during astigmatism search
D	Min tilt to detect tilt handedness	10	Tilt angles with absolute values lower than this number will not be used for handedness estimation
D	Max tilt to detect tilt handedness	50	Tilt angles with absolute values higher than this number will not be used for handedness estimation

- **Geometry-based**
  - Surface segmentation
  - Constrained picking
- **Size-based**
  - Suitable for globular proteins
- **Neural network-based**
  - Sparse labels
  - Training on GPU
- **Manual**
  - User-friendly
  - Scalable
- **Import coordinates**
  - From other programs
  - IMOD format
- **Template-search**
  - pytom-match-pick
  - Requires GPU to run

Session 1 › Particle picking › Size-based & NN-based



/nfs/bartesaghilab/nextpyp/work...
123 file(s)

3 tilt-series

19,972 particles

B10-Particle-Picking

Filter tomograms: (None)

☐ Show advanced options

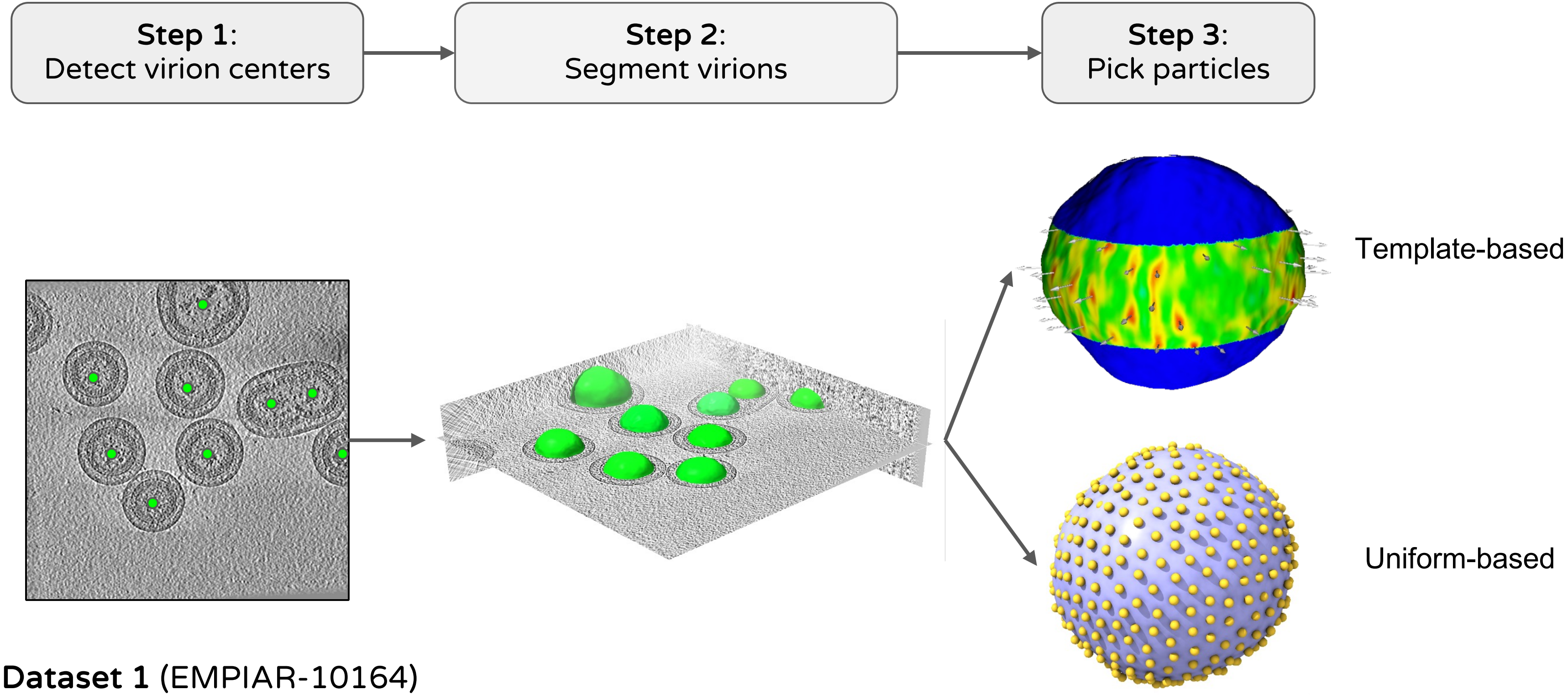
Detection method: none — Strategy to use for particle detection

- none
- size-based
- import
- manual
- virions
- template-search (pytom)

Particle radius (A): 50 — Size to use for particle detection

Dataset 2 (EMPIAR-10499)
Dataset 3 (EMPIAR-10987)

Session 1 › Particle picking › Geometry-based



Step 1:
Detect virion centers

Step 2:
Segment virions

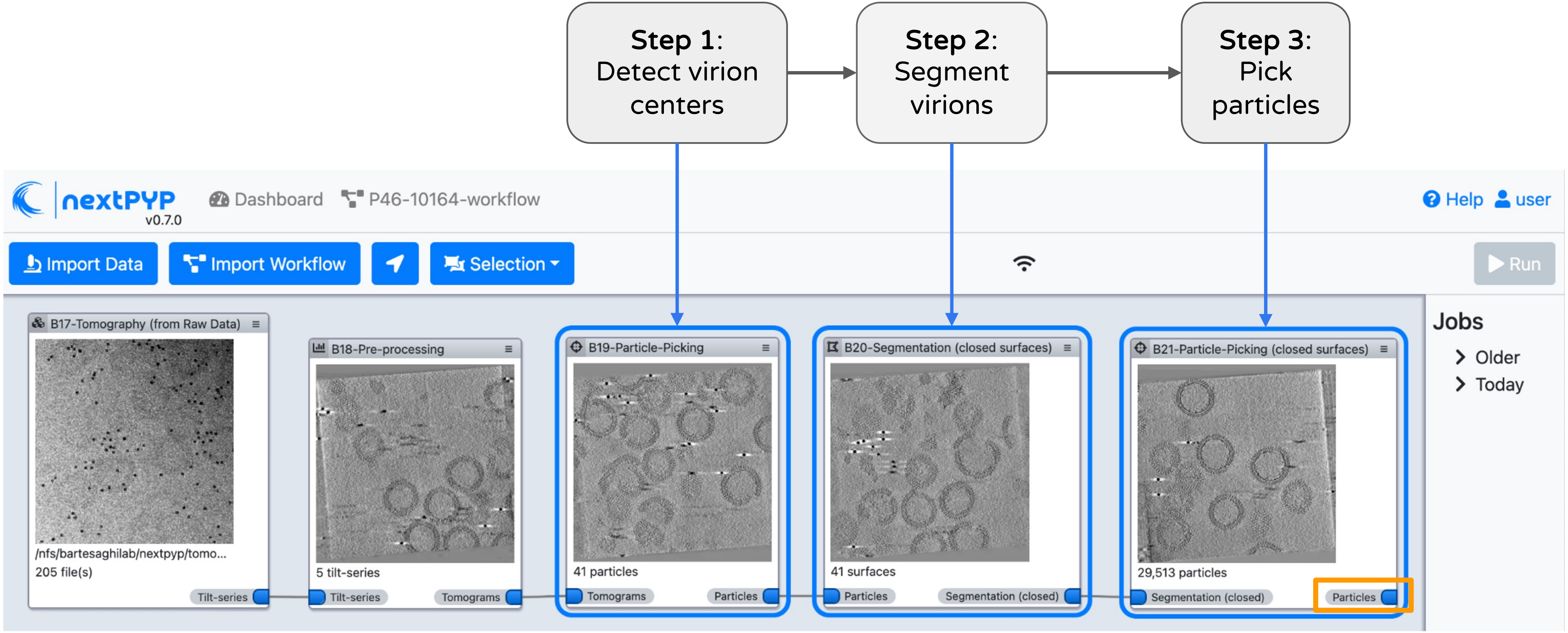
Step 3:
Pick particles

Template-based

Uniform-based

Dataset 1 (EMPIAR-10164)

Session 1 › Particle picking › Geometry-based

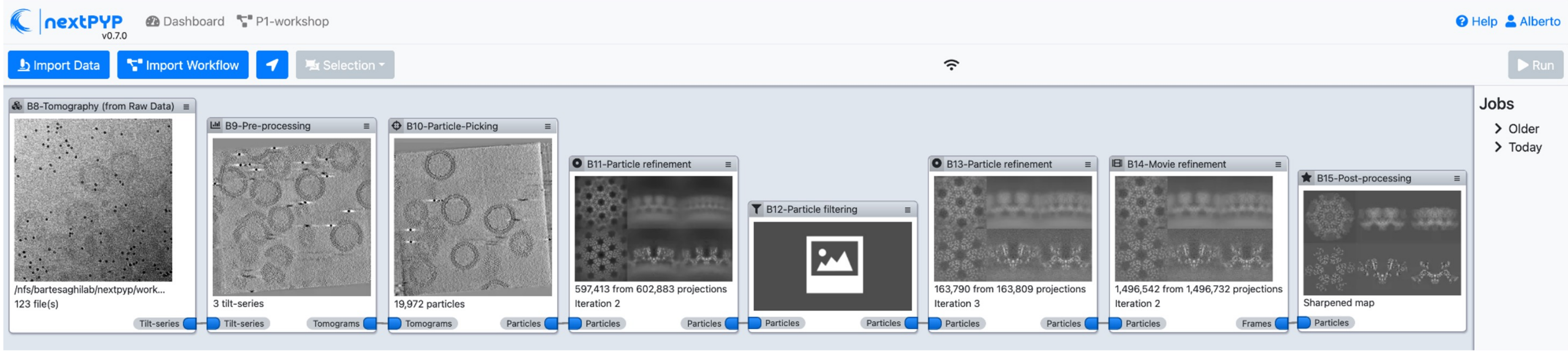


Step 1: Detect virion centers

Step 2: Segment virions

Step 3: Pick particles

Session 2 — 3D refinement - Overview



B8-Tomography (from Raw Data)
/nfs/bartesaghilab/nextpyp/work...
123 file(s)
Tilt-series

B9-Pre-processing
3 tilt-series
Tilt-series
Tomograms

B10-Particle-Picking
19,972 particles
Tomograms
Particles

B11-Particle refinement
597,413 from 602,883 projections
Iteration 2
Particles
Particles

B12-Particle filtering
Particles
Particles

B13-Particle refinement
163,790 from 163,809 projections
Iteration 3
Particles
Particles

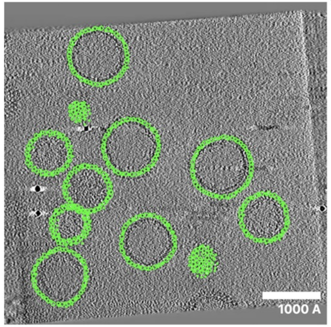
B14-Movie refinement
1,496,542 from 1,496,732 projections
Iteration 2
Particles
Frames

B15-Post-processing
Sharpened map
Particles

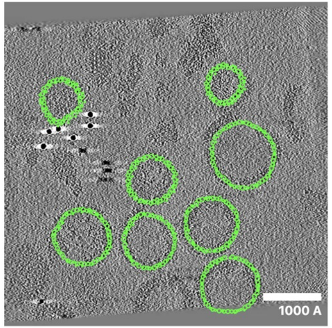
Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
Import coordinates	Import alignments	Particle filtering	Region-based refinement	Movie-frame Refinement	Post-Processing

Import/export particle coordinates

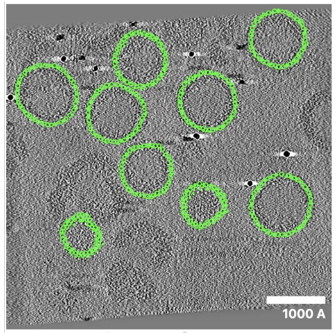
- 19,972 particles exported in IMOD format (*.spk files)
- Import coordinates to ensure consistency during refinement



TS_01



TS_43

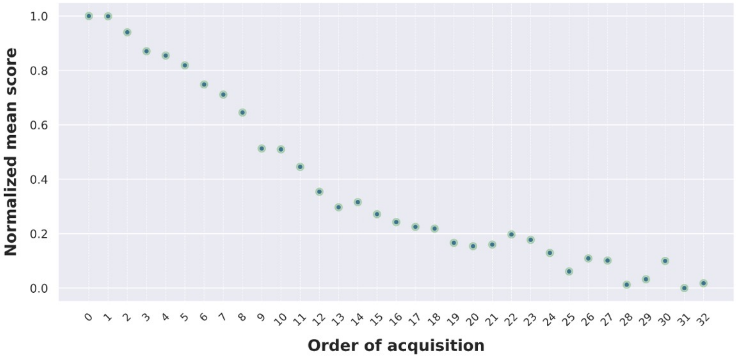


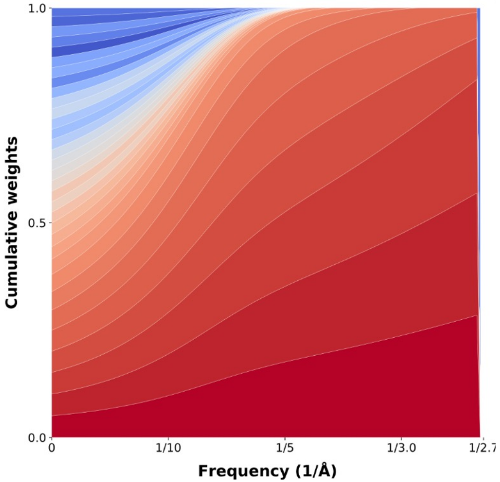
TS_45

~6,500 particles
per tilt-series

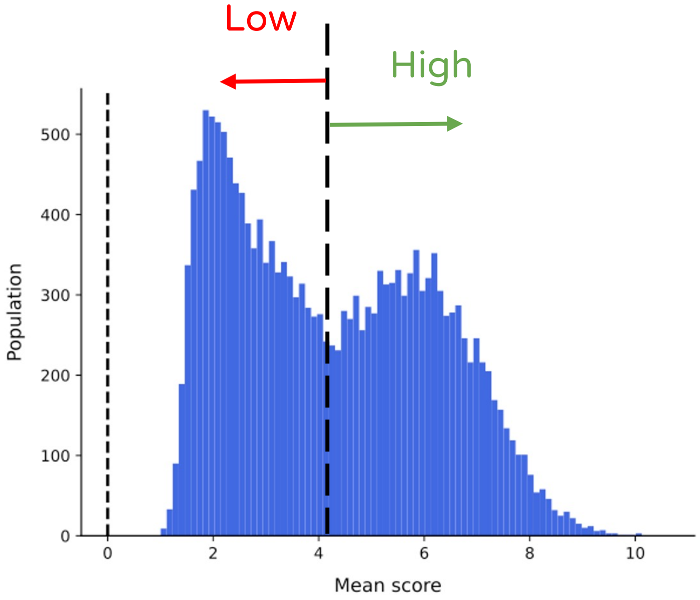
Import alignments from reference-based refinement

- Create 3D reconstruction from particle projections
- Data-driven dose-weighting





- Remove particles with low scores
- Remove duplicates

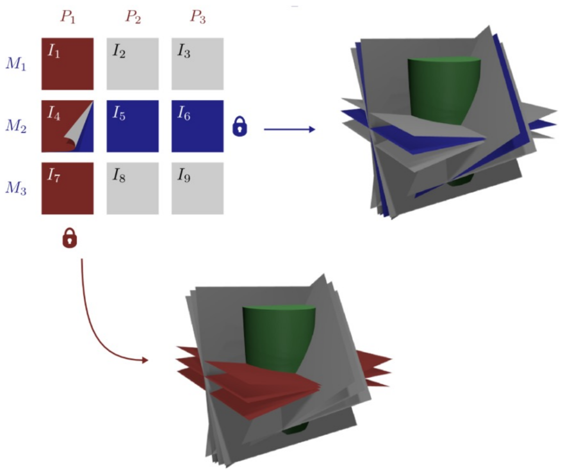


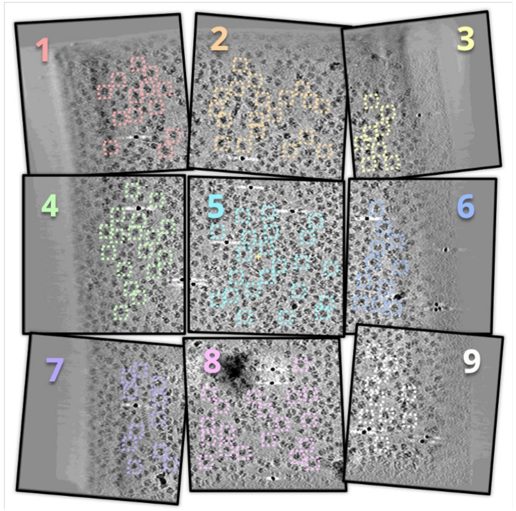
Zhou, Y.et al. (2019) **Unsupervised particle sorting for high-resolution single-particle cryo-EM**, *Inverse Problems,* 36, 044002.

Constrained refinement

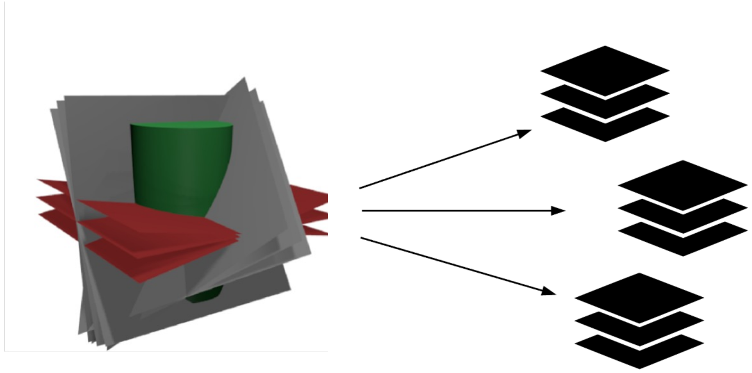
- Refine particle orientations
- Refine tilt-geometry using particle data

- Region-based refinement

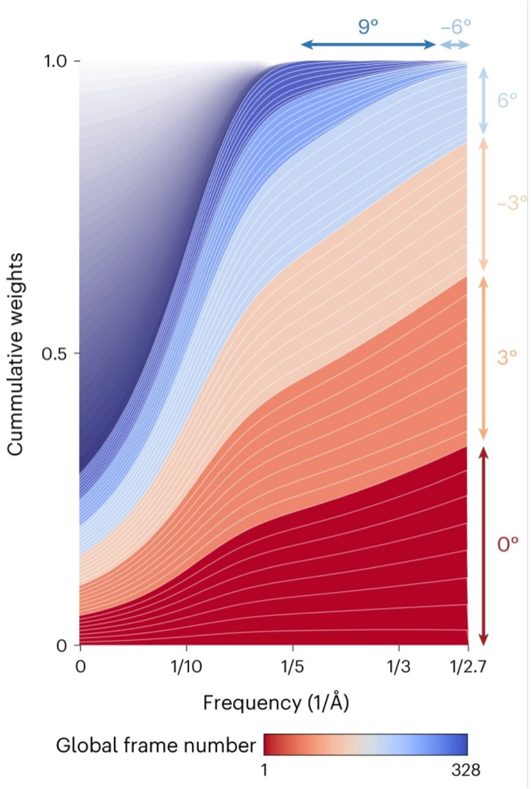




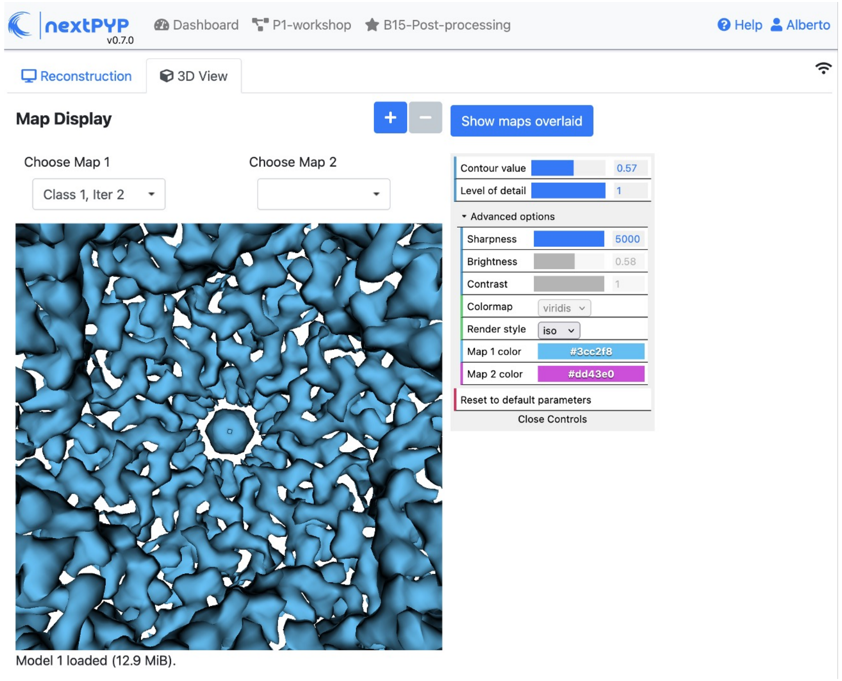
Session 2 › Step 5 › Movie-frame refinement



- 5,000 particles
- 41 tilts x 8 frames = 328 projections/particle
- ~1.5M particle projections



Session 2 Step 6 Post-processing



Choose Map 1: Class 1, Iter 2

Choose Map 2:

Contour value: 0.57

Level of detail: 1

Sharpness: 5000

Brightness: 0.58

Contrast: 1

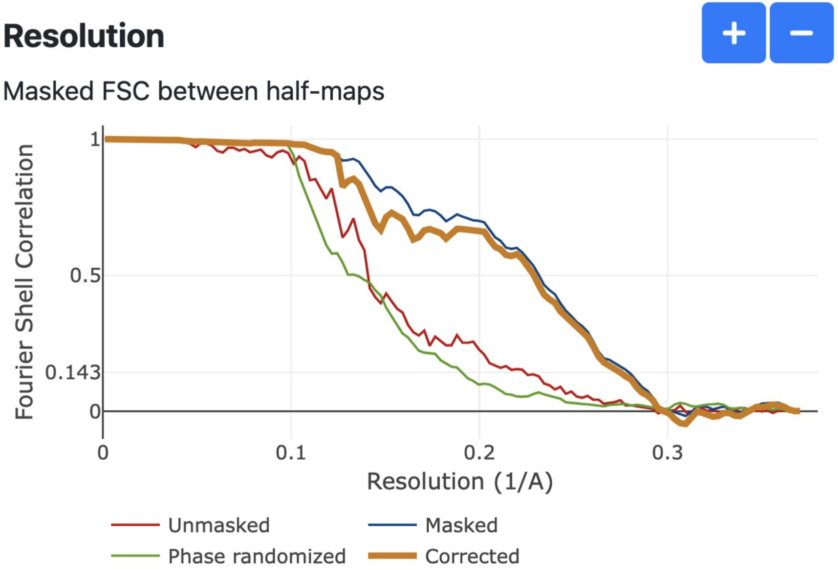
Colormap: viridis

Render style: iso

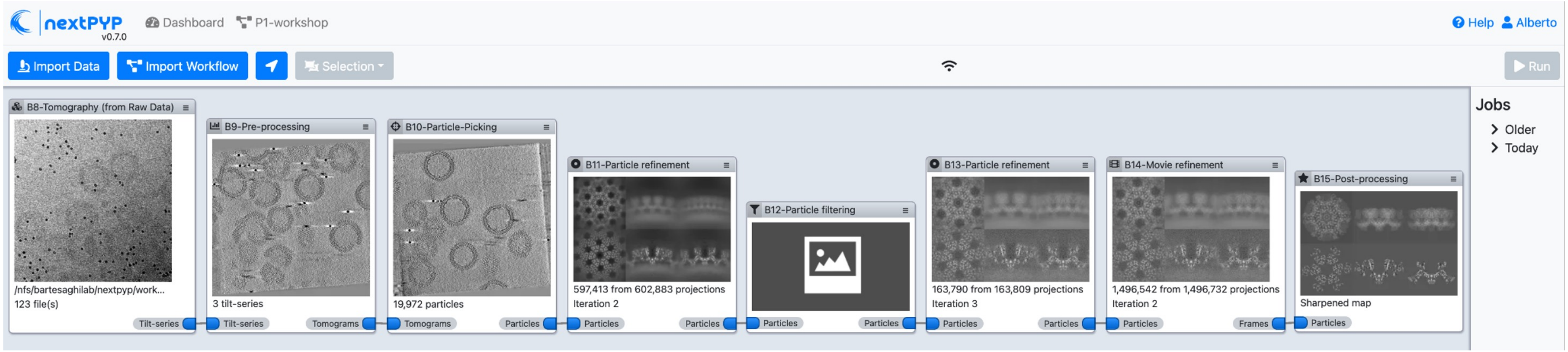
Map 1 color: #3cc2f8

Map 2 color: #dd43e0

Model 1 loaded (12.9 MiB).



Session 2 Overview



B8-Tomography (from Raw Data)
/nfs/bartesaghilab/nextpyp/work...
123 file(s)

B9-Pre-processing
3 tilt-series

B10-Particle-Picking
19,972 particles

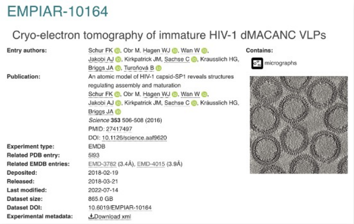
B11-Particle refinement
597,413 from 602,883 projections
Iteration 2

B12-Particle filtering

B13-Particle refinement
163,790 from 163,809 projections
Iteration 3

B14-Movie refinement
1,496,542 from 1,496,732 projections
Iteration 2

B15-Post-processing
Sharpened map



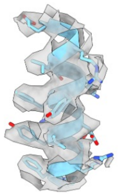
EMPIAR-10164

Cryo-electron tomography of immature HIV-1 dMACANC VLPs

Entry authors:	Schur FK, Obr M, Hagen WJ, Wan W, Jakobi AJ, Kirkpatrick JM, Sachse C, Krausslich HG, Briggs JA, Turoňová B
Publication:	An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation Schur FK, Obr M, Hagen WJ, Wan W, Jakobi AJ, Kirkpatrick JM, Sachse C, Kräusslich HG, Briggs JA *Science* **353** 506-508 (2016) PMID: 27417497 DOI: 10.1126/science.aaf9620
Experiment type:	EMDB
Related PDB entry:	5l93
Related EMDB entries:	EMD-3782 (3.4Å), EMD-4015 (3.9Å)
Deposited:	2018-02-19
Released:	2018-03-21
Last modified:	2022-07-14
Dataset size:	865.0 GB
Dataset DOI:	10.6019/EMPIAR-10164
Experimental metadata:	Download xml

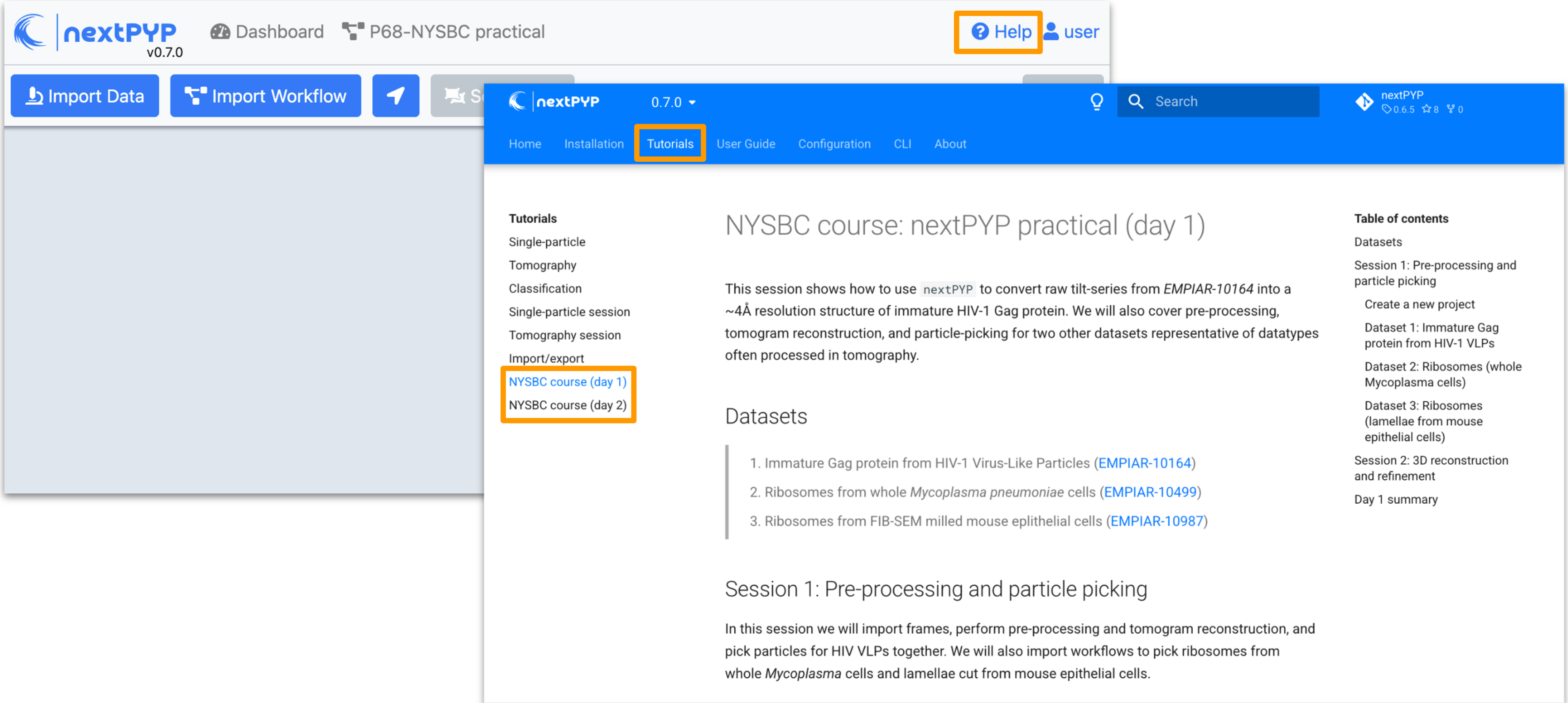
Contains: micrographs

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
Import coordinates	Import alignments	Particle filtering	Region-based refinement	Movie-frame Refinement	Post-Processing
1 min	3 min	5 min	20 min	20 min	1 min



~3.9 Å

Instructions for the practical



NYSBC course: nextPYP practical (day 1)

This session shows how to use `nextPYP` to convert raw tilt-series from *EMPIAR-10164* into a ~4Å resolution structure of immature HIV-1 Gag protein. We will also cover pre-processing, tomogram reconstruction, and particle-picking for two other datasets representative of datatypes often processed in tomography.

Datasets

> 1. Immature Gag protein from HIV-1 Virus-Like Particles (EMPIAR-10164)
> 2. Ribosomes from whole *Mycoplasma pneumoniae* cells (EMPIAR-10499)
> 3. Ribosomes from FIB-SEM milled mouse eplithelial cells (EMPIAR-10987)

Session 1: Pre-processing and particle picking

In this session we will import frames, perform pre-processing and tomogram reconstruction, and pick particles for HIV VLPs together. We will also import workflows to pick ribosomes from whole *Mycoplasma* cells and lamellae cut from mouse epithelial cells.

nextPYP

Q&A

Bartesaghi Lab

Duke
UNIVERSITY

nextPYP

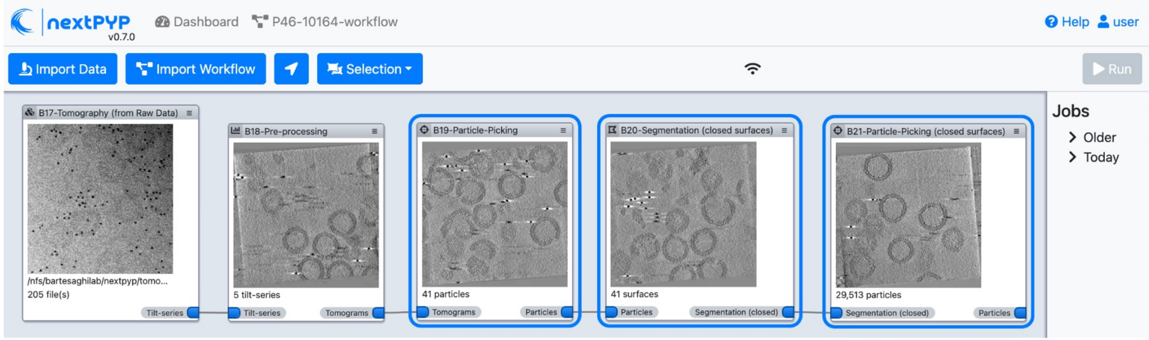
New features and advanced topics

Bartesaghi Lab

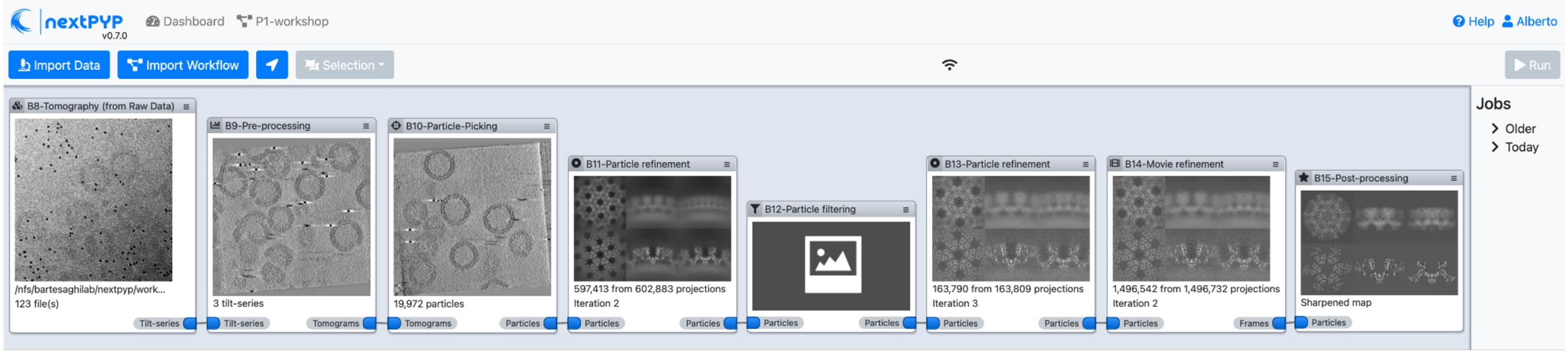
Duke UNIVERSITY

Recap

Session 1

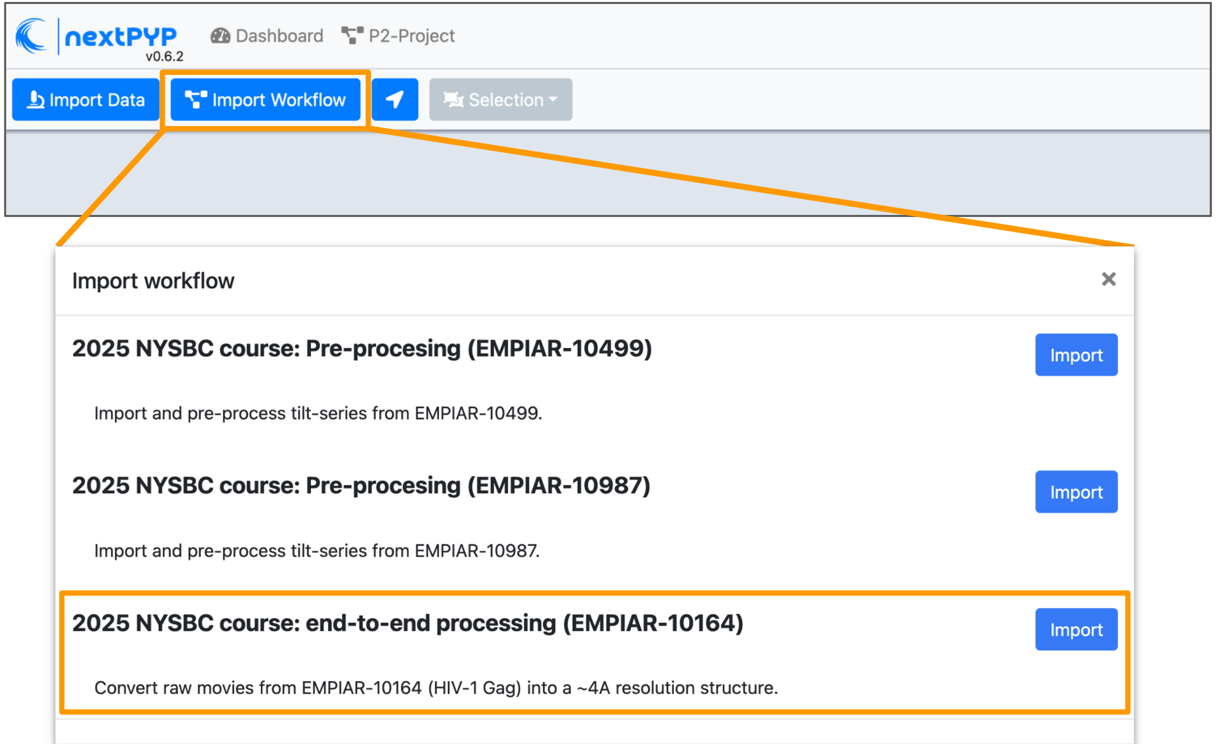


Session 2



Workflows

Pre-defined block sequences with pre-set parameters



Import workflow

2025 NYSBC course: Pre-procesing (EMPIAR-10499)

Import and pre-process tilt-series from EMPIAR-10499.

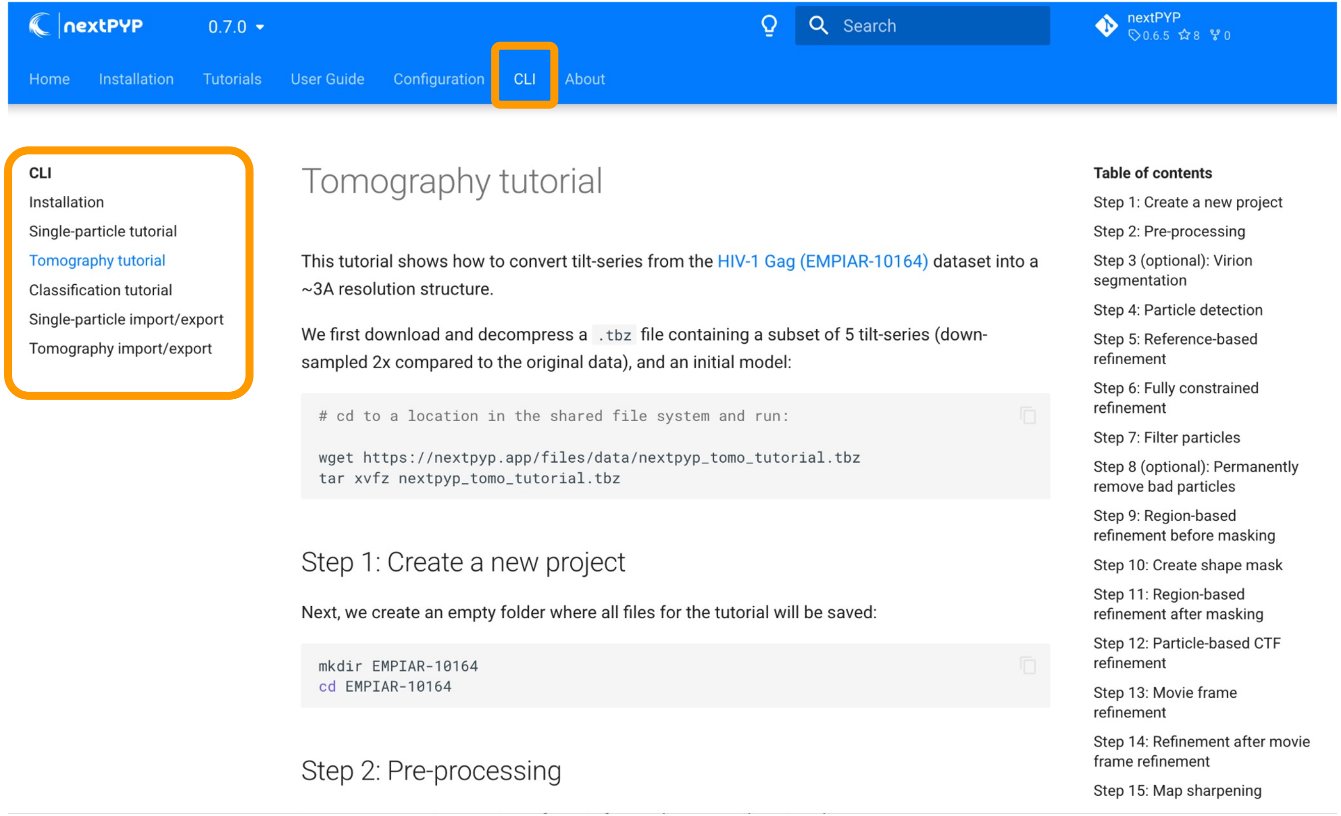
2025 NYSBC course: Pre-procesing (EMPIAR-10987)

Import and pre-process tilt-series from EMPIAR-10987.

2025 NYSBC course: end-to-end processing (EMPIAR-10164)

Convert raw movies from EMPIAR-10164 (HIV-1 Gag) into a ~4A resolution structure.

PYP: command line interface (CLI)



Tomography tutorial

This tutorial shows how to convert tilt-series from the HIV-1 Gag (EMPIAR-10164) dataset into a ~3A resolution structure.

We first download and decompress a `.tbz` file containing a subset of 5 tilt-series (down-sampled 2x compared to the original data), and an initial model:

```
# cd to a location in the shared file system and run:

wget https://nextpyp.app/files/data/nextpyp_tomo_tutorial.tbz
tar xvfz nextpyp_tomo_tutorial.tbz
```

Step 1: Create a new project

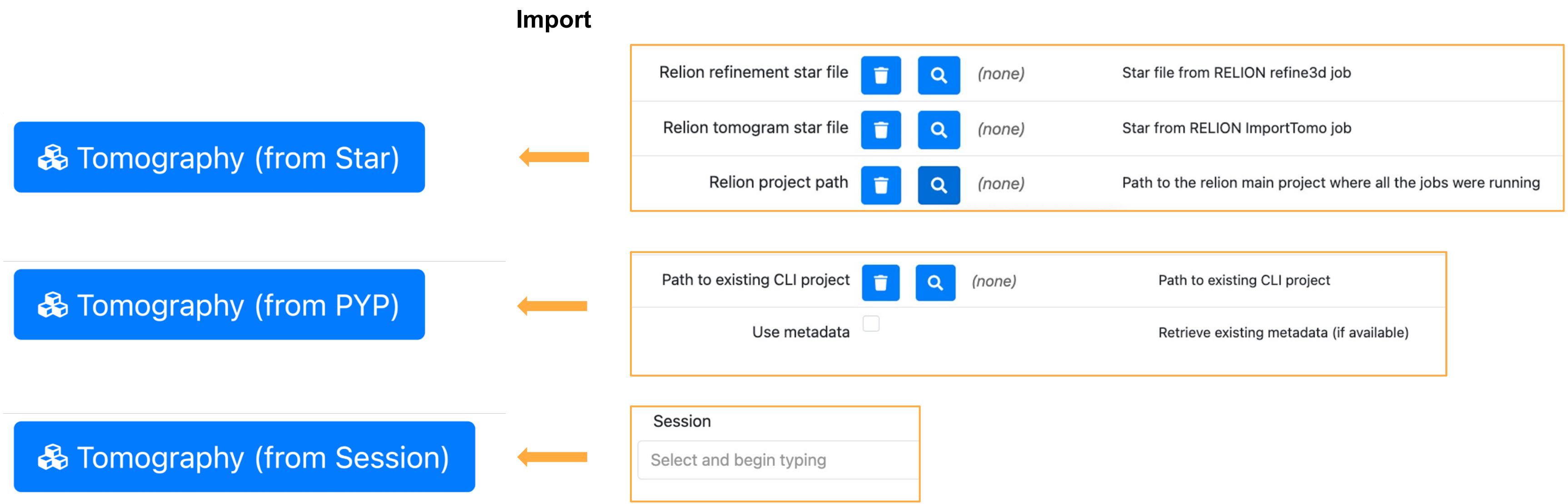
Next, we create an empty folder where all files for the tutorial will be saved:

```
mkdir EMPIAR-10164
cd EMPIAR-10164
```

Step 2: Pre-processing

Table of contents

Importing data into nextPYP



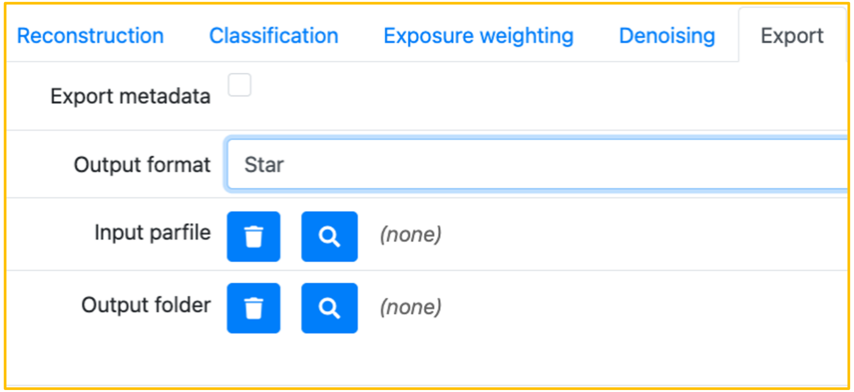
Relion refinement star file	*(none)*	Star file from RELION refine3d job
Relion tomogram star file	*(none)*	Star from RELION ImportTomo job
Relion project path	*(none)*	Path to the relion main project where all the jobs were running

Path to existing CLI project	*(none)*	Path to existing CLI project
Use metadata	☐	Retrieve existing metadata (if available)

Session

Select and begin typing

Exporting data to external programs



Export metadata ☐

Output format: Star

Input parfile: *(none)*

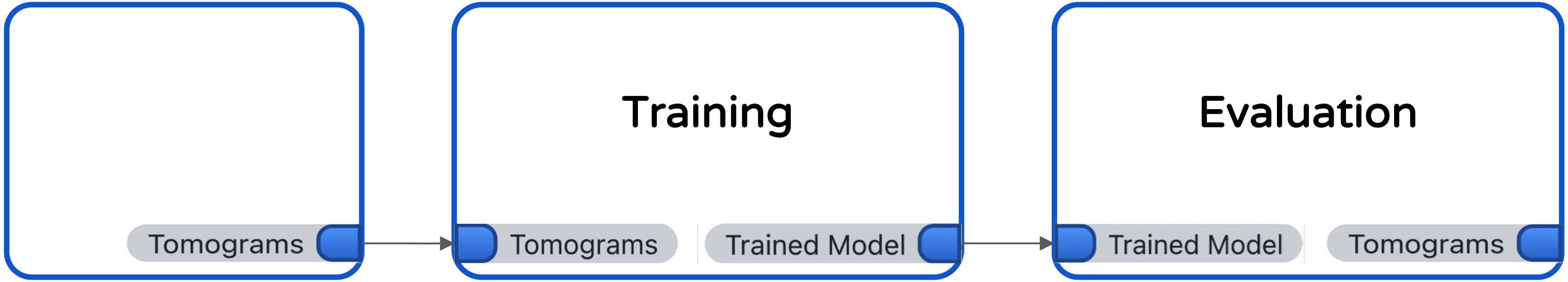
Output folder: *(none)*

nextPYP particle refinement block

Export →

```
tomograms.star
particles.star
project_folders/Movies
```

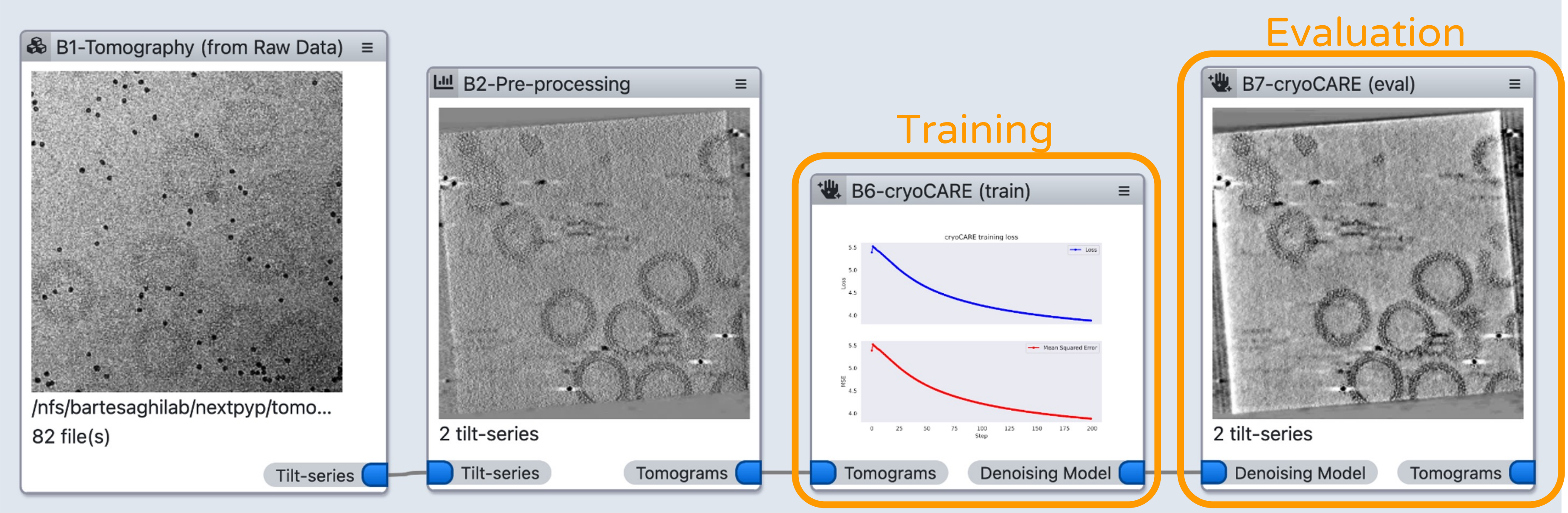
Blocks for ML-based operations



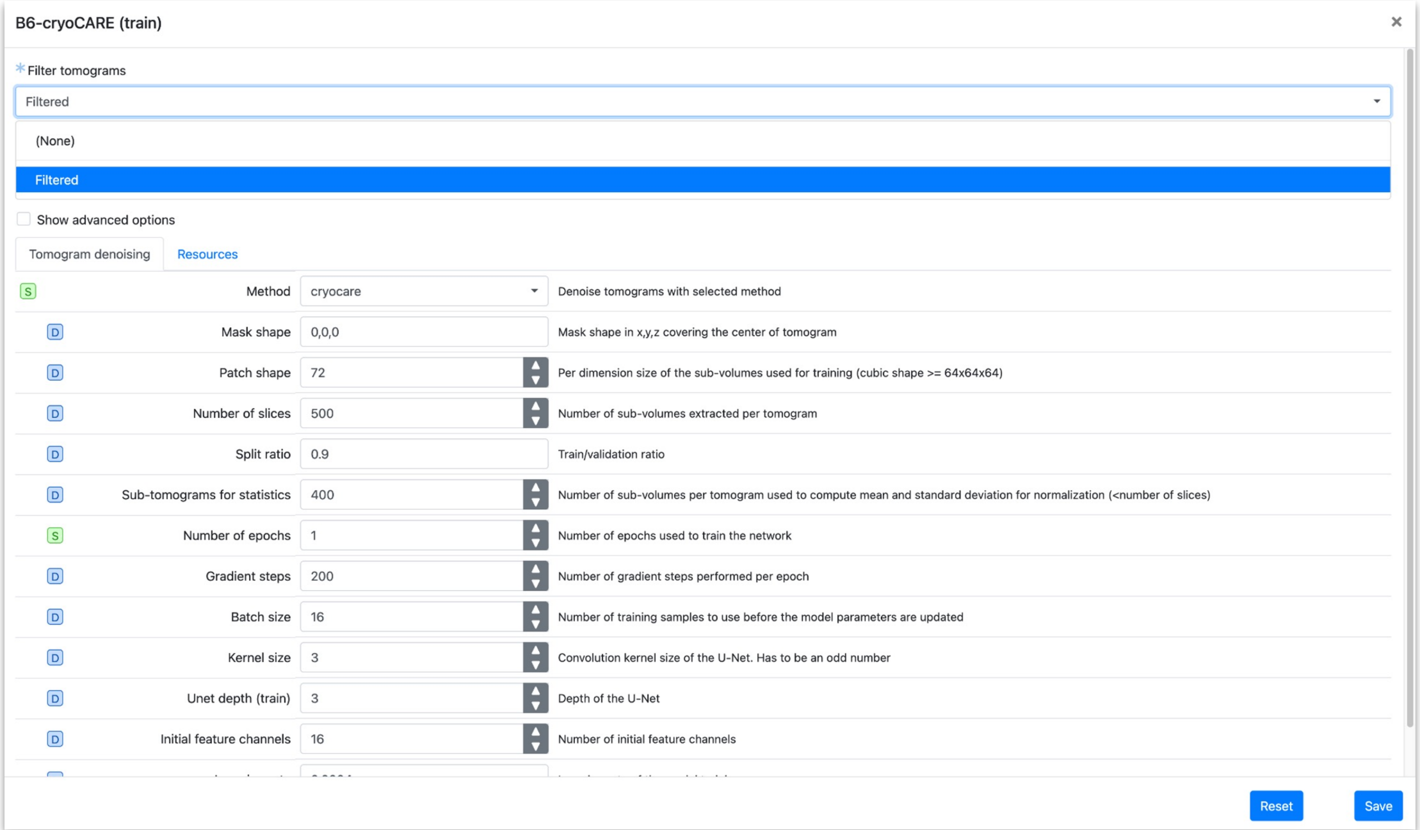
- Denoising (Topaz, IsoNet, cryoCARE)
- Particle picking
- Segmentation (membrain-seg, eval only)
- Continuous heterogeneity (tomoDRGN)

Example: NN-based tomogram denoising

cryoCARE



Setting training/evaluation parameters



B6-cryoCARE (train)

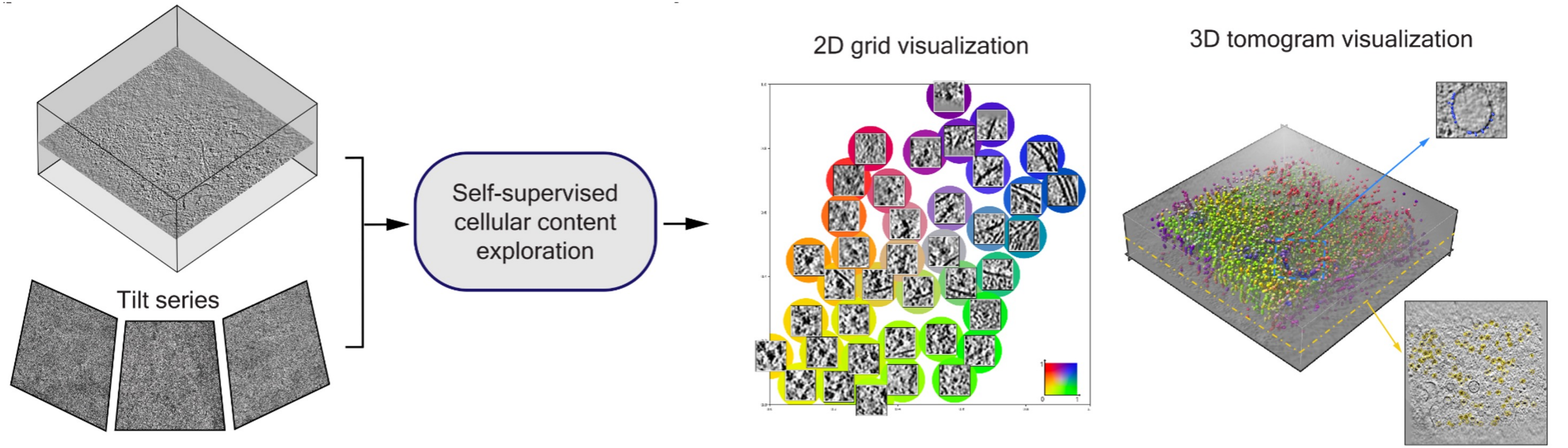
* Filter tomograms: Filtered

- (None)
- Filtered

☐ Show advanced options

S	Method	cryocare	Denoise tomograms with selected method
D	Mask shape	0,0,0	Mask shape in x,y,z covering the center of tomogram
D	Patch shape	72	Per dimension size of the sub-volumes used for training (cubic shape >= 64x64x64)
D	Number of slices	500	Number of sub-volumes extracted per tomogram
D	Split ratio	0.9	Train/validation ratio
D	Sub-tomograms for statistics	400	Number of sub-volumes per tomogram used to compute mean and standard deviation for normalization (<number of slices)
S	Number of epochs	1	Number of epochs used to train the network
D	Gradient steps	200	Number of gradient steps performed per epoch
D	Batch size	16	Number of training samples to use before the model parameters are updated
D	Kernel size	3	Convolution kernel size of the U-Net. Has to be an odd number
D	Unet depth (train)	3	Depth of the U-Net
D	Initial feature channels	16	Number of initial feature channels

MiLoPYP: cellular pattern mining



Tilt series

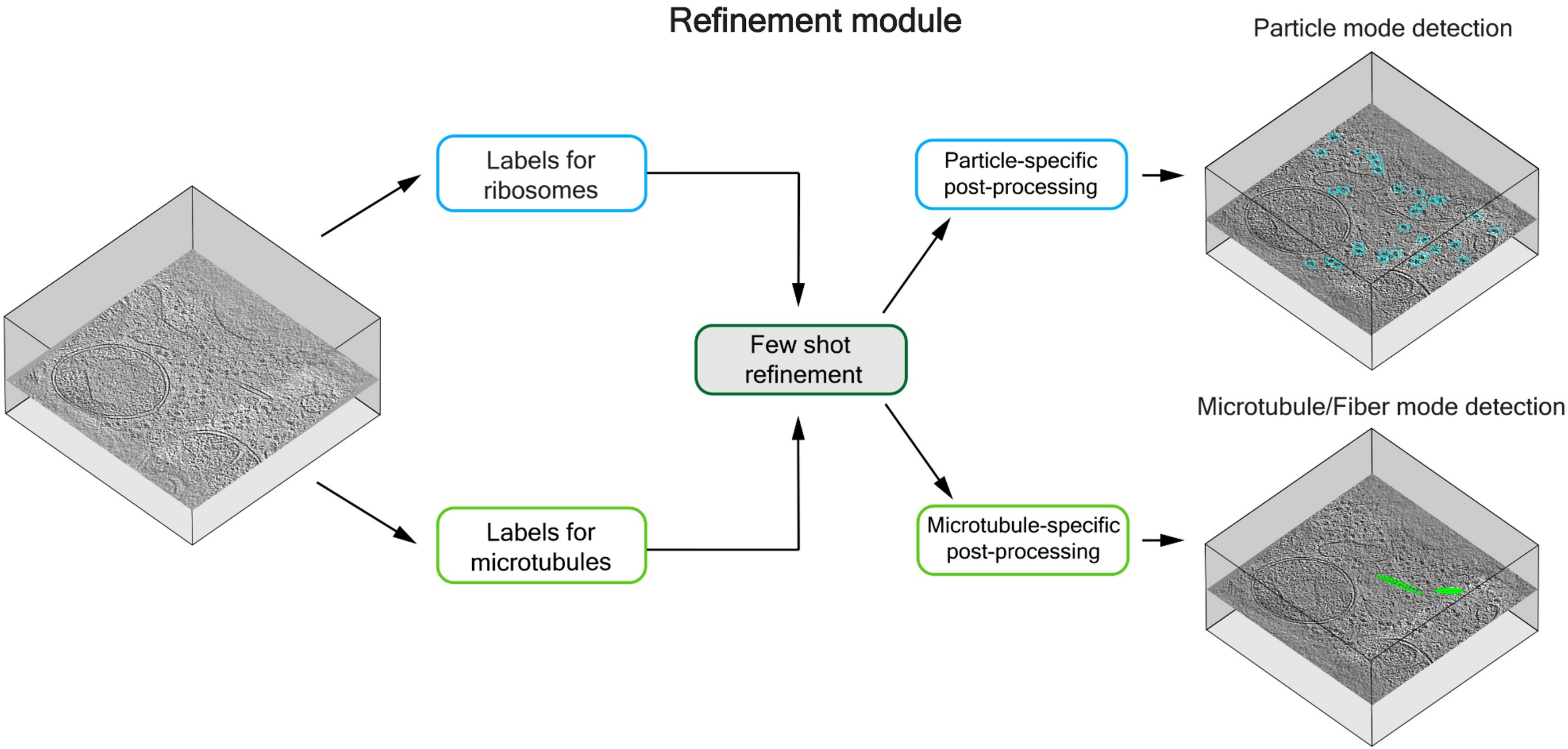
Self-supervised cellular content exploration

2D grid visualization

3D tomogram visualization

Huang, Q. et al. (2024) **MiLoPYP: self-supervised molecular pattern mining and particle localization in situ**. *Nat Meth*, 21:1863–1872.

MiLoPYP: particle refinement module



Refinement module

Labels for ribosomes

Labels for microtubules

Few shot refinement

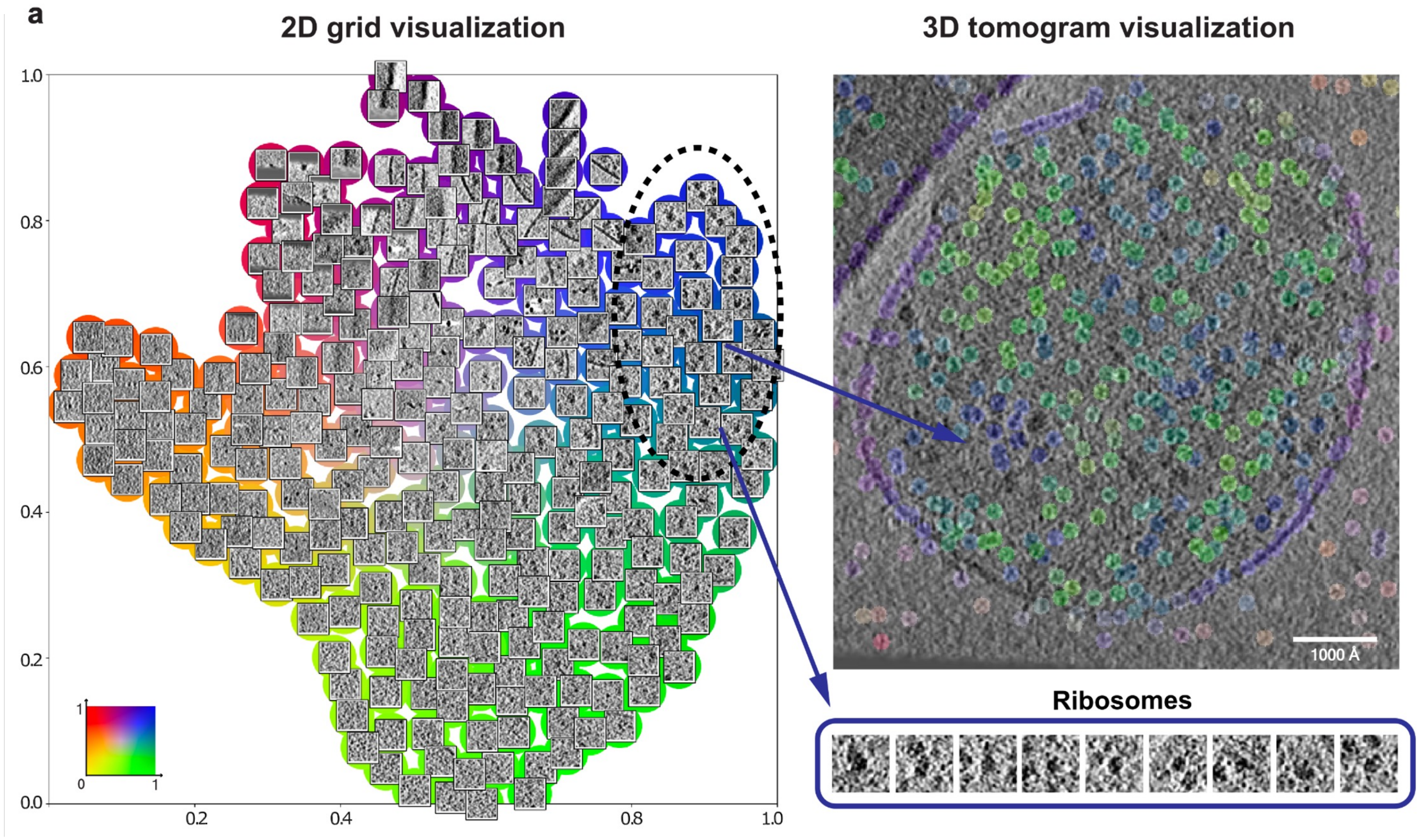
Particle-specific post-processing

Microtubule-specific post-processing

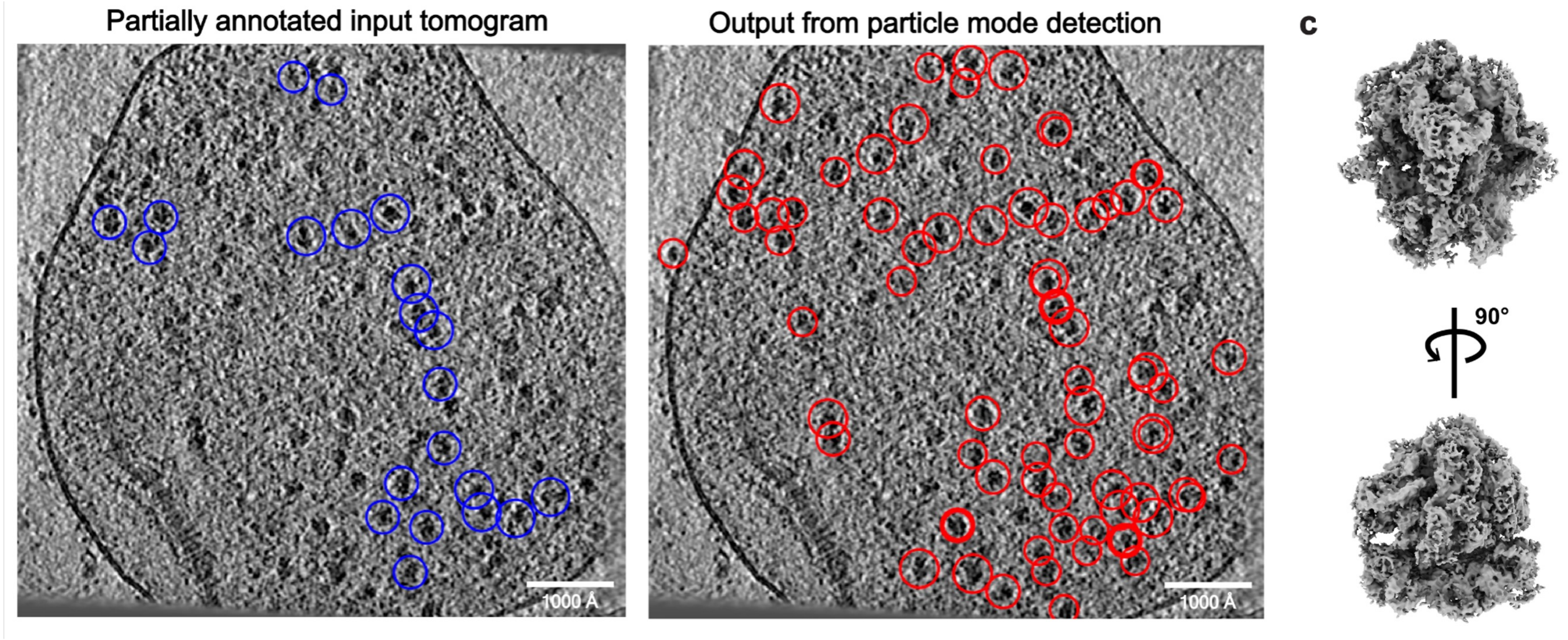
Particle mode detection

Microtubule/Fiber mode detection

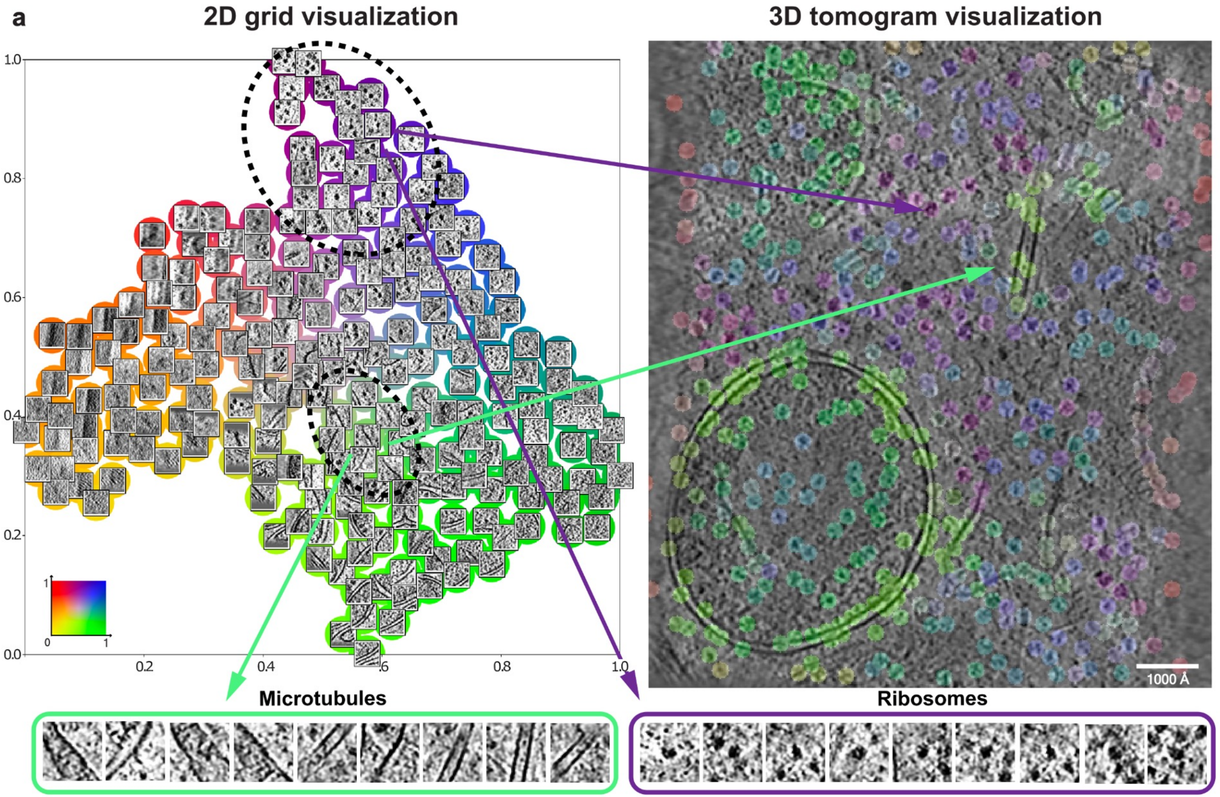
70S M. pneumoniae cells (EMPIAR-10499)



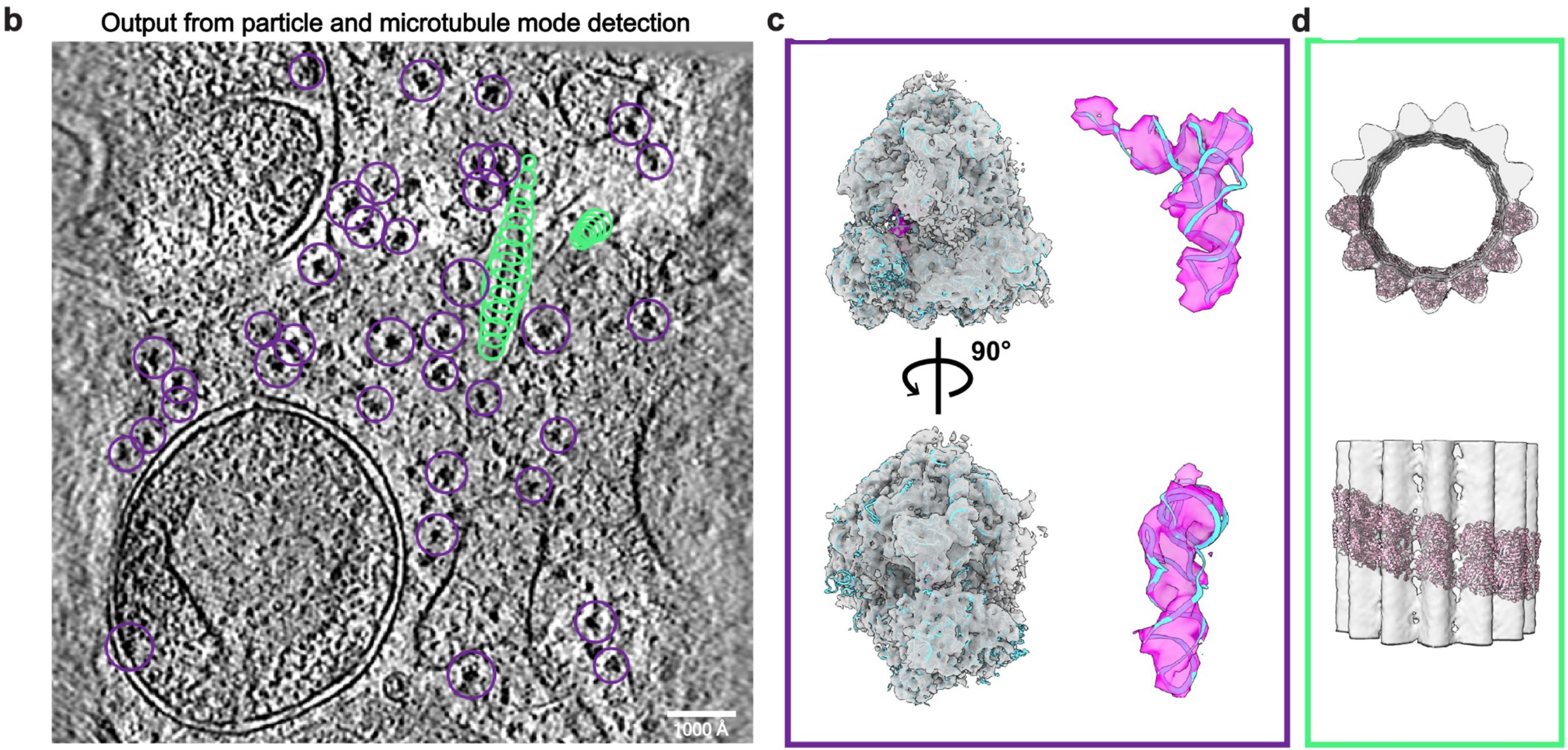
70S M. pneumoniae cells (EMPIAR-10499)



Mouse epithelial cells (EMPIAR-10987)



Detection and refinement of multiple species

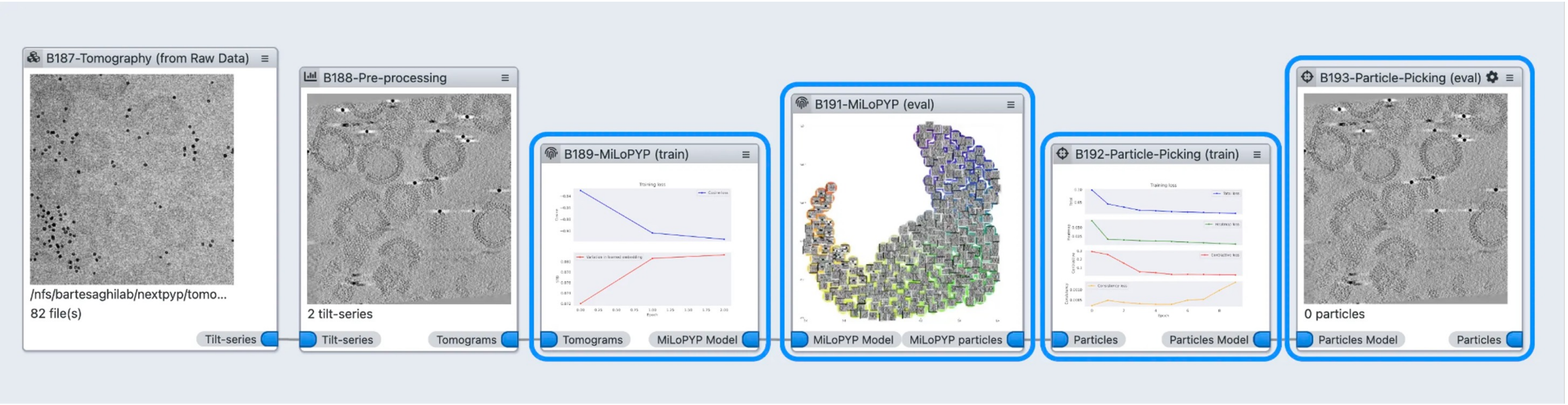


Blocks for running MiLoPYP

Exploration

Localization

Step 1: Train → Step 2: Eval → Step 3: Train → Step 4: Eval



B187-Tomography (from Raw Data)
/nfs/bartesaghilab/nextpyp/tomo...
82 file(s)
Tilt-series

B188-Pre-processing
2 tilt-series
Tilt-series
Tomograms

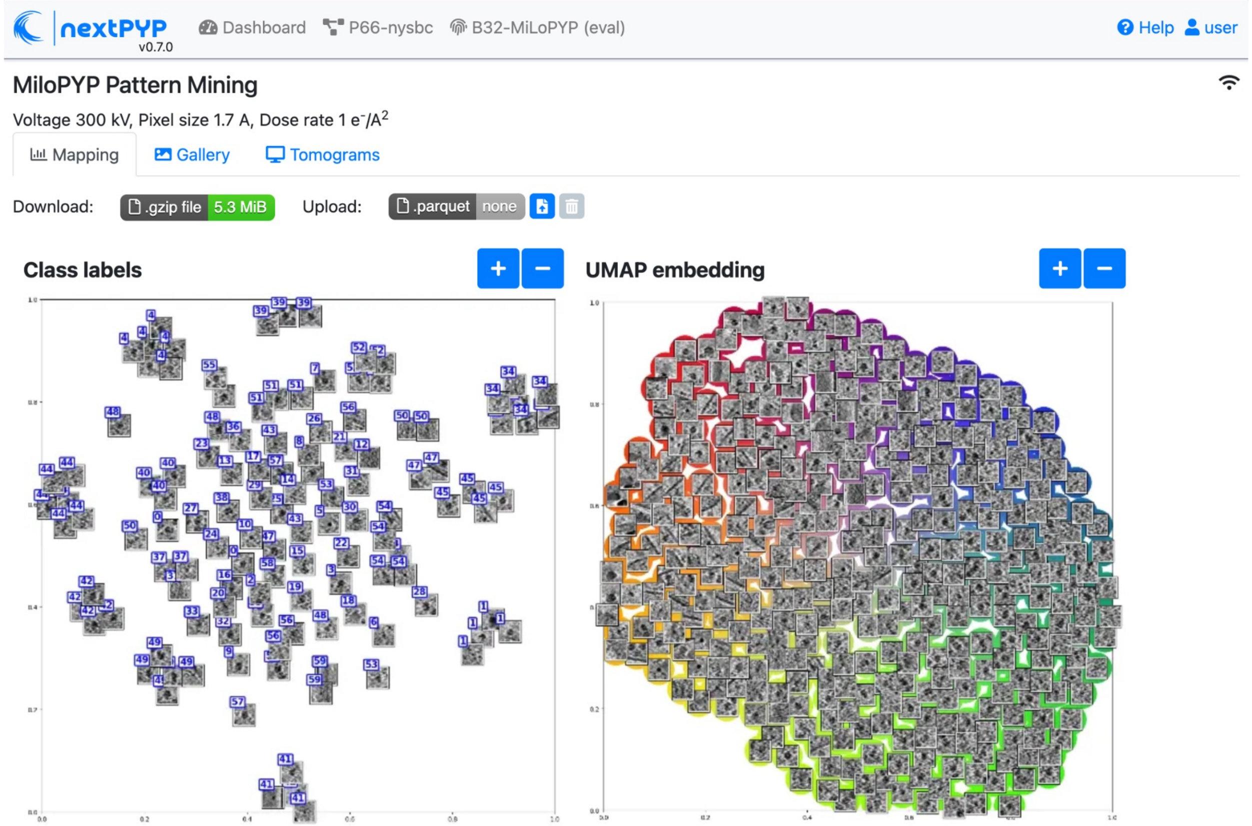
B189-MiLoPYP (train)
Tomograms
MiLoPYP Model

B191-MiLoPYP (eval)
MiLoPYP Model
MiLoPYP particles

B192-Particle-Picking (train)
Particles
Particles Model

B193-Particle-Picking (eval)
0 particles
Particles Model
Particles

UI for cellular pattern mining

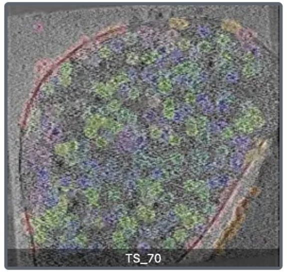


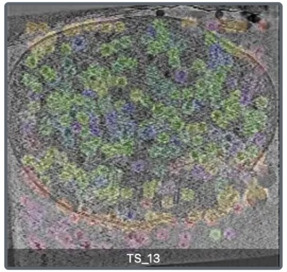
MiloPYP Pattern Mining

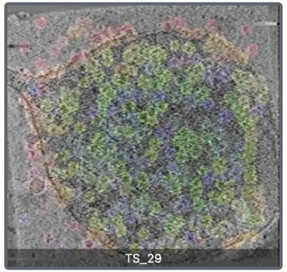
Voltage 300 kV, Pixel size 1.7 A, Dose rate 1 e$^-$/A^2

Download: .gzip file 5.3 MiB

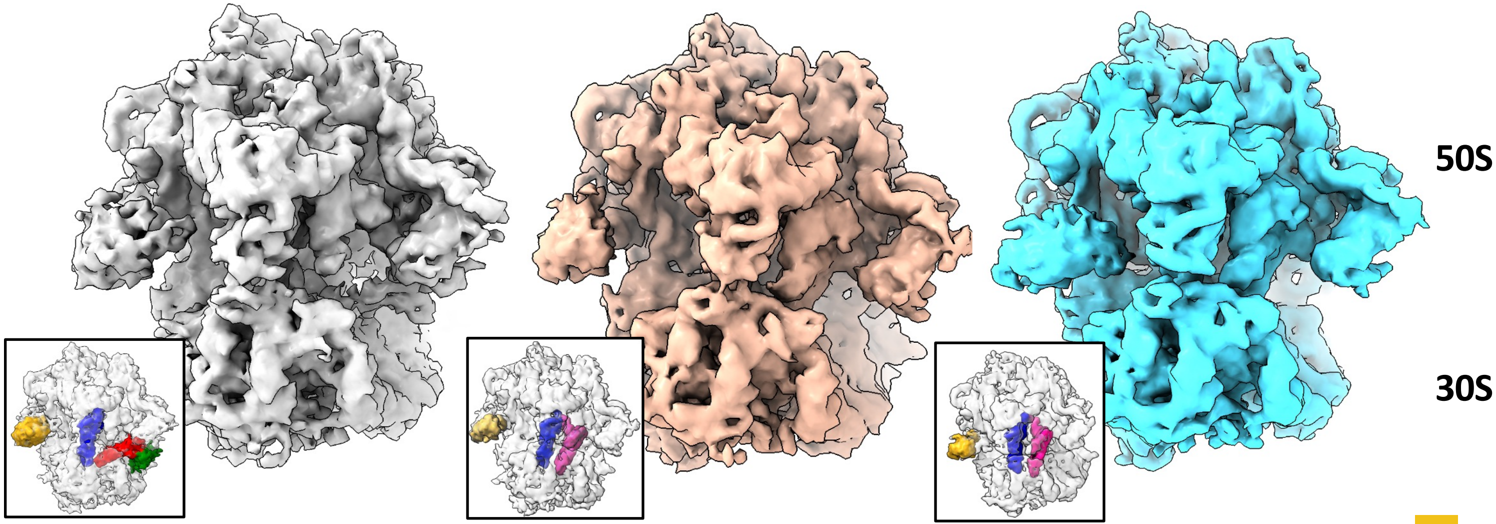
Upload: .parquet none


TS_70


TS_13


TS_29

Tools to study conformational heterogeneity


Class 1

Class 2

Class 3

Class 1 (28%) 5,086 particles
EF-Tu-tRNA P-site tRNA
L1 stalk open
Decoding
(PDB ID 7PAK)

Class 2 (30%) 5,449 particles
A- and P-site tRNAs
L1 stalk open
Classical, non-rotated
(PDB ID 7PAL)

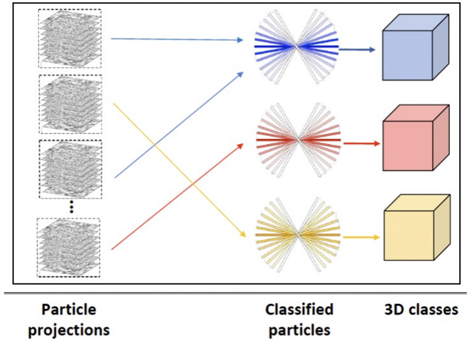
Class 3 (42%) 7,629 particles
A- and P-site tRNAs
L1 stalk intermediate
Classical, non-rotated
(PDB ID 7PAL)

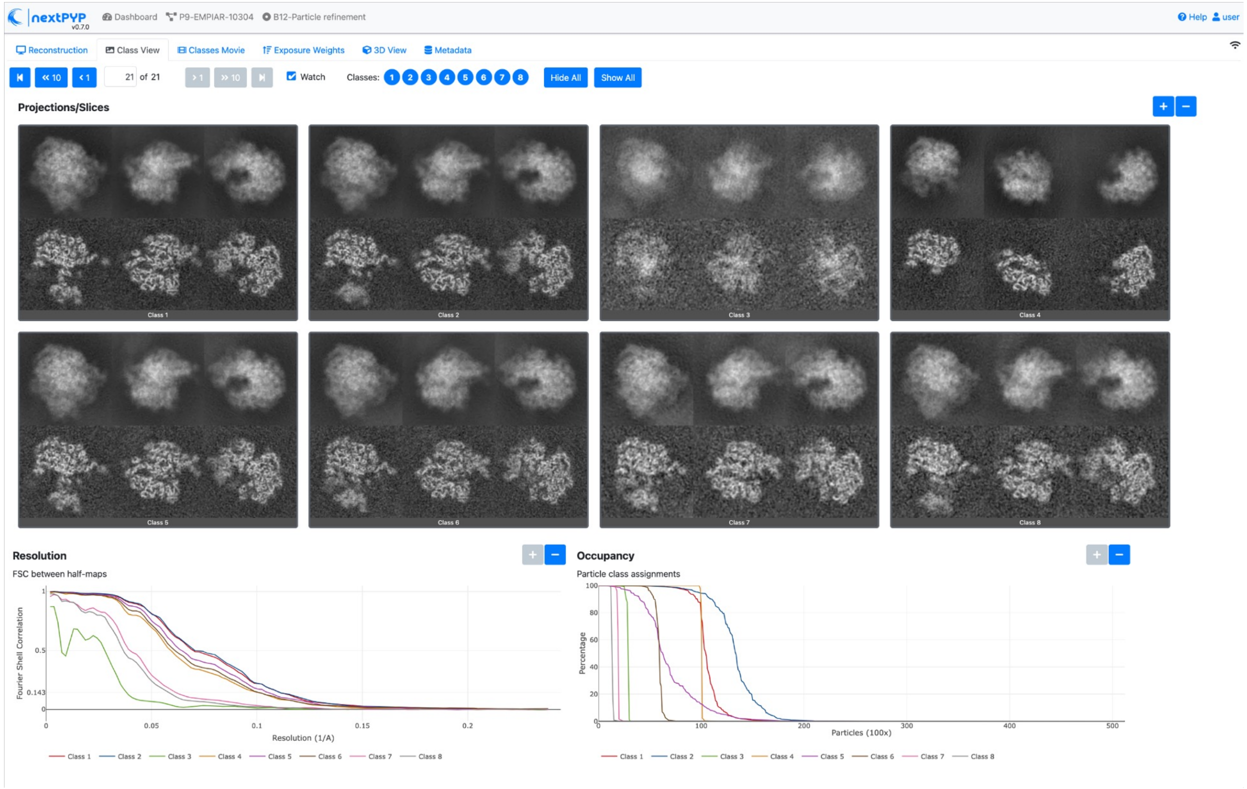
🟨 L1 stalk
🟦 P-tRNA
🟩 EF-Tu-tRNA 🟥
🟪 A-tRNA

Classification of 70S ribosomes from intact cells (EMPIAR-10499)

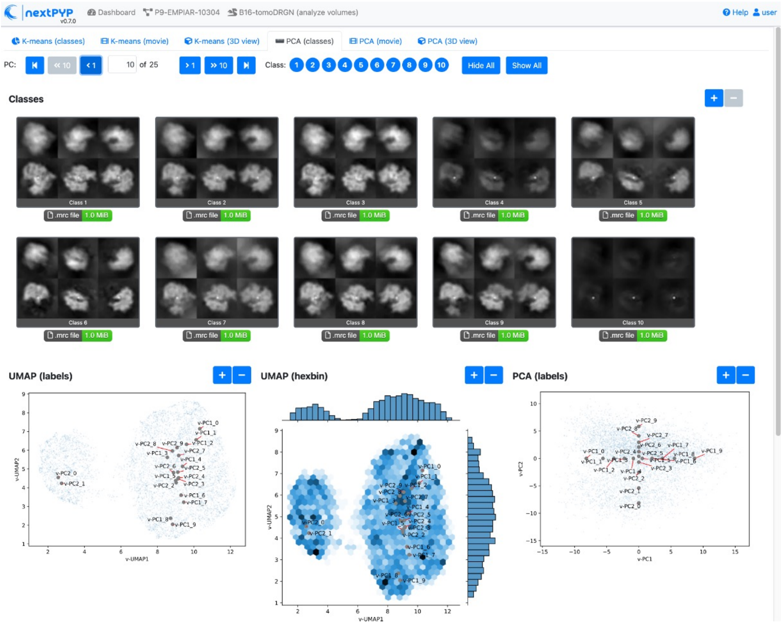
Discrete heterogeneity (3D classification)

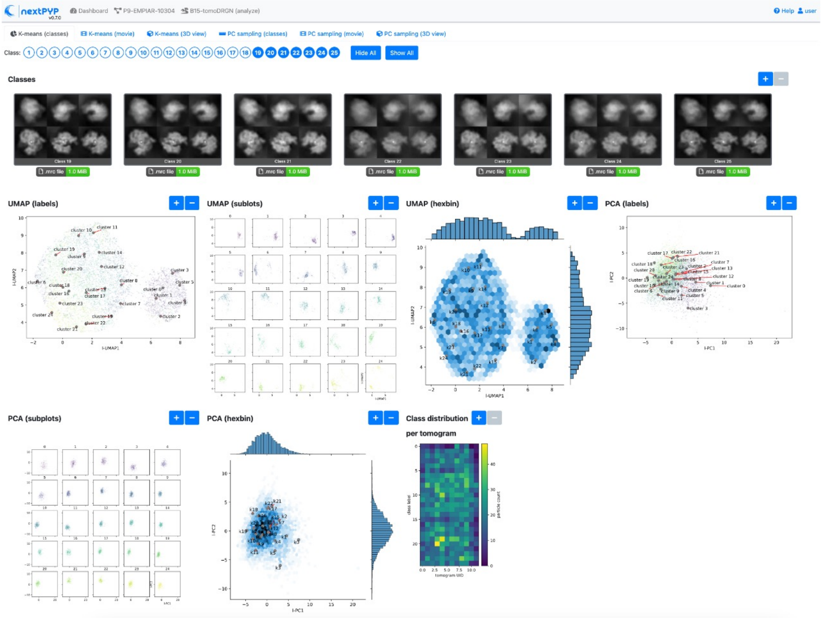
Constrained classification



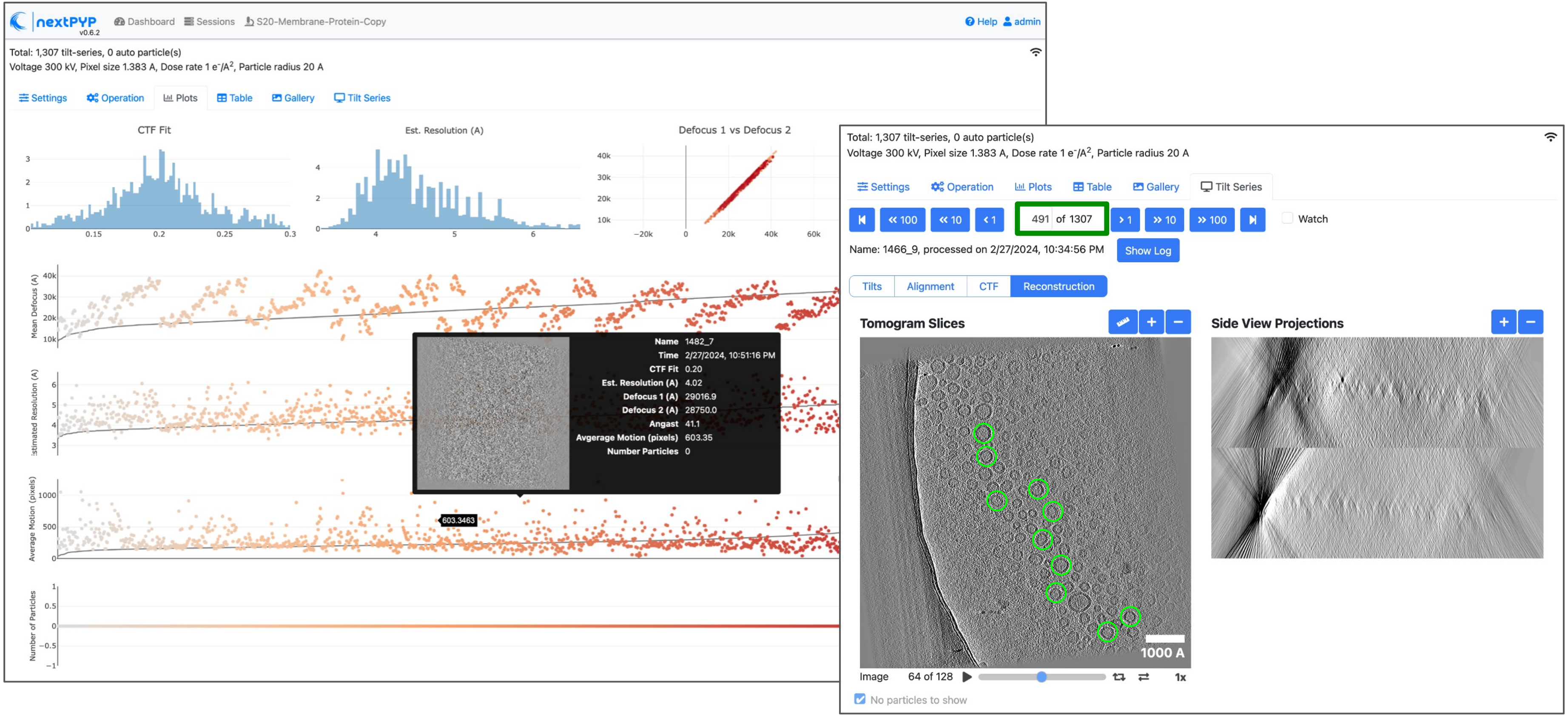


Continuous heterogeneity (tomoDRGN)





On-the-fly pre-processing



Total: 1,307 tilt-series, 0 auto particle(s)
Voltage 300 kV, Pixel size 1.383 A, Dose rate 1 e^-/A^2, Particle radius 20 A

Name	1482_7
Time	2/27/2024, 10:51:16 PM
CTF Fit	0.20
Est. Resolution (A)	4.02
Defocus 1 (A)	29016.9
Defocus 2 (A)	28750.0
Angast	41.1
Avgerage Motion (pixels)	603.35
Number Particles	0

Total: 1,307 tilt-series, 0 auto particle(s)
Voltage 300 kV, Pixel size 1.383 A, Dose rate 1 e^-/A^2, Particle radius 20 A

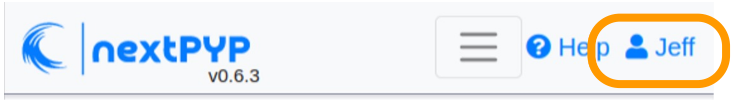
491 of 1307

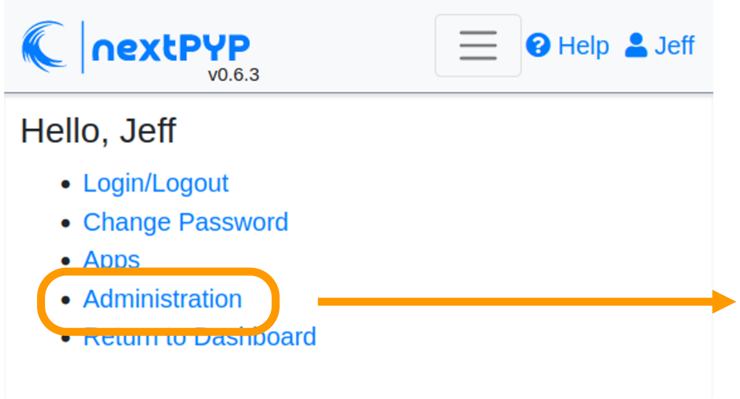
Name: 1466_9, processed on 2/27/2024, 10:34:56 PM

☑ No particles to show

Administration

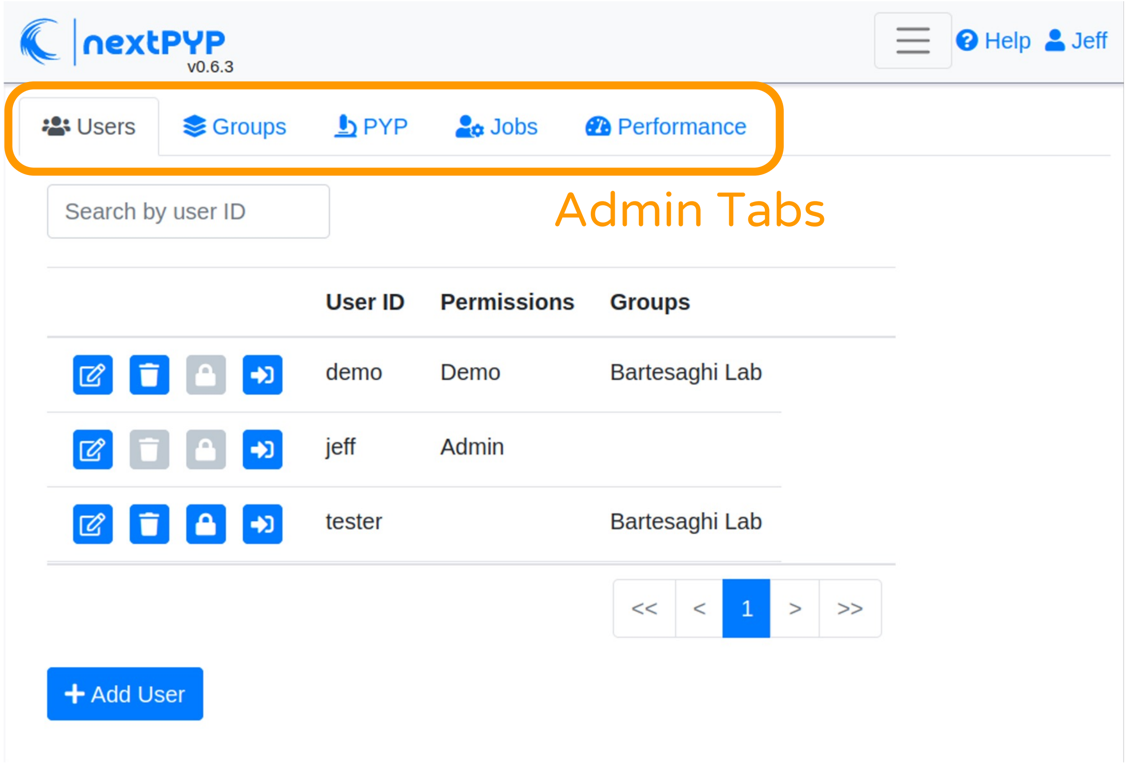
Management





Hello, Jeff

- Login/Logout
- Change Password
- Apps
- Administration
- Return to Dashboard



Admin Tabs

User ID	Permissions	Groups
demo	Demo	Bartesaghi Lab
jeff	Admin	
tester		Bartesaghi Lab

Getting help

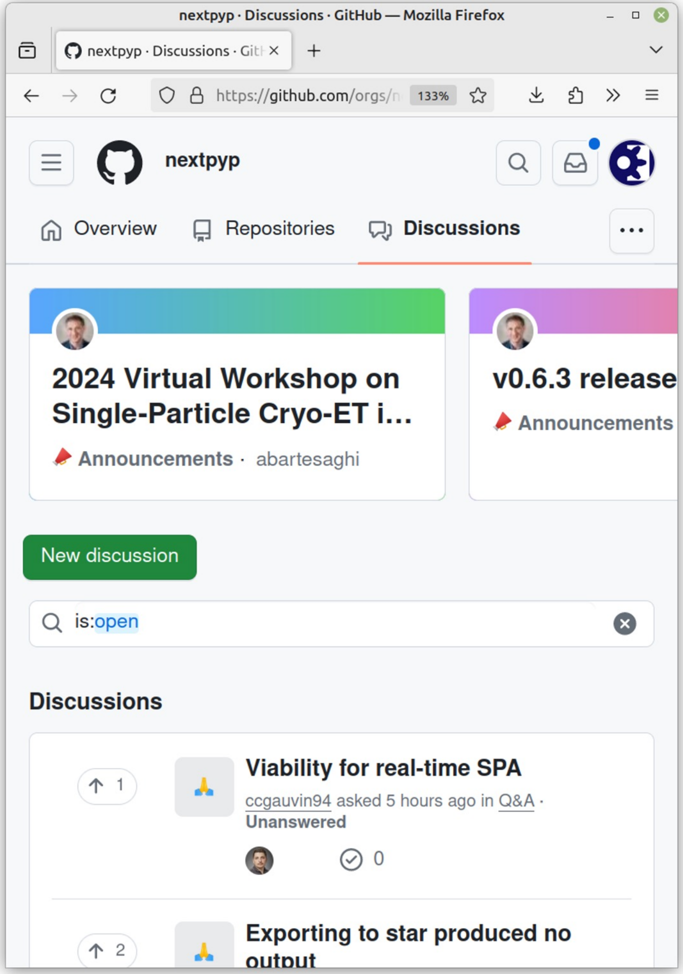
Github discussions

Announcements, new releases

Ask a question, report a new issue

See answers to previous questions, search for solutions

https://github.com/nextpyp



2024 Virtual Workshop on Single-Particle Cryo-ET i...
Announcements · abartesaghi

v0.6.3 release
Announcements

Viability for real-time SPA
ccgauvin94 asked 5 hours ago in Q&A · **Unanswered**
1 · 0

Exporting to star produced no output
2

nextPYP

Q&A

Bartesaghi Lab

Duke
UNIVERSITY