Neural network particle picking and denoising with

TOPAZ

Pick more of anything

Alex J. Noble

Topaz Workshop at SEMC!

Mar 8, 2023

Topaz Picking: Nature Methods

Topaz Denoising: Nature Communications

http://cb.csail.mit.edu/cb/topaz

Tristan Bepler

Massachusetts Institute of Technology

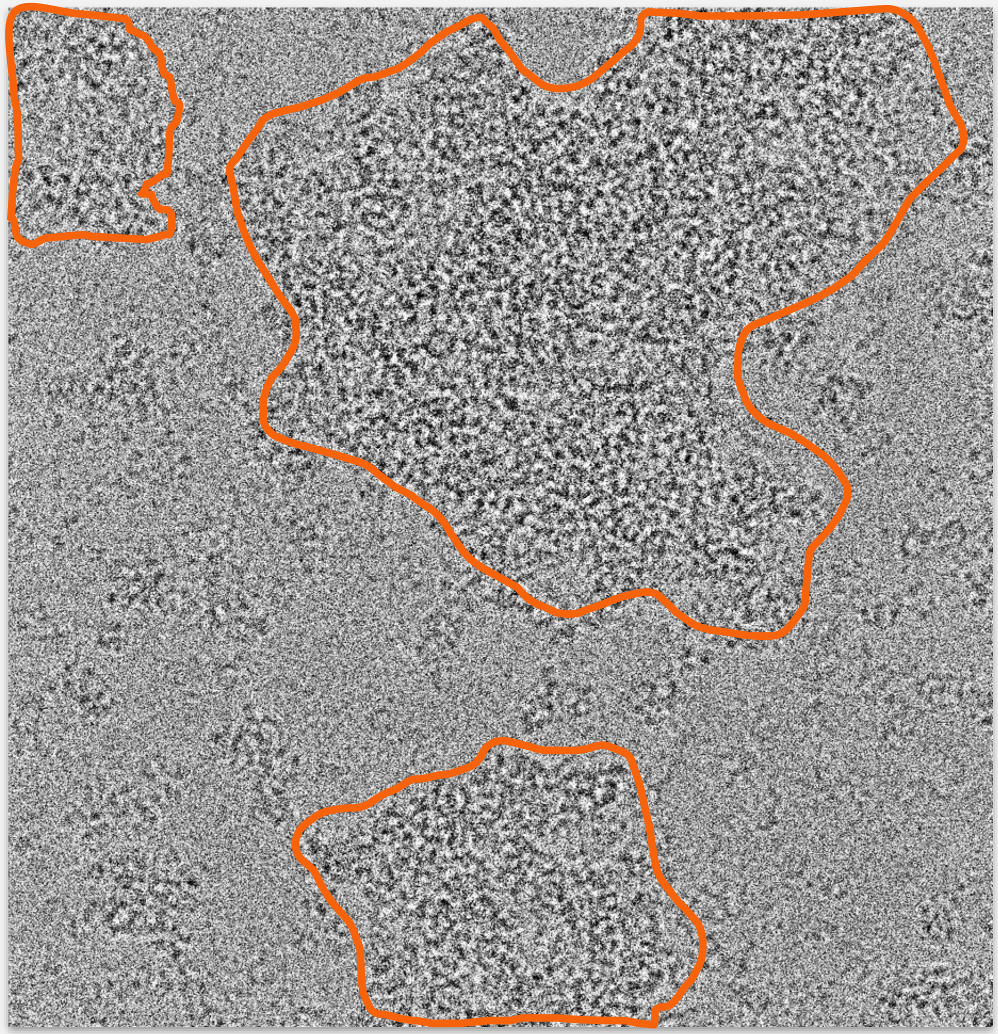
SMLC SIMONS MACHINE LEARNING CENTER

SIMONS ELECTRON MICROSCOPY CENTER

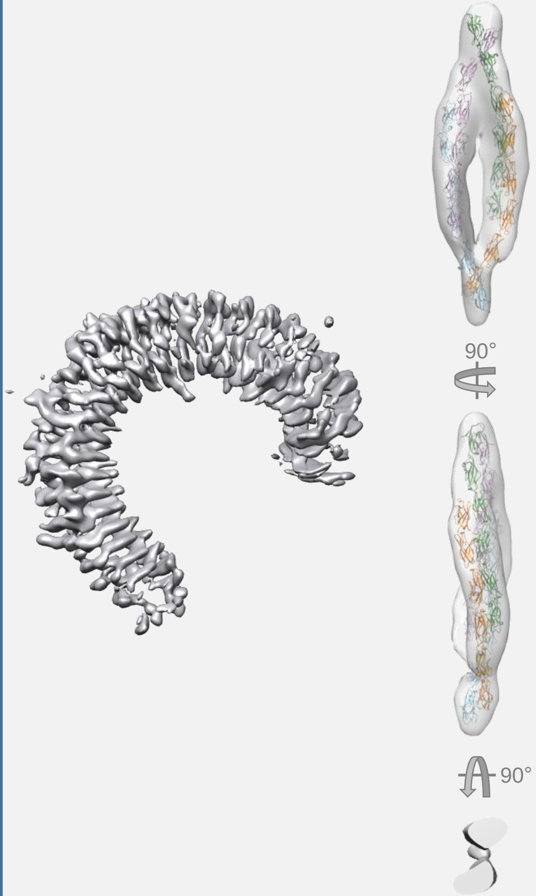
Workshop Goals=)

1. **Understand** how **Topaz** can help you
2. Get **hands-on Topaz experience** so you can get **better cryoEM structures**!

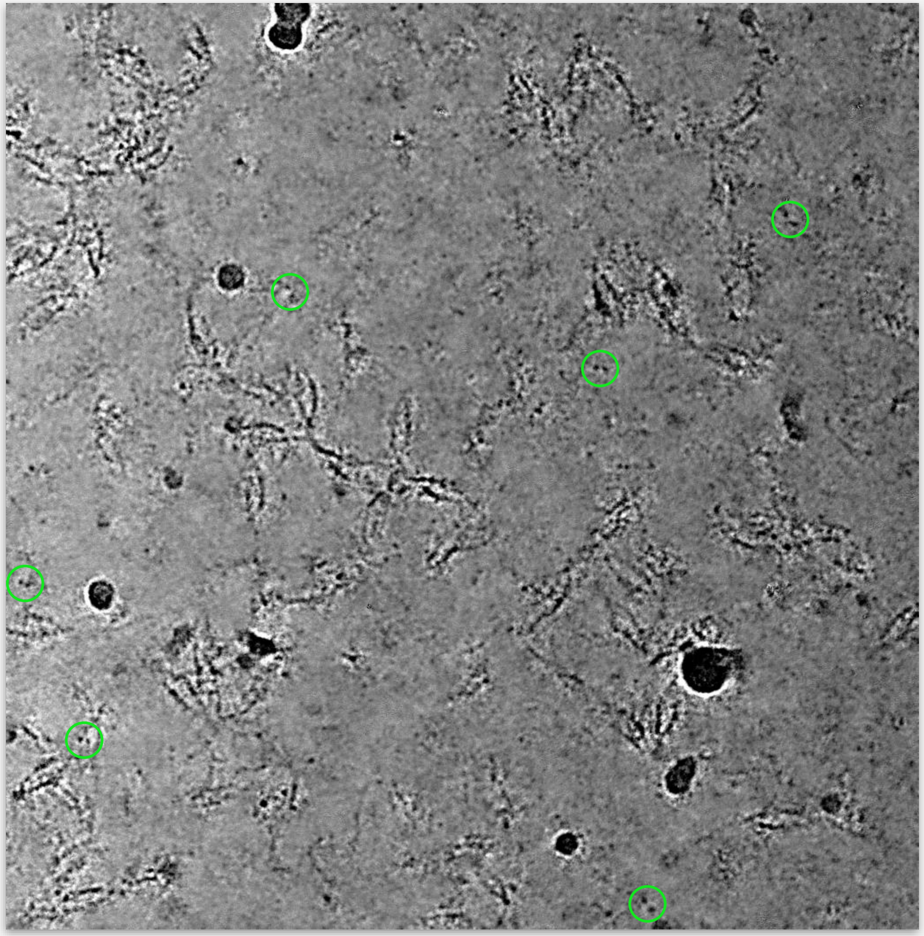
Common problems in cryoEM



aggregation

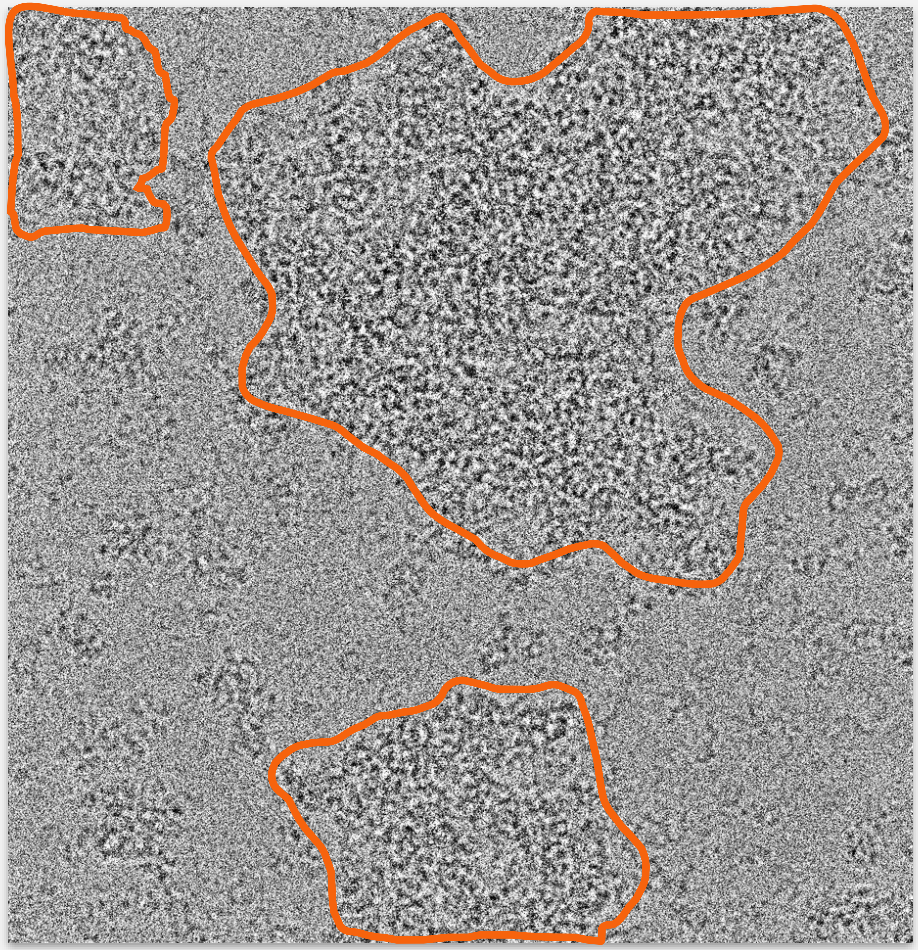


weird-shaped proteins

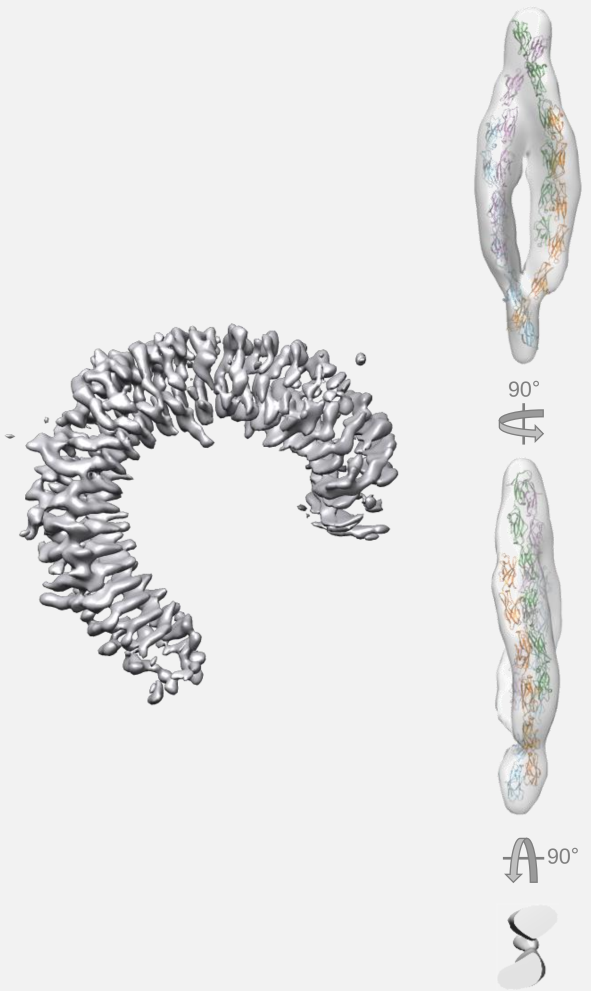


small particles

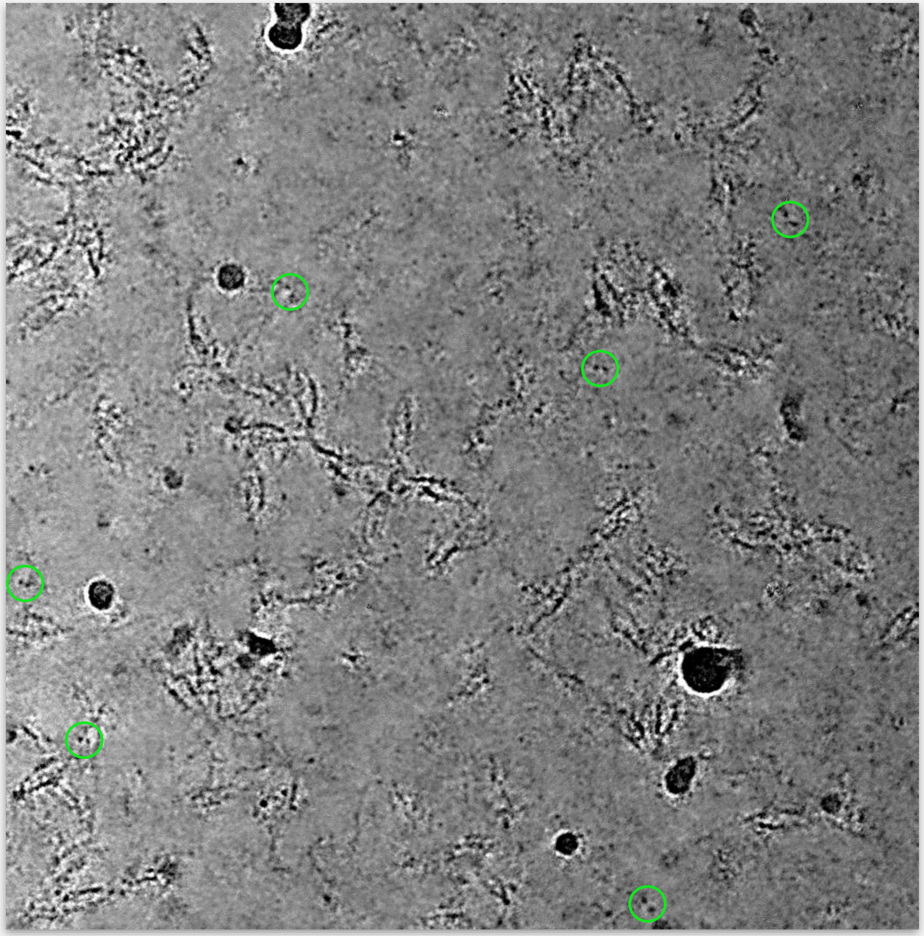
Common problems in cryoEM



aggregation

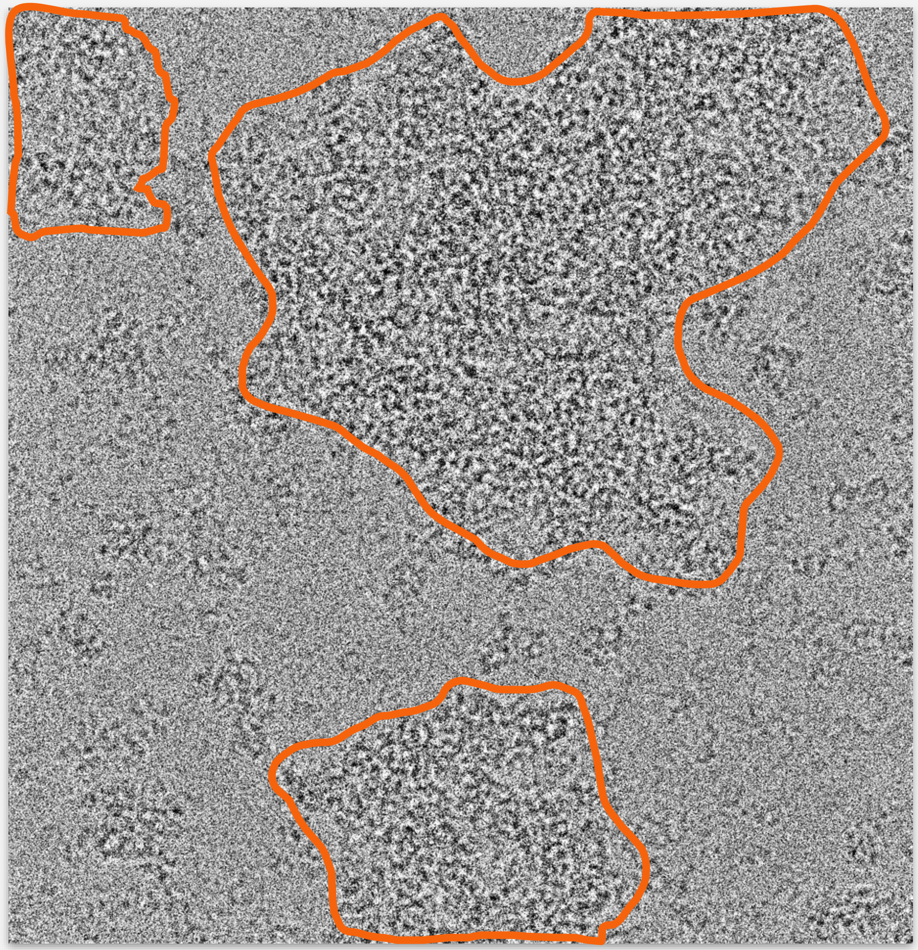


weird-shaped proteins

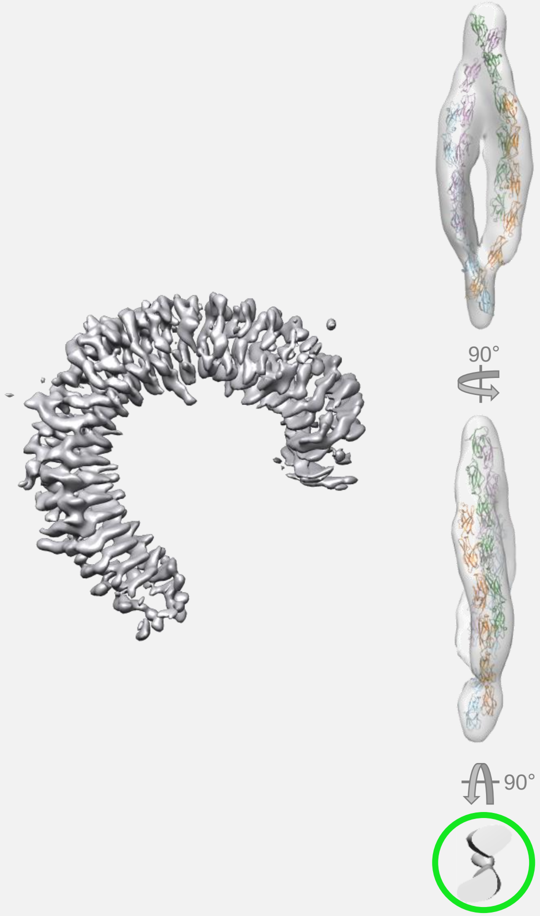


small particles

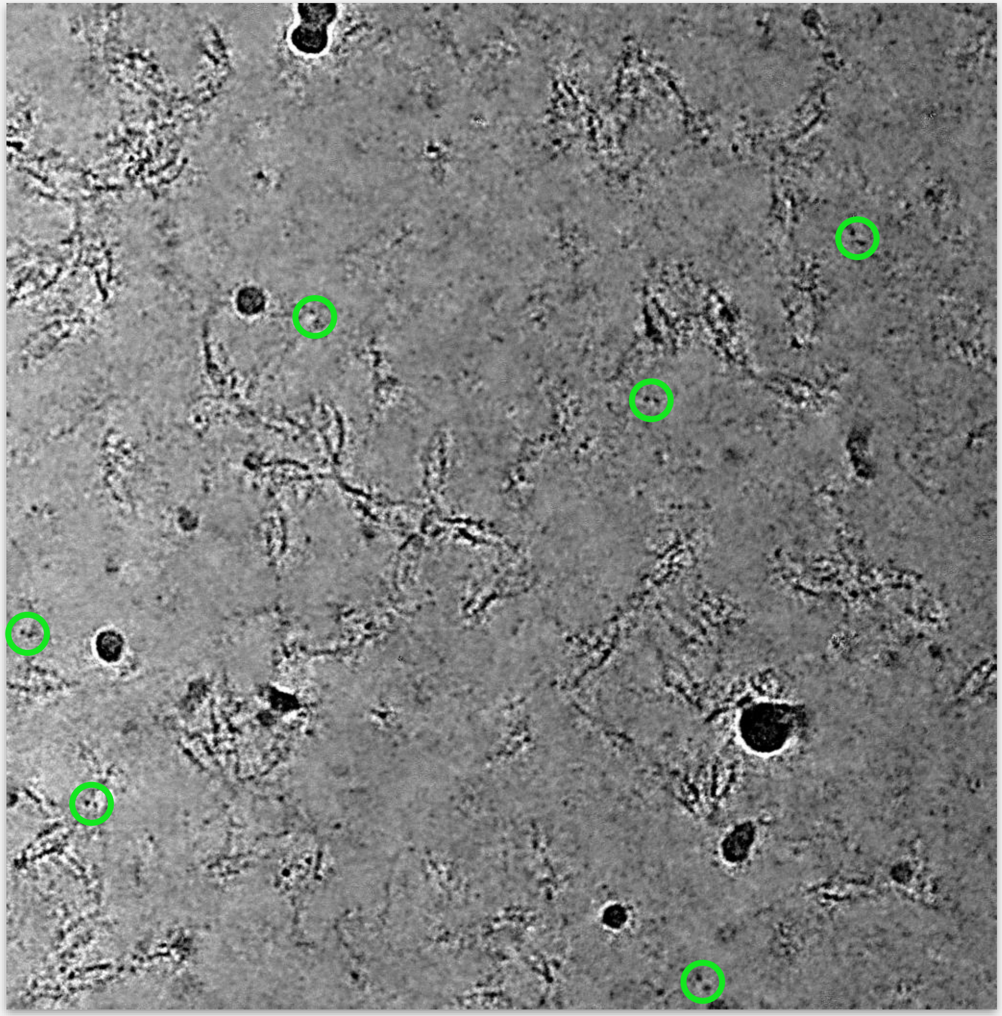
Common problems in cryoEM



aggregation



weird-shaped proteins

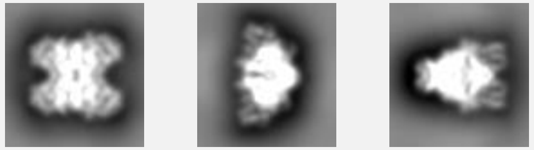
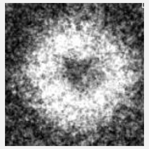
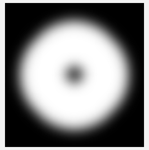


small particles

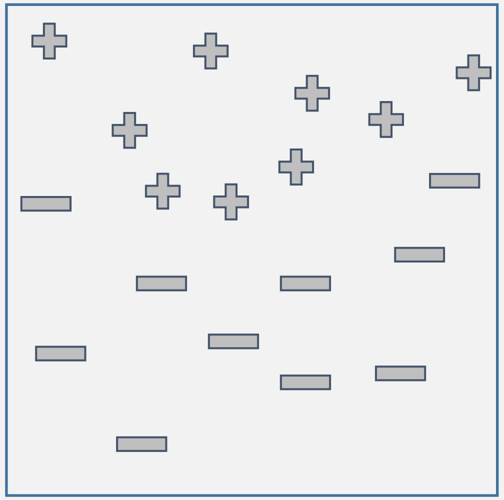
Conceptual Topaz Overview

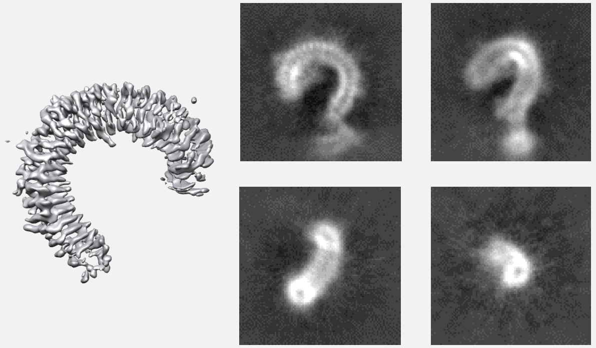
- How TOPAZ *picking* is **beneficially different**
- What **problems** TOPAZ **solves**
- Examples
- How can *denoising* **help you?**

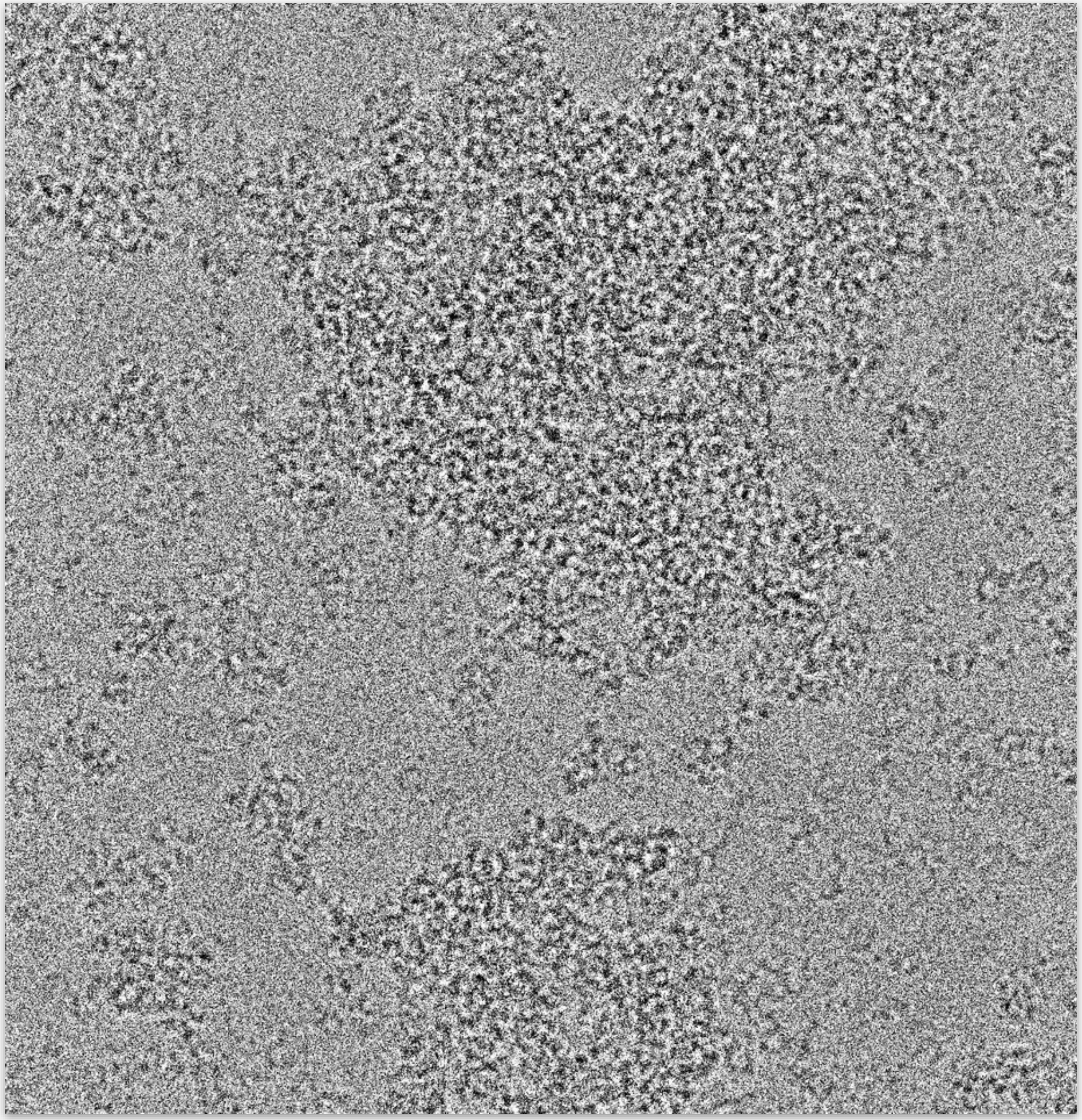
Current particle picking methods

- Historically, particle picking =
  - Template matching 
  - Rotational averaging 
  - Difference of Gaussians 
- Newer methods use **positive-negative** convolutional neural networks (e.g. crYOLO, Warp, DeepPicker)
- TOPAZ is different – **positive-unlabeled**

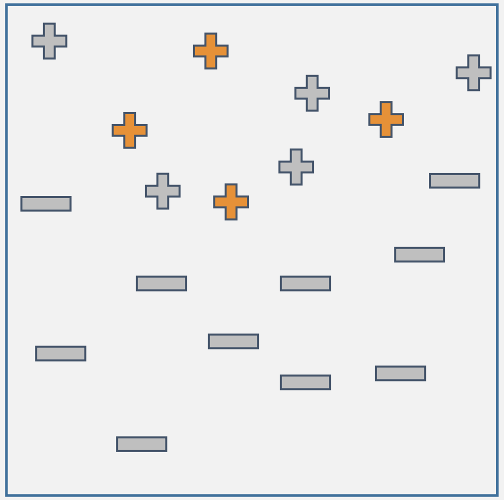
What is a *positive-unlabeled* framework?

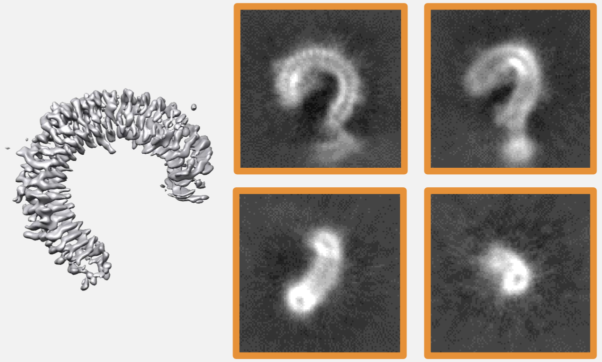


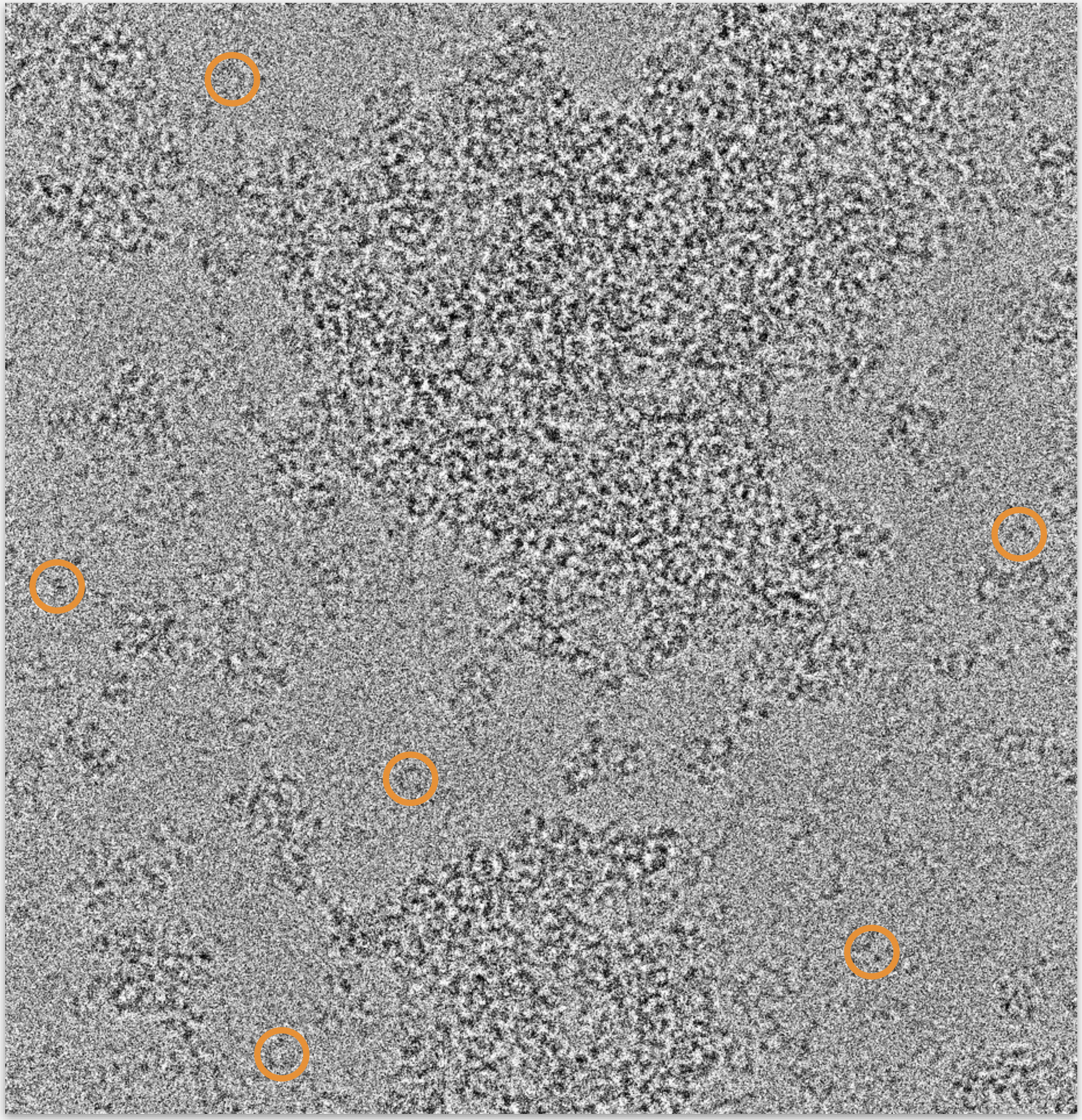




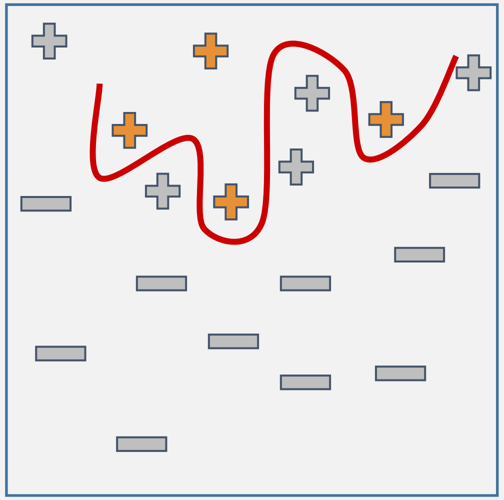
What is a *positive-unlabeled* framework?

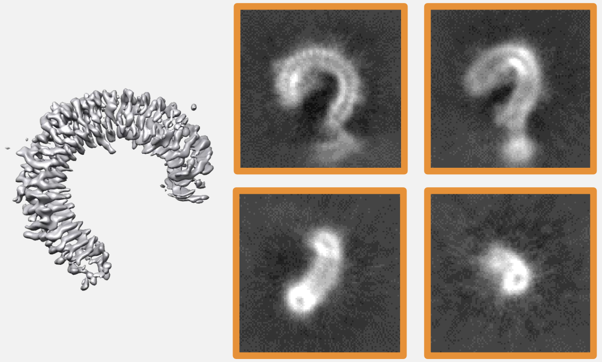


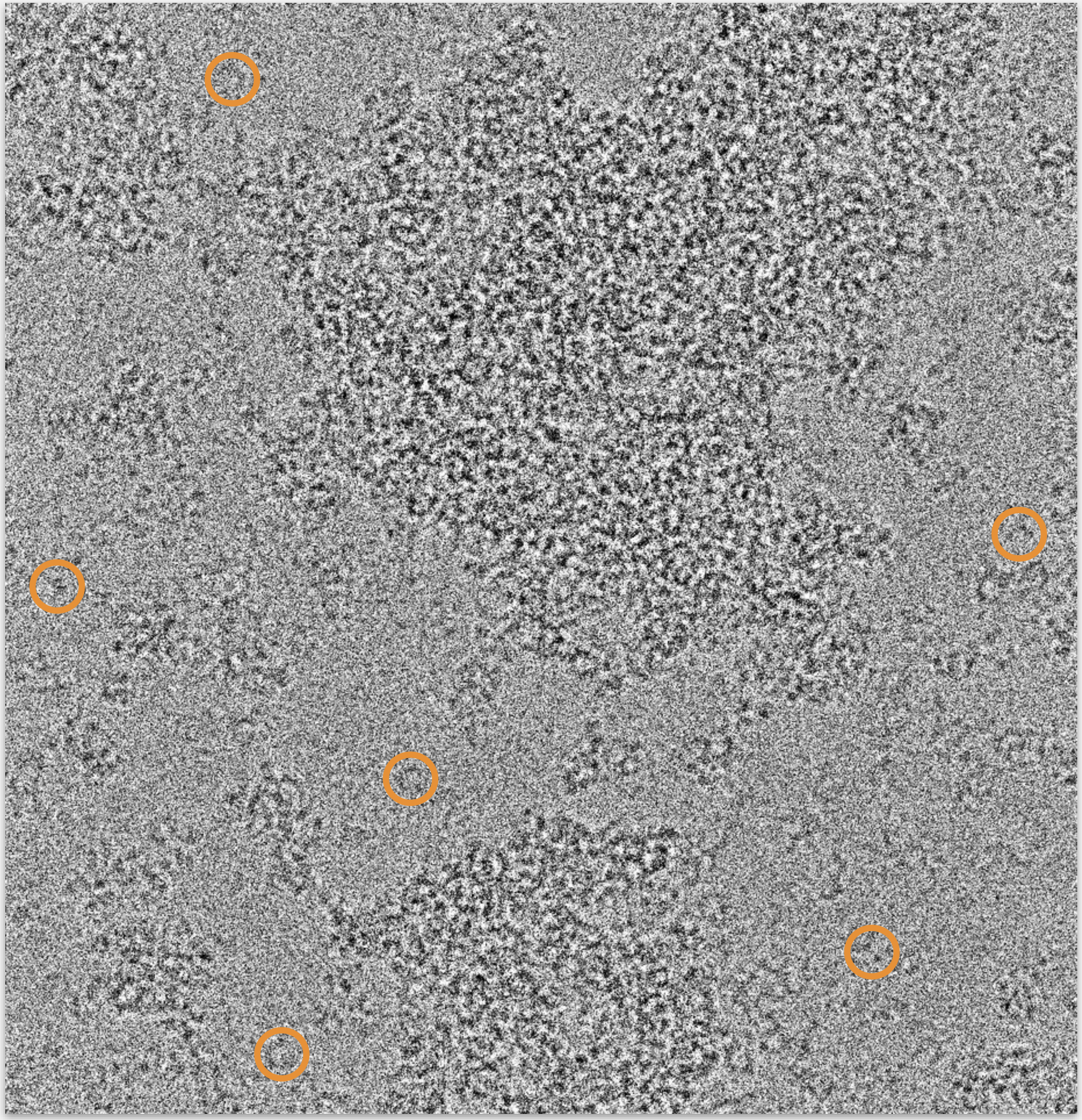




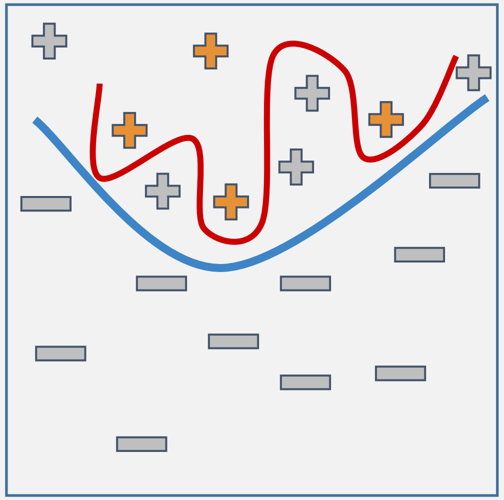
What is a *positive-unlabeled* framework?

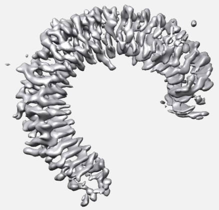


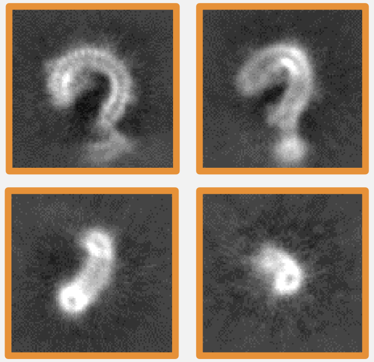


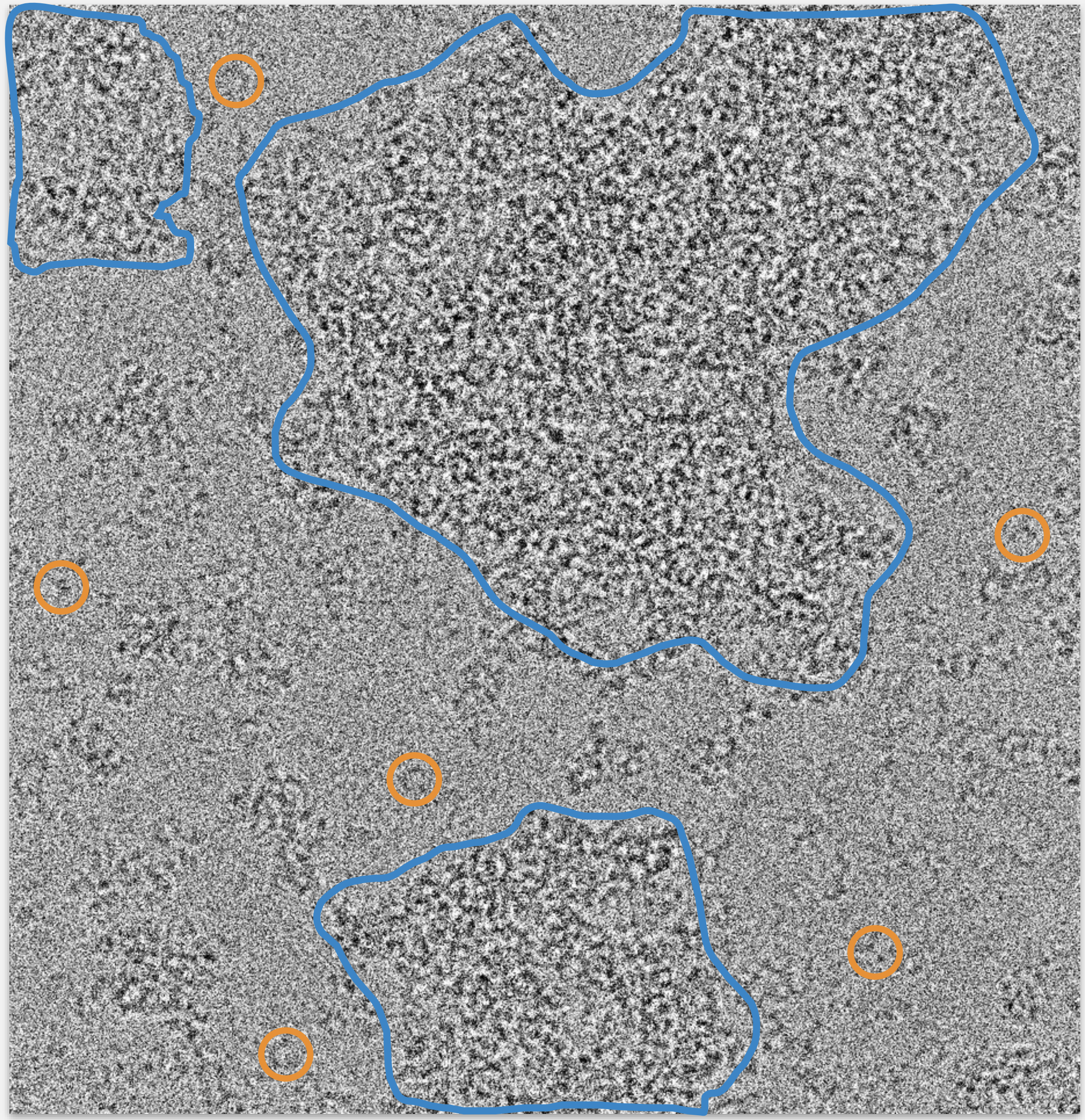


What is a *positive-unlabeled* framework?









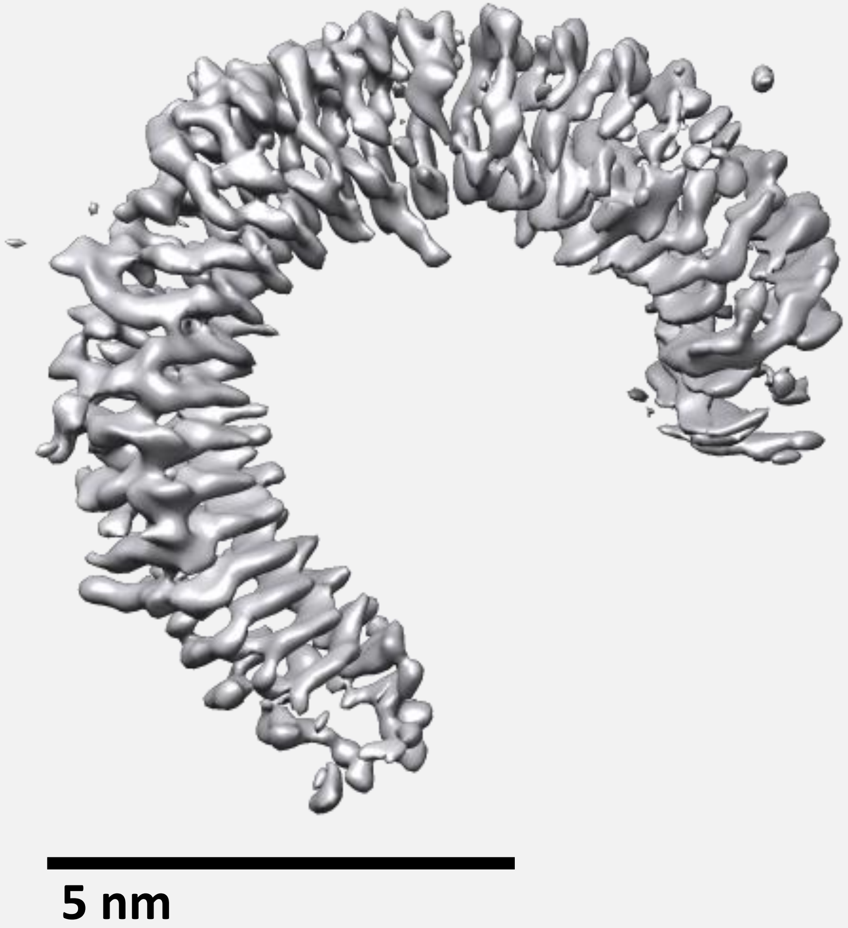
Problems TOPAZ solves

- Assumes **sparsely labeled** training data (positive-unlabeled)
- Pick **any size & shape particles**
- Picks **more real particles** than other pickers
- Very **high ratio** of **true positives to false positives**
  - This *reduces bias* in classification filtering
- **Centers particles well** (surprisingly important!)

Ok great, how well does TOPAZ *actually work*??

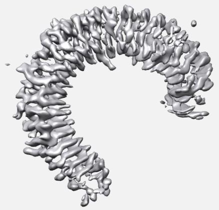
TOPAZ picks all shapes and sizes

- 105 kDa
- non-globular
- asymmetric
- aggregates

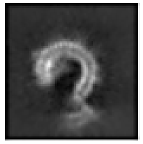
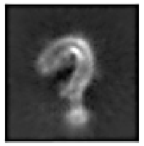
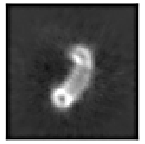
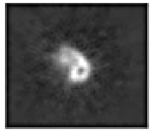
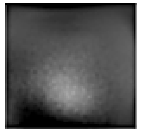


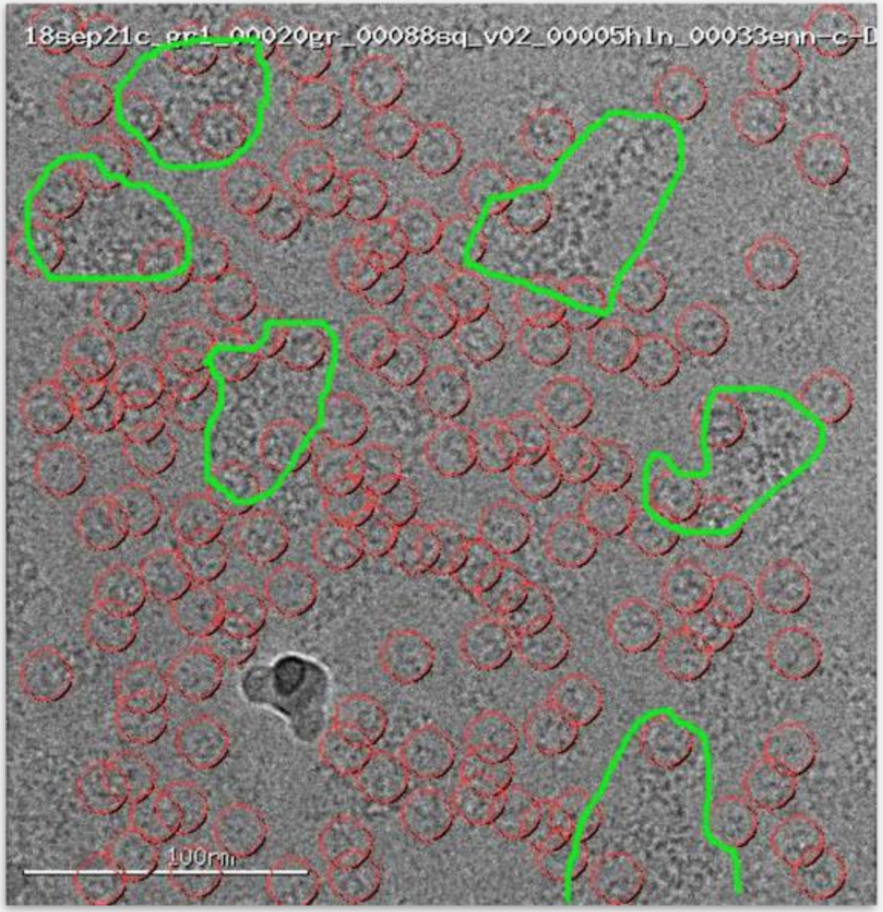

Micah Rapp
Columbia

TOPAZ picks all shapes and sizes



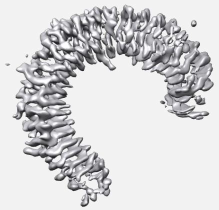
- 105 kDa
- non-globular
- asymmetric
- aggregates

	Front	Oblique	Side	Top	Junk
					
Topaz	43% (435k)	20% (195k)	14% (144k)	23% (231k)	0
Template	64% (401k)	14% (86k)	14% (90k)	3% (17k)	5% (33k)
DoG	59% (456k)	14% (104k)	7% (50k)	6% (43k)	14% (118k)

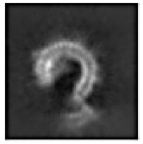
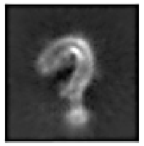
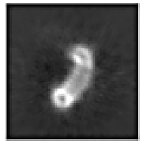
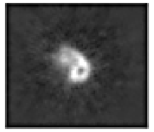
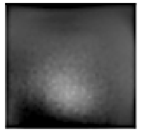


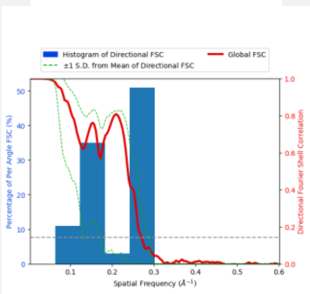
Micah Rapp
Columbia

TOPAZ picks all shapes and sizes

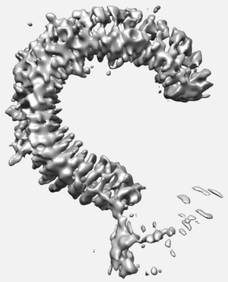


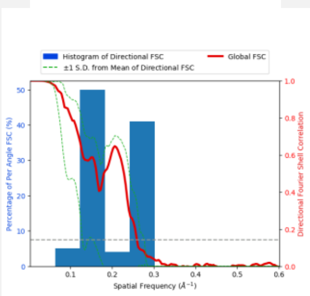
- 105 kDa
- non-globular
- asymmetric
- aggregates

	Front	Oblique	Side	Top	Junk
					
Topaz	43% (435k)	20% (195k)	14% (144k)	23% (231k)	0
Template	64% (401k)	14% (86k)	14% (90k)	3% (17k)	5% (33k)
DoG	59% (456k)	14% (104k)	7% (50k)	6% (43k)	14% (118k)

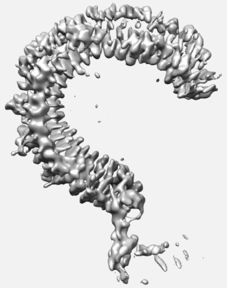


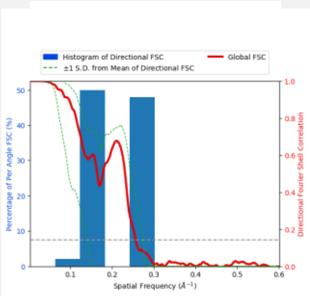
Topaz
- **3.70 Å**
- **1,010,937 particles**





DoG
- 3.86 Å
- 770,263 particles

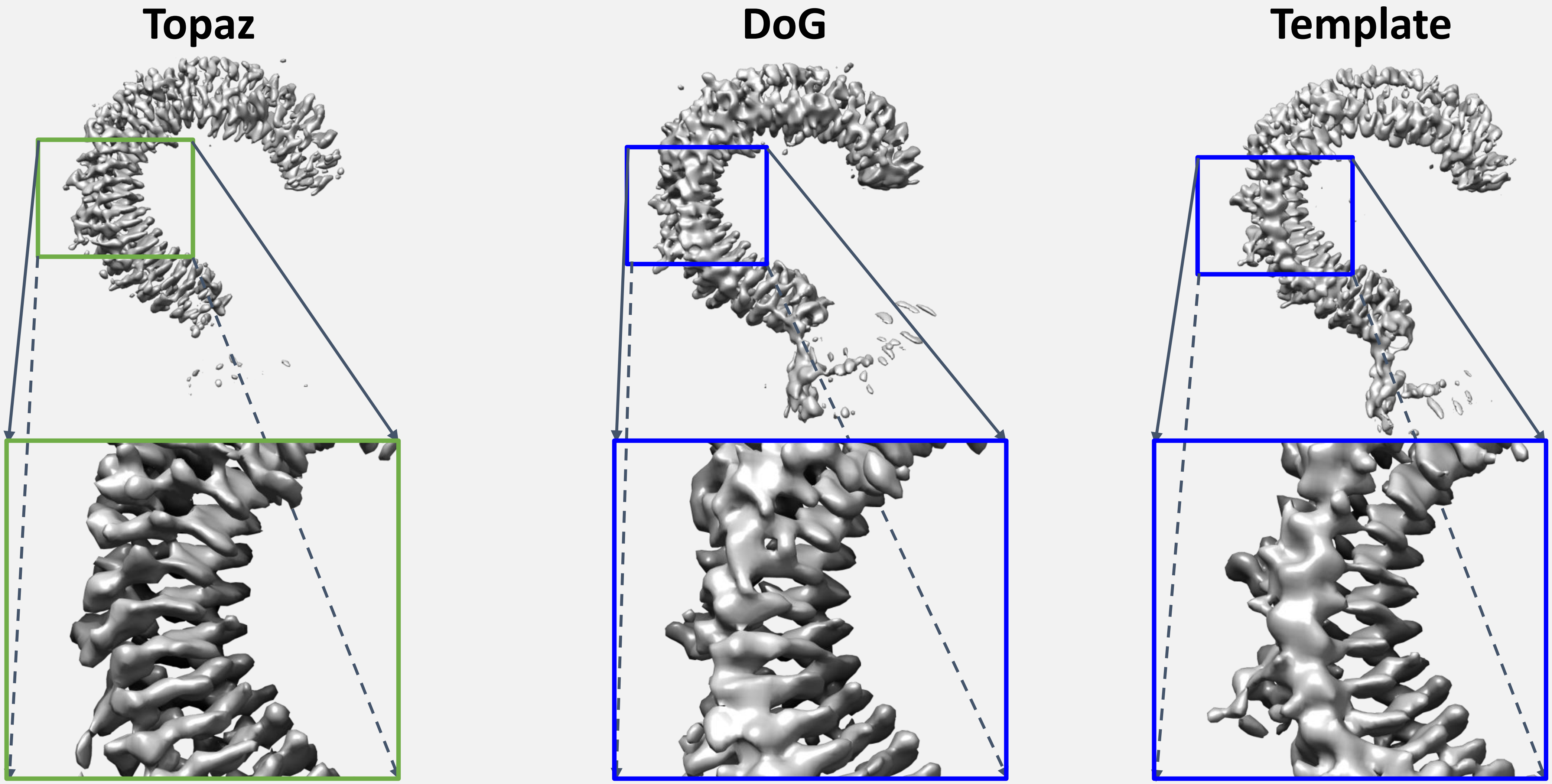




Template
- 3.92 Å
- 627,533 particles

Micah Rapp
Columbia

TOPAZ picks all shapes and sizes



Topaz particles *uniquely* resolve secondary structure *w/o curation*

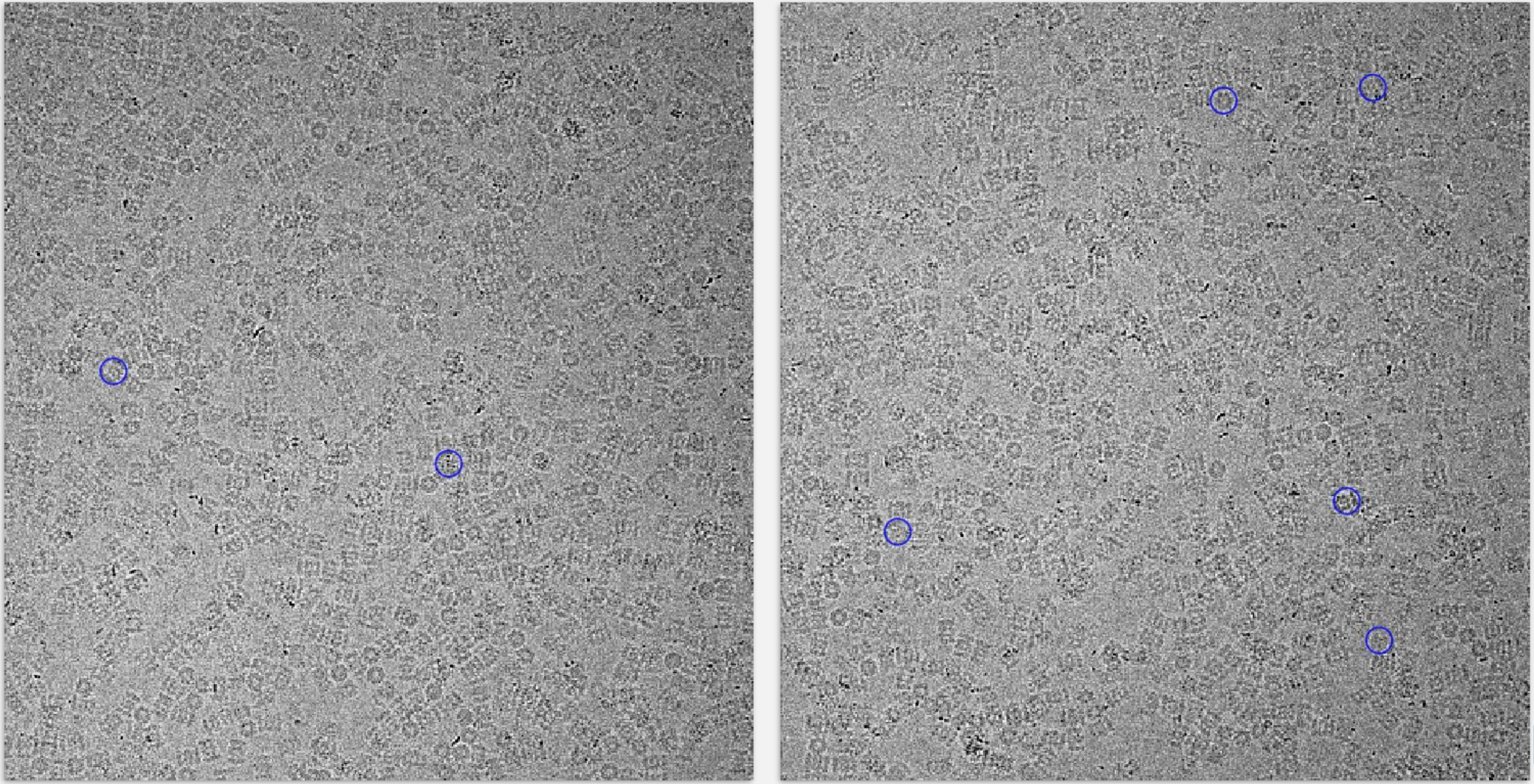
Micah Rapp
Columbia

How **sparse** can we **pick for training?**

- Pick ~100 – 1,000 particles this sparsely:

= *labeled* particles

T20S Proteasome

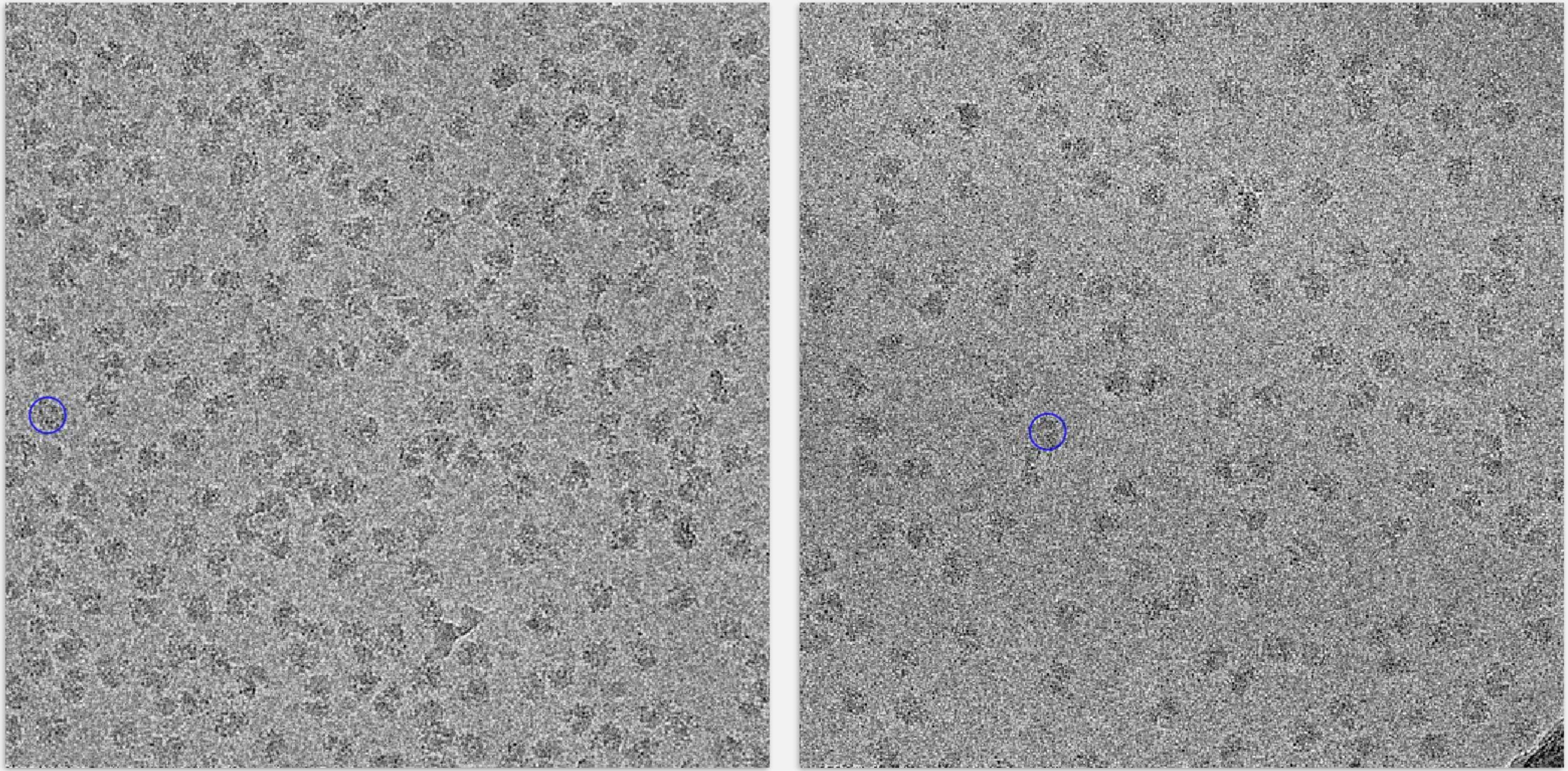


How **sparse** can we **pick for training?**

- Pick ~100 – 1,000 particles this sparsely:

= *labeled* particles

80S Ribosome

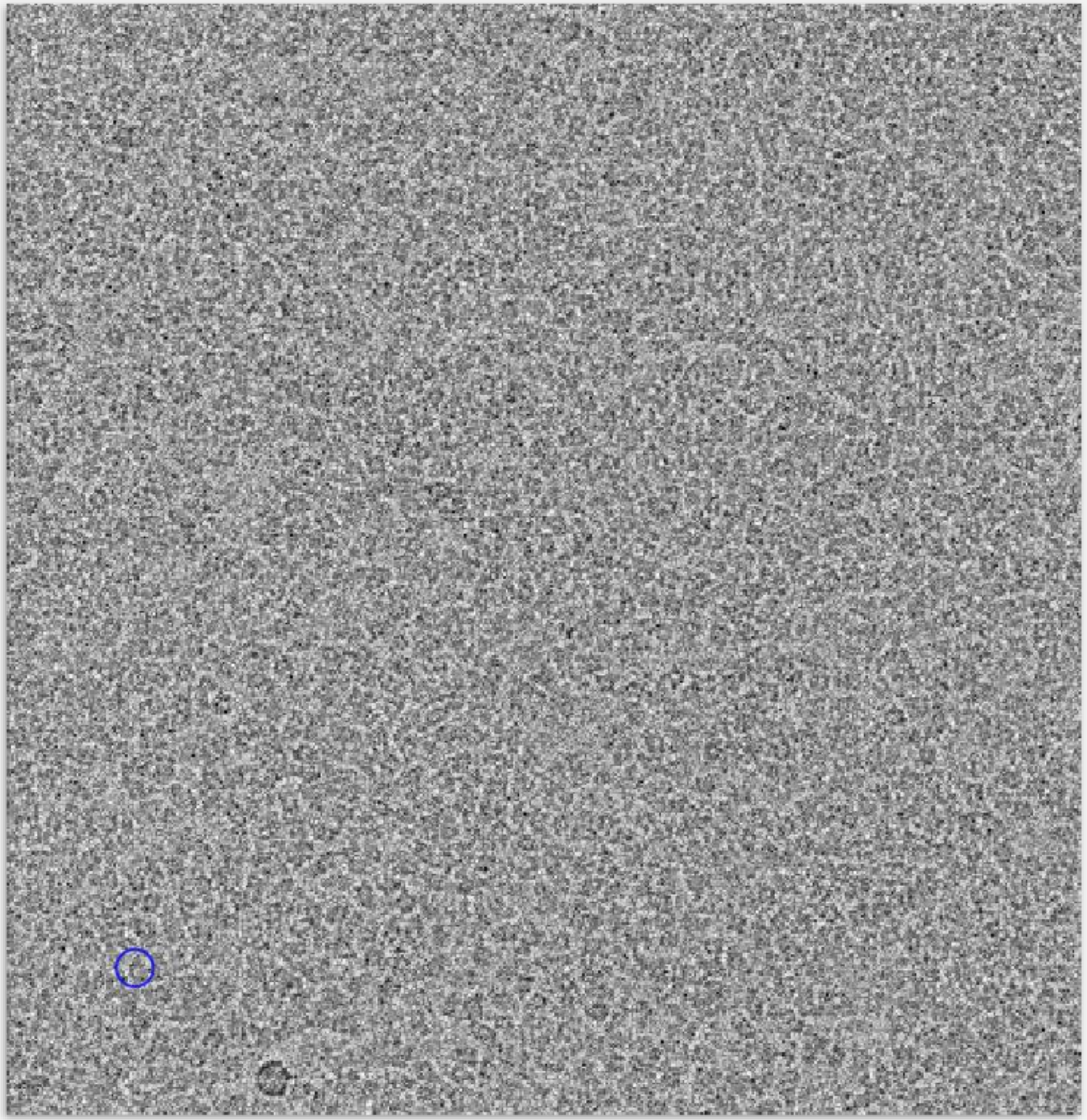


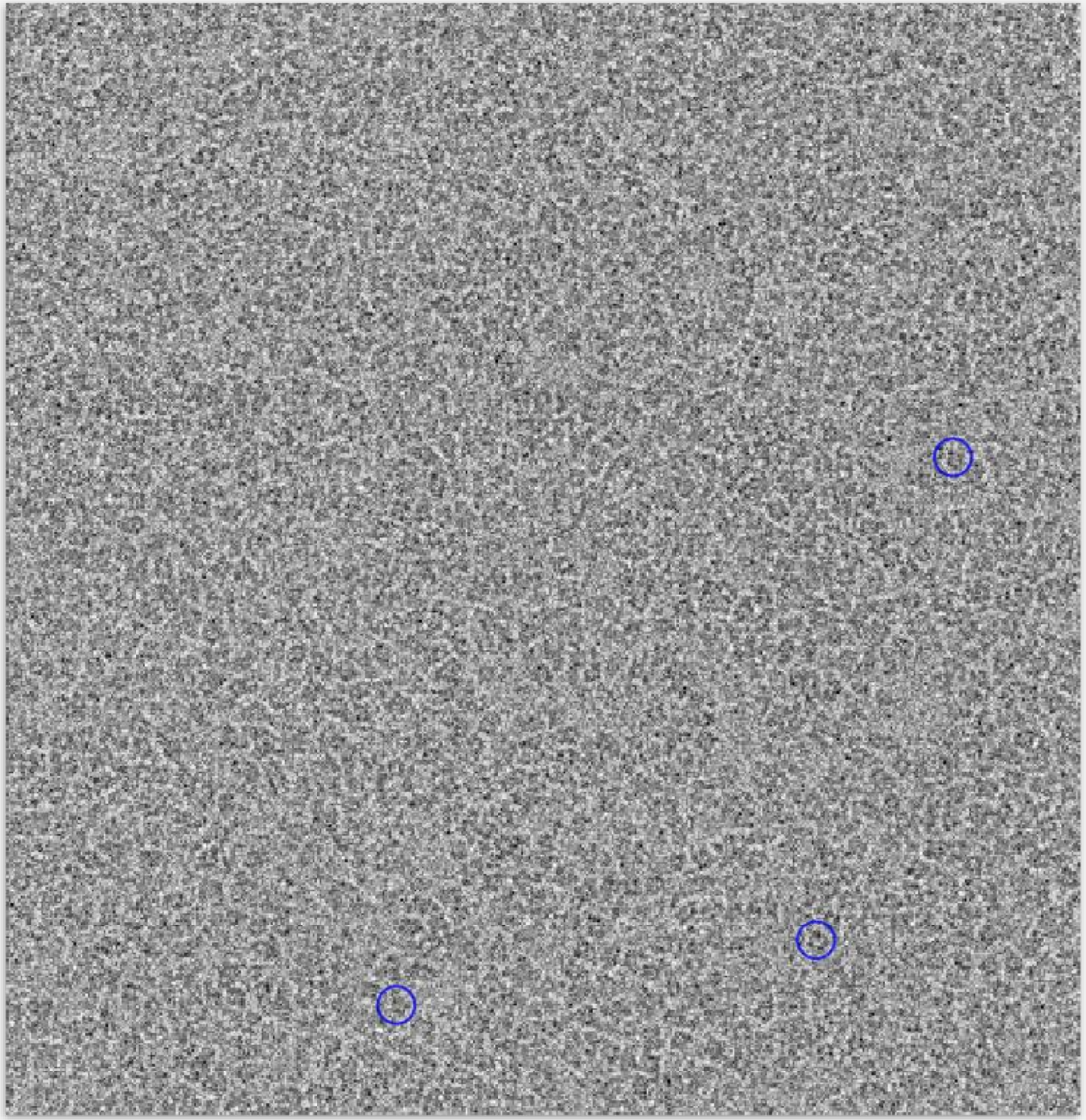
How **sparse** can we **pick for training?**

- Pick ~100 – 1,000 particles this sparsely:

= *labeled* particles

Aldolase

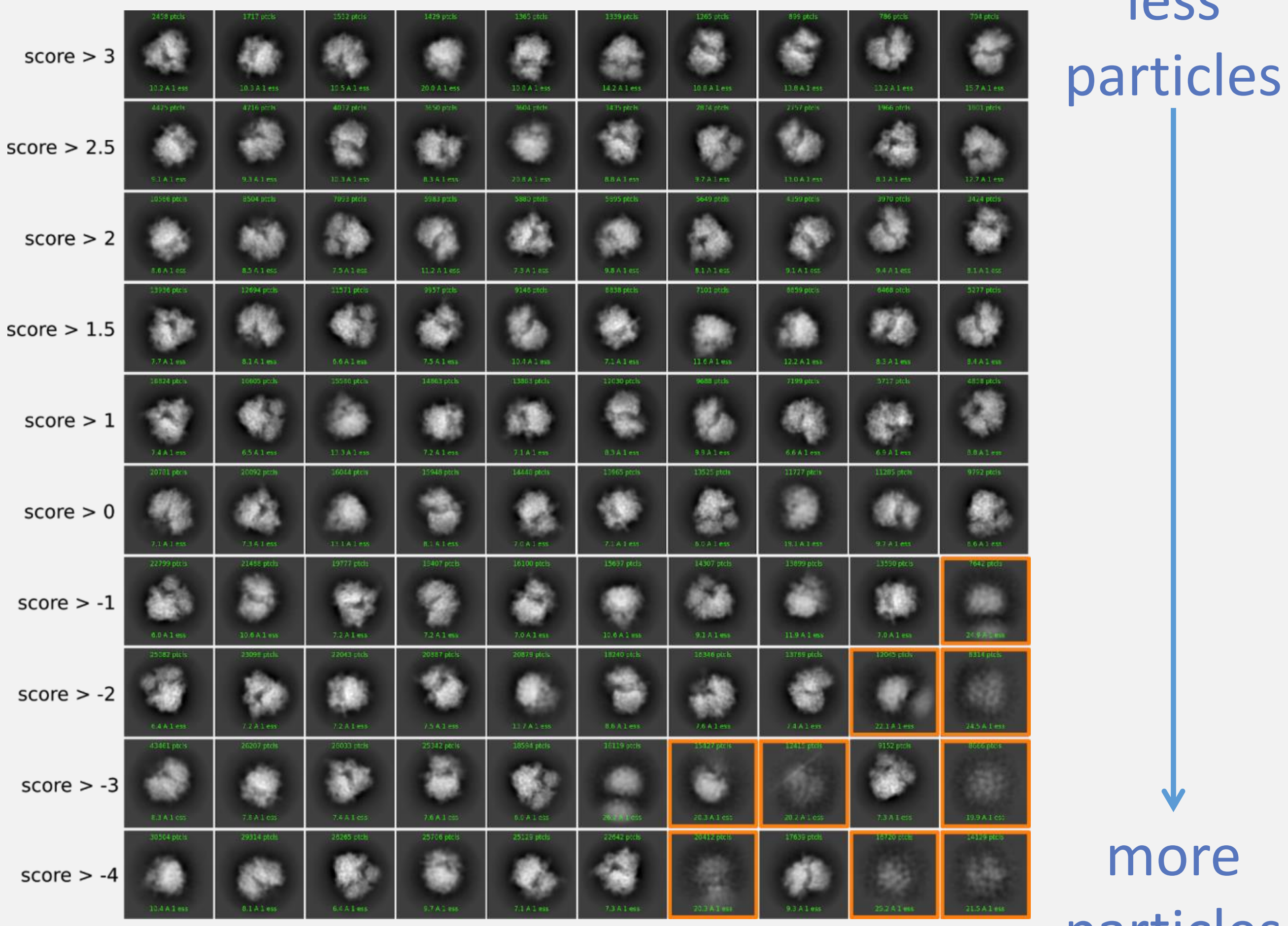




TOPAZ ranks particles well, avoids junk

(2D classes vs. picking threshold)

80S
Ribosome

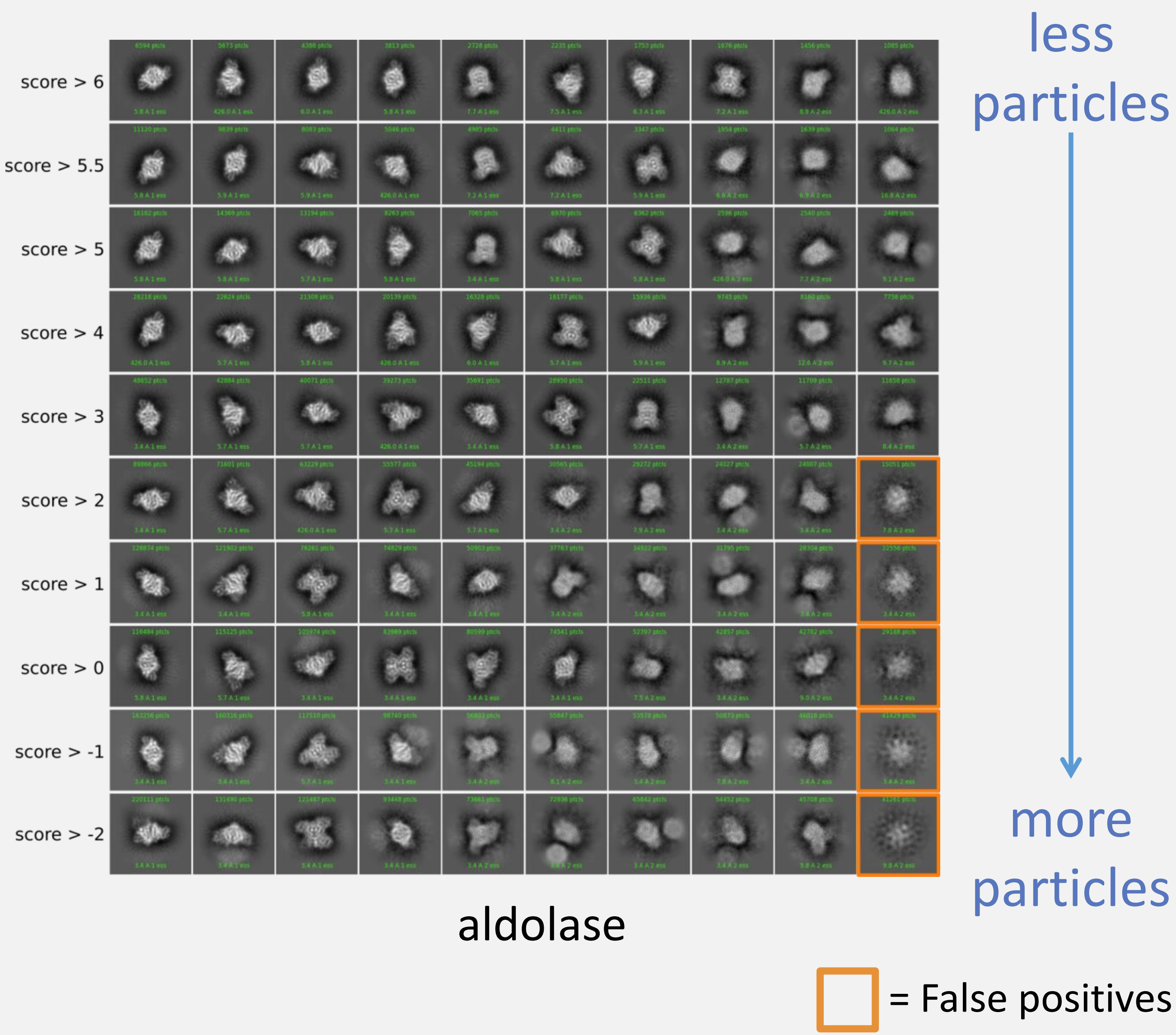


EMPIAR-10025

TOPAZ ranks particles well, avoids junk

(2D classes vs. picking threshold)

Aldolase



aldolase

TOPAZ ranks particles well, avoids junk

(2D classes vs. picking threshold)

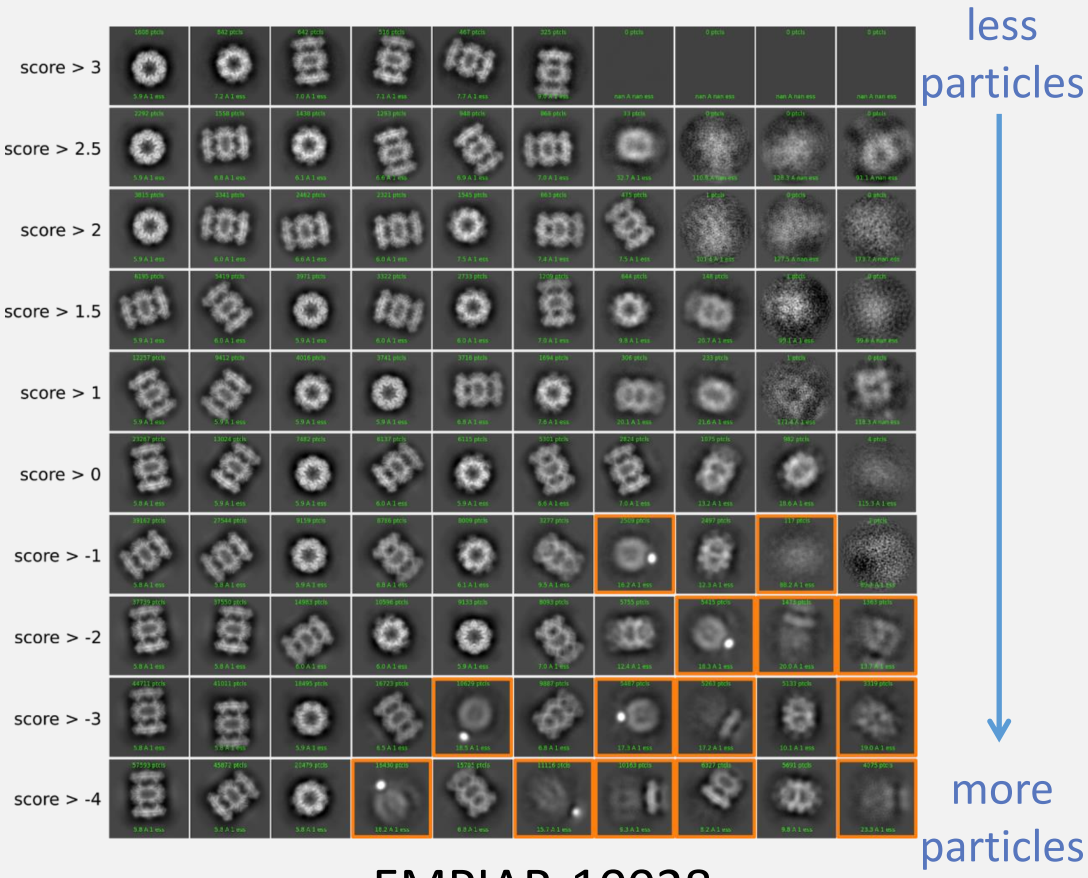
T20S Proteasome

EMPIAR-10028

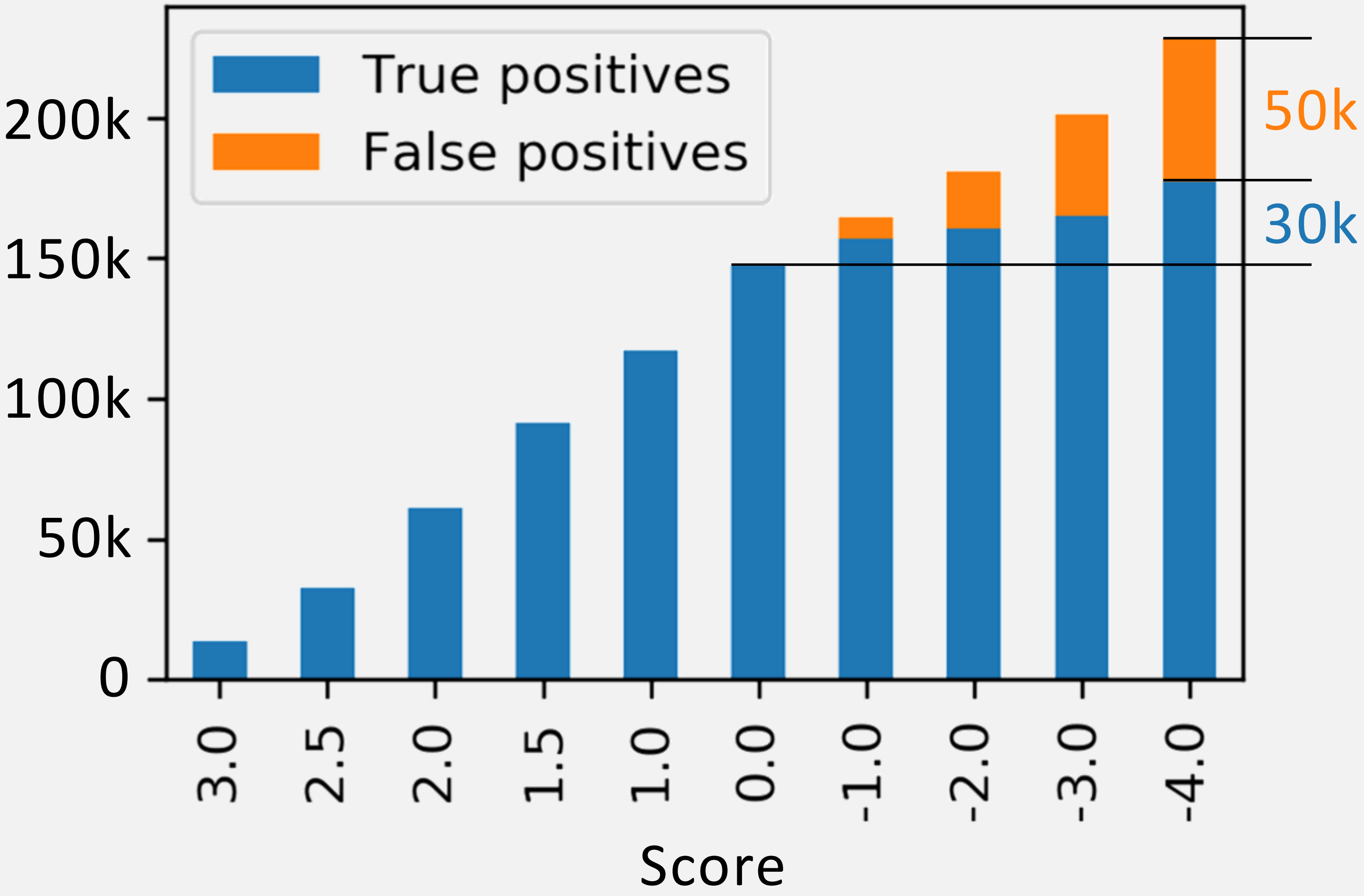
= False positives

TOPAZ ranks particles well, avoids junk

(Some good particles have negative scores)



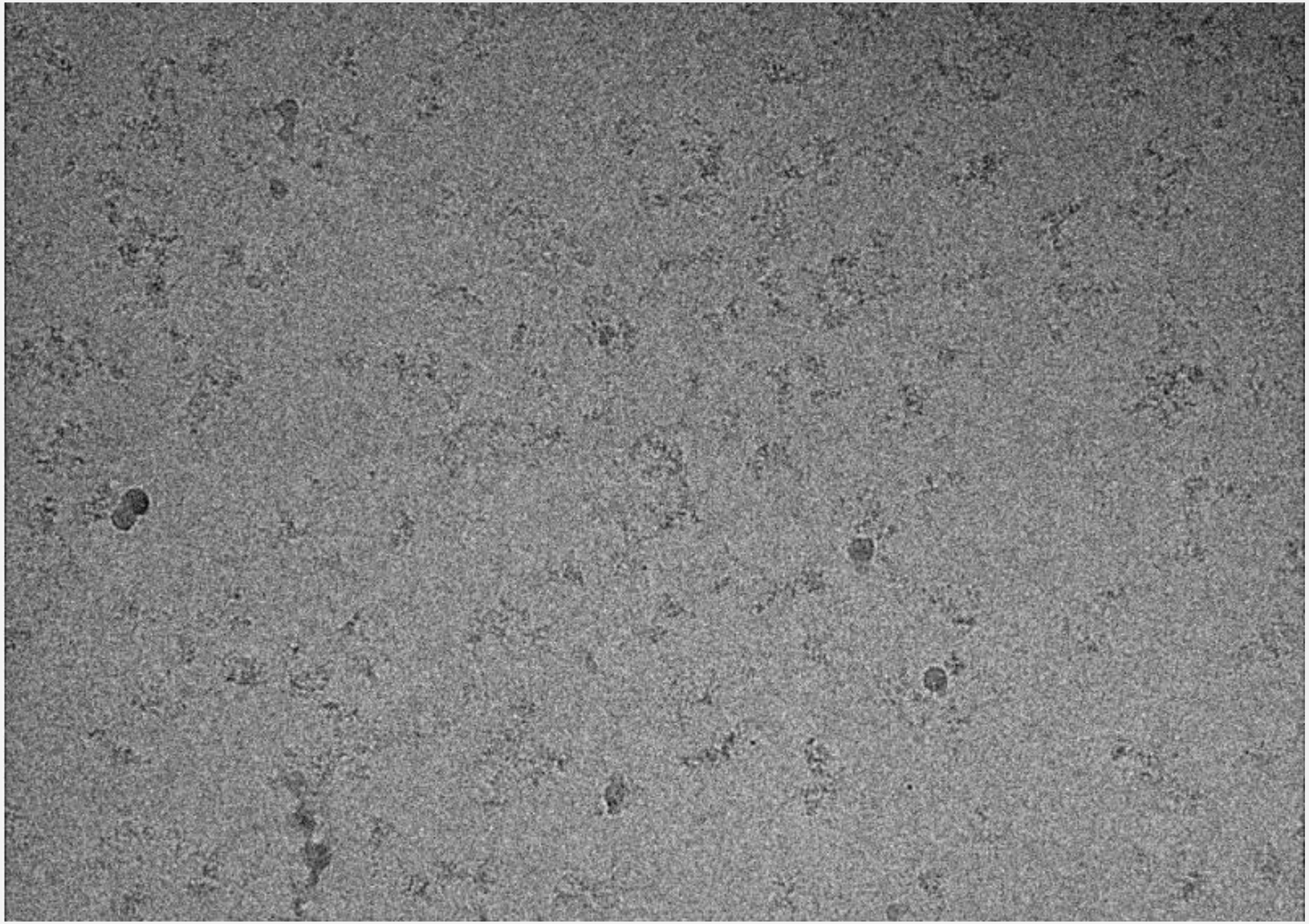
EMPIAR-10028



= False positives

COVID-19 picking successes

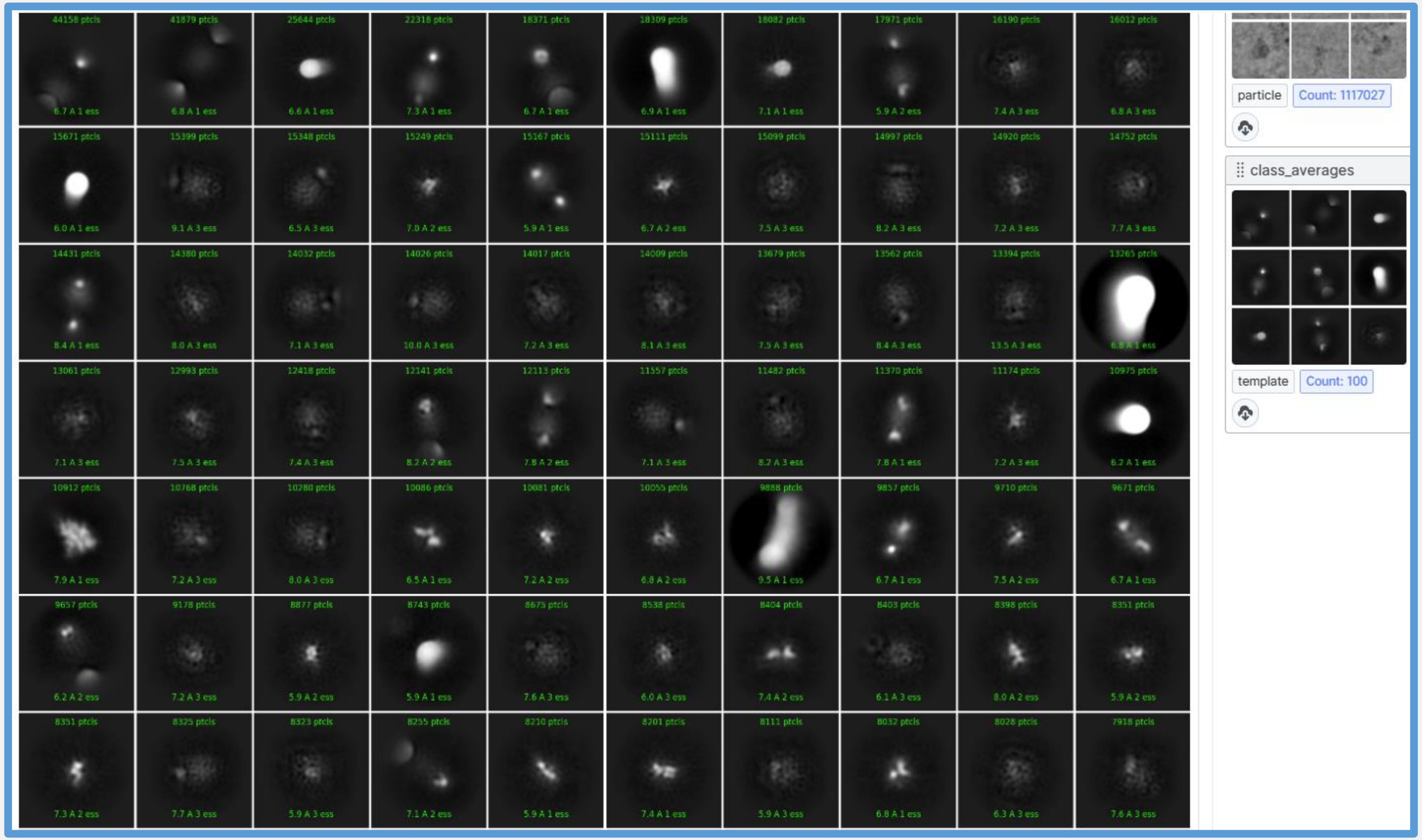
Topaz improves **COVID-19** spike **Eulers**

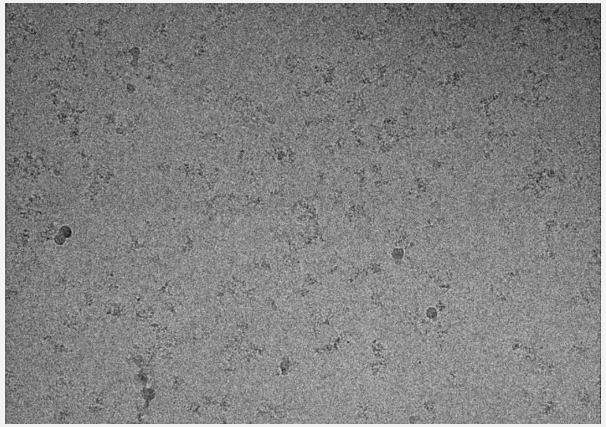


In collaboration with the Shapiro lab (Columbia)

Topaz improves **COVID-19** spike **Eulers**

Template & blob picking
1.1m initial particles

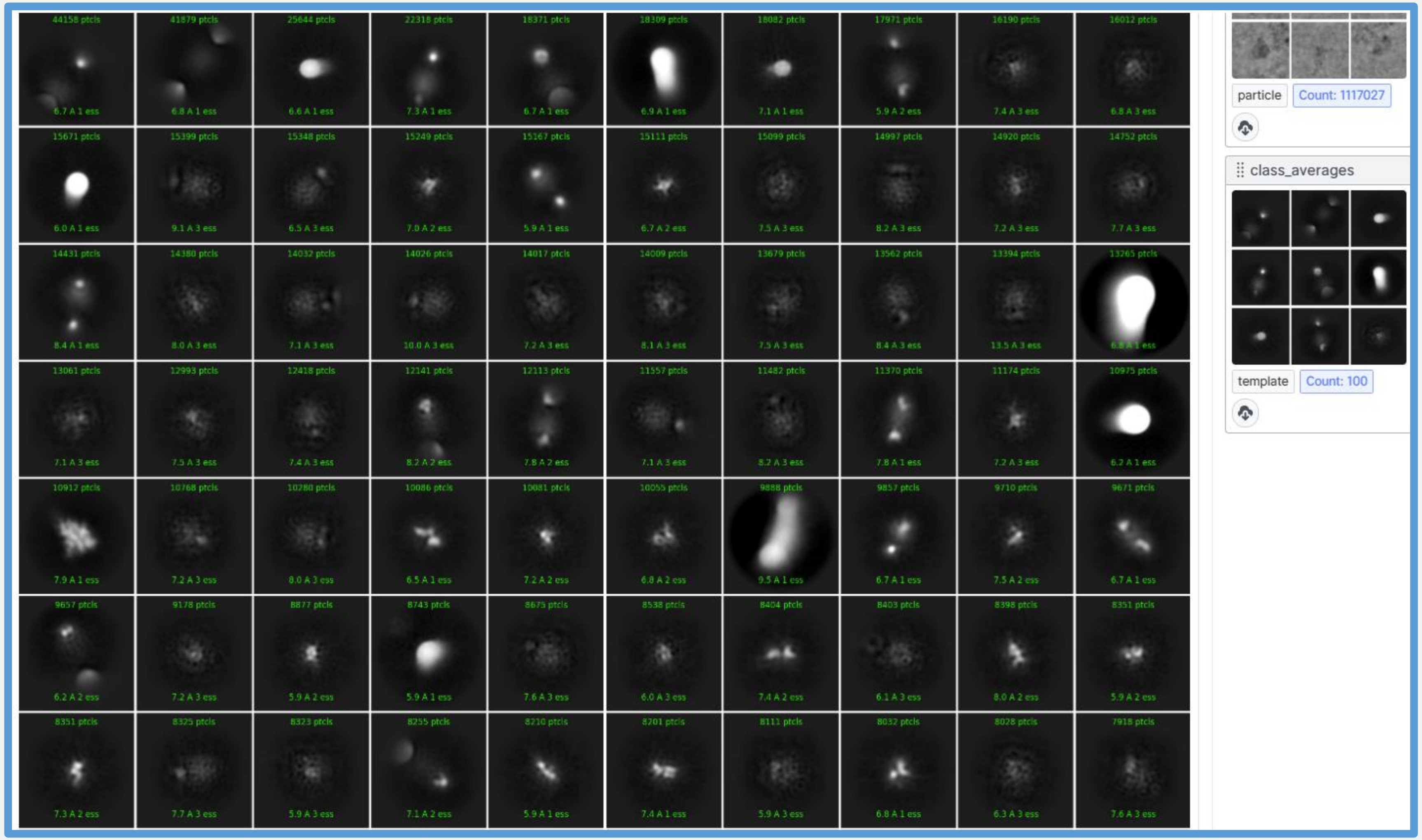


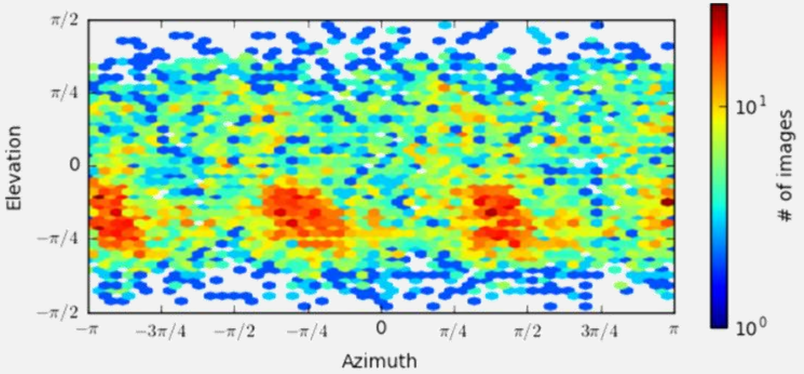


In collaboration with the Shapiro lab (Columbia)

Topaz improves **COVID-19** spike **Eulers**

Template & blob picking
1.1m initial particles



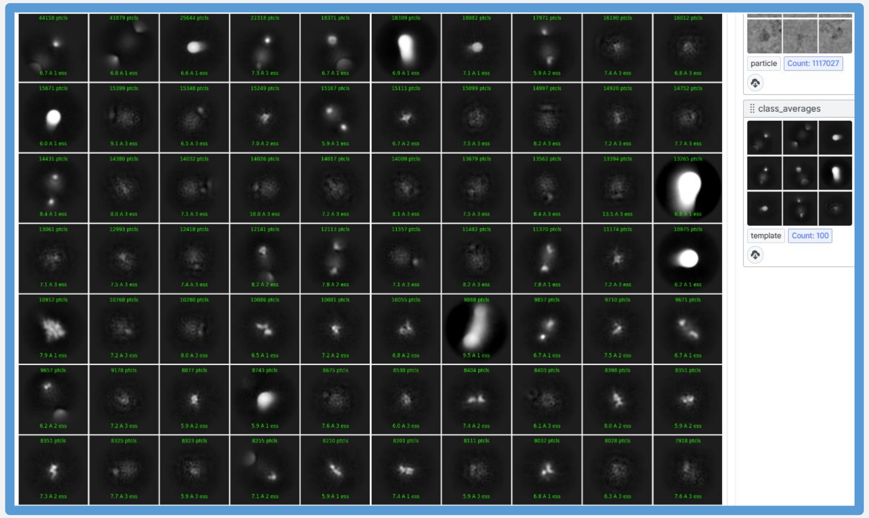


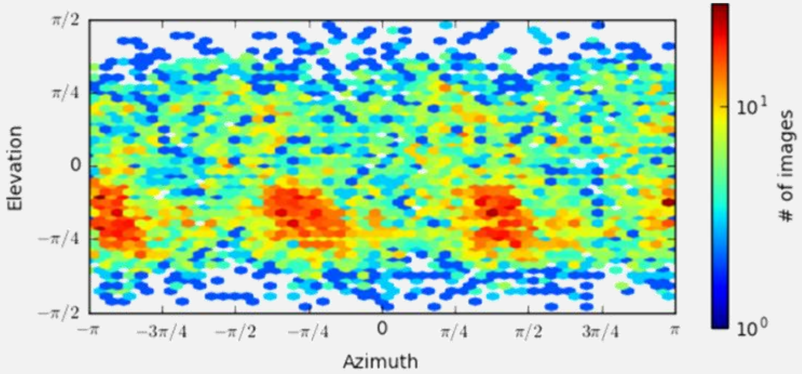
13k final particles

In collaboration with the Shapiro lab (Columbia)

Topaz improves **COVID-19** spike **Eulers**

Template & blob picking
1.1m initial particles





13k final particles

↓

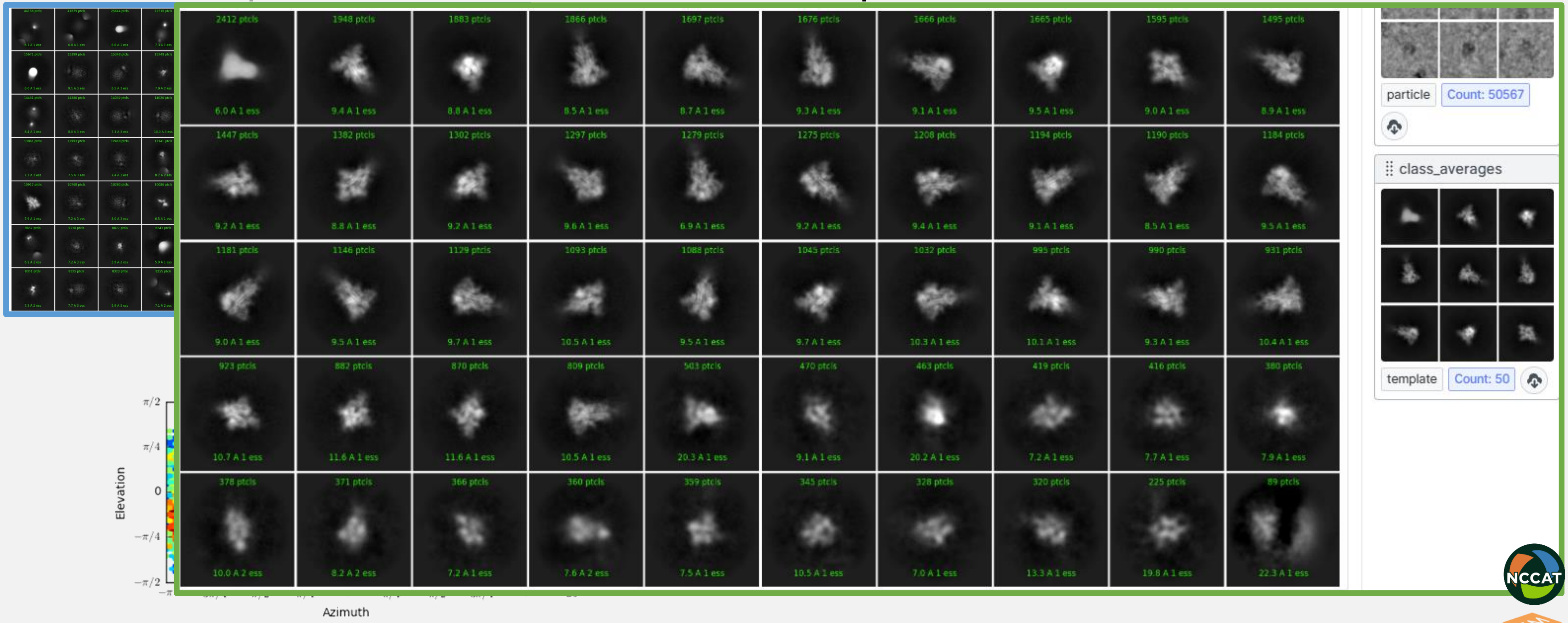
Train **Topaz**

In collaboration with the Shapiro lab (Columbia)

Topaz improves **COVID-19** spike **Eulers**

Template & blob picking
1.1m initial particles

Topaz picking
50k initial particles

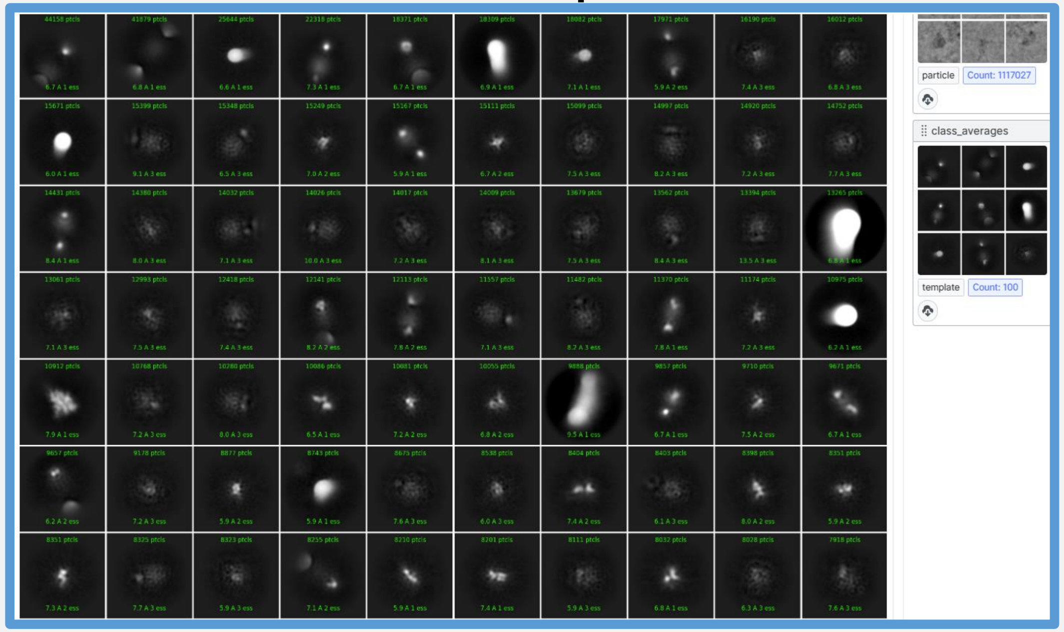


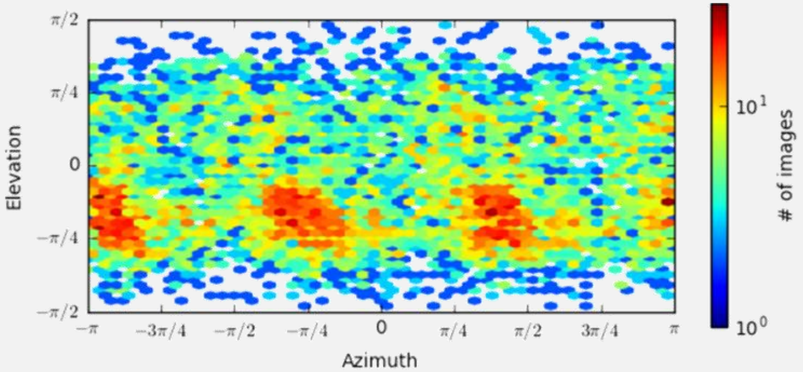
13k final particles

In collaboration with the Shapiro lab (Columbia)

Topaz improves **COVID-19** spike **Eulers**

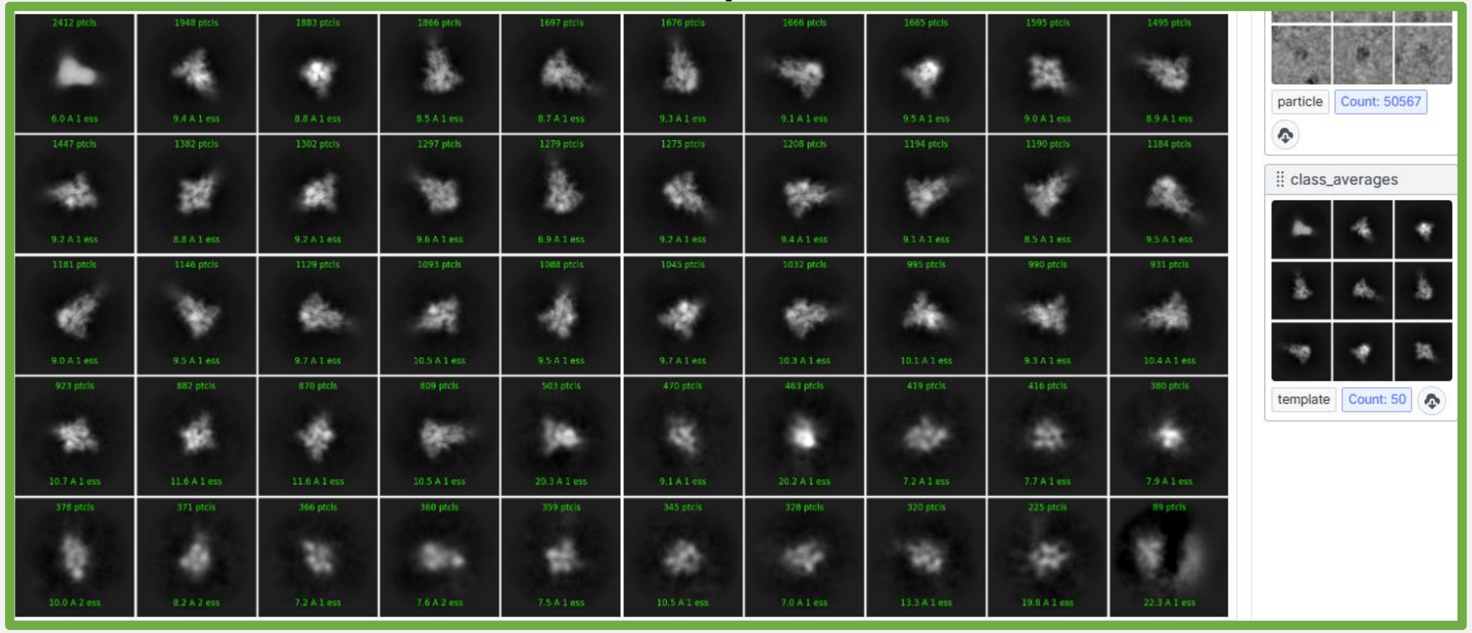
Template & blob picking
1.1m initial particles

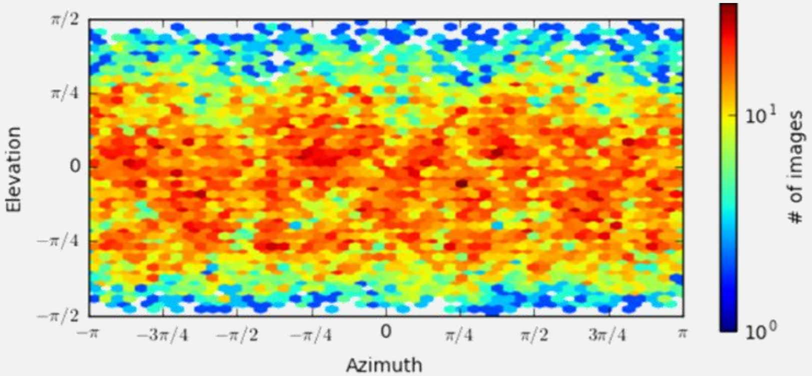




13k final particles

Topaz picking
50k initial particles





29k final particles

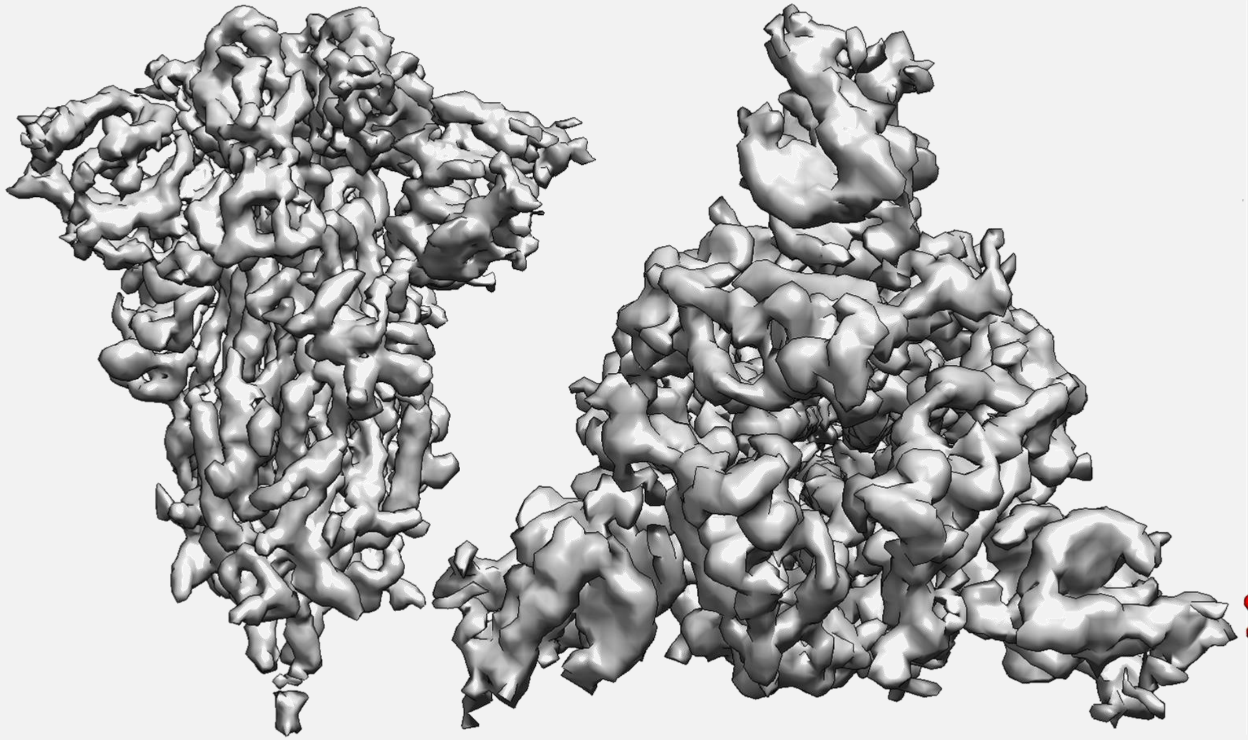
In collaboration with the Shapiro lab (Columbia)

NCCAT

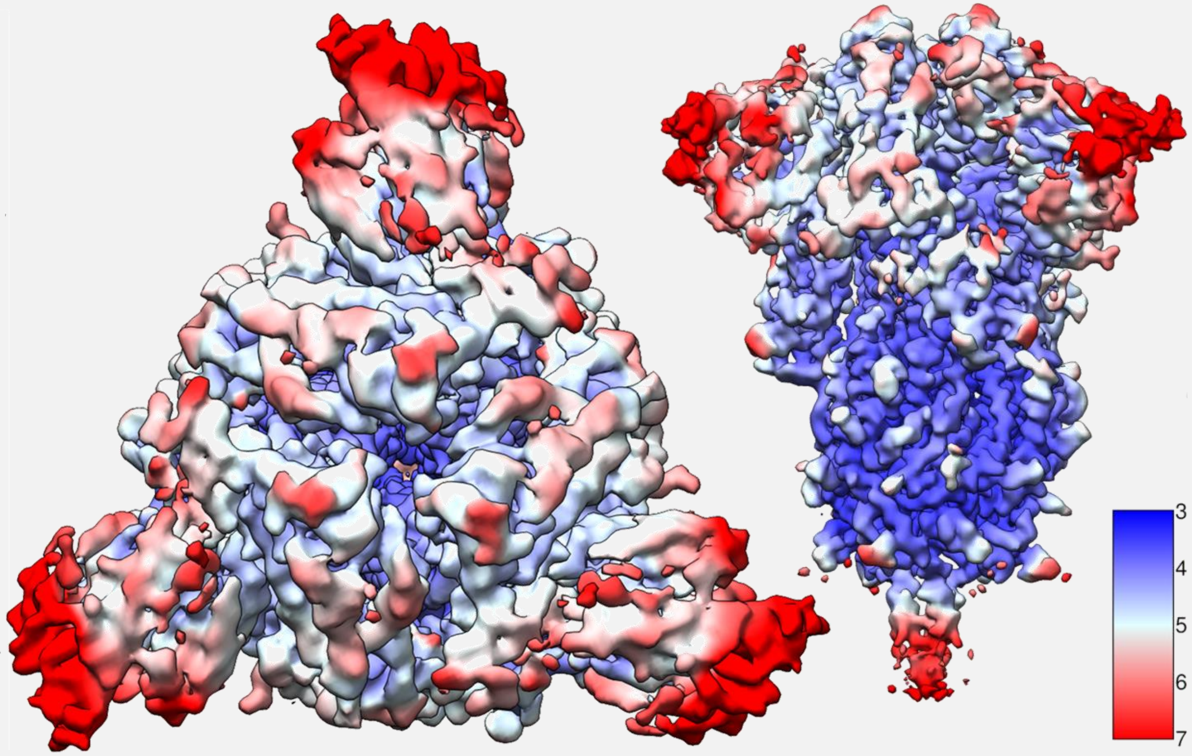
NRAMM

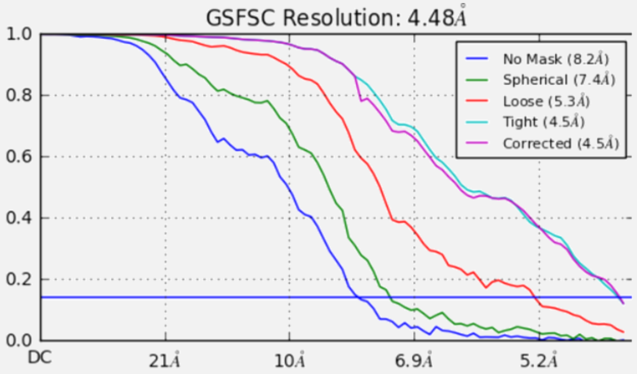
Topaz improved spike resolution by **1Å**

Template & blob picking



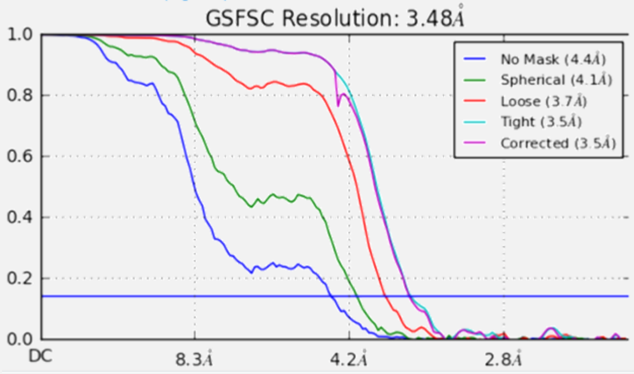
Topaz picking





4.48Å

3.48Å



In collaboration with the Shapiro lab (Columbia)

Can we optimize this workflow *further*?

It all depends on **training particles**

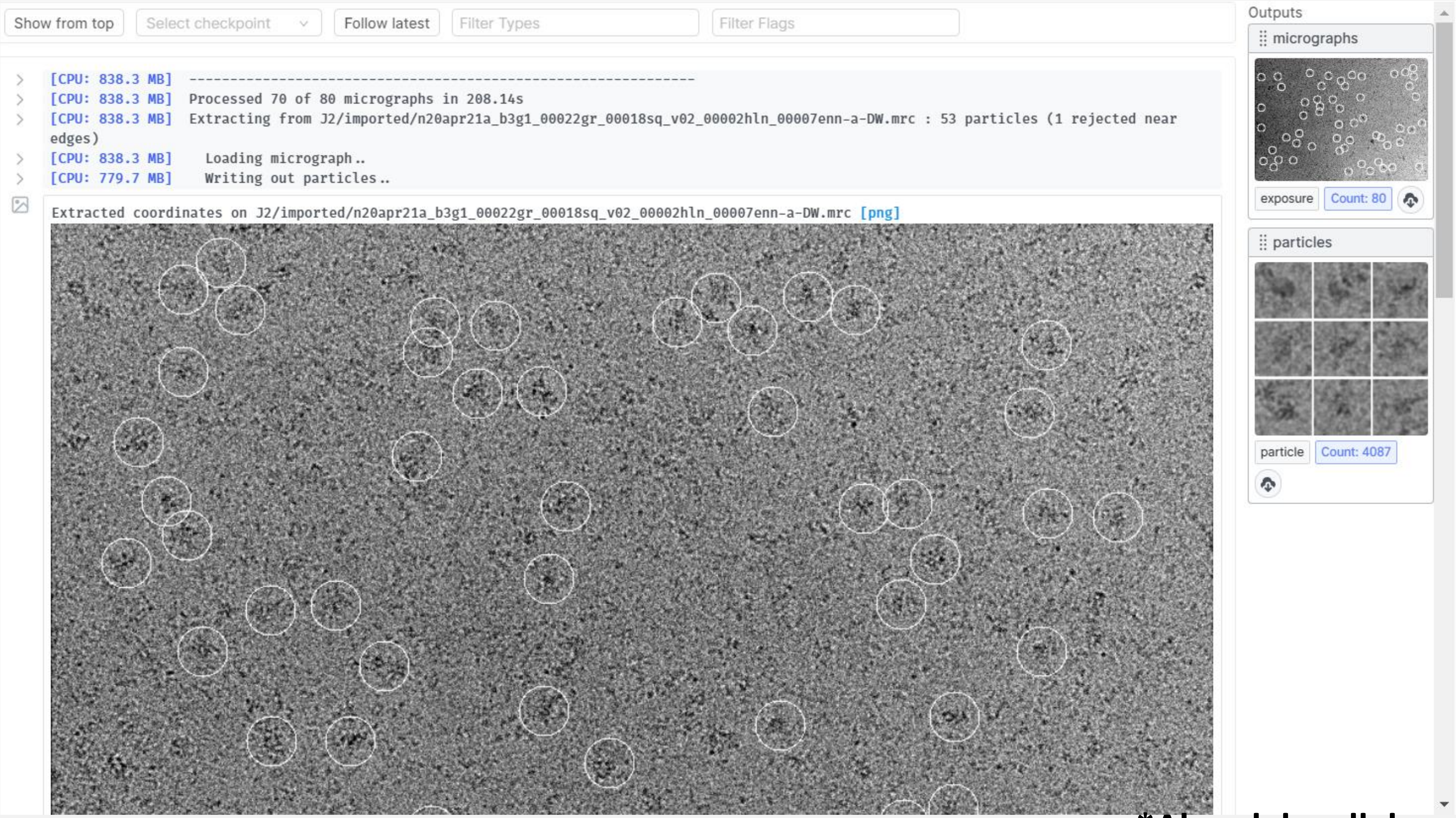
First workflow

Blob/template pick >> classify >> train Topaz

Better workflow?

Manually curate picks >> classify >> train Topaz

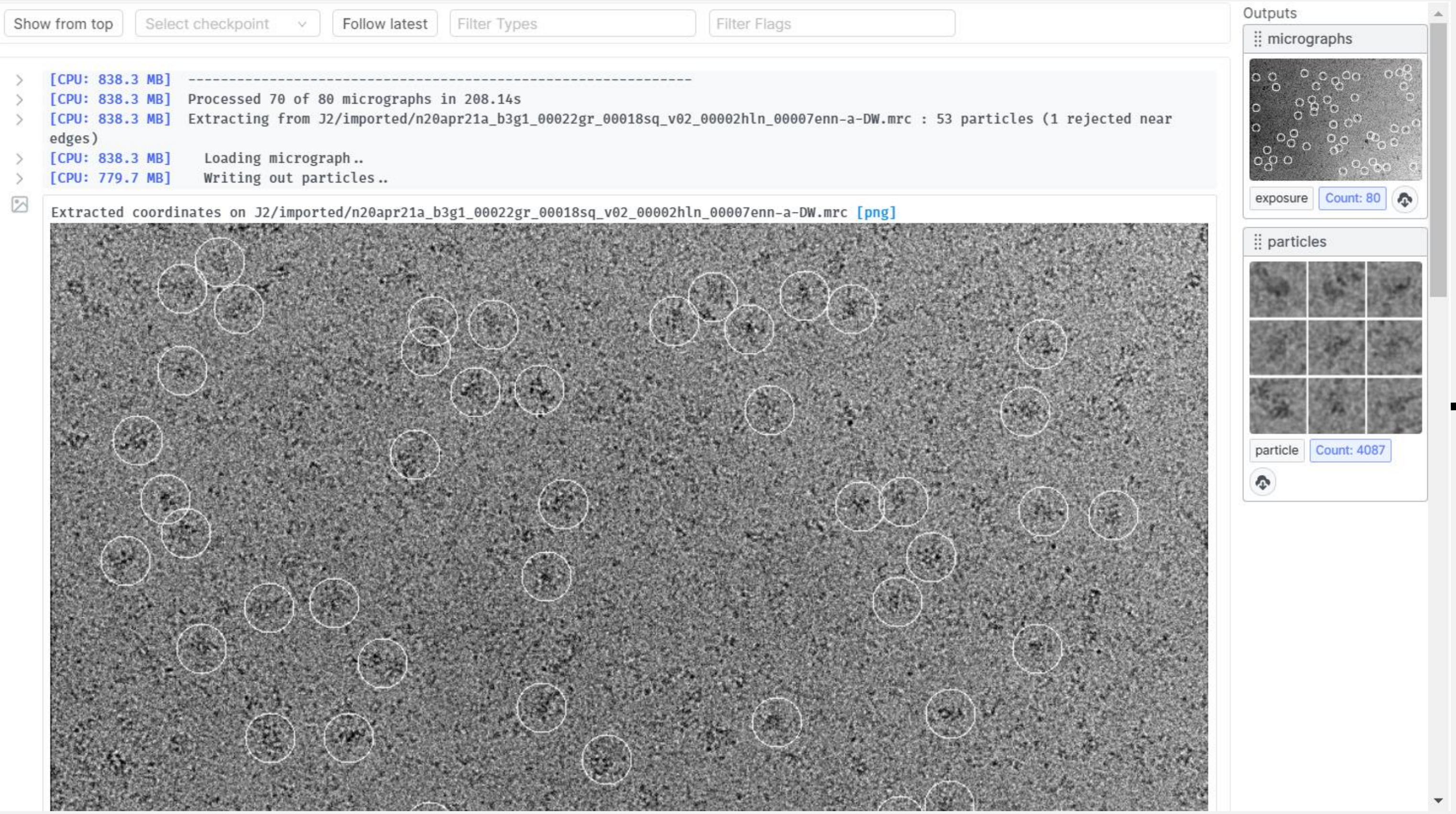
How well does manual curation work?

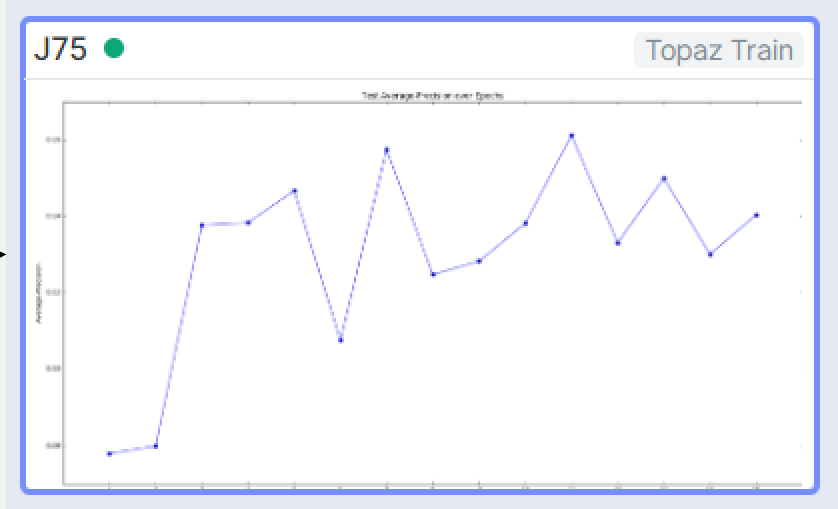


***Almo lab collaboration**

Herrera et al., ACS Omega 2020

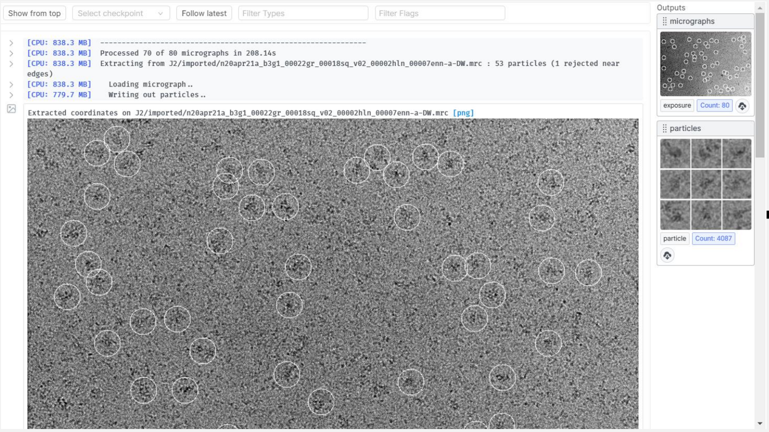
How well does manual curation work?



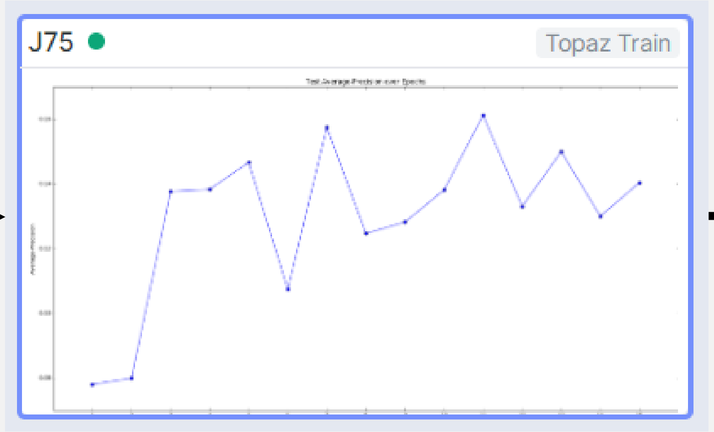


***Almo lab collaboration**
Herrera et al., ACS Omega 2020

How well does manual curation work?



Manual curation

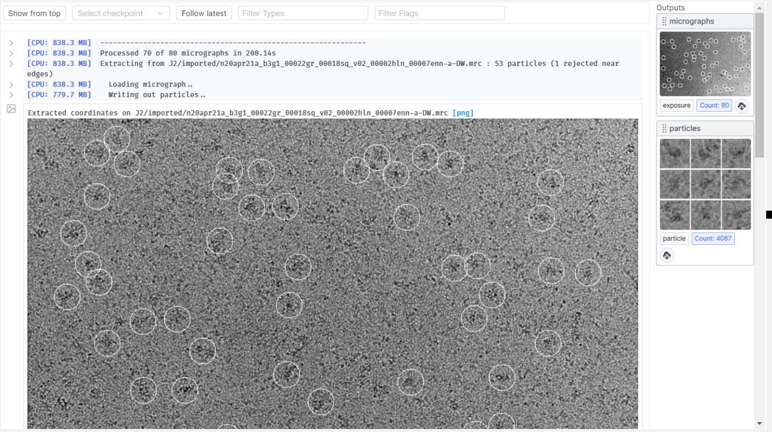


Train model

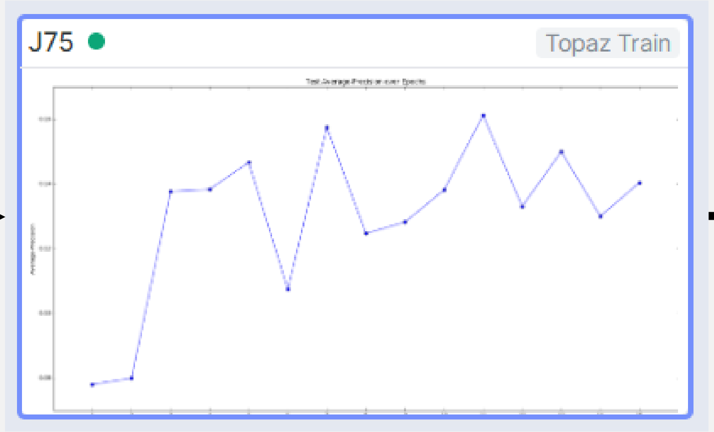
Pick 75k particles

***Almo lab collaboration**
Herrera et al., ACS Omega 2020

How well does manual curation work?

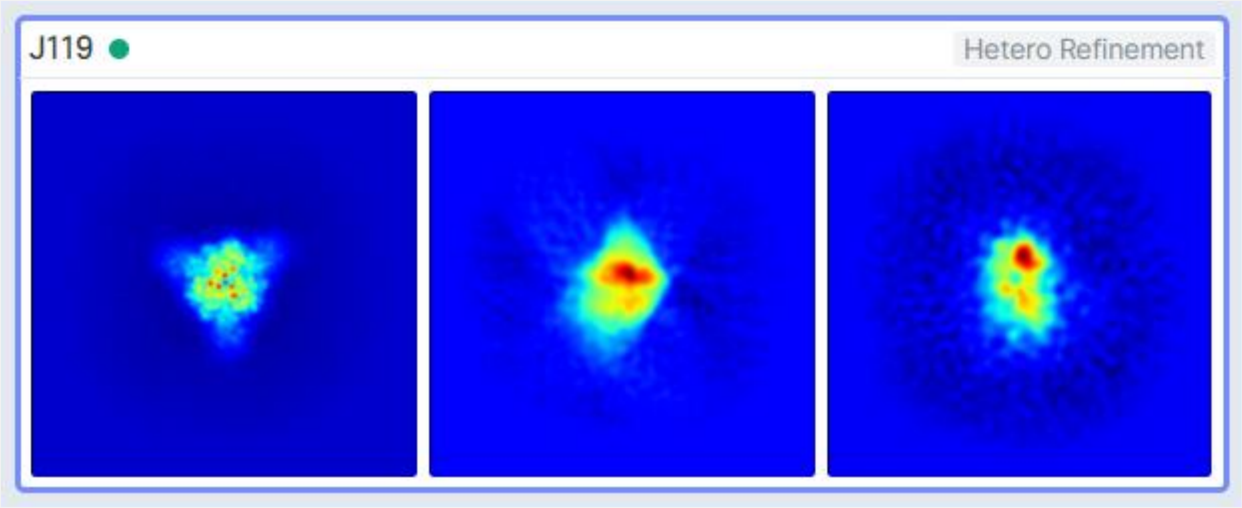


Manual curation



Train model

Pick 75k particles

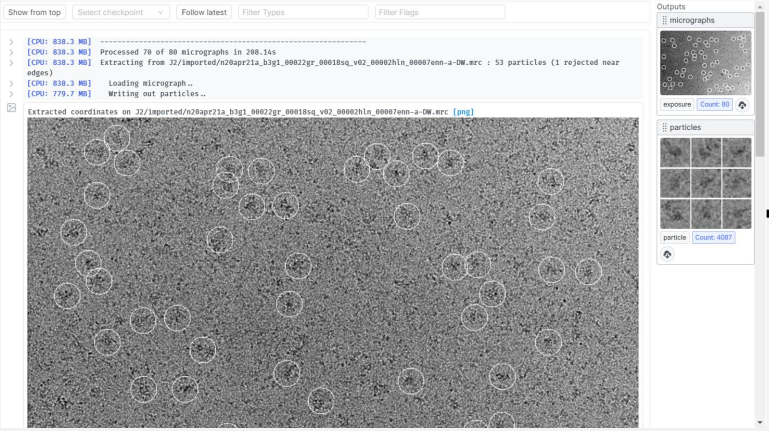


3D classification

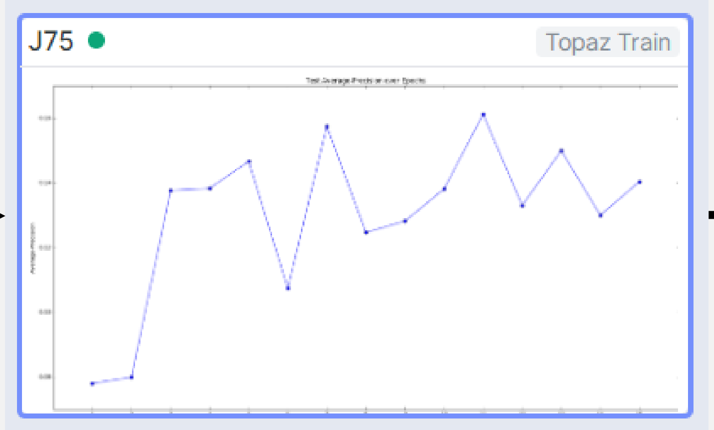
***Almo lab collaboration**

Herrera et al., ACS Omega 2020

How well does manual curation work?

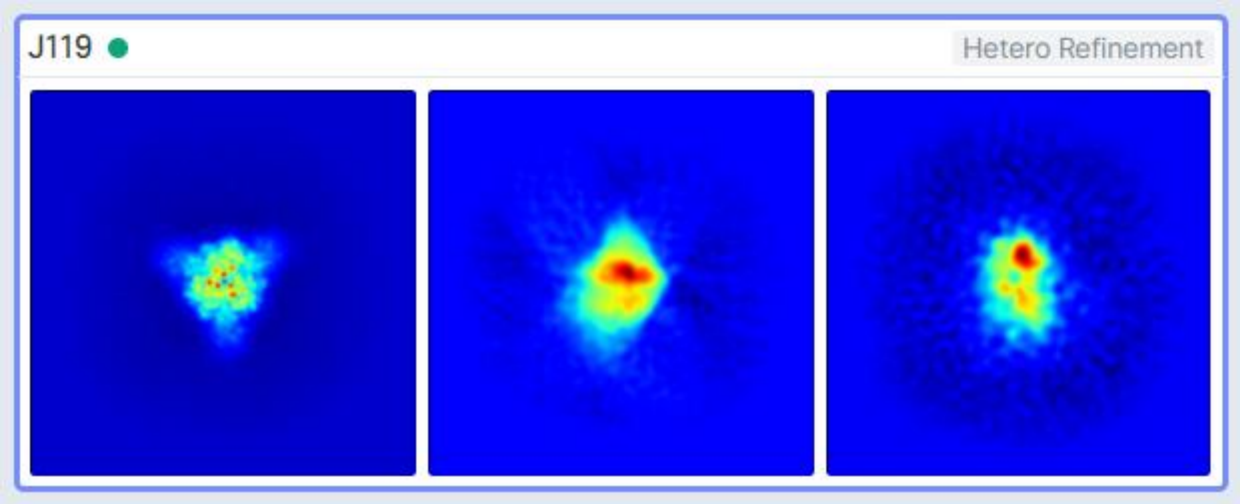


Manual curation

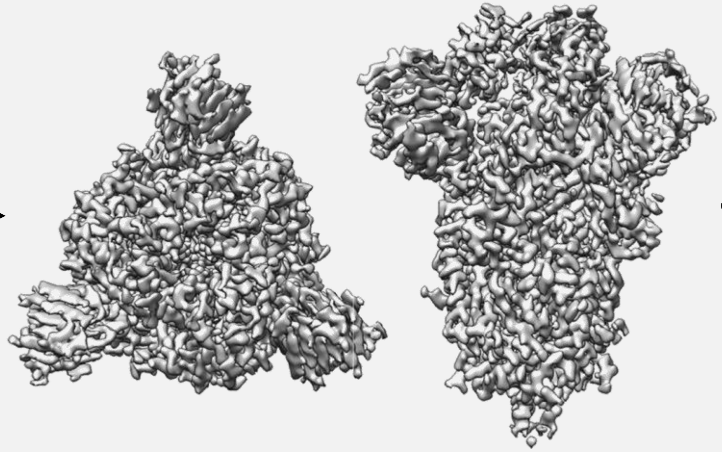


Train model

Pick 75k particles



3D classification

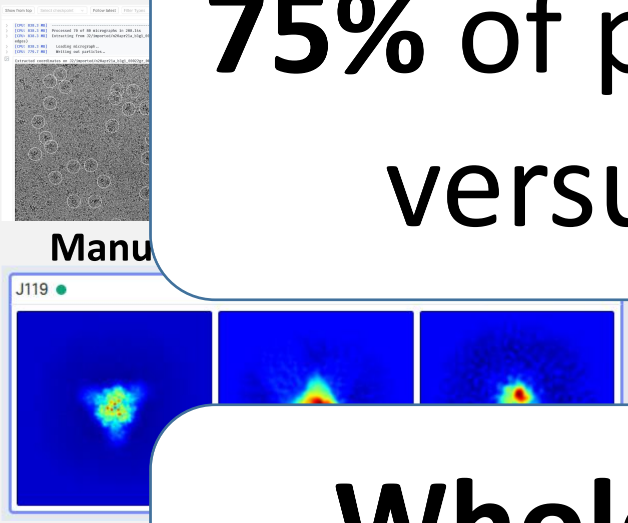


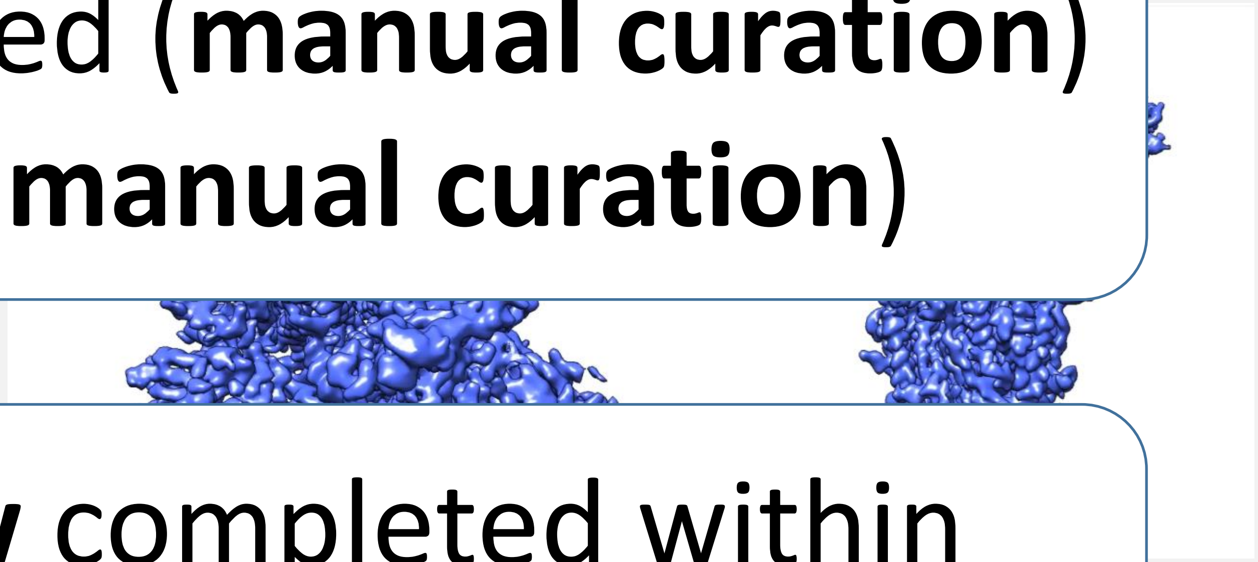
56k particles
3.2 Å

***Almo lab collaboration**
Herrera et al., ACS Omega 2020

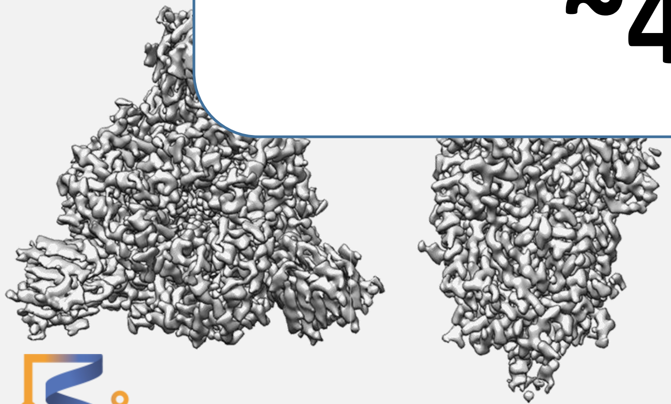
How well does manual curation work?

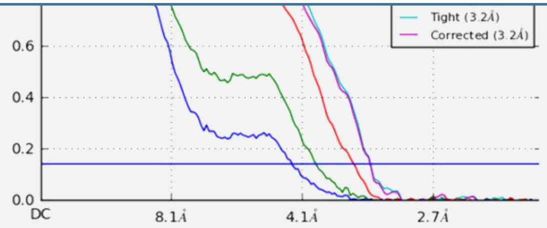
75% of particles used (**manual curation**) versus 60% (**no manual curation**)





Whole workflow completed within **~48 hours** of data collection





56k particles
3.2 Å

***Almo lab collaboration**
Herrera et al., ACS Omega 2020

It all depends on **training particles**

First workflow

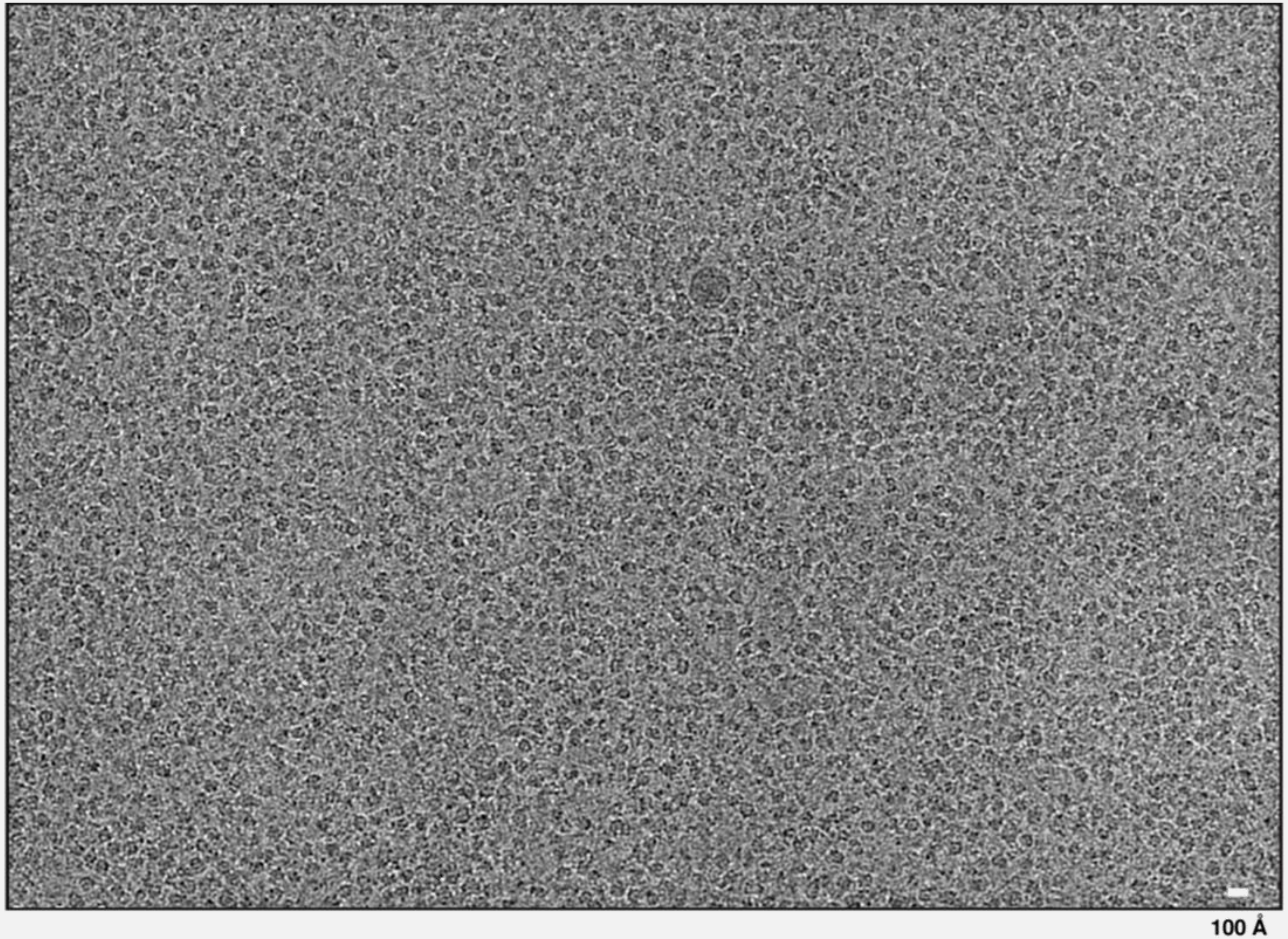
Blob/template pick >> classify >> train Topaz

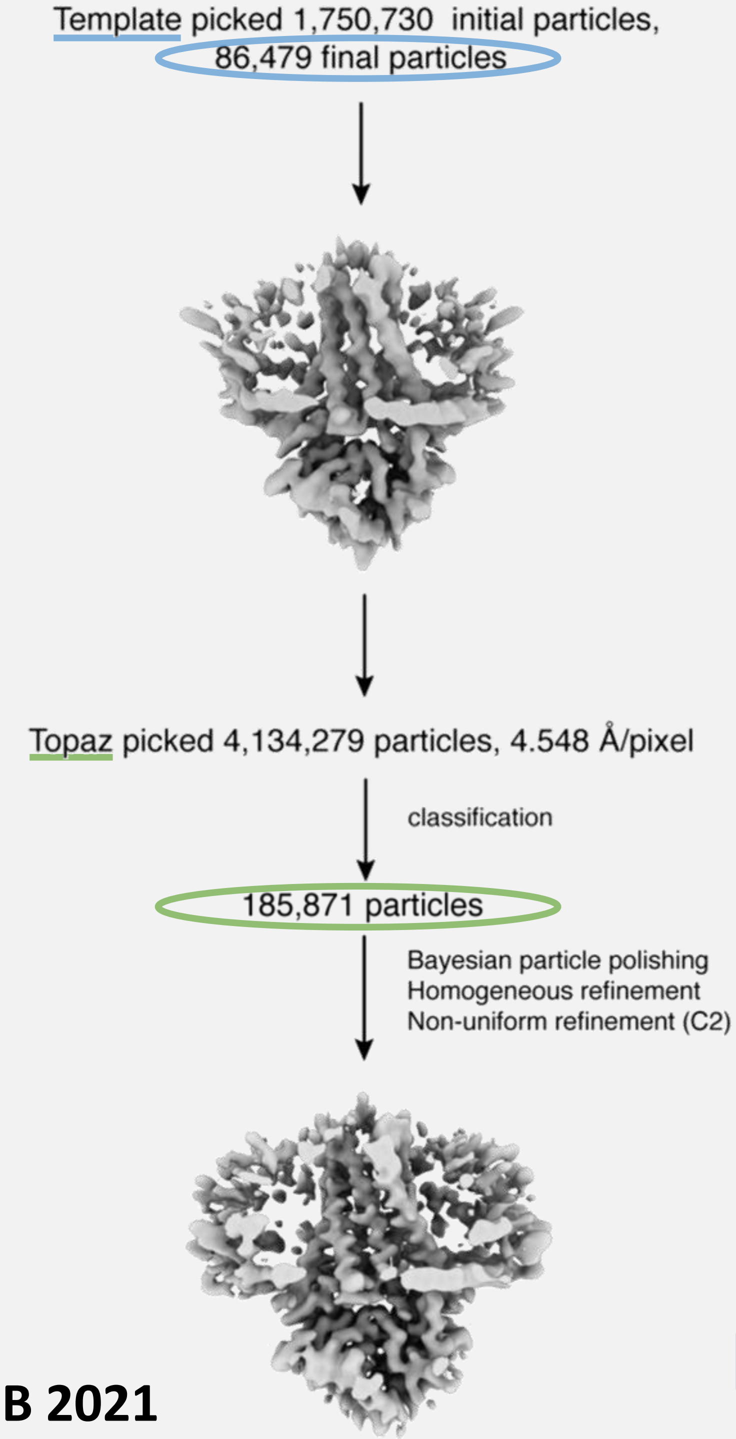
Better workflow ✓

Manually curate picks >> classify >> train Topaz

Some Topaz use examples in the literature

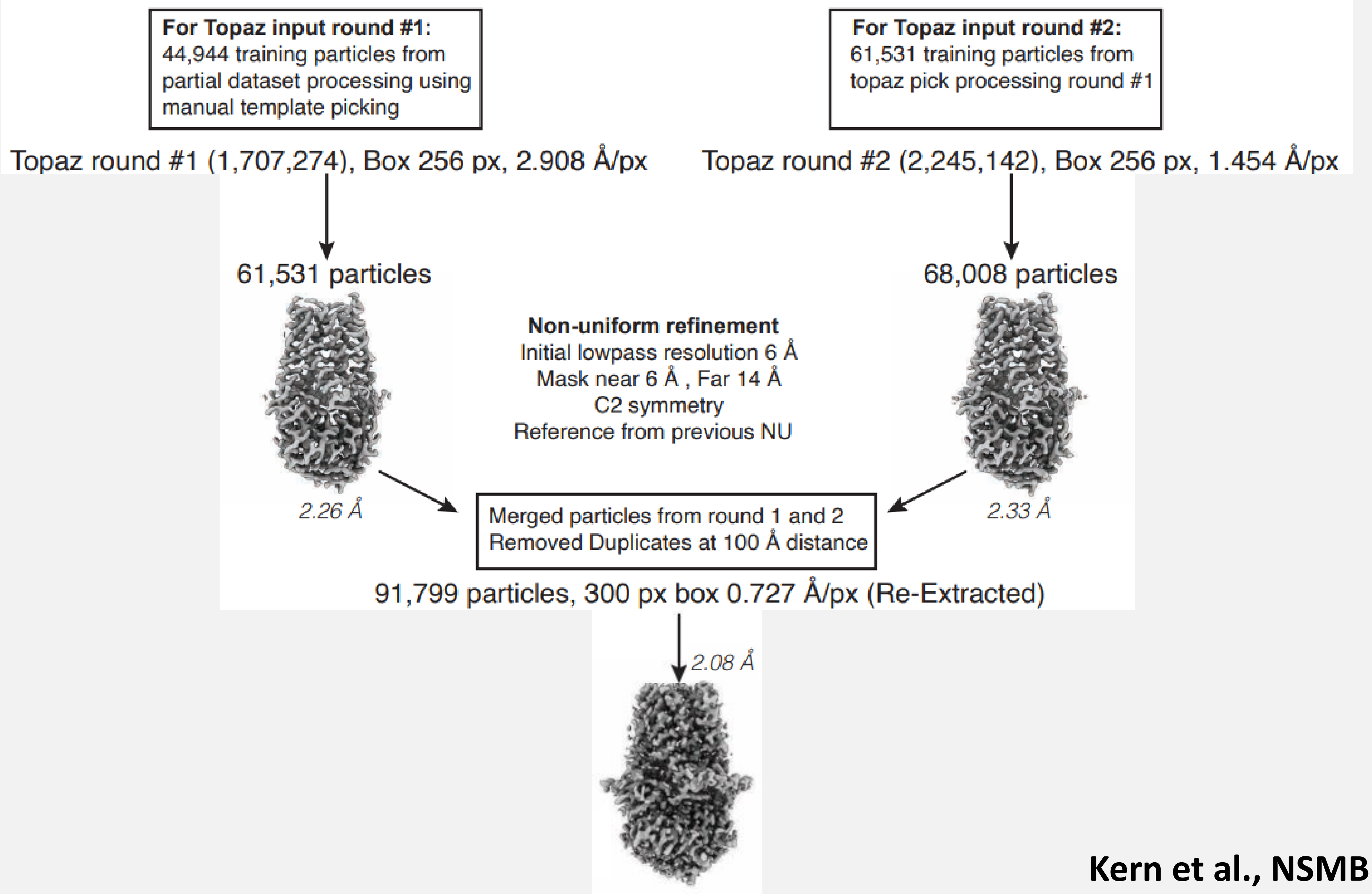
SARS-CoV-2 in nanodisc





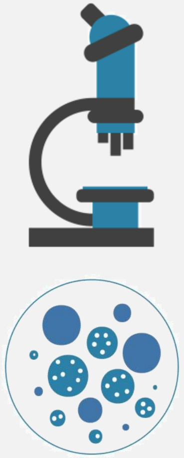
Kern et al., NSMB 2021

SARS-CoV-2 in nanodisc

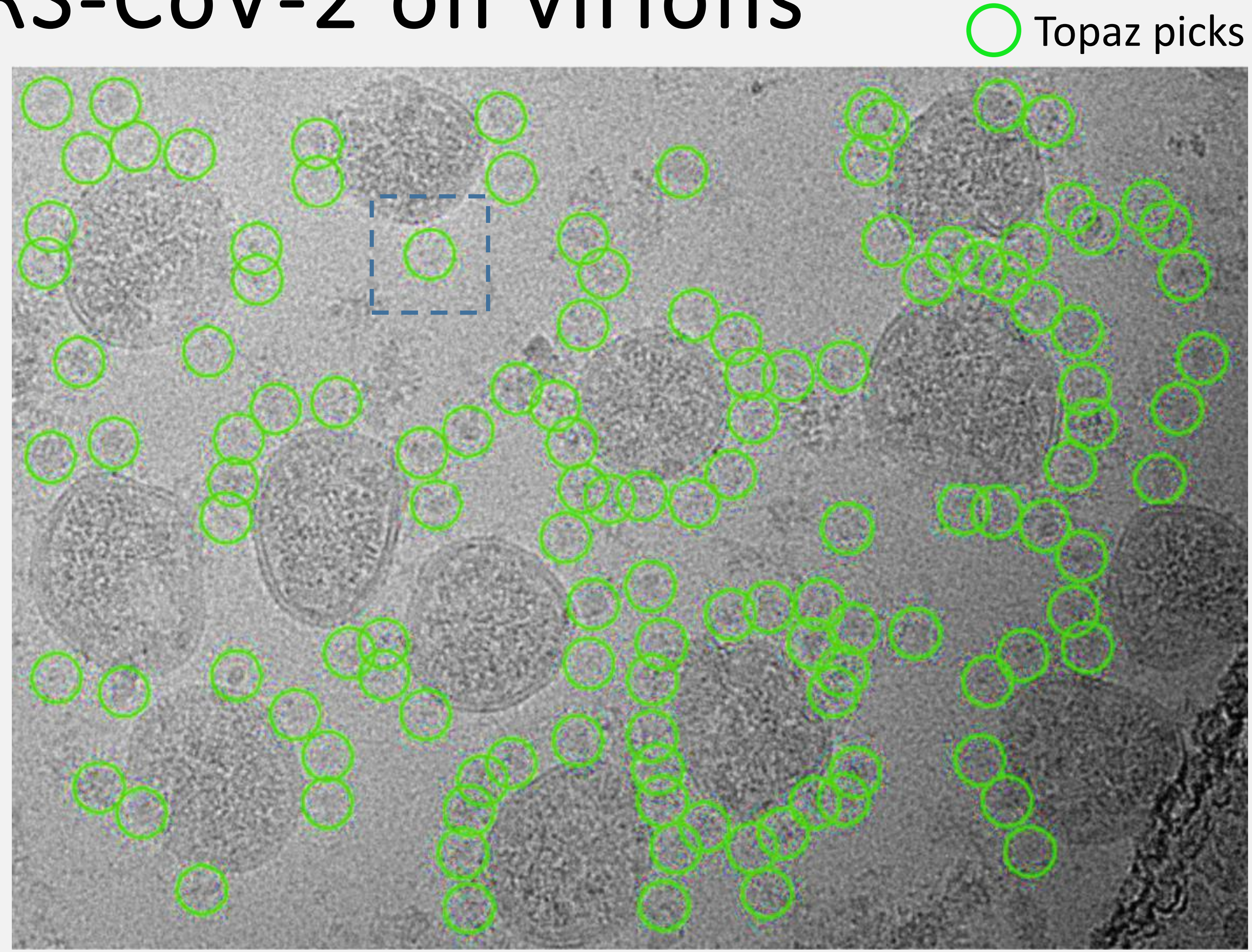


Kern et al., NSMB 2021

SARS-CoV-2 on virions

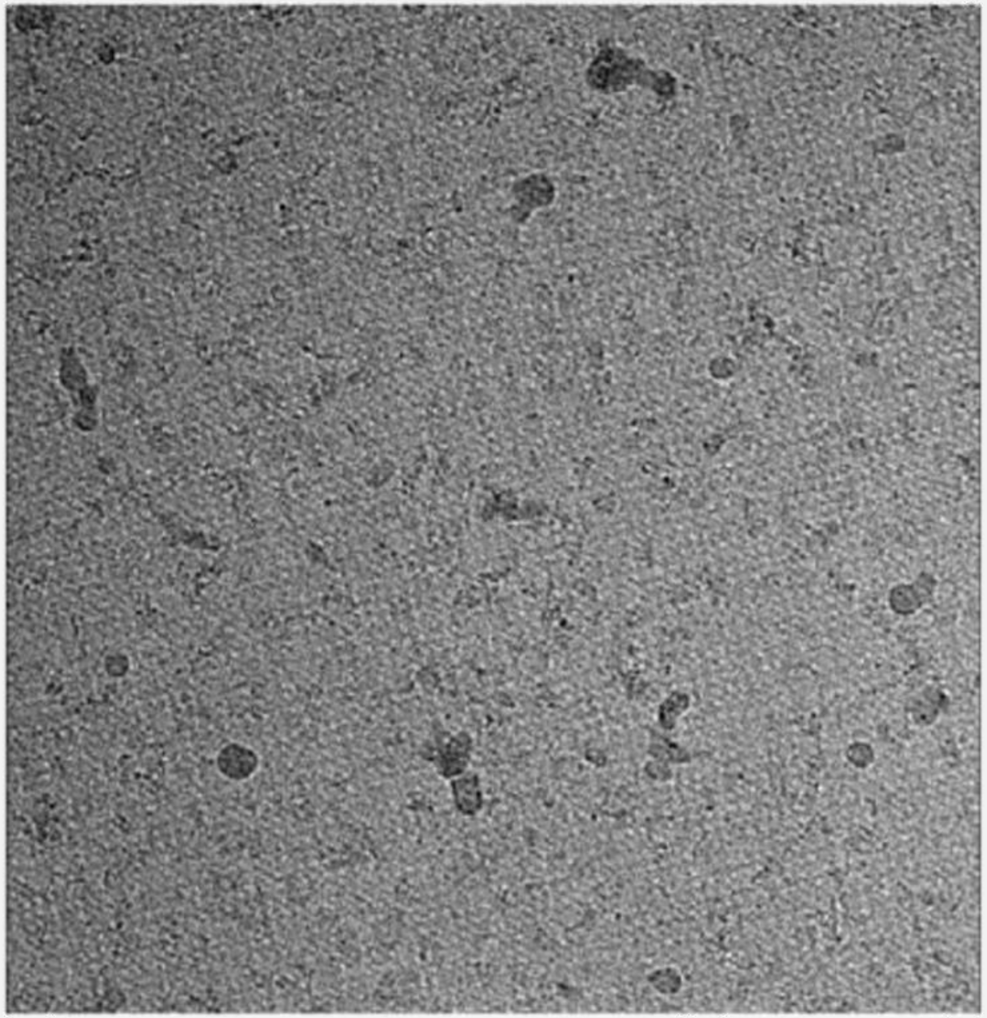


After manually curating 100 micrographs:



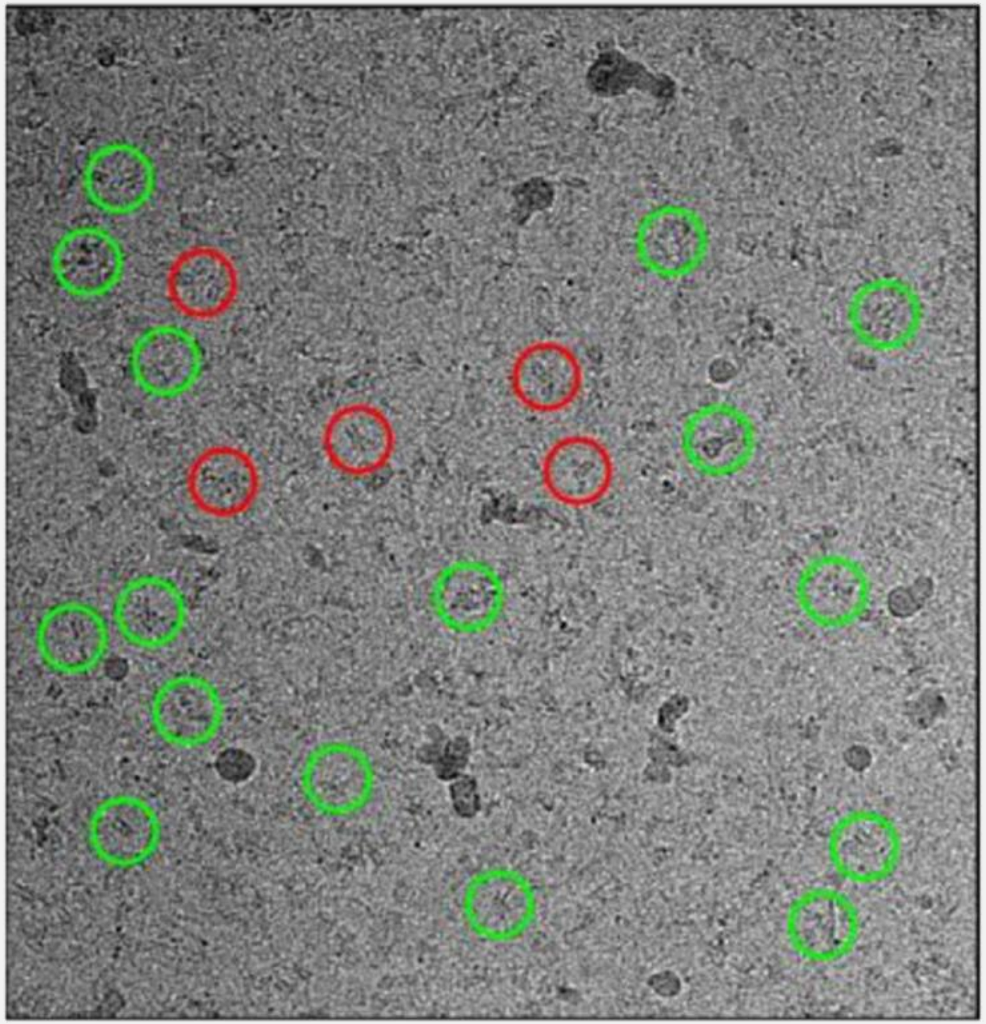
Ke et al., Nature 2020

Polycomb complex



Streptavidin + sample

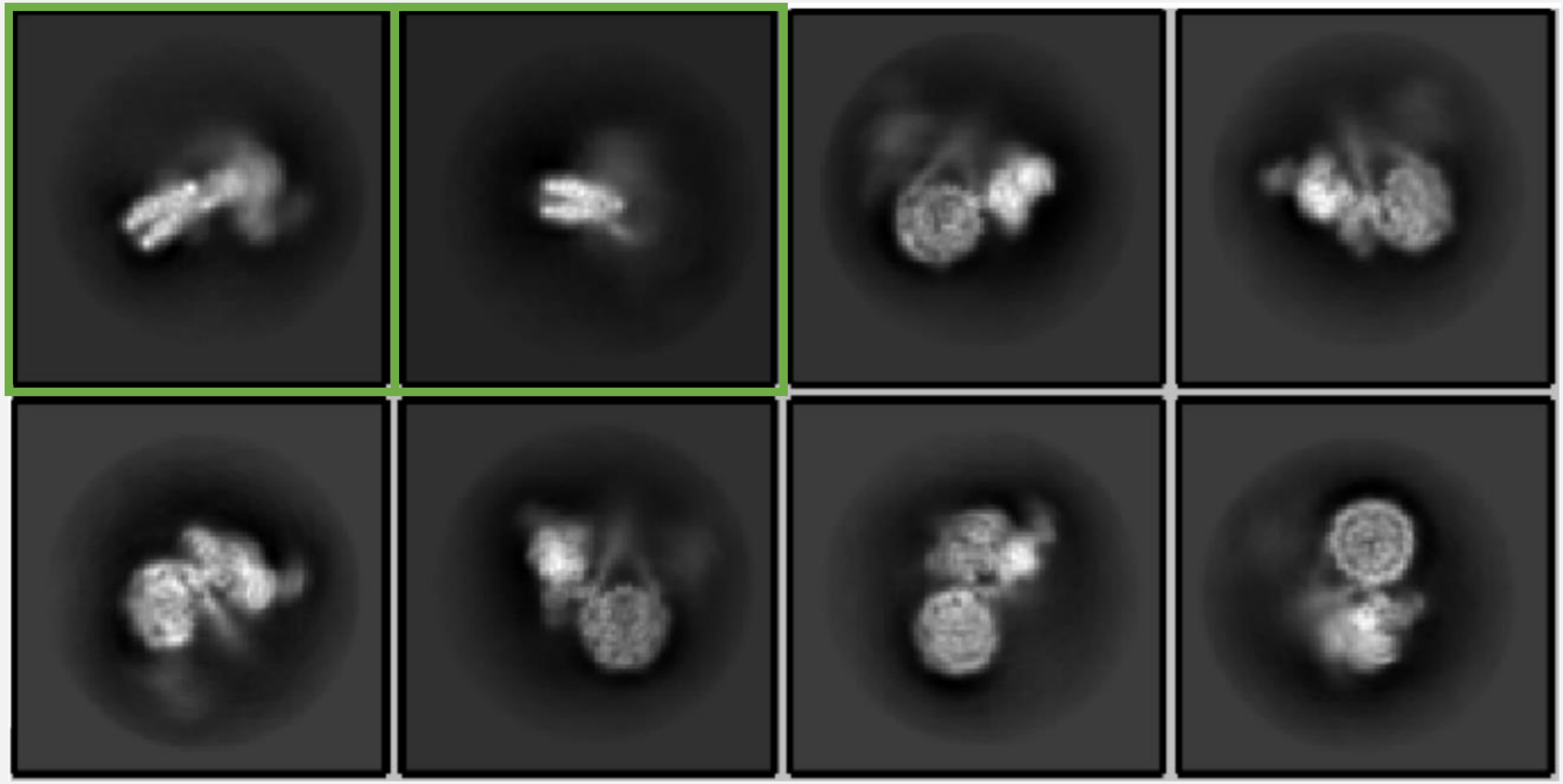
Lattice
Subtraction

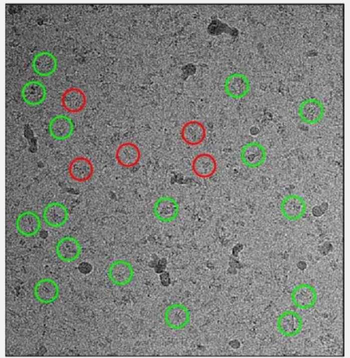


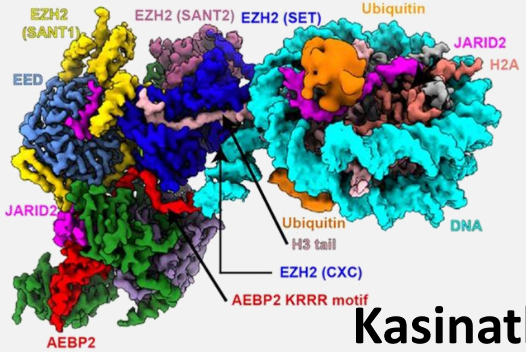
Kasinath et al., Science 2021

Polycomb complex

After trying Relion picking, EMAN neural picking, and Topaz, **only Topaz picked** **these views well**



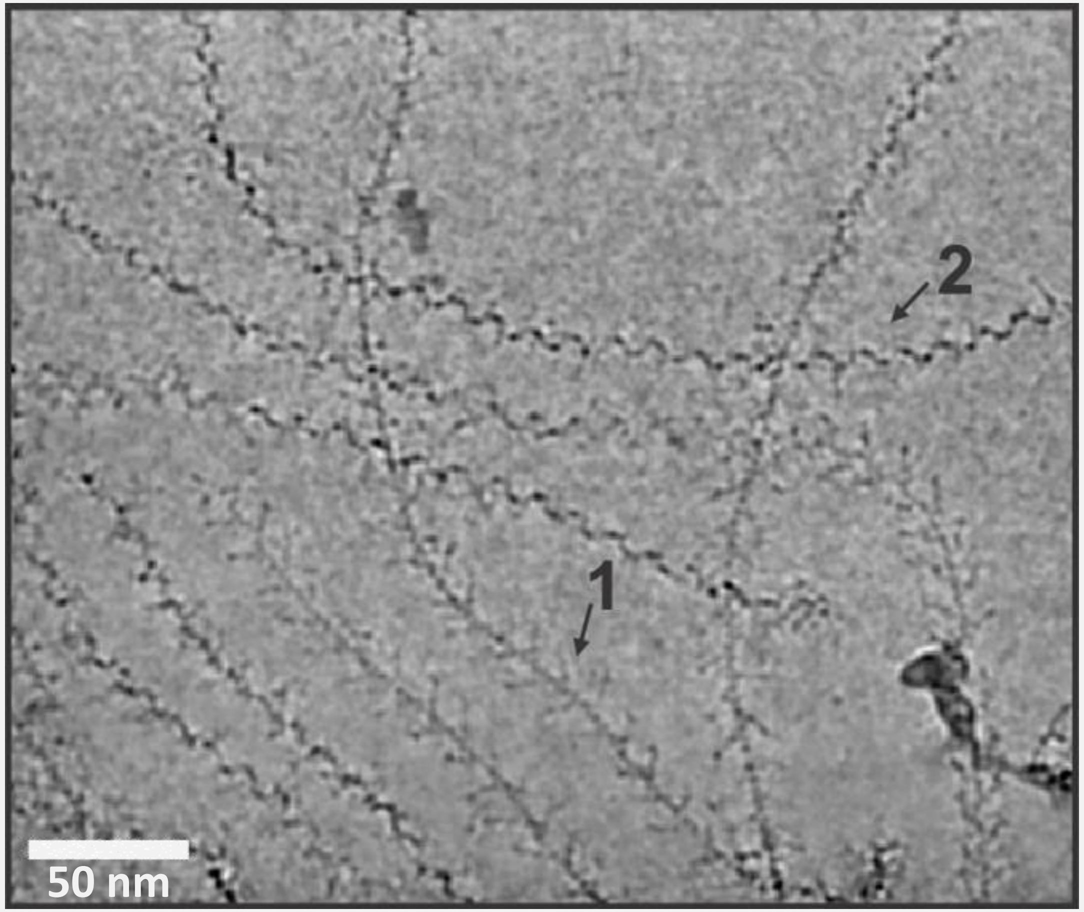




3.5 Å

Kasinath et al., Science 2021

Uromodulin filament core



Trained Topaz with 600 manually picked particles

↓

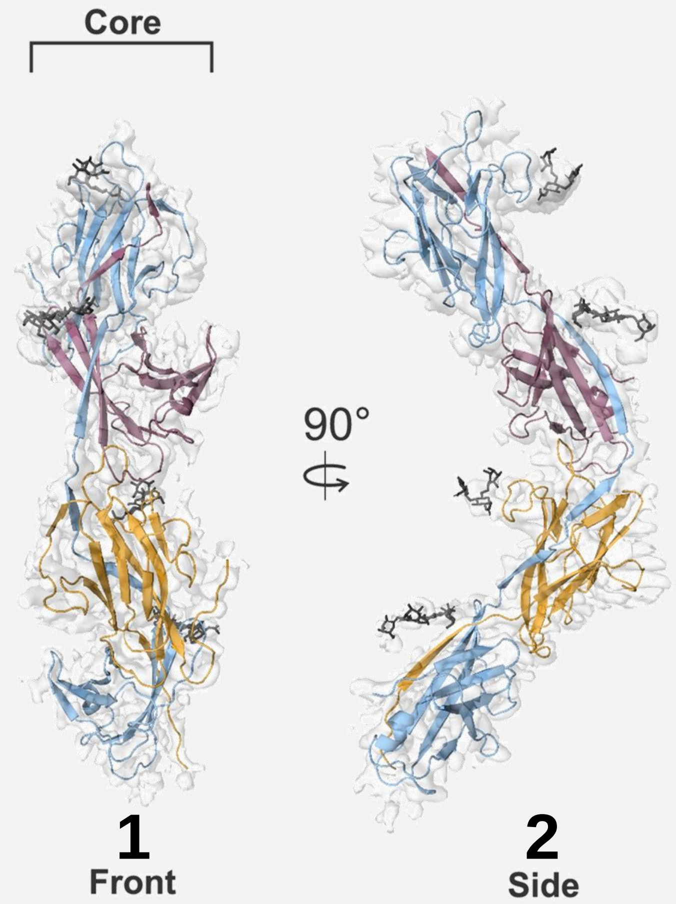
1.3M Topaz picks

↓

723k good picks

↓

best class



Stanisich et al., eLife 2020

Interphotoreceptor retinoid-binding protein

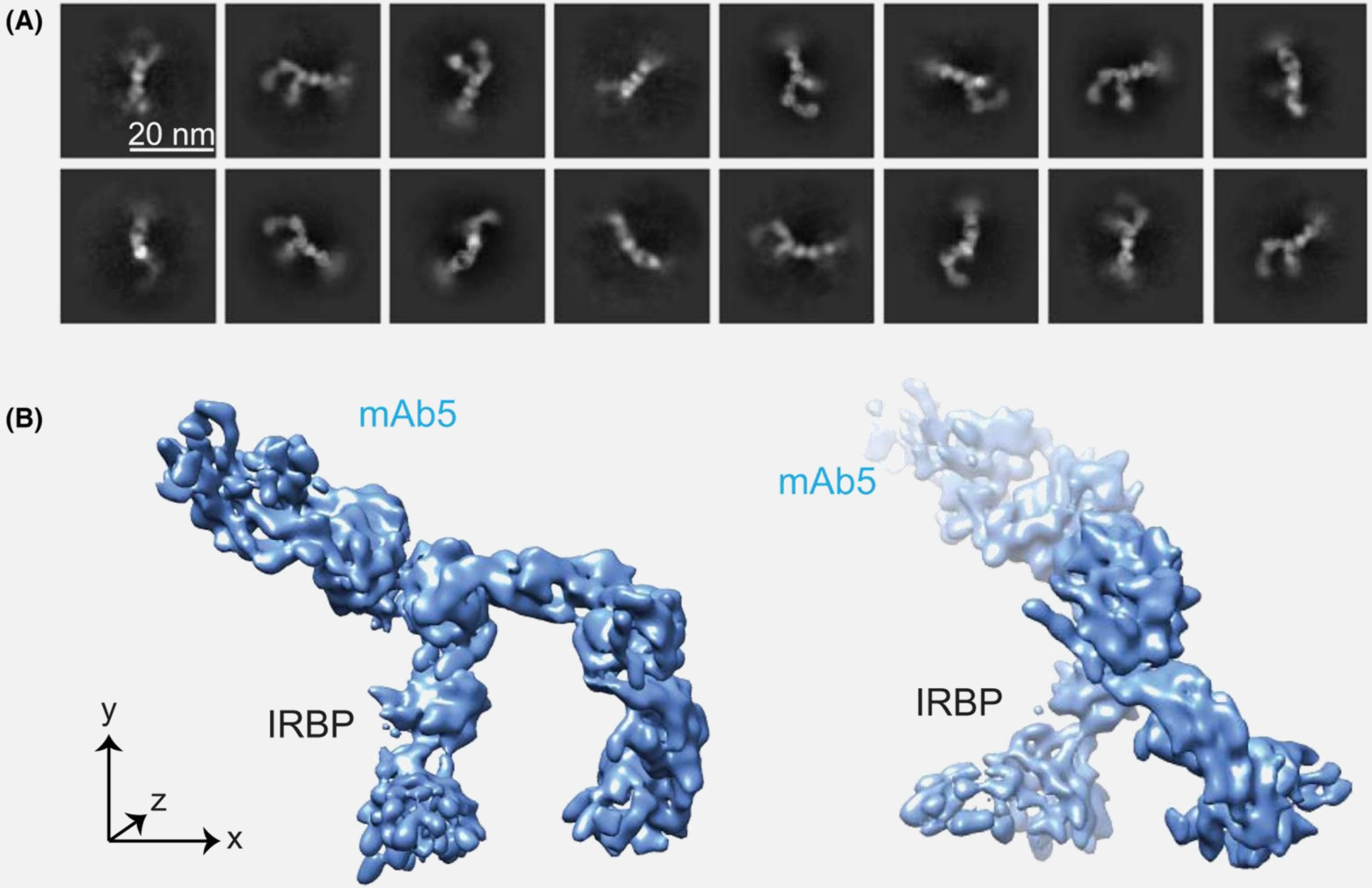
Picked 109k particles using templates

↓

2D/3D curation down to 17k particles

↓

Trained Topaz to pick the best 13k particles →



Sears et al., FASEB 2020

Human citrate transporter from 20° tilted images

Picked 113k ‘good’ particles from 652 images with Warp

↓

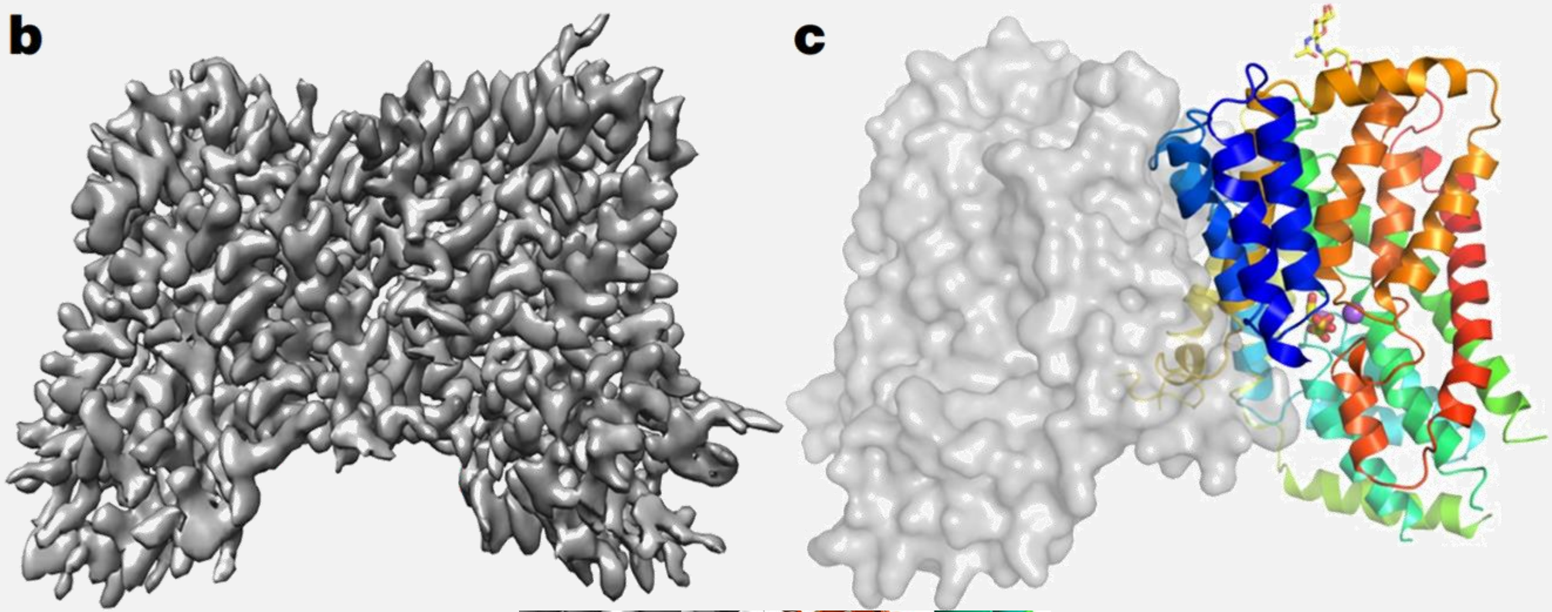
Trained Topaz

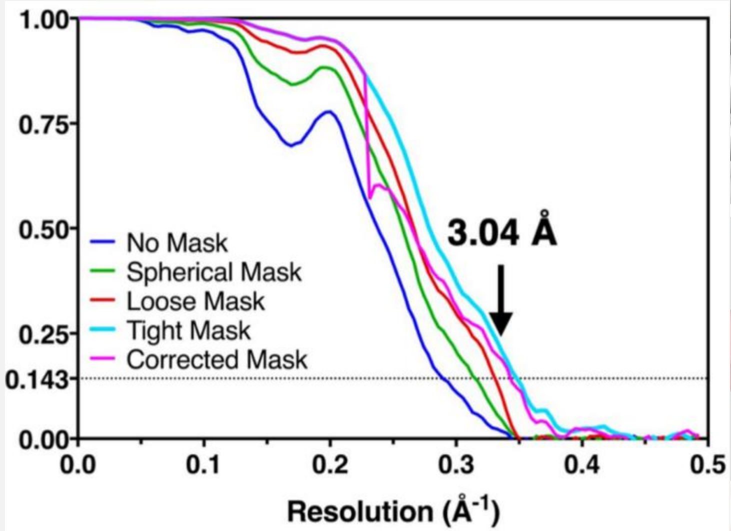
↓

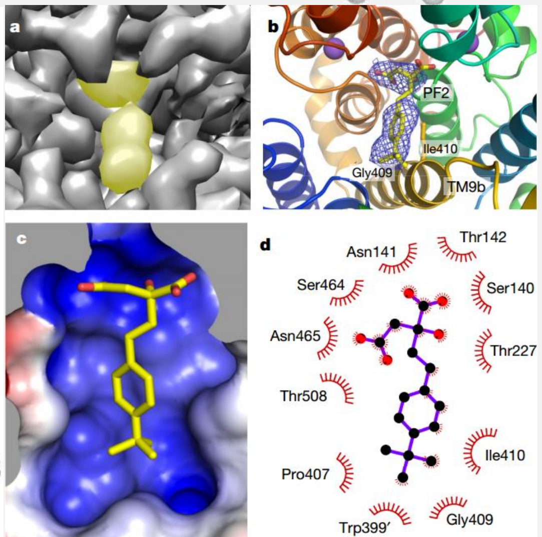
Picked 1.15m particles from 1712 images

↓

564k particles (90% more than Warp)

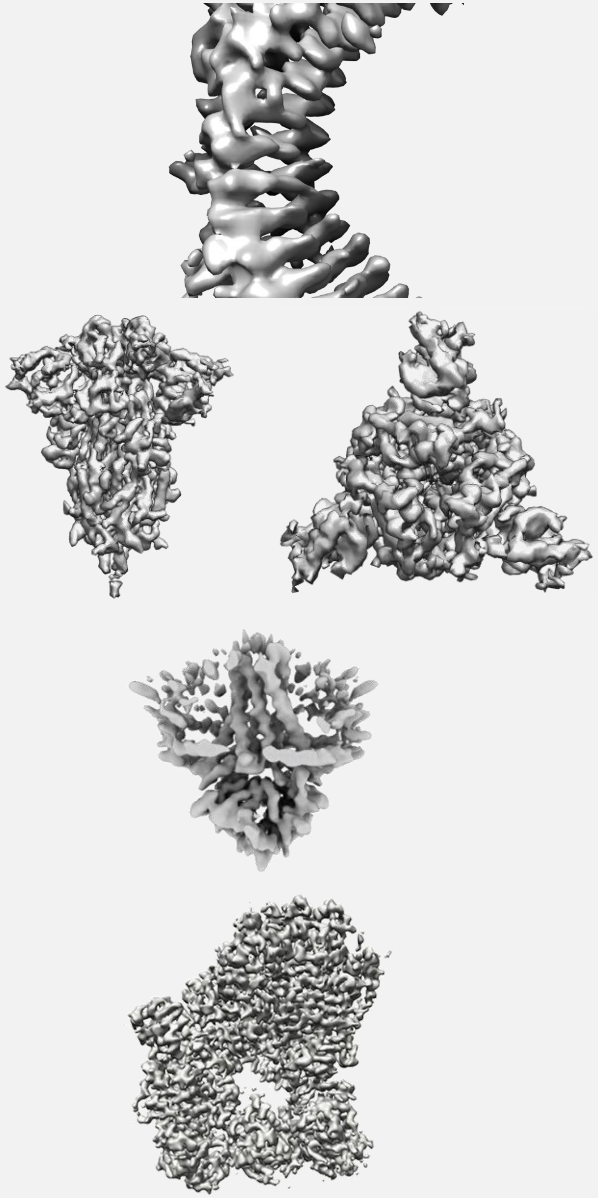




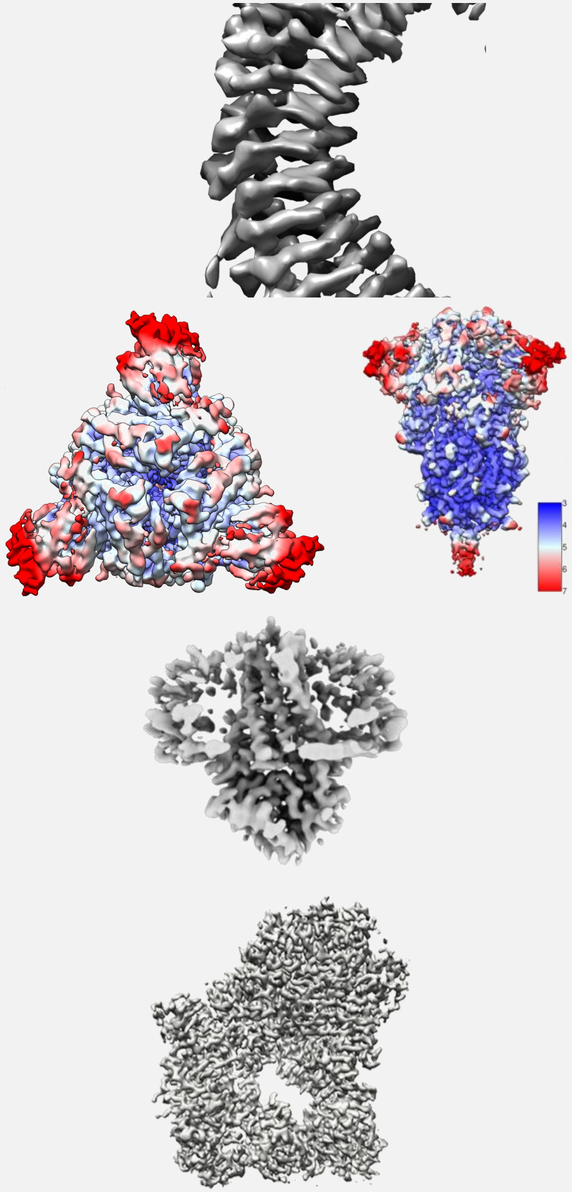


Sauer et al., Nature 2021

Without Topaz



With Topaz



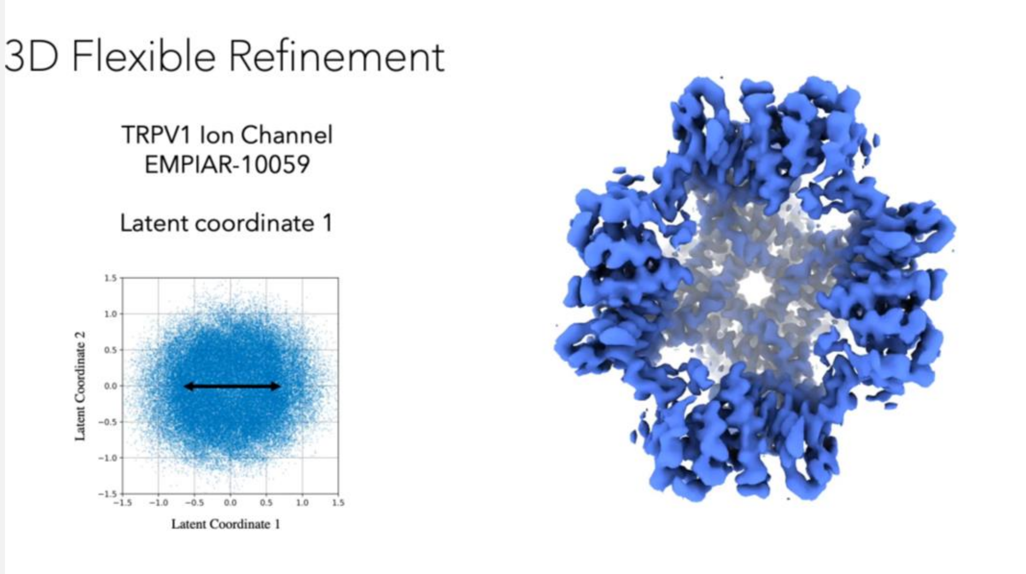
More particles

Higher resolution

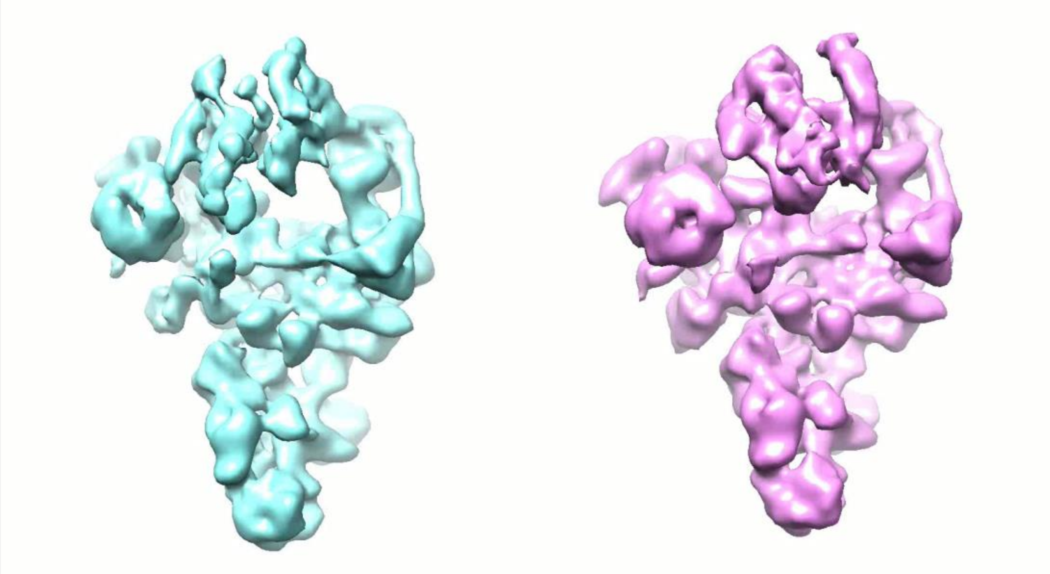
Better angular distribution

Better classification

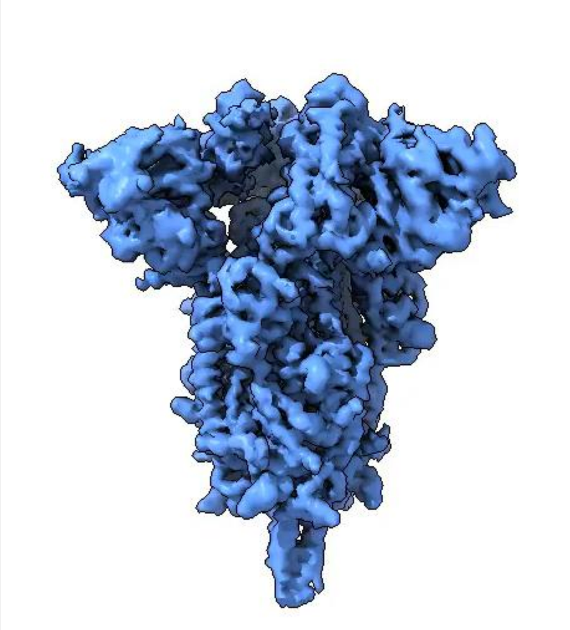
More particles = better flexibility analysis



CryoSPARC 3DFlex



EMAN2

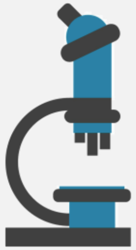


CryoDRGN

Next: Neural network

Denoising

Topaz-Denoise

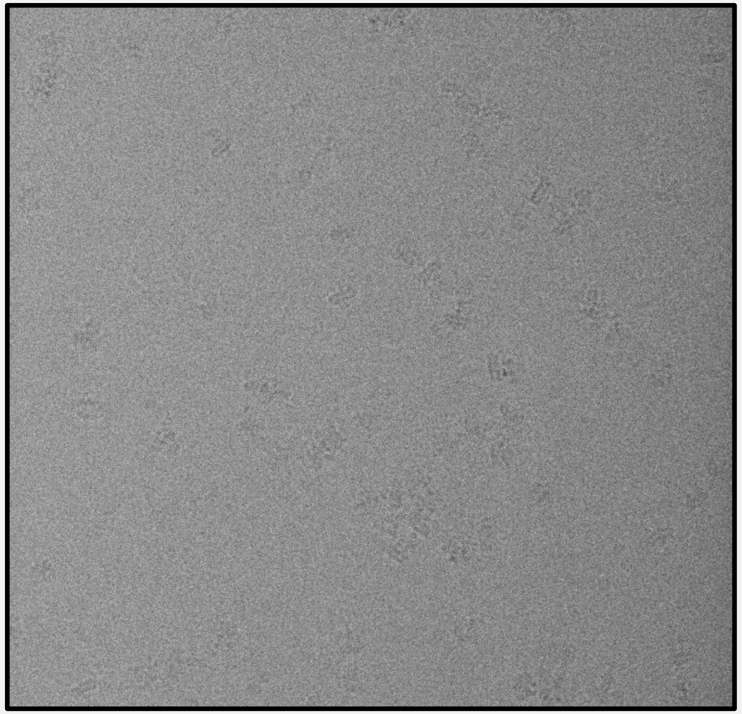


- Noise2Noise* models **trained on dozens of cryoEM datasets**

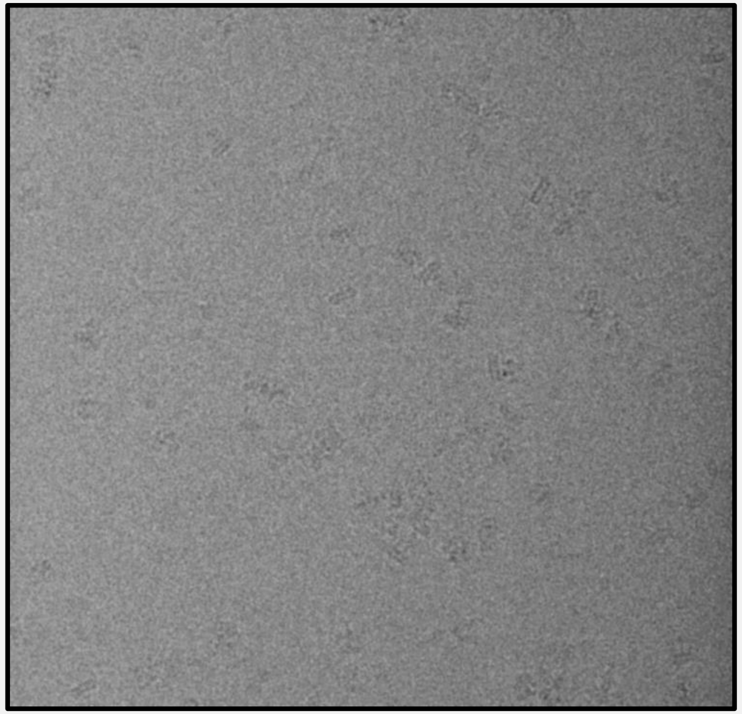
- **Train your own** denoising model
- **Directly visualize difficult proteins** *with confidence*

*Lehtinen et al., 2018

How does Noise2Noise work?



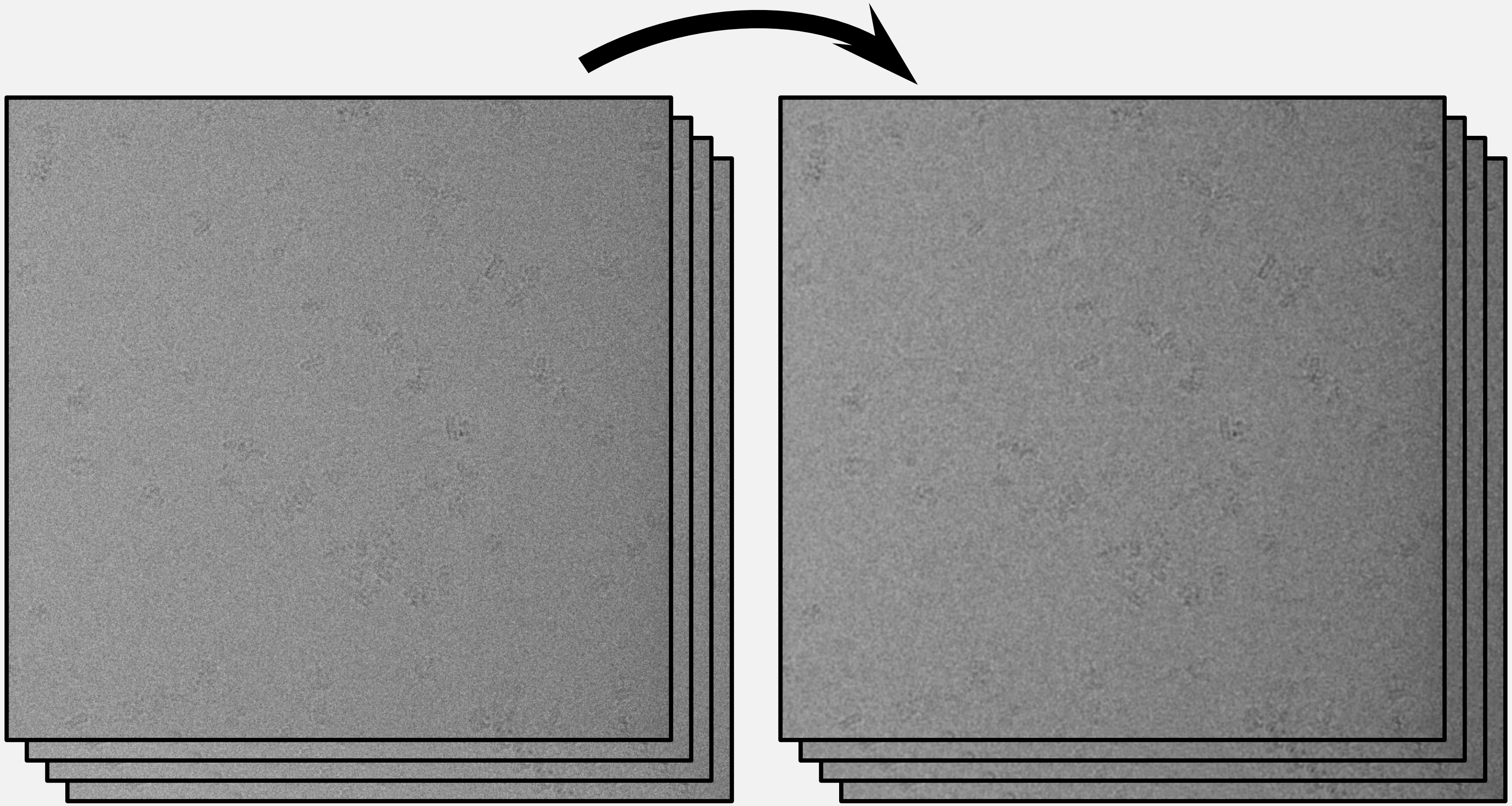
Noisy image 1



Noisy image 2

*same field of view

How does Noise2Noise work?



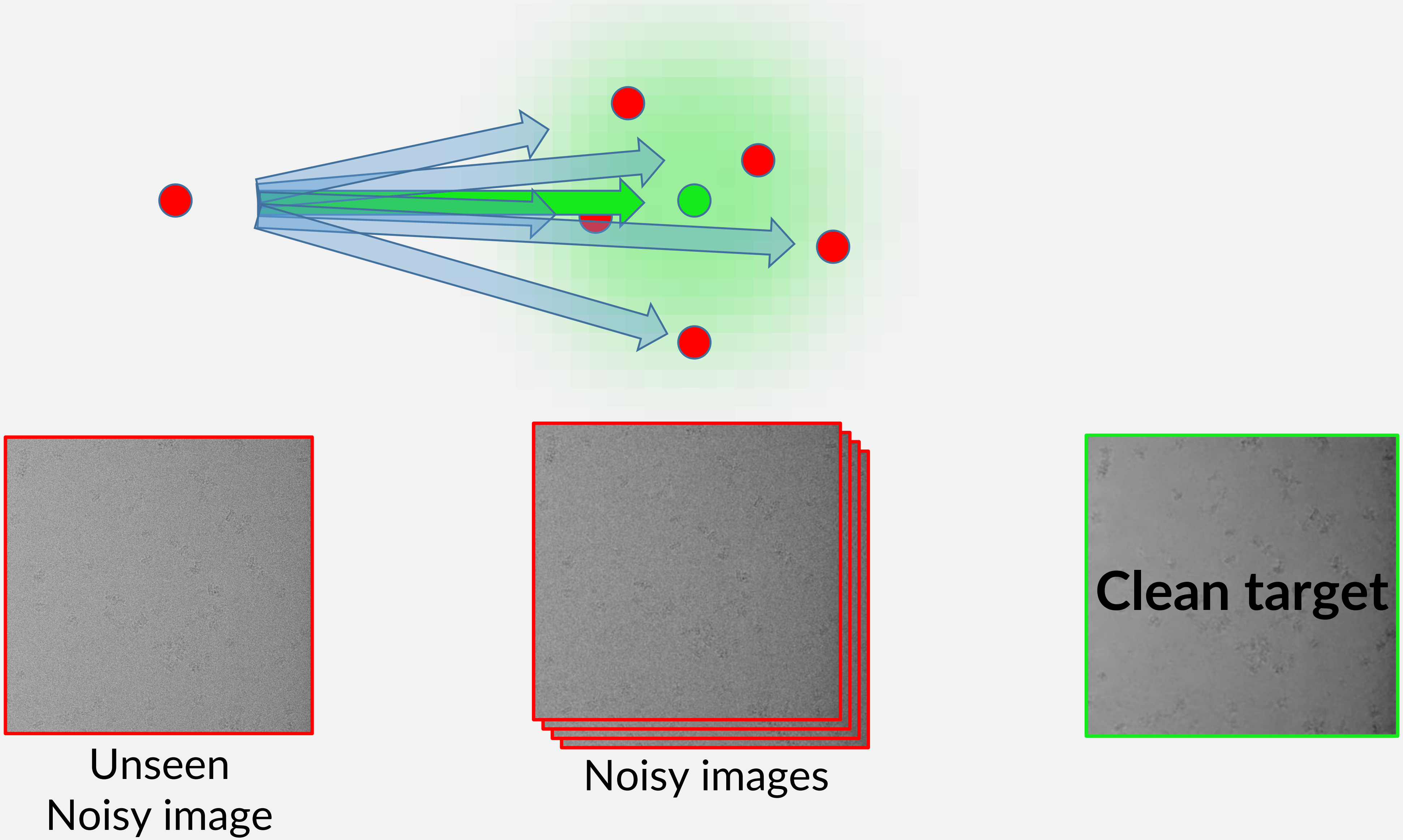
Noisy images 1

Noisy images 2

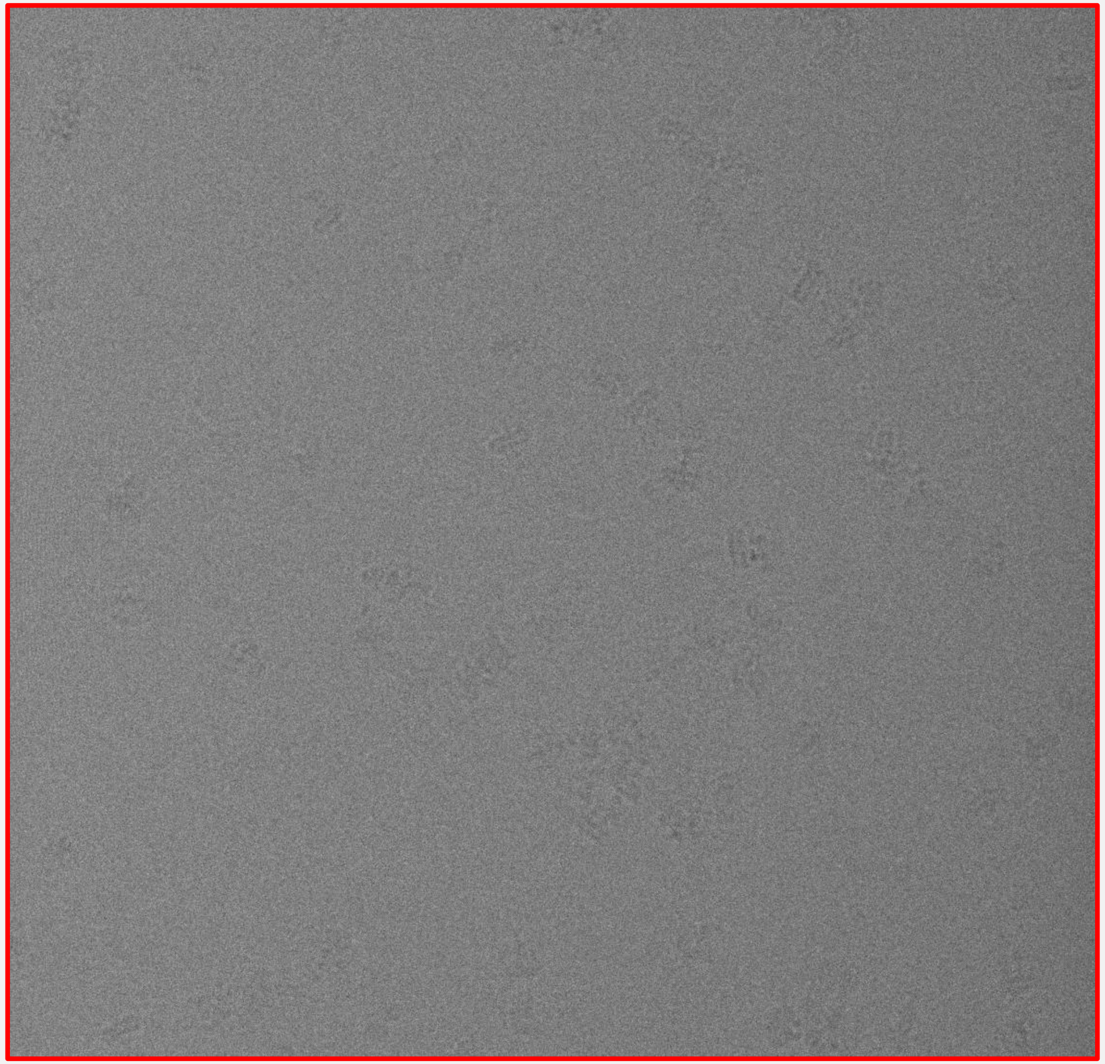
*same fields of view

**several microscope/ camera/ collection parameters

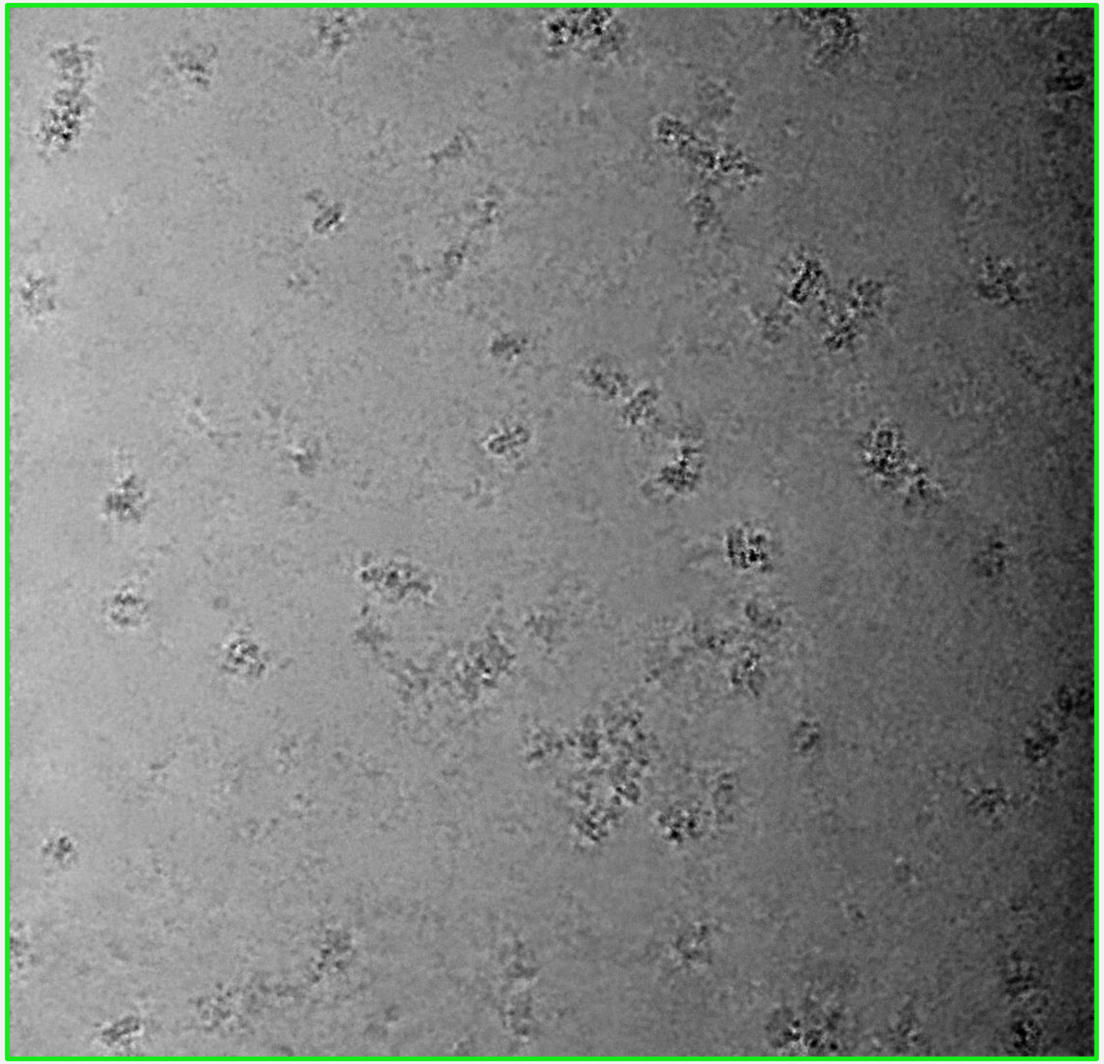
How does Noise2Noise work?



How well does the Topaz Denoise pre-trained model work?



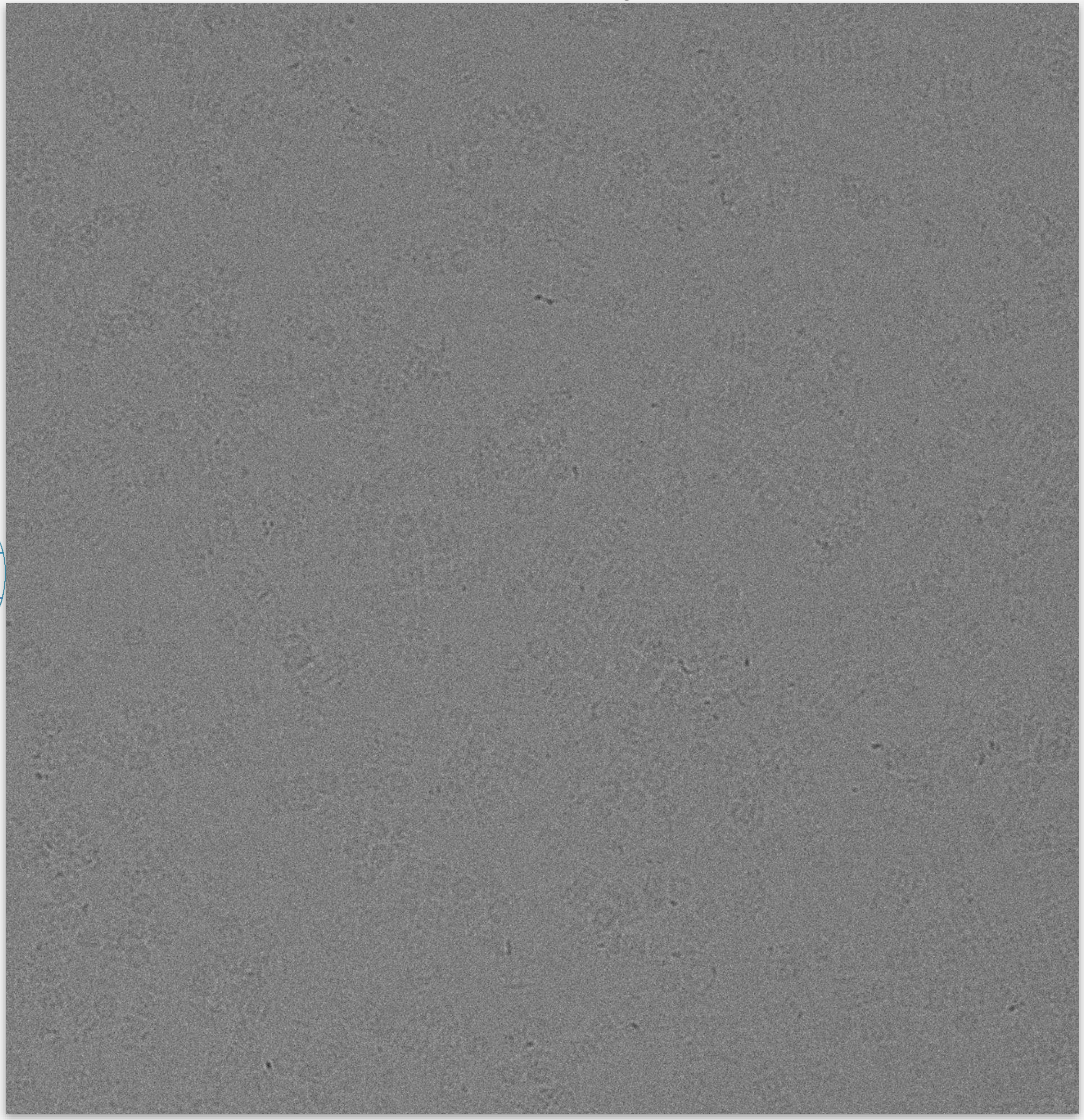
Raw, noisy image



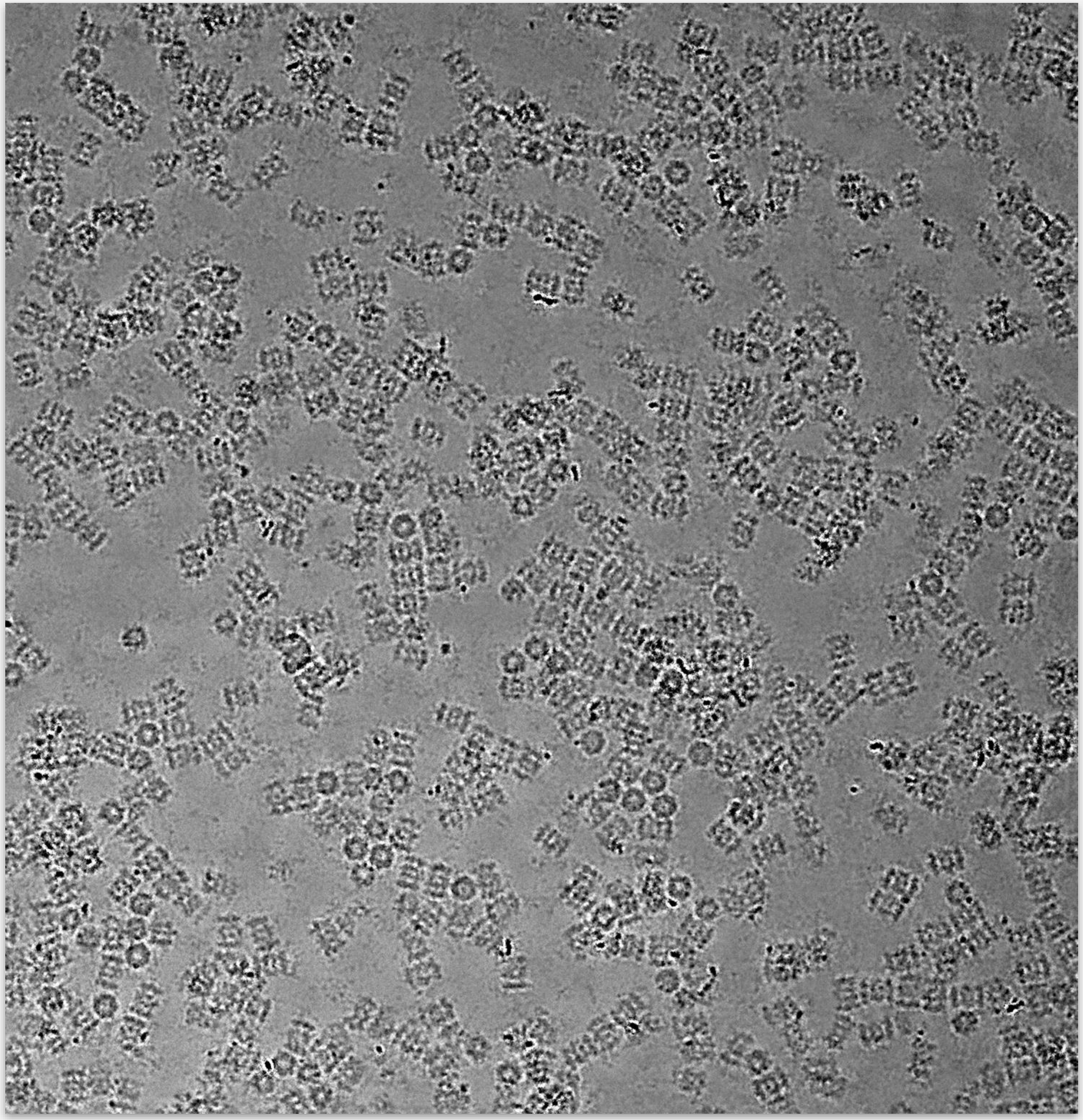
Denoised

Protein in nanodisc

How well does the Topaz Denoise pre-trained model work?



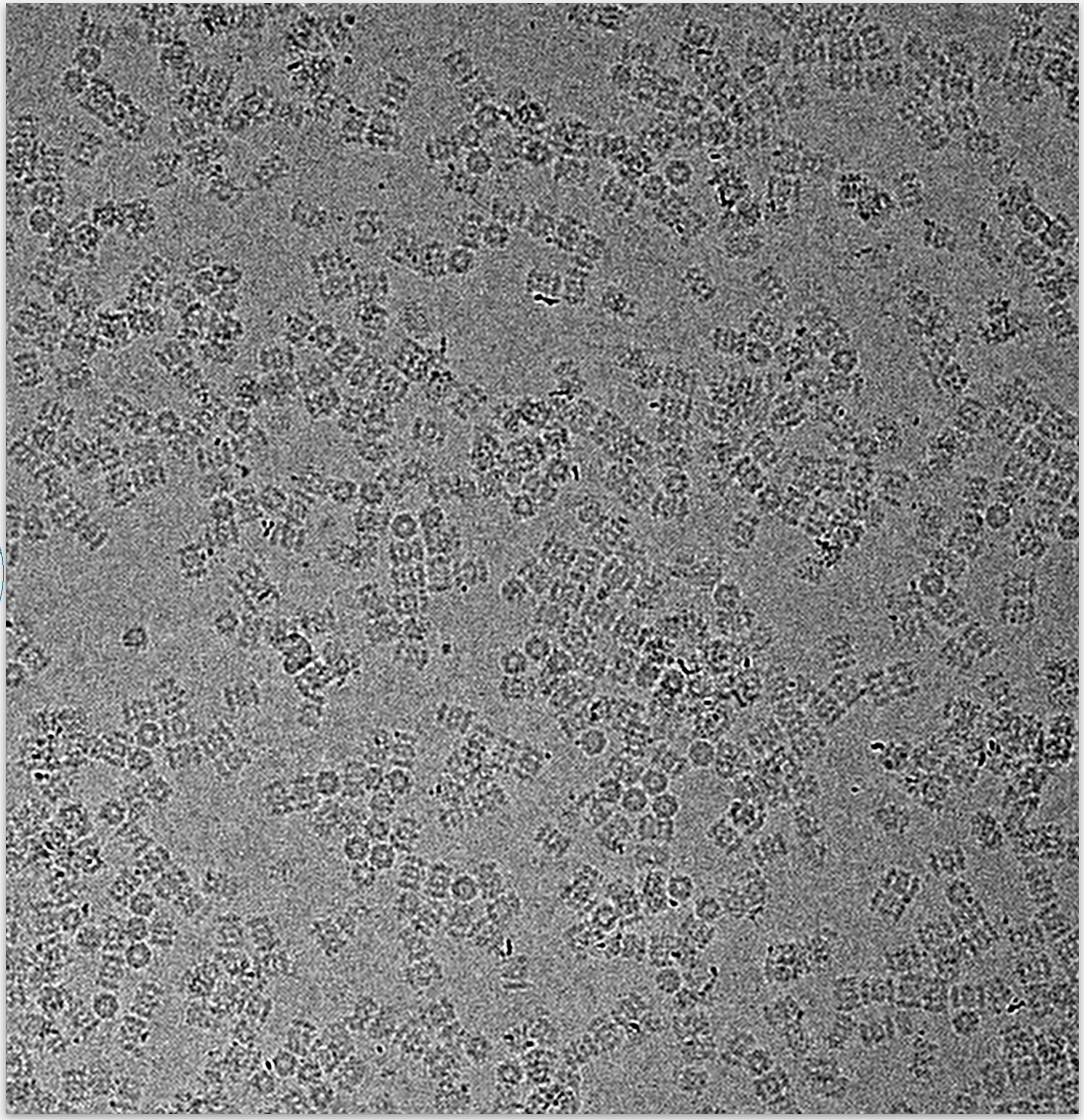
Raw



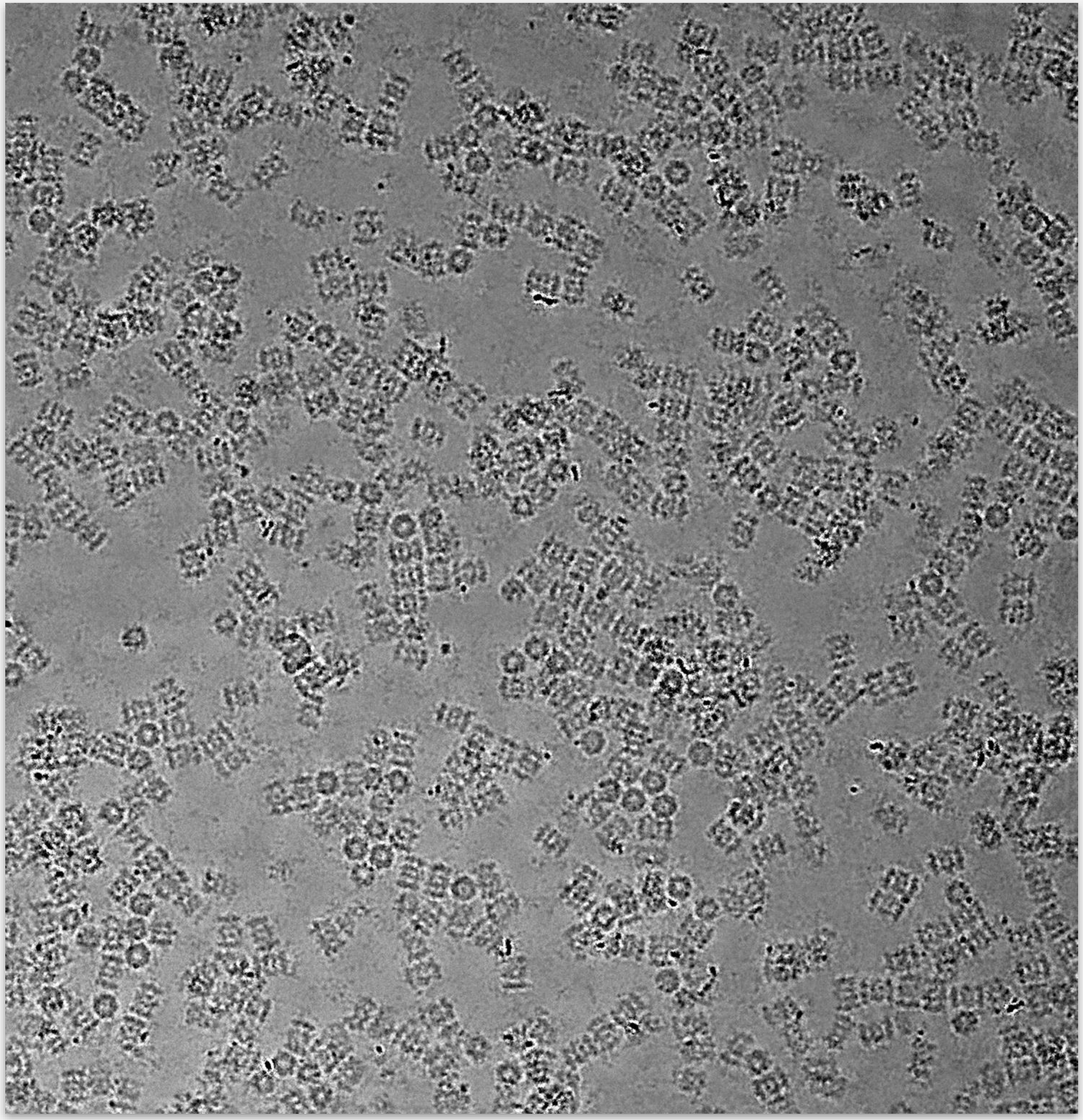
Denoised

T20S Proteasome

How well does the Topaz Denoise pre-trained model work?



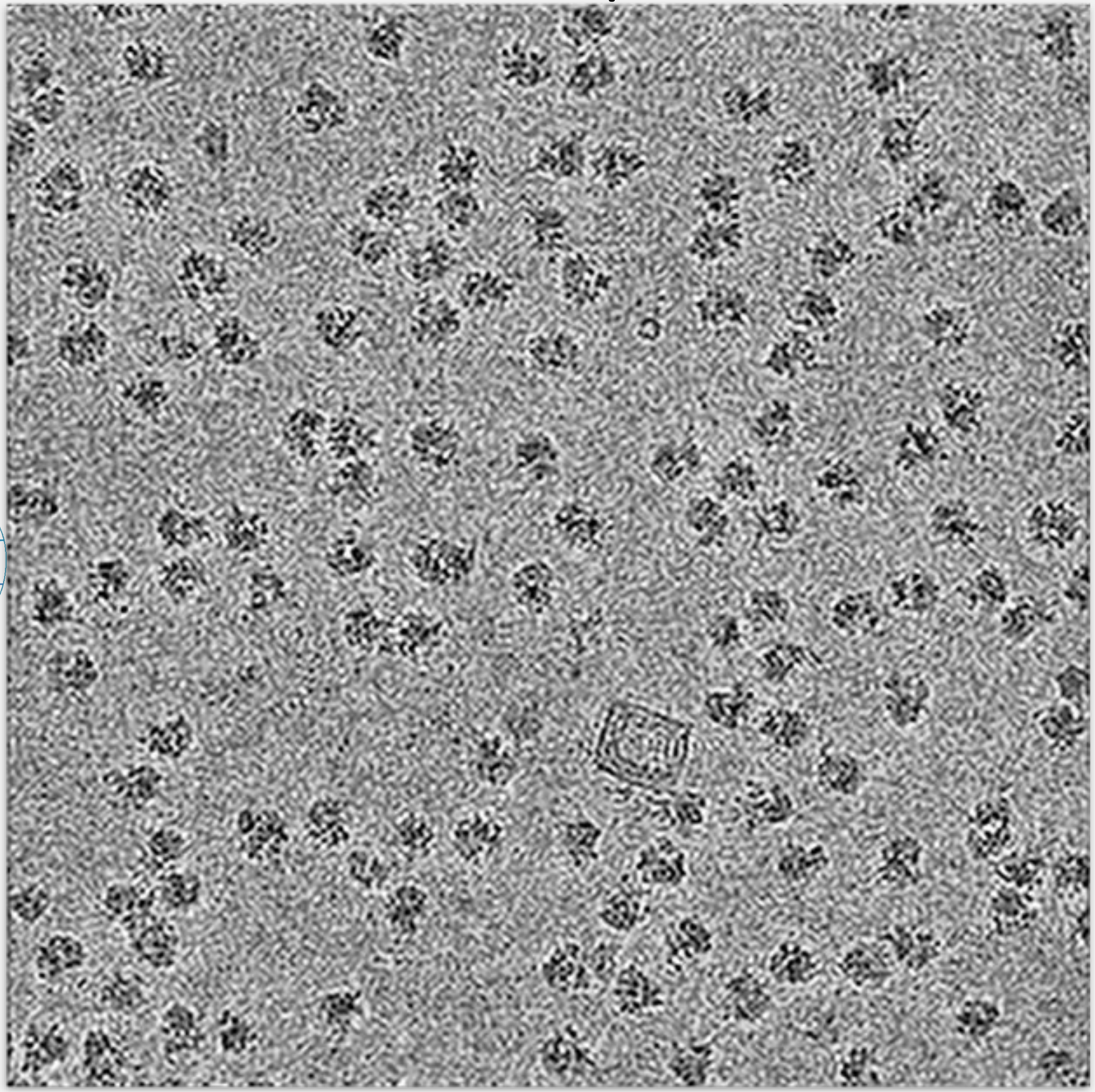
'Optimally' lowpassed



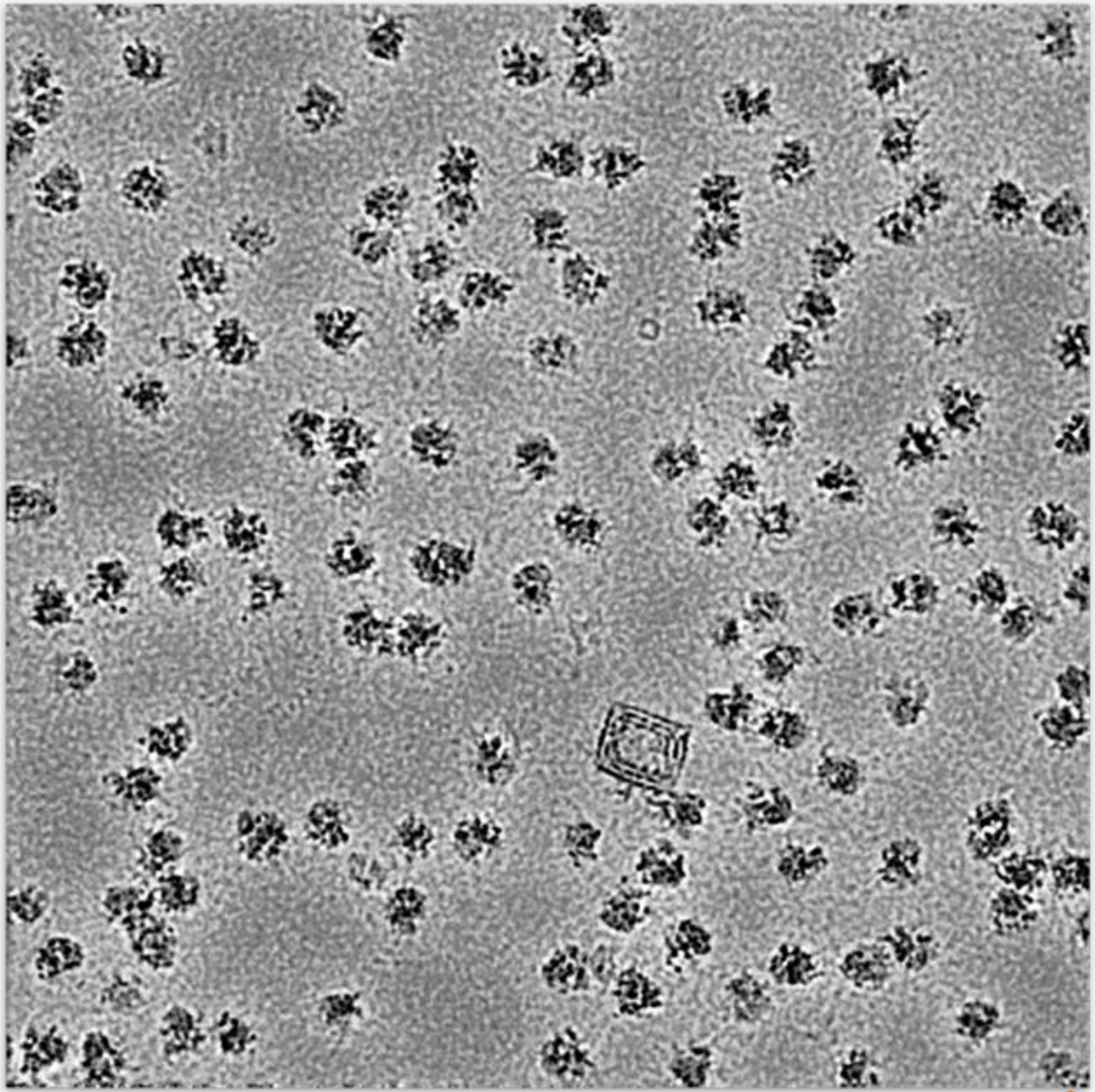
Denoised

T20S Proteasome

How well does the Topaz Denoise pre-trained model work?



'Optimally' lowpassed



Denoised

Ribosome

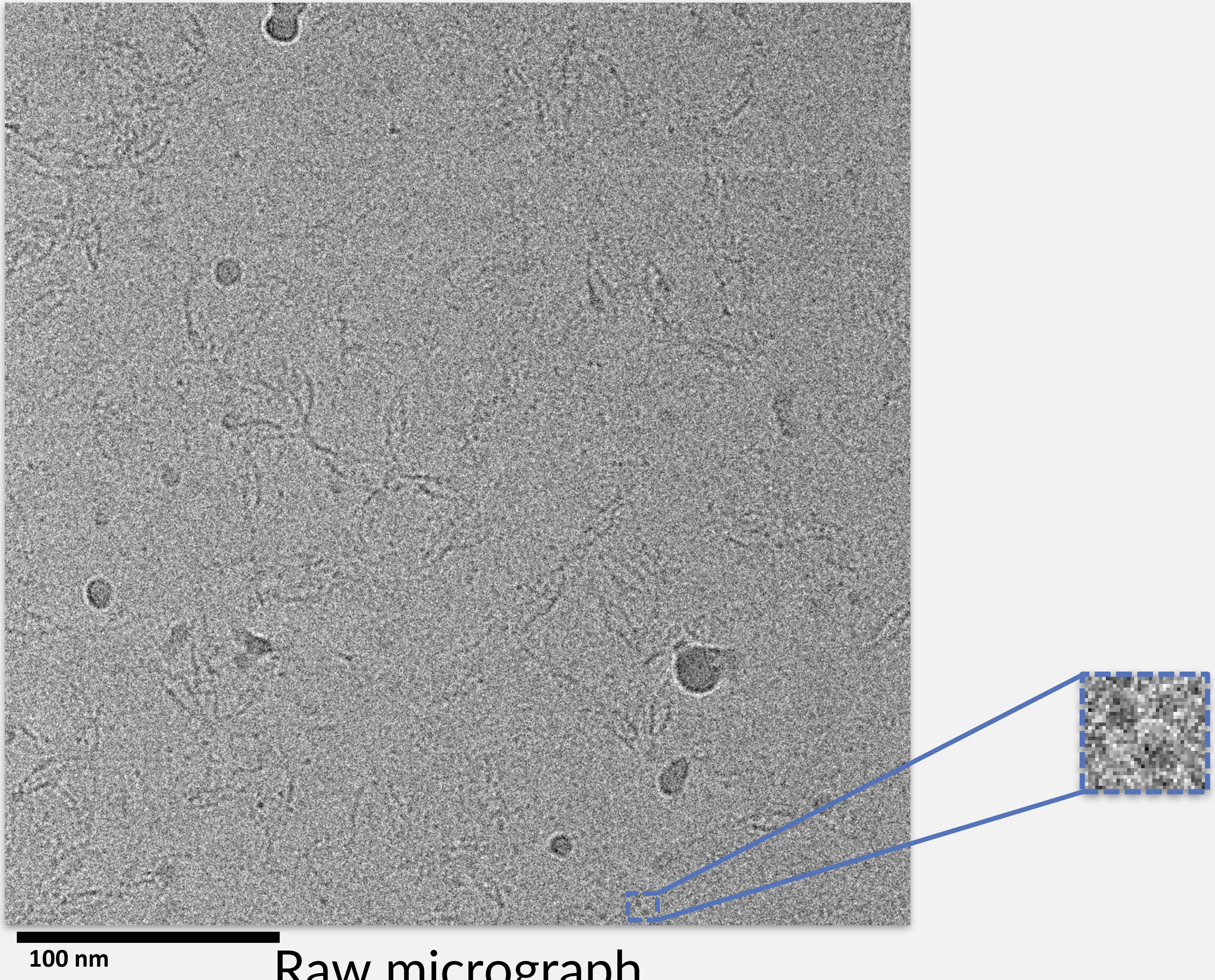
Ok, that’s a **nice**

Anti-Squinting Device

…

But **is it useful?**

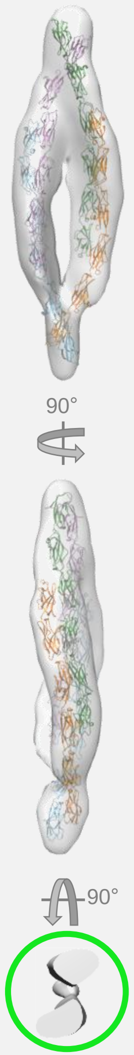
Can we identify small particle views?

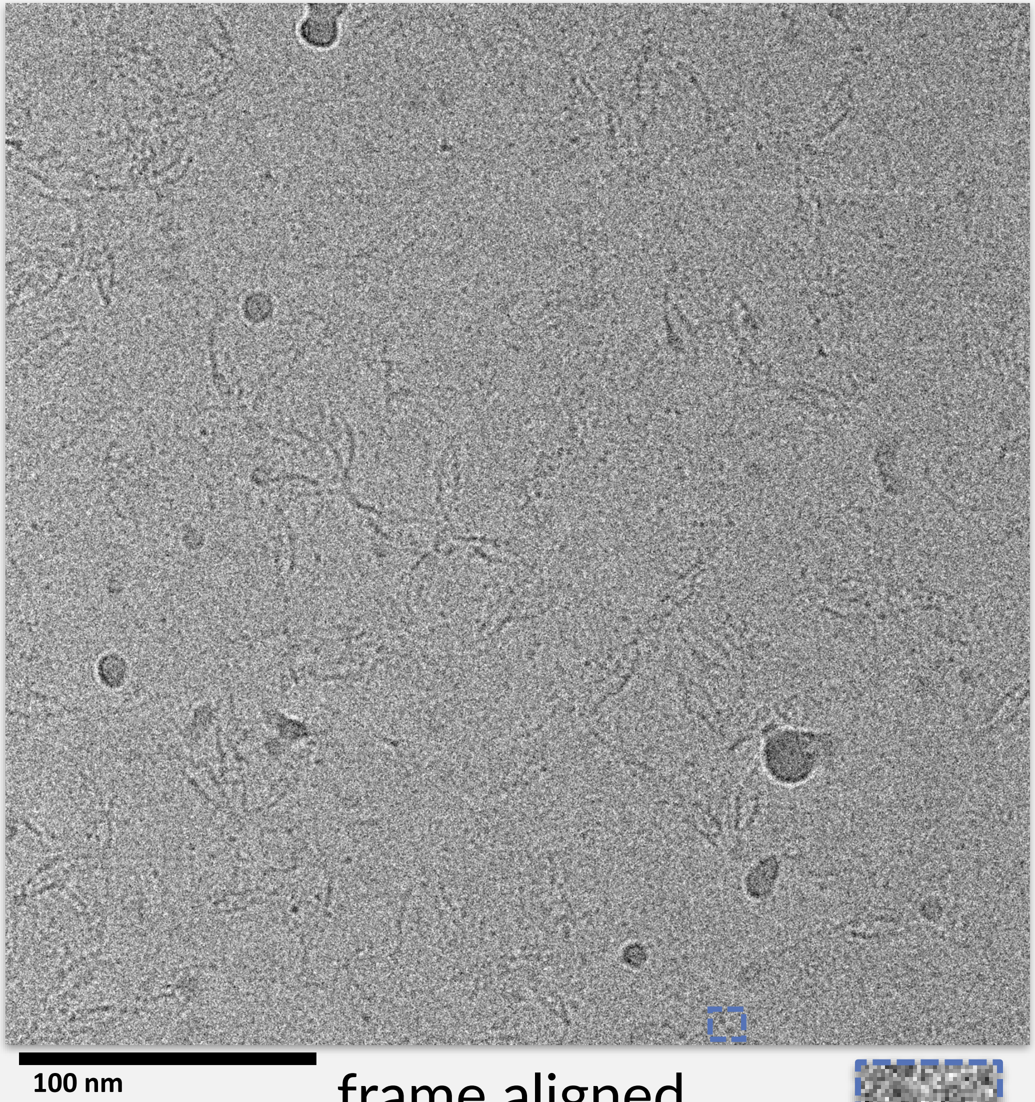


Raw micrograph

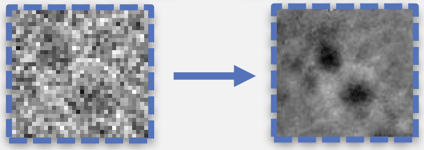
Is this a top view??

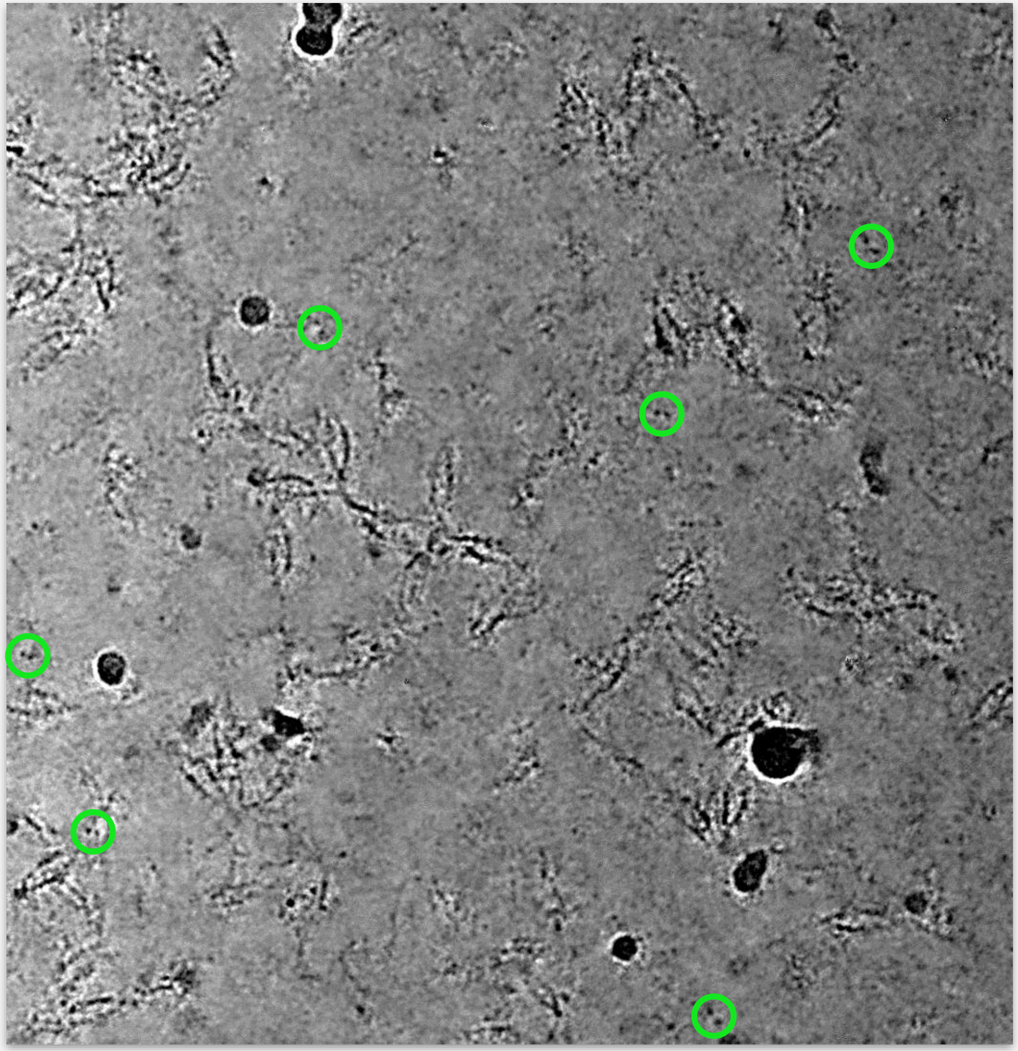
Can we identify small particle views?





frame aligned,
dose-weighted

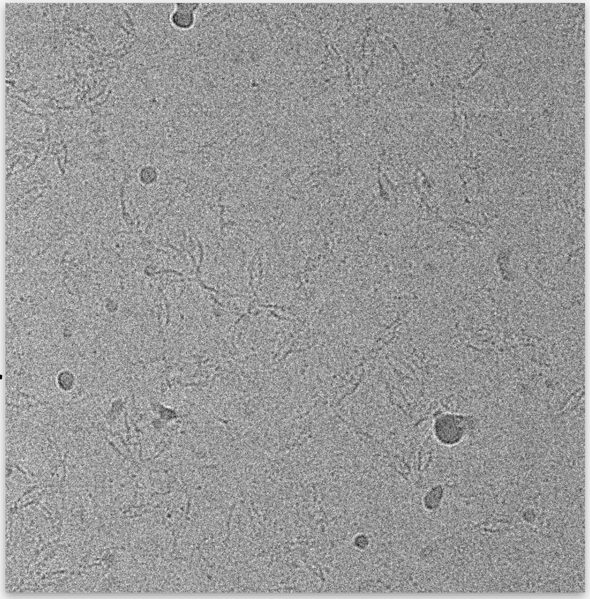




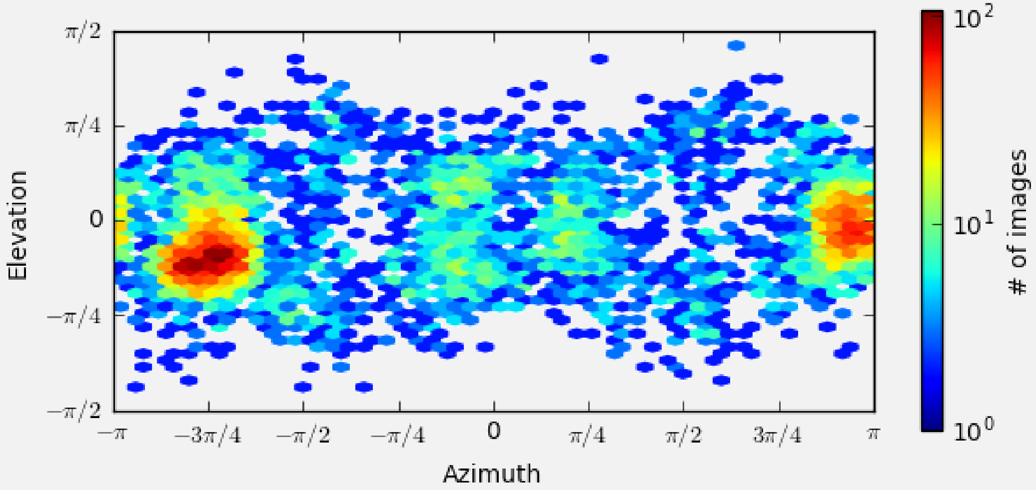
Topaz-Denoise

Yes! Denoising gives 215% more particles

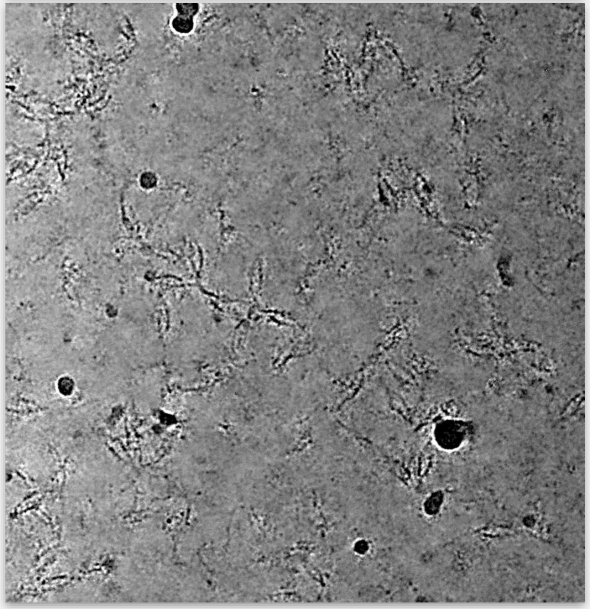
1.5k initial particles picked by data owner



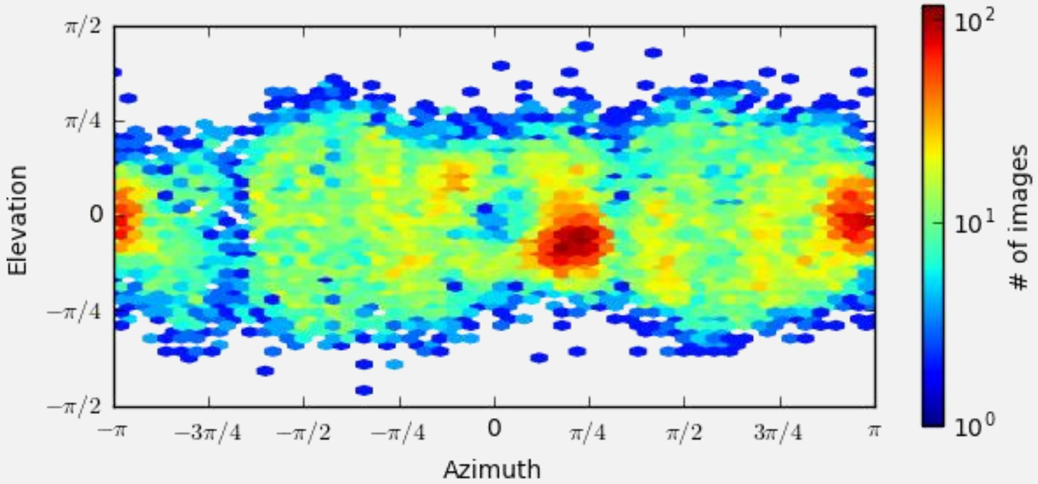
→10k ptcls



1k initial particles picked by non-data owner (me)

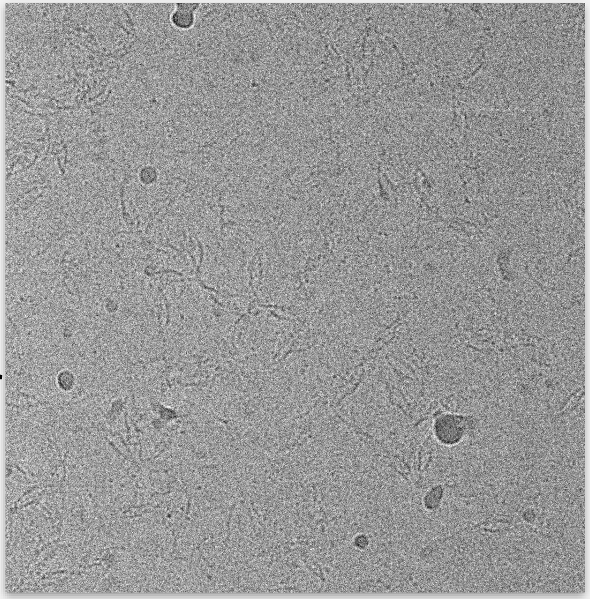


→21.5k ptcls

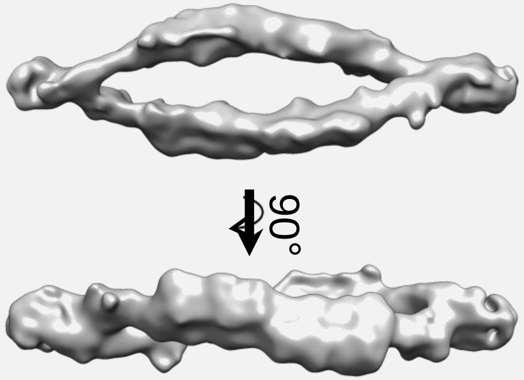


Yes! Denoising gives 215% more particles

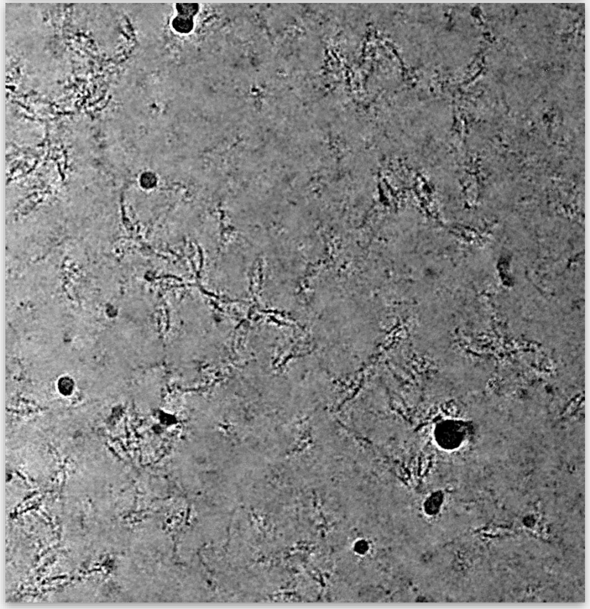
1.5k initial particles picked by data owner



→10k ptcls

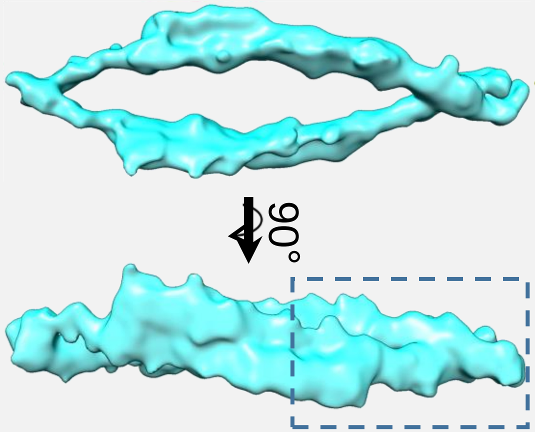


1k initial particles picked by non-data owner (me)

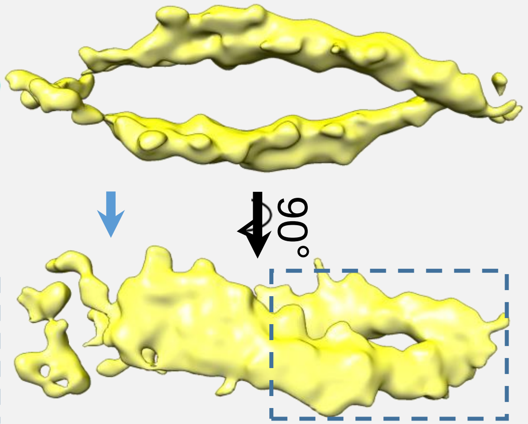


→21.5k ptcls

13.4k particles

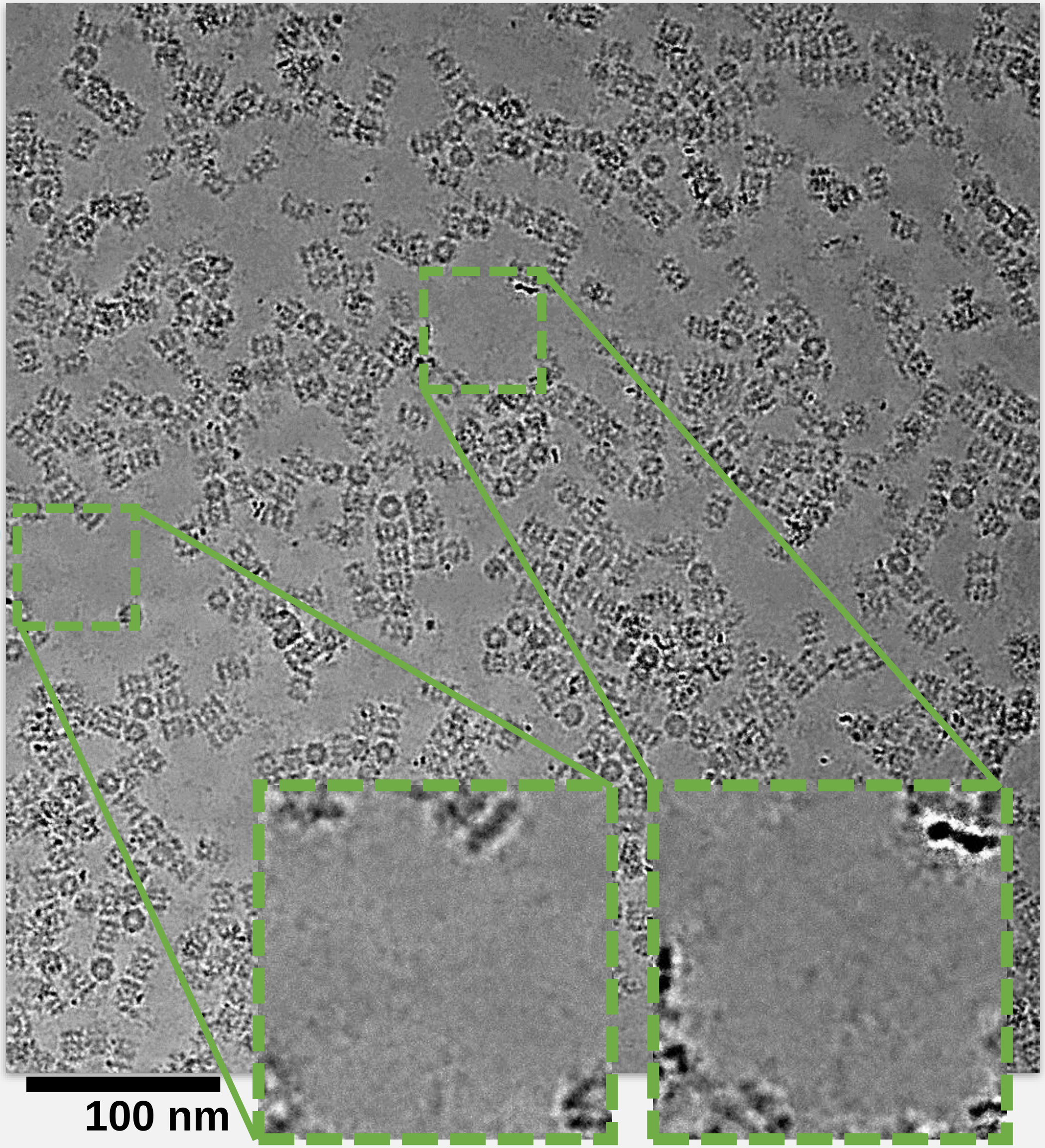


8.1k particles

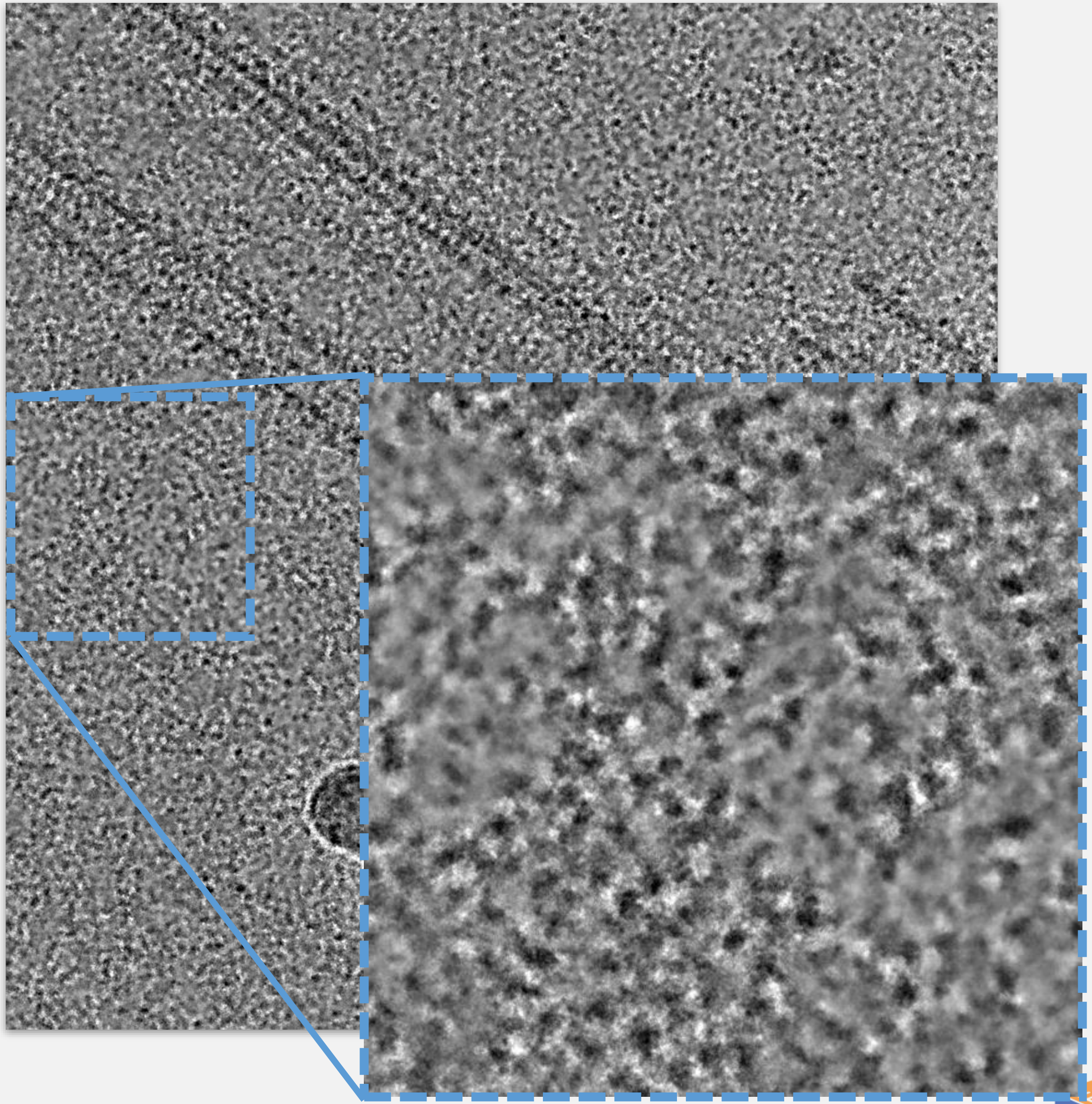


Can we be more **confident**
about what is in
image/sample ***backgrounds***?

What is in your background?

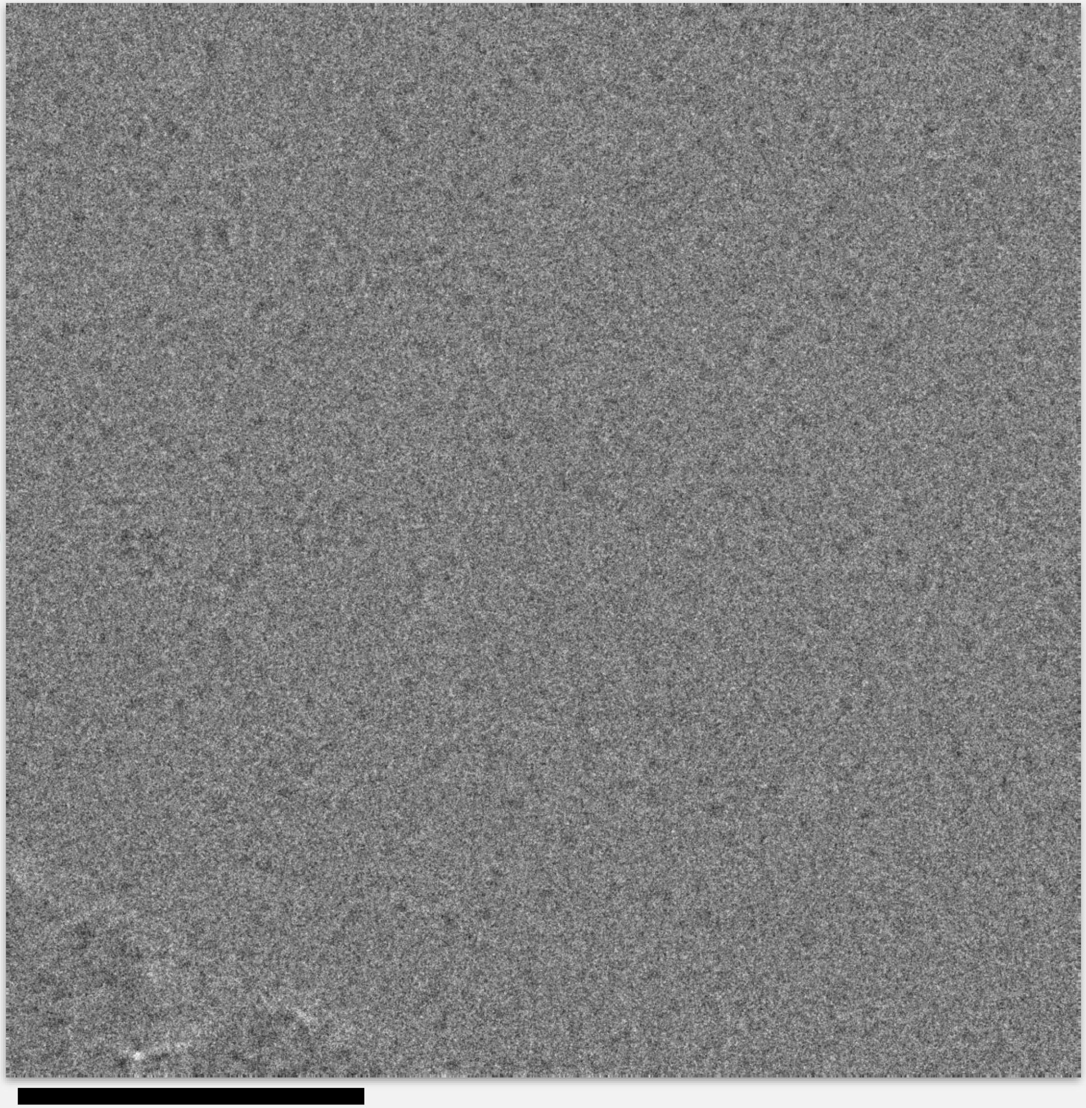


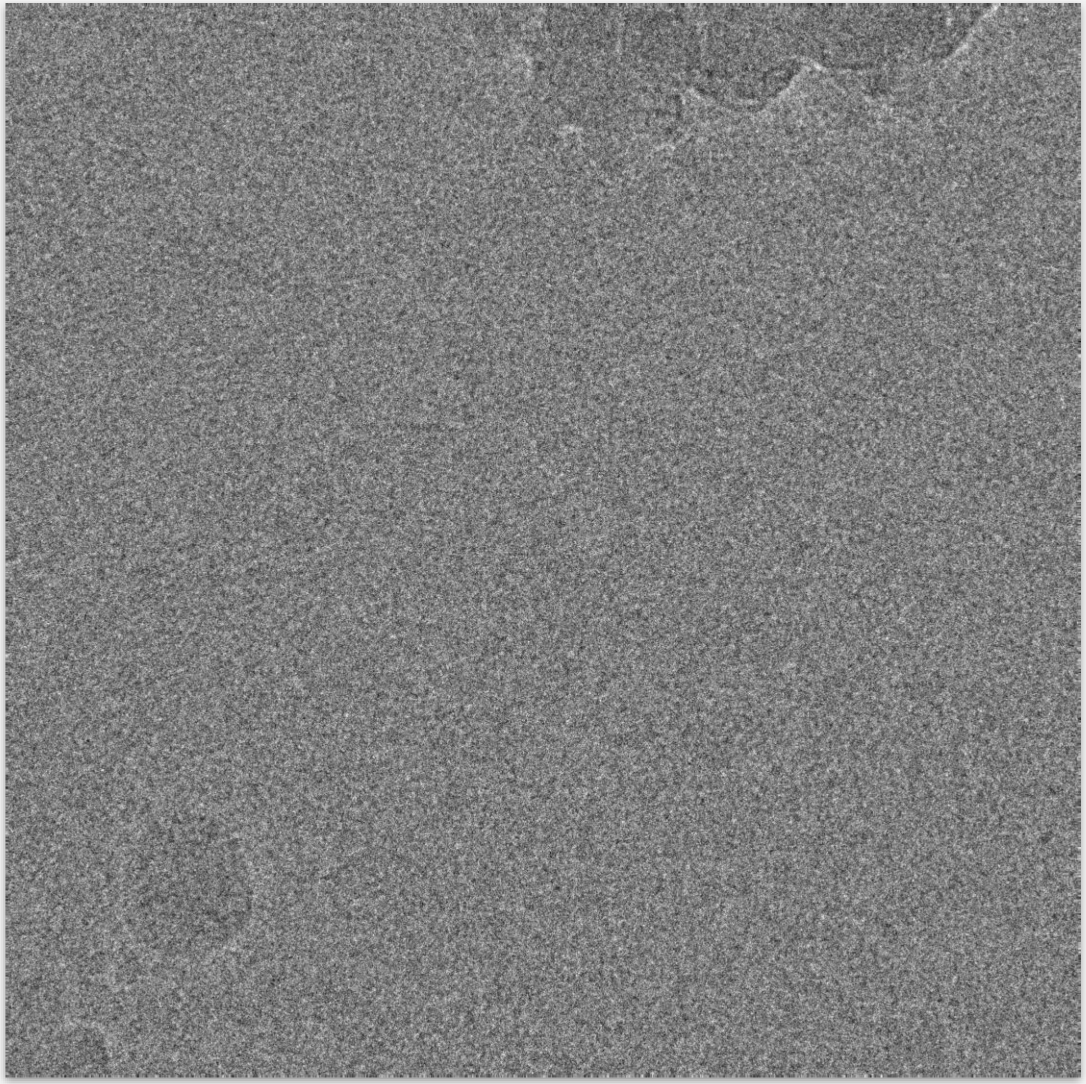
Negative control



Positive control

What is in your background?

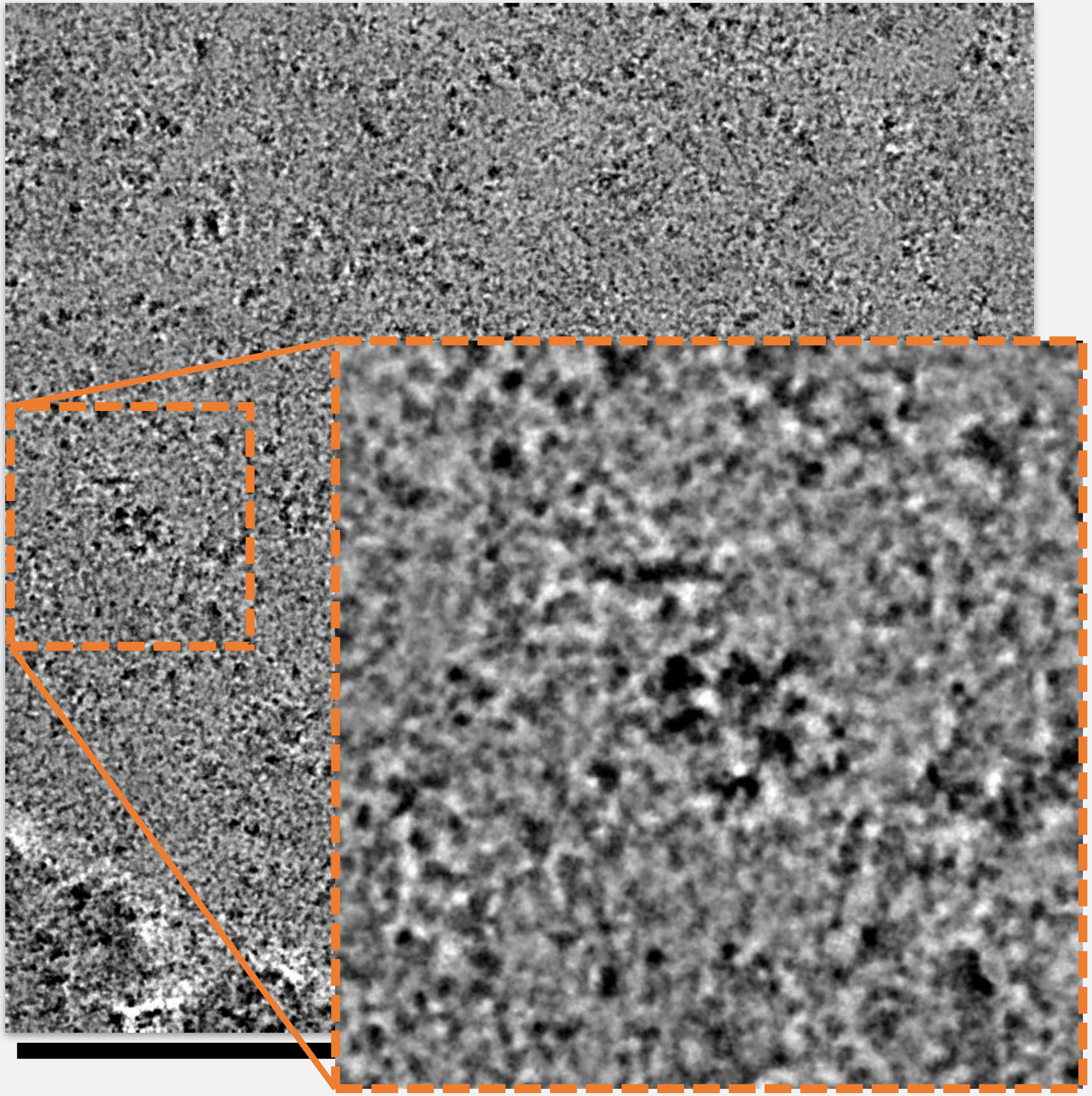


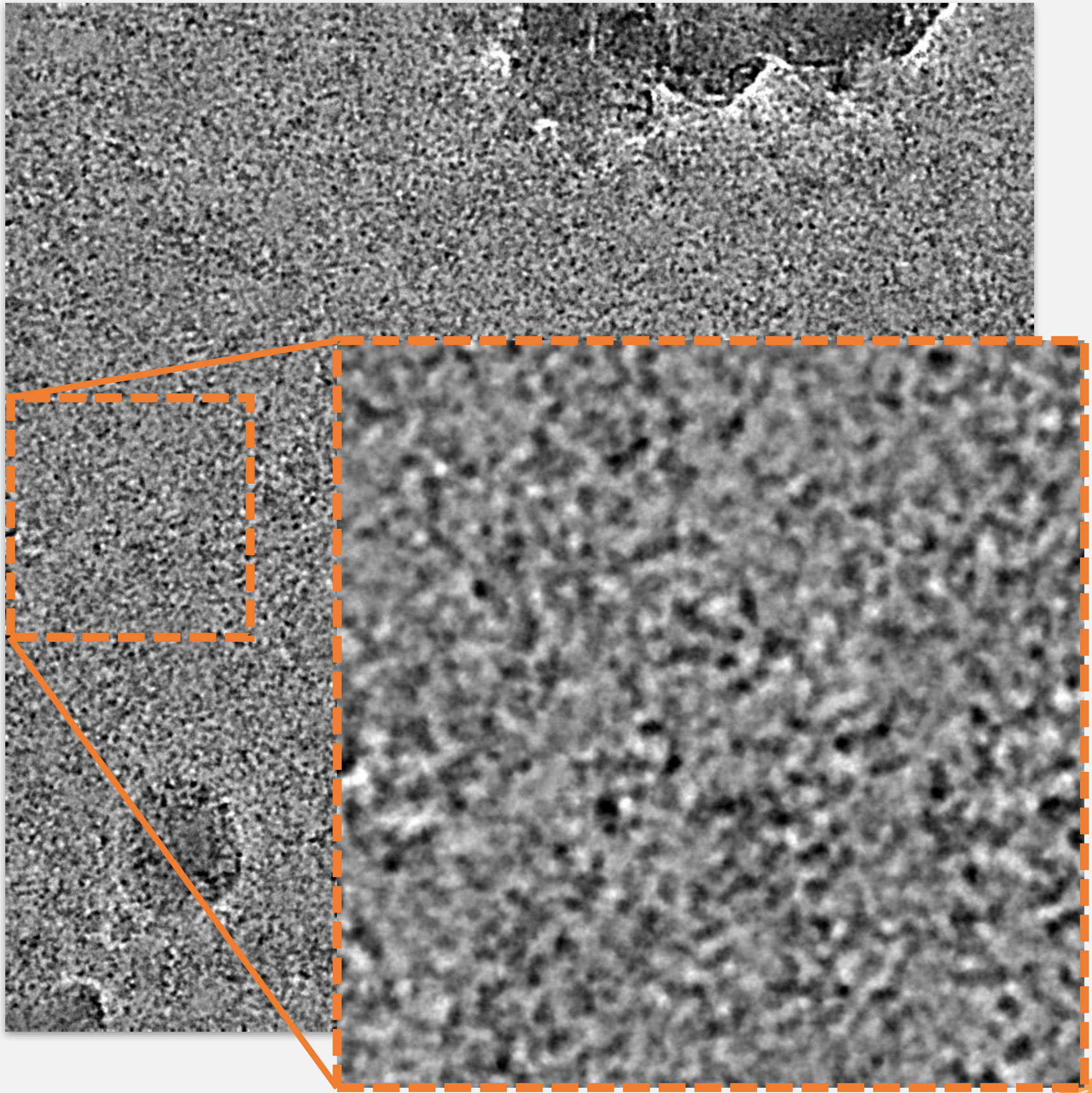


Raw images

Mao et al., PNAS, 2013

What is in your background?



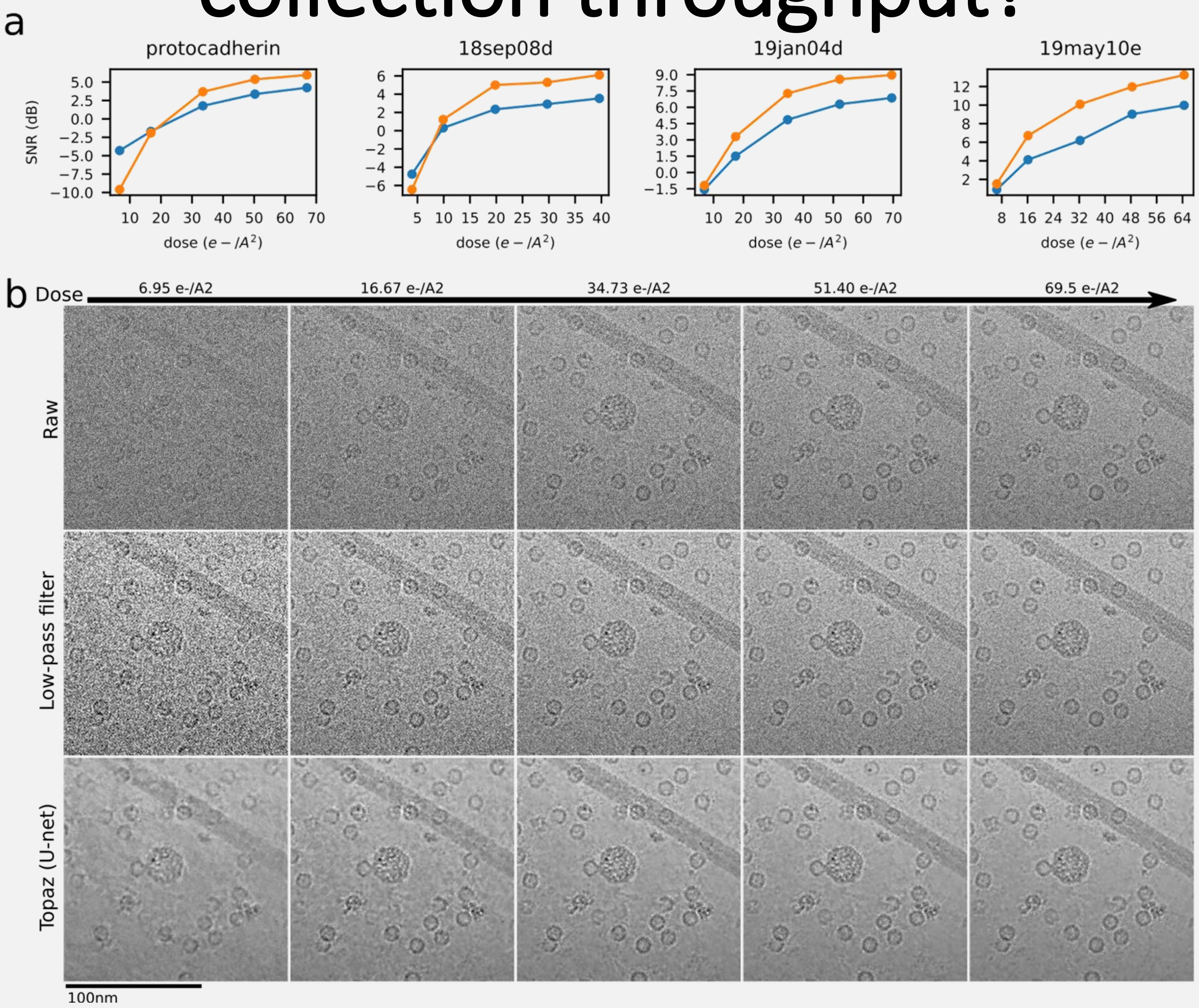


Topaz Denoise

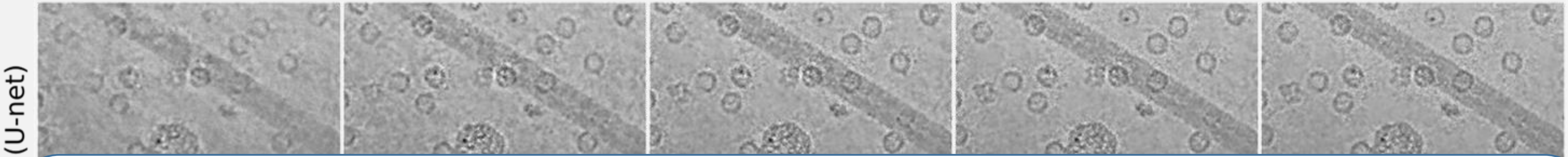
Mao et al., PNAS, 2013

Can denoising help collection throughput?

Can Topaz-Denoise help collection throughput?

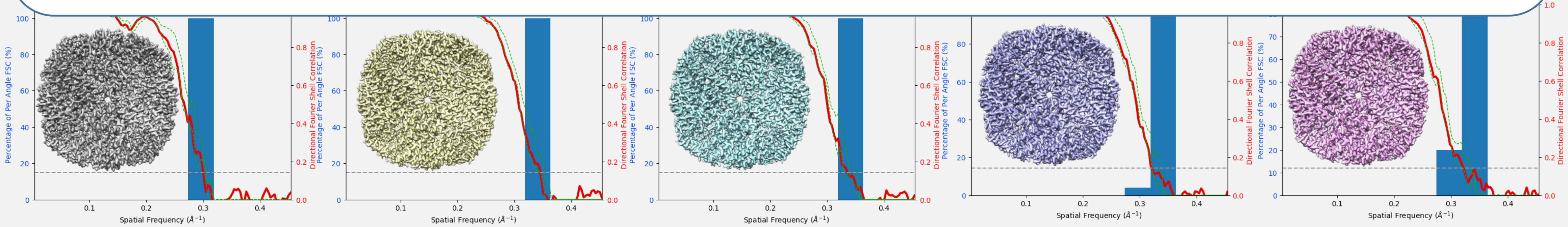


Can Topaz-Denoise help collection throughput?



Yes! You don't need more than ~20 e-/Å²

(for medium-large particles)



7 e-/Å²	17 e-/Å²	35 e-/Å²	51 e-/Å²	70 e-/Å²
3.35 Å	2.90 Å	3.0 Å	3.13 Å	3.13 Å

Ok, give me Topaz please!

github.com/tbepler/topaz

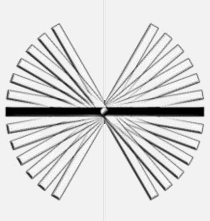
- **Pre-trained picking and denoising models**
- **Full tutorial** and explanations are included
- A universal HTML-based **GUI** is included
- Integrated into **CryoSparc, Relion**, Scipion, and Appion.

Summary

- **Pick more** and **more accurately** with Topaz positive-unlabeled picking

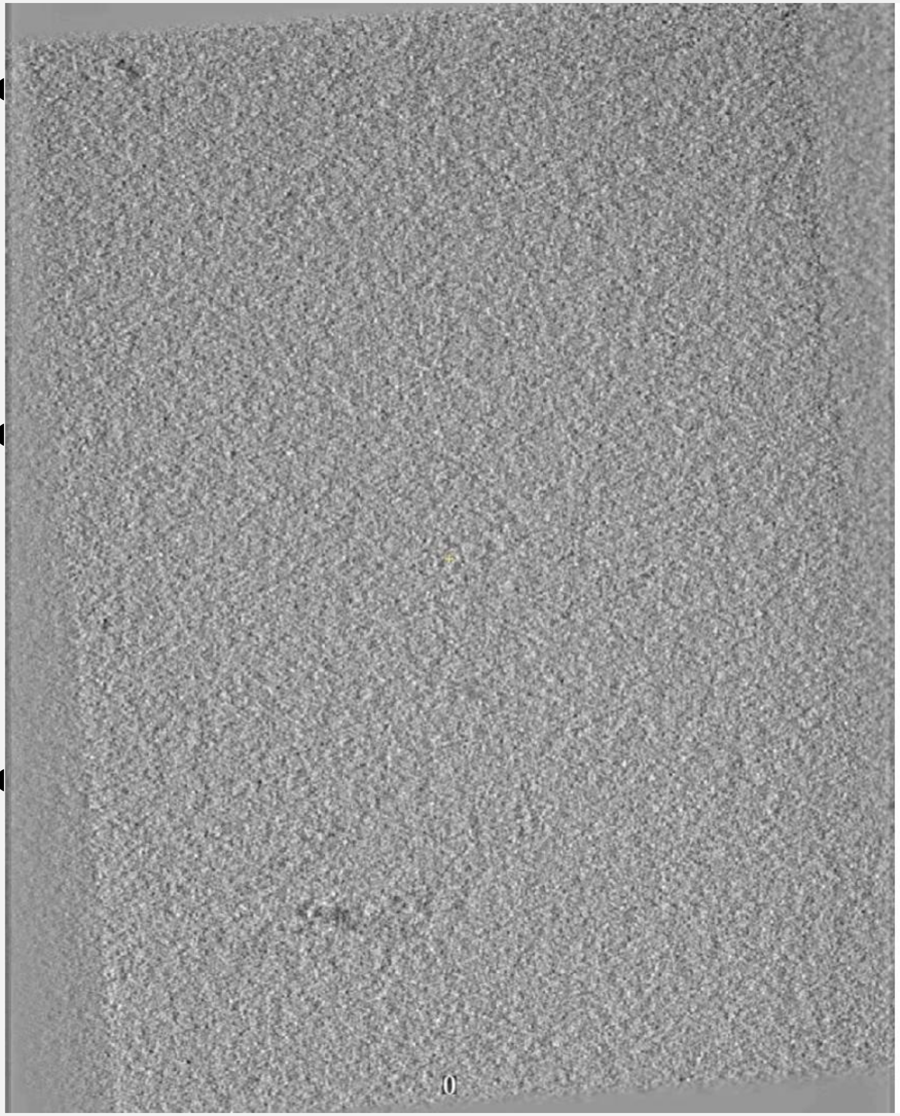
- **Visualize** and **identify proteins** more **confidently** with Topaz Denoising

- Note: **If you don't know** what your particle looks like, **do tomography** first!



doi: 10.1038/s41592-019-0575-8 (Picking) 10.1038/s41467-020-18952-1 (Denoising)

Summary



doi: [10.1038/s41592-019-0575-8 (Picking)](https://doi.org/10.1038/s41592-019-0575-8) [10.1038/s41467-020-18952-1 (Denoising)](https://doi.org/10.1038/s41467-020-18952-1)

Summary

- In practice, people often publish with **multiple pickers – combine picks**
- **If you run Topaz iteratively**, you must run 2D/3D **classification** and **manually check** some micrographs so you **don't get Einstein from noise**!

doi: 10.1038/s41592-019-0575-8 (Picking) 10.1038/s41467-020-18952-1 (Denoising)

Acknowledgements

Massachusetts Institute of Technology

- **Tristan Bepler**

- **Bonnie Berger**

- Andrew Morin
- Hoon Cho
- Tommi Jaakkola

CryoSparc Team

- **Jay Yoo**

Simons Electron Microscopy Center

- **Kotaro Kelley**
- Ed Eng
- Daija Bobe
- Misha Kopylov
- Bridget Carragher
- Clint Potter

Columbia University

- **Micah Rapp**
- **Julia Brasch**
- Larry Shapiro

CSAIL SF NIH National Institutes of Health

CSBi NRAMM NCCAT COLUMBIA UNIVERSITY

Massachusetts Institute of Technology

SMLC SIMONS MACHINE LEARNING CENTER

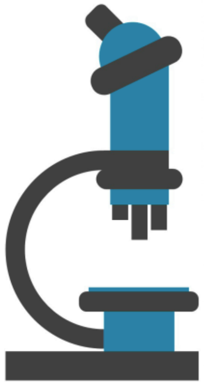
SIMONS ELECTRON MICROSCOPY CENTER

SEMC/SRAMM

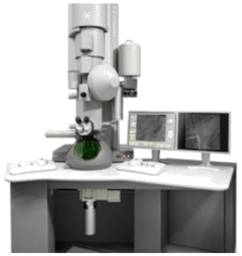
SRAMM

NCCAT









NCITU






Massachusetts Institute of Technology

SMLC
SIMONS MACHINE LEARNING CENTER

SIMONS ELECTRON MICROSCOPY CENTER

SEMC/SRAMM

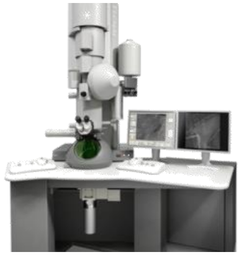
SRAMM

NCCAT

Apply for time & training (FREE!)

CryoEM: https://www.cryoemcenters.org

CryoET: https://www.cryoetportal.org

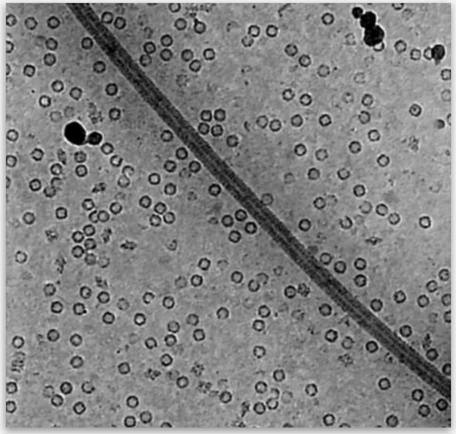


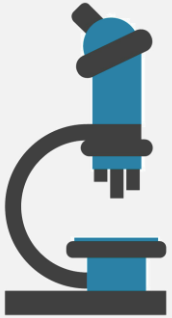
NCITU

Massachusetts Institute of Technology

SMLC
SIMONS MACHINE LEARNING CENTER

SIMONS ELECTRON MICROSCOPY CENTER

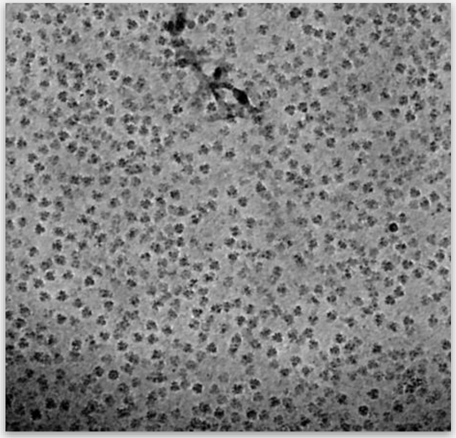




TOPAZ

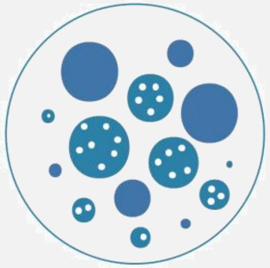
Pick more of anything

http://cb.csail.mit.edu/cb/topaz

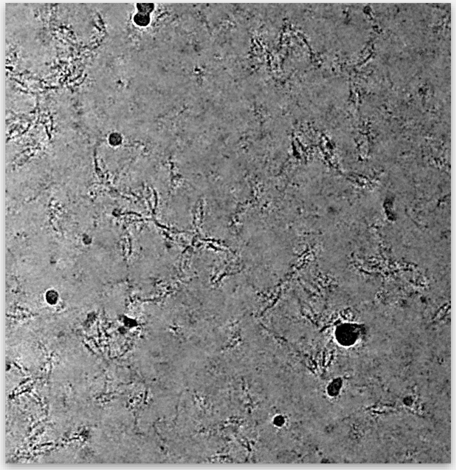


Thank you!

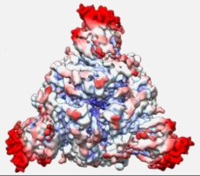
Any questions?



tw: **@alexjamesnoble**



Denoising
Nature Communications
bioRxiv: 10.1101/838920



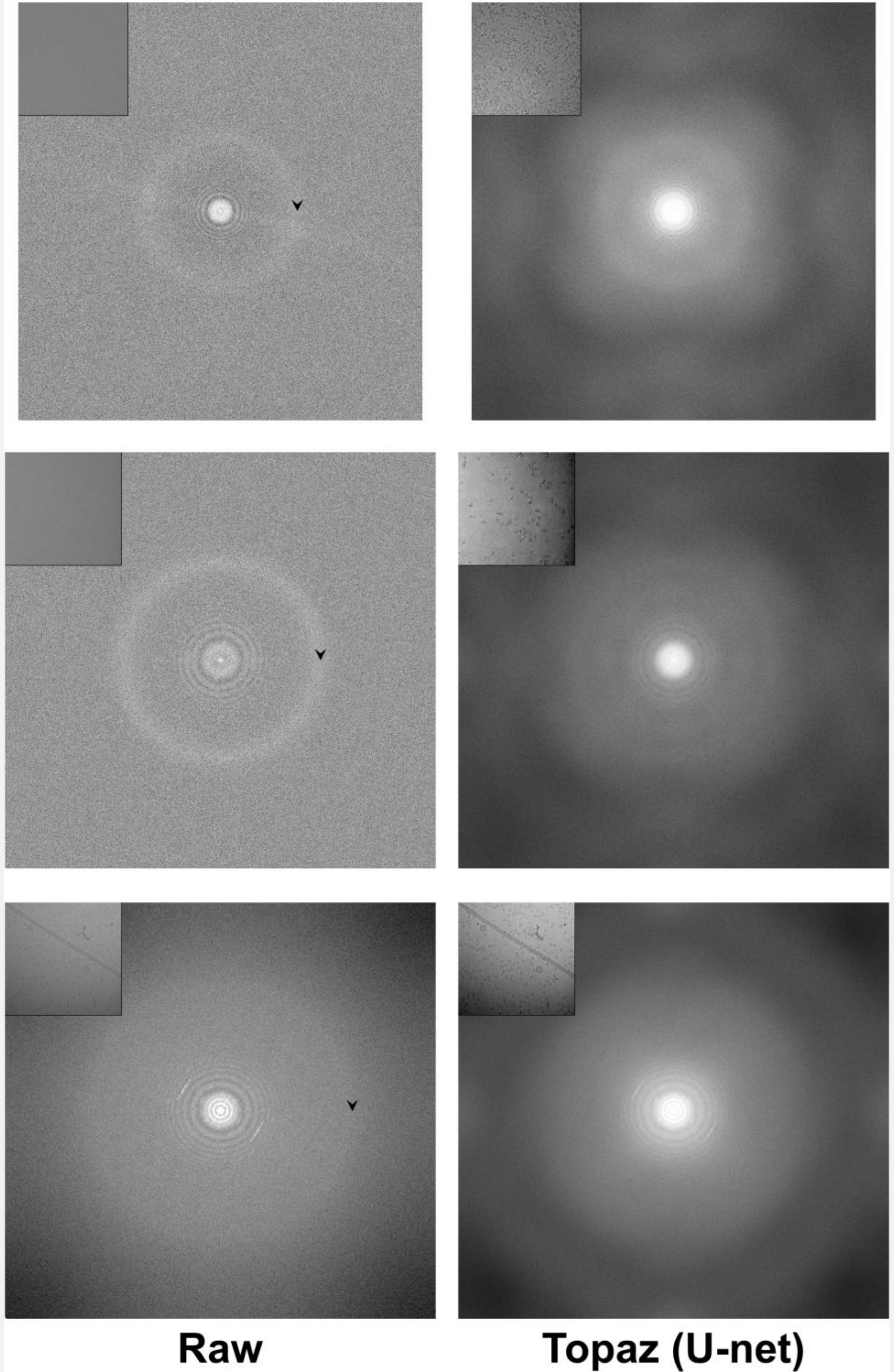
Particle picking
Nature Methods
Preprint on arXiv: 1803.08207

Massachusetts Institute of Technology

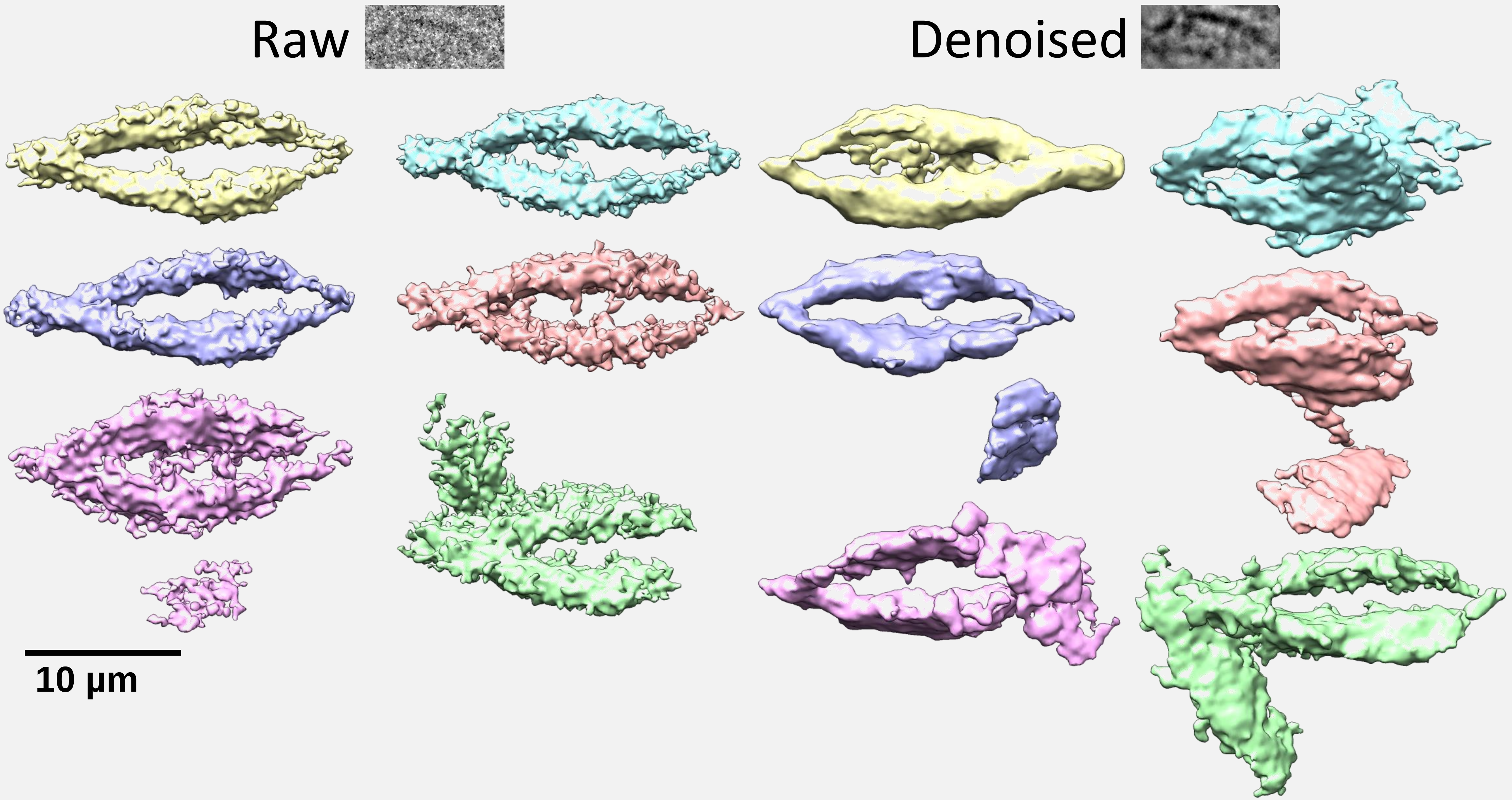
SMLC
SIMONS MACHINE LEARNING CENTER

SIMONS ELECTRON MICROSCOPY CENTER

FFTs look cool!

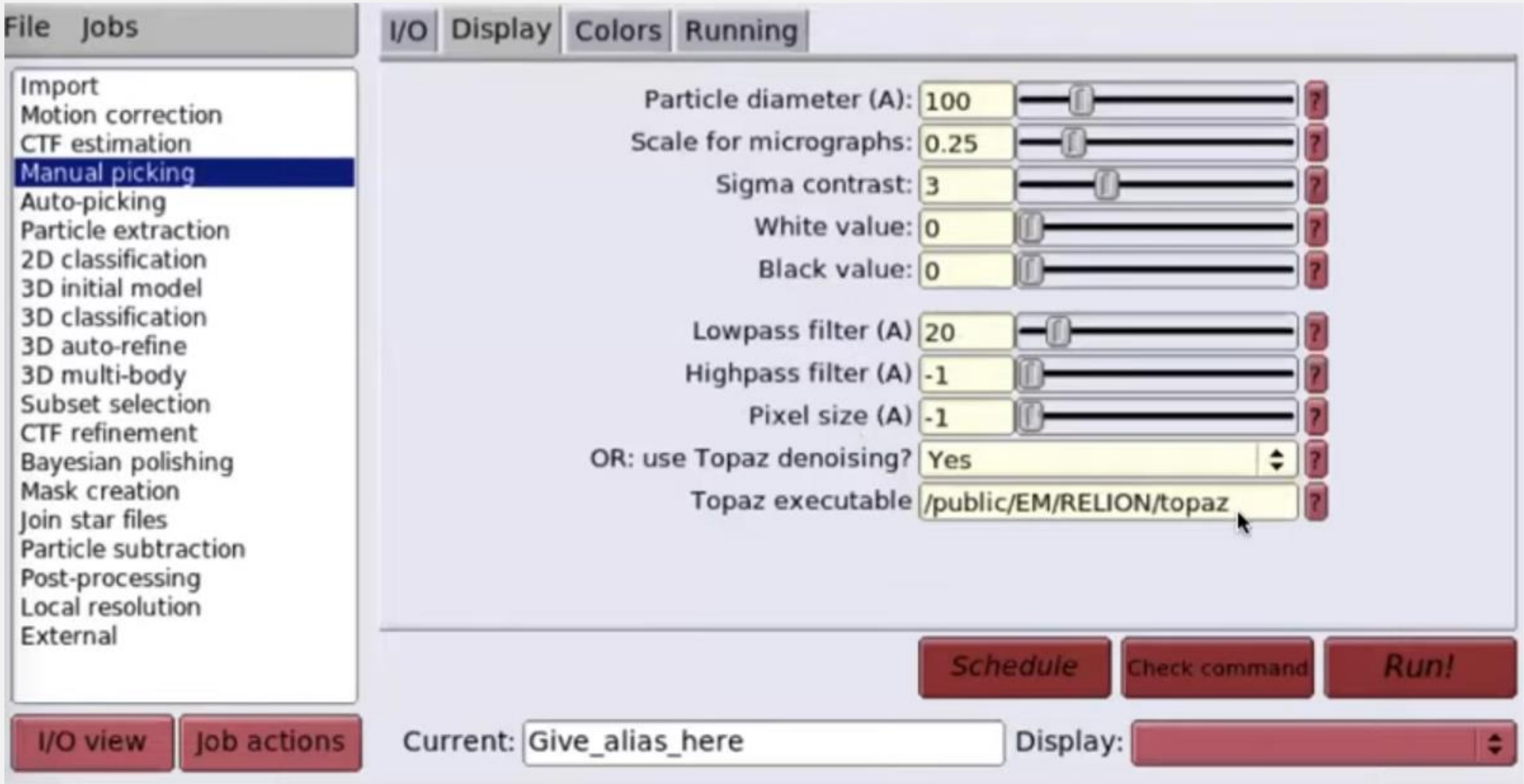


Denoised ab-initio models are less reliable



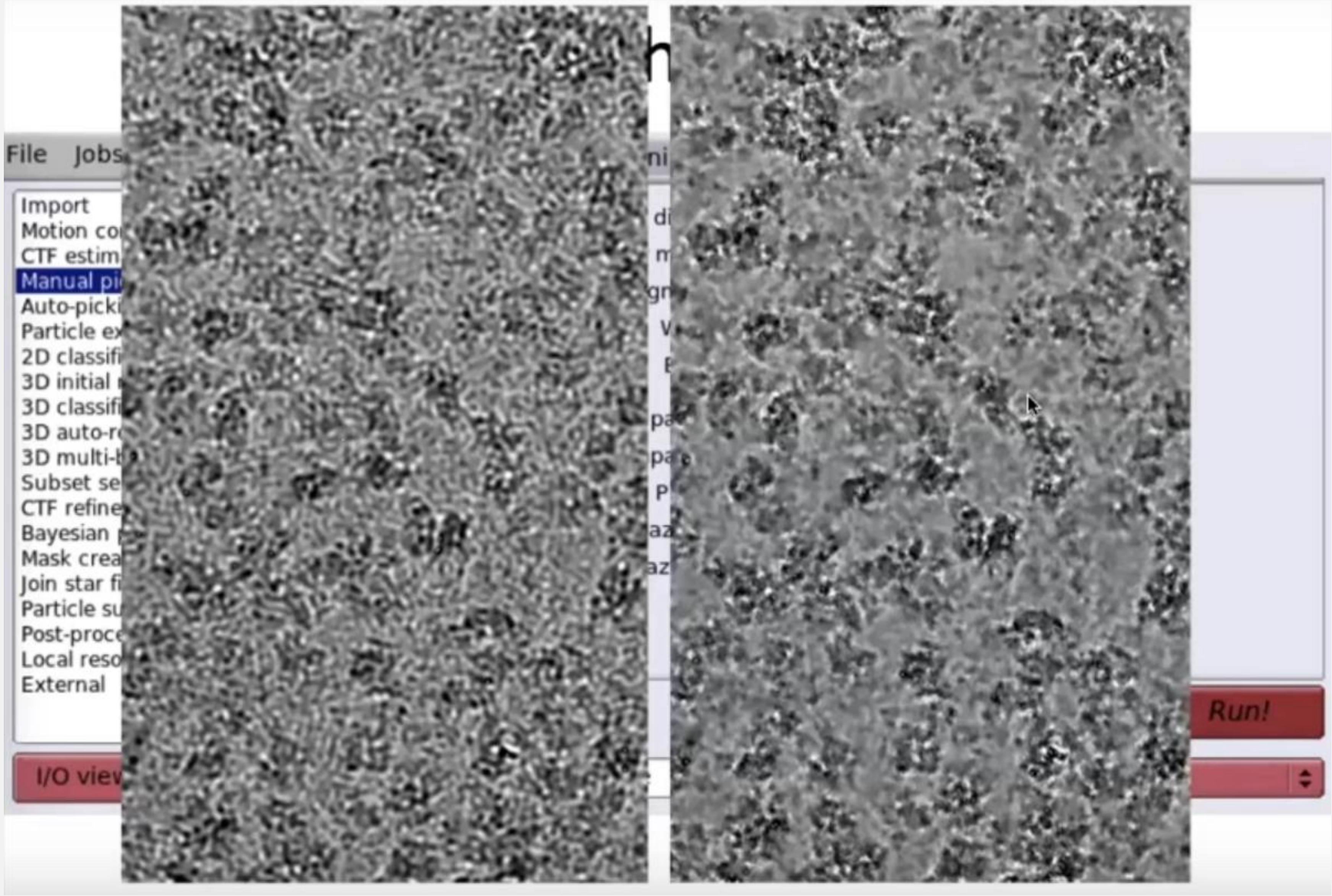
Quick look at Relion4 integration

Denoising in Relion4



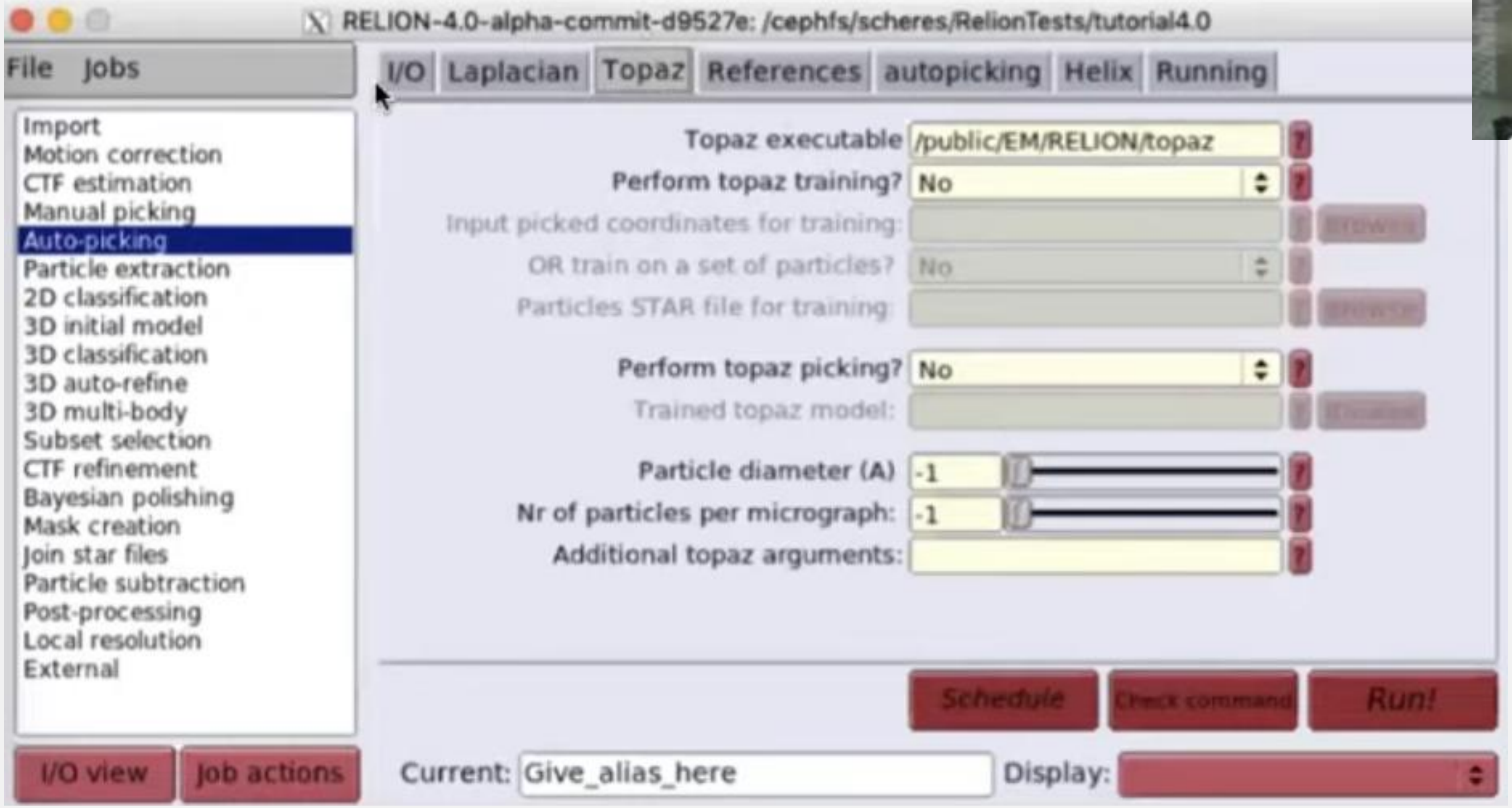
https://youtu.be/NA6i23qmWfw

Denoising in Relion4



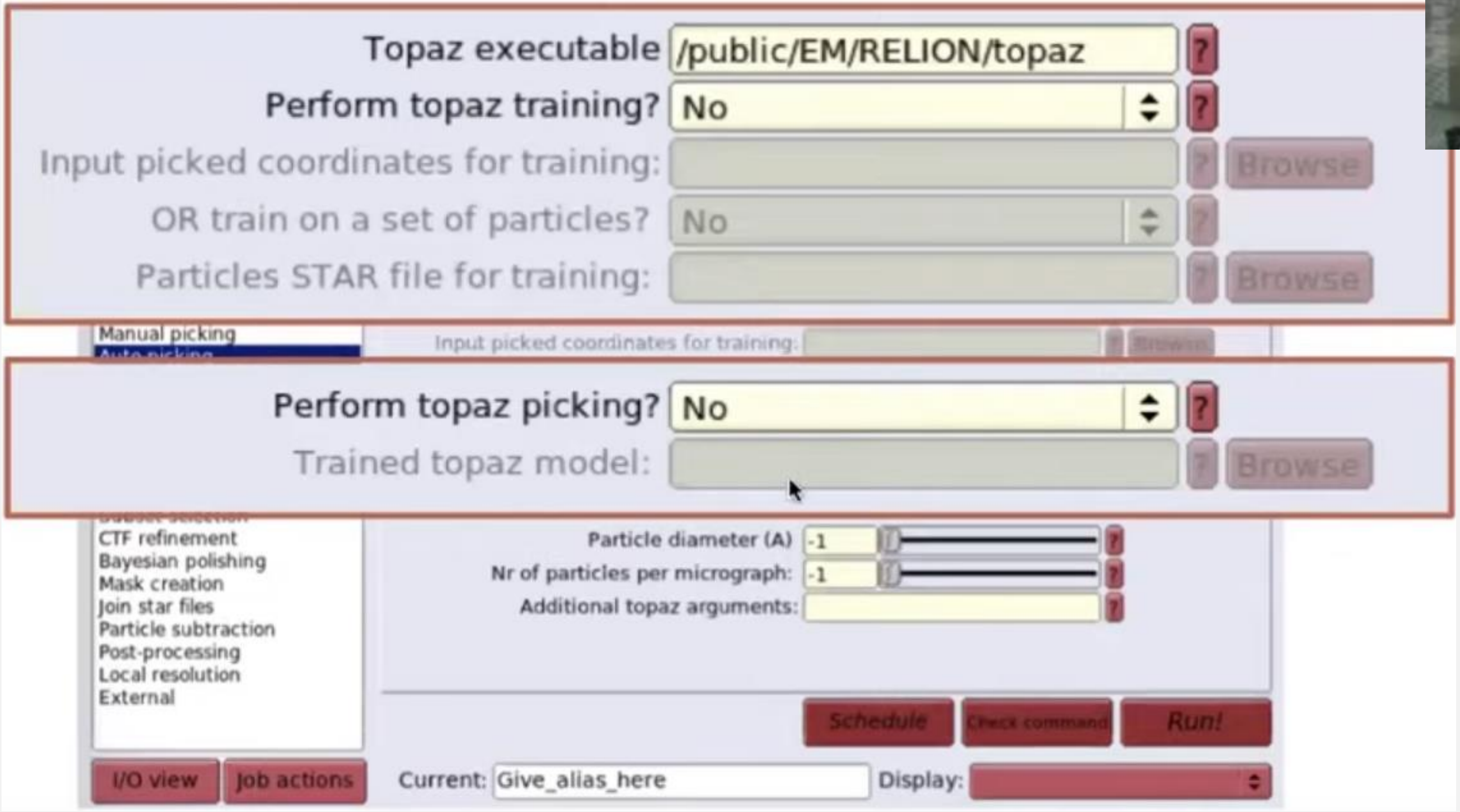
https://youtu.be/NA6i23qmWfw

Topaz autopicking in Relion4



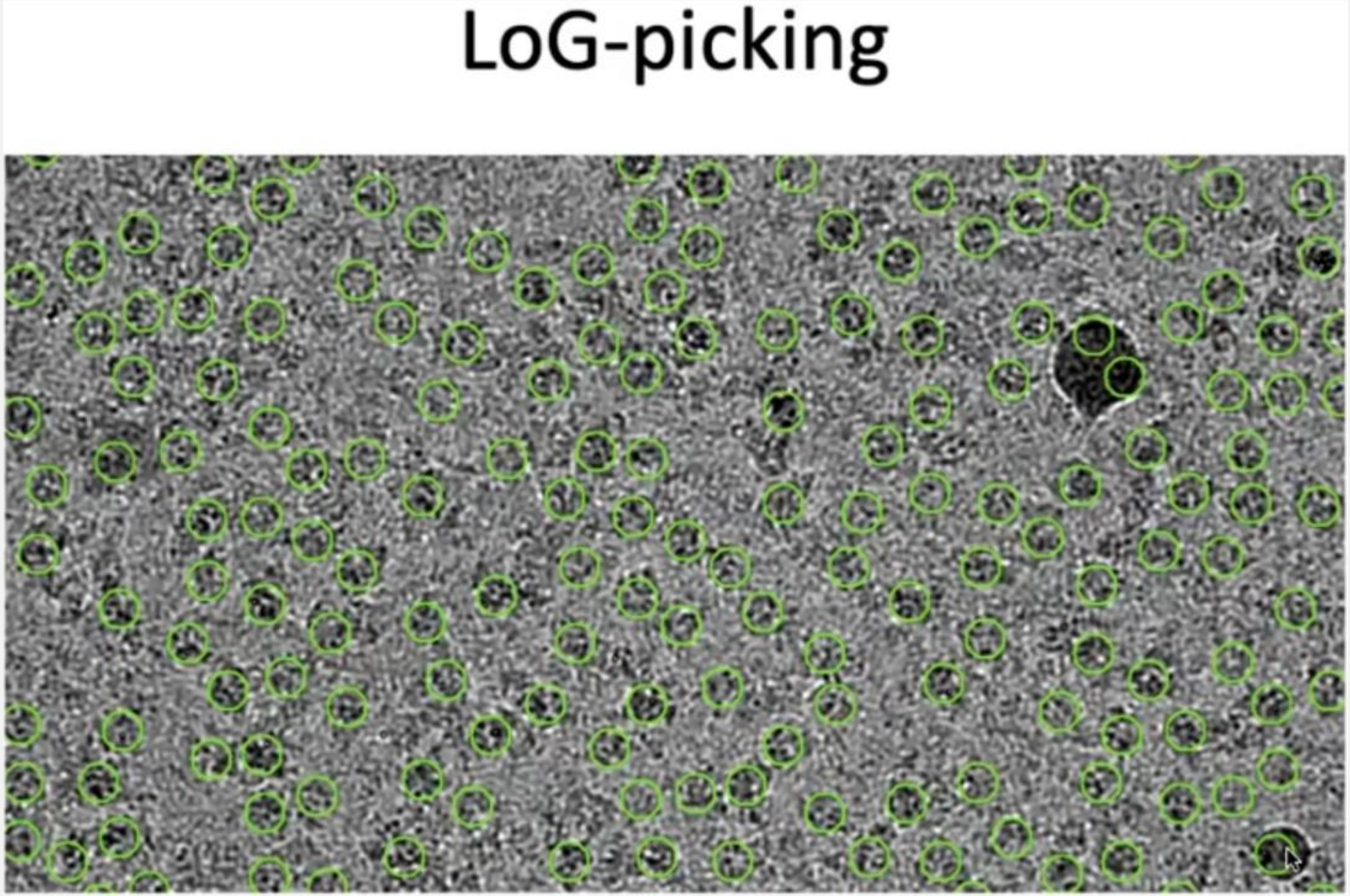
https://youtu.be/NA6i23qmWfw

Topaz autopicking in Relion4



https://youtu.be/NA6i23qmWfw

Topaz autopicking in Relion4



https://youtu.be/NA6i23qmWfw

Topaz autopicking in Relion4

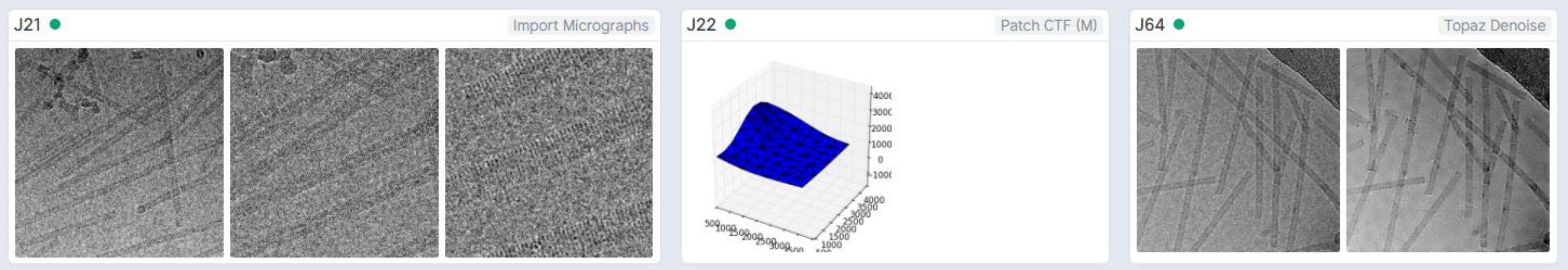
https://youtu.be/NA6i23qmWfw

Best-practices Topaz Workflow (within CryoSparc)

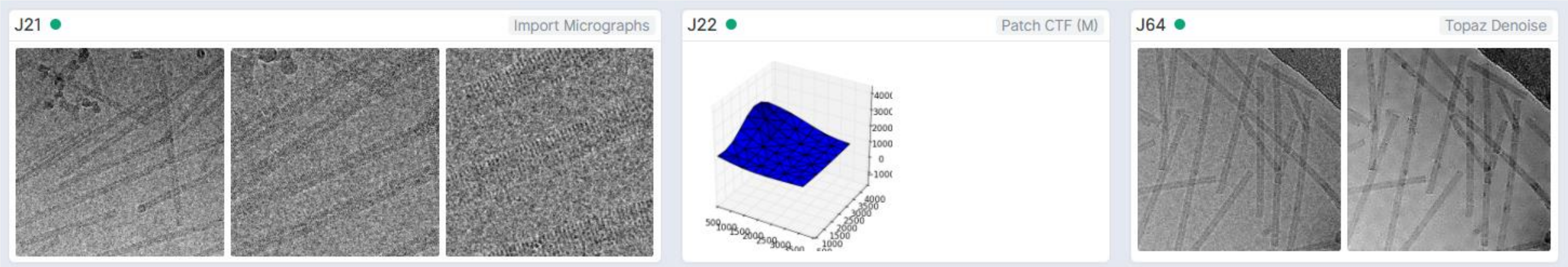
General Recommended Workflow

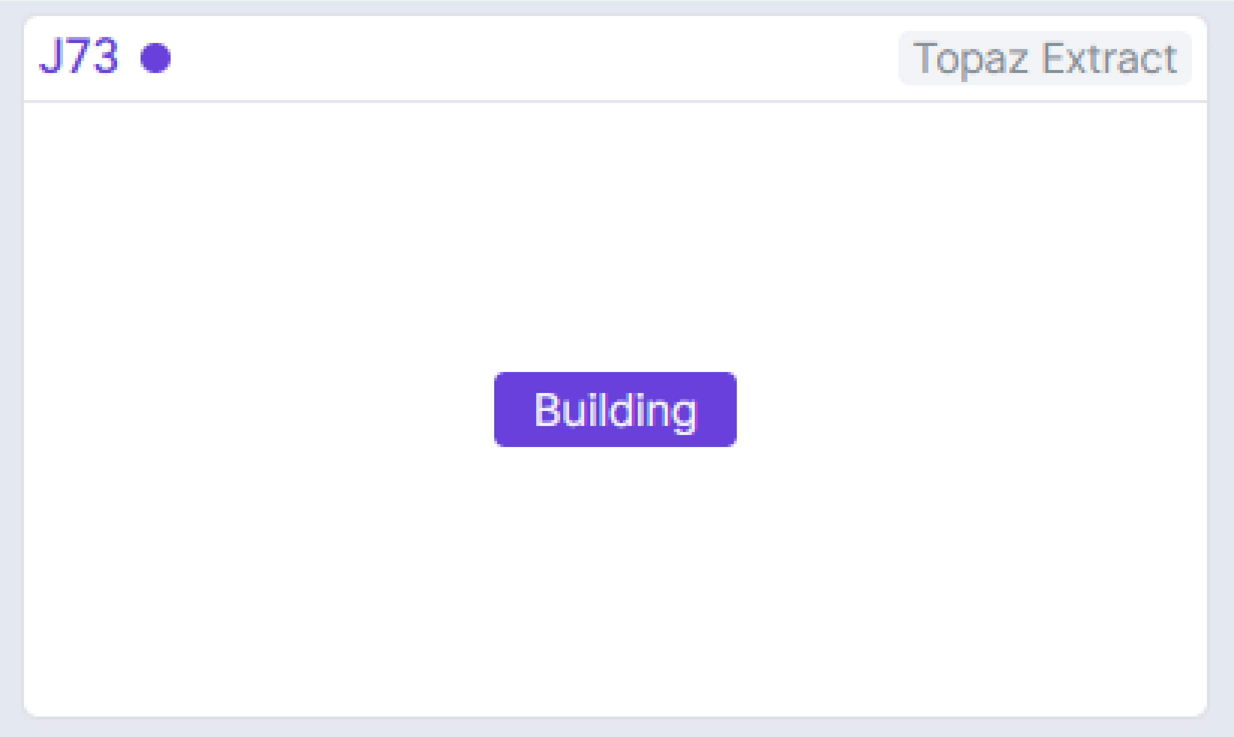
shown in CryoSPARC

EMPIAR-10022: TMV

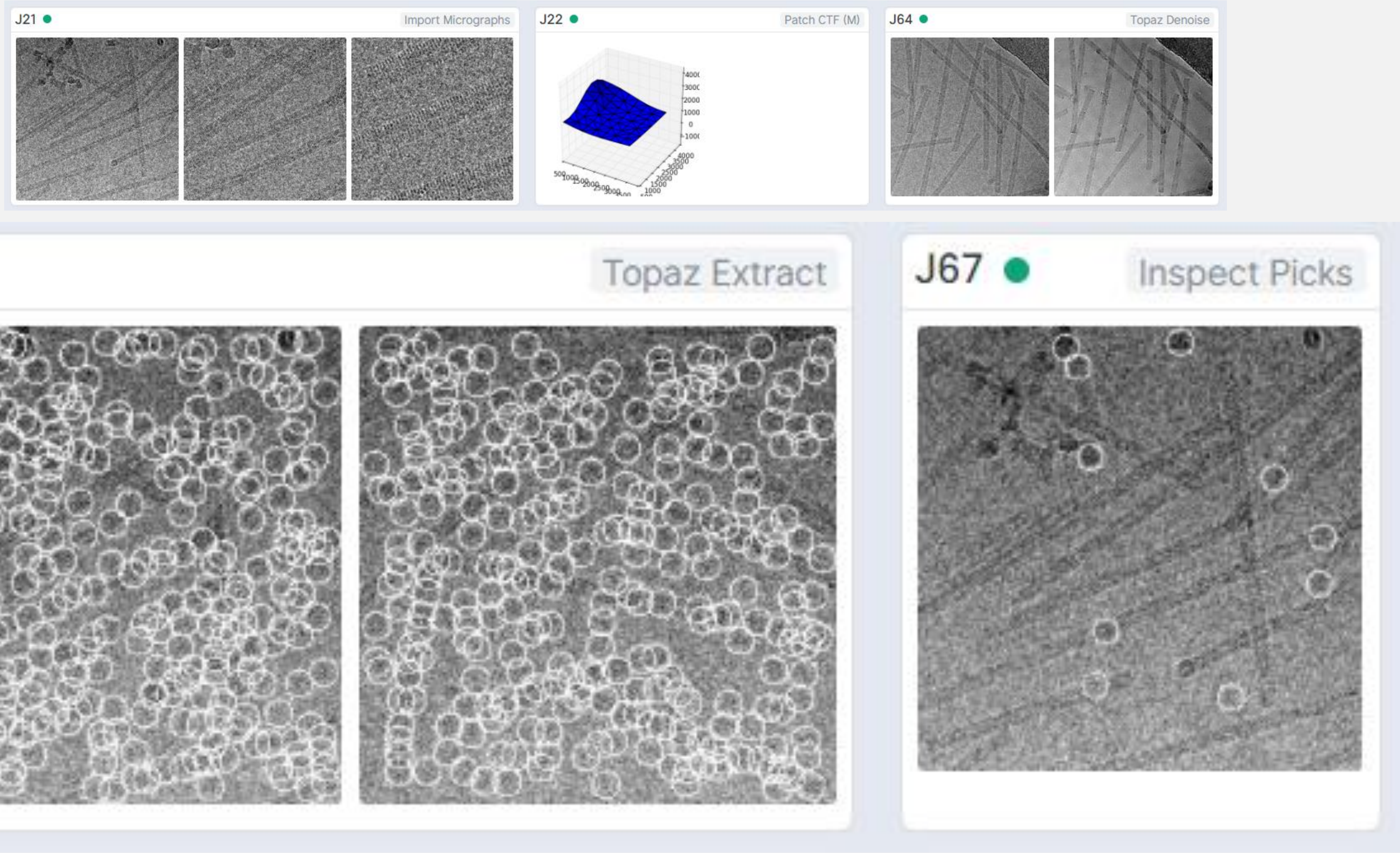


General Recommended Workflow
shown in CryoSPARC

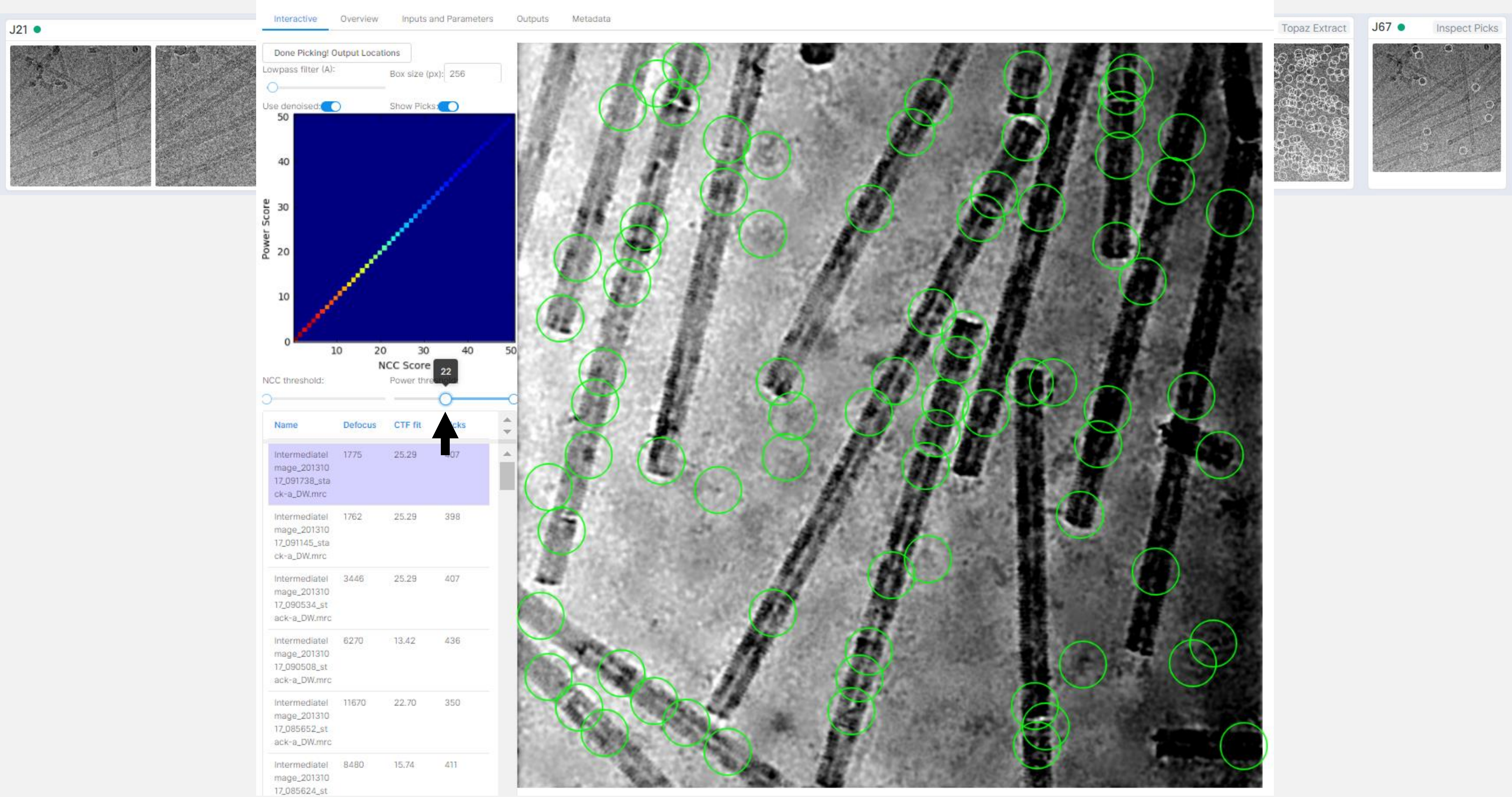




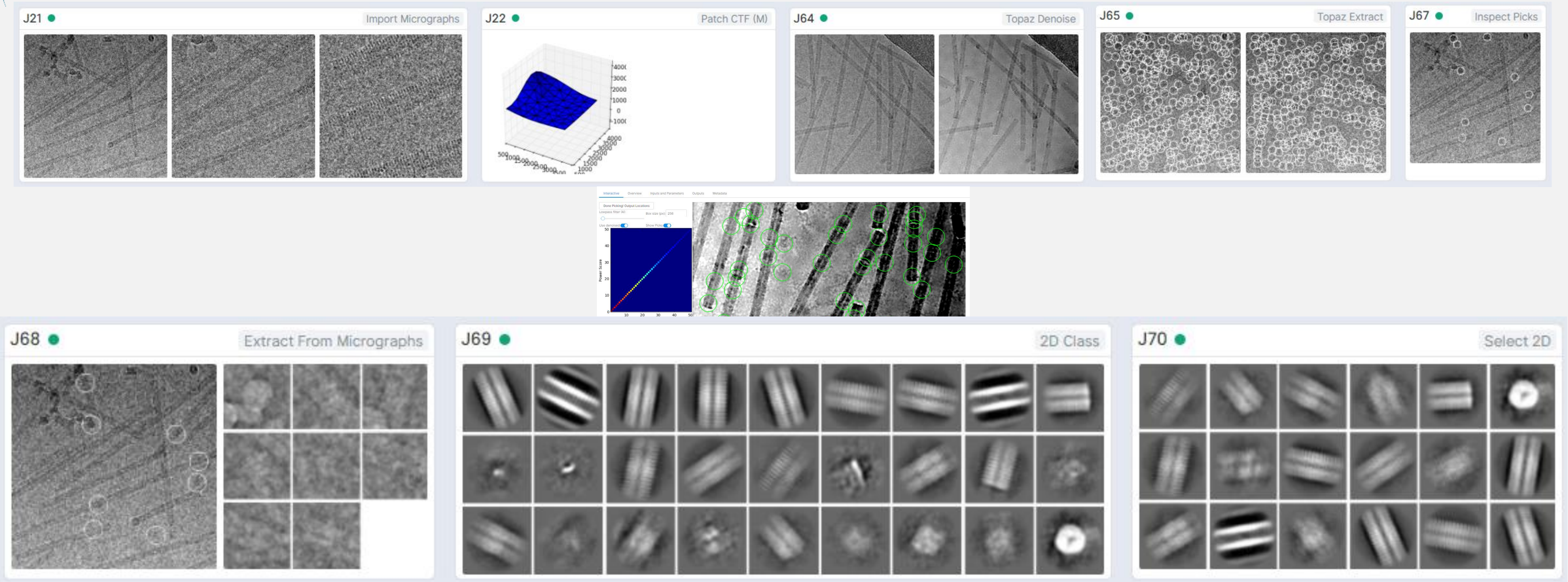
General Recommended Workflow
shown in CryoSPARC



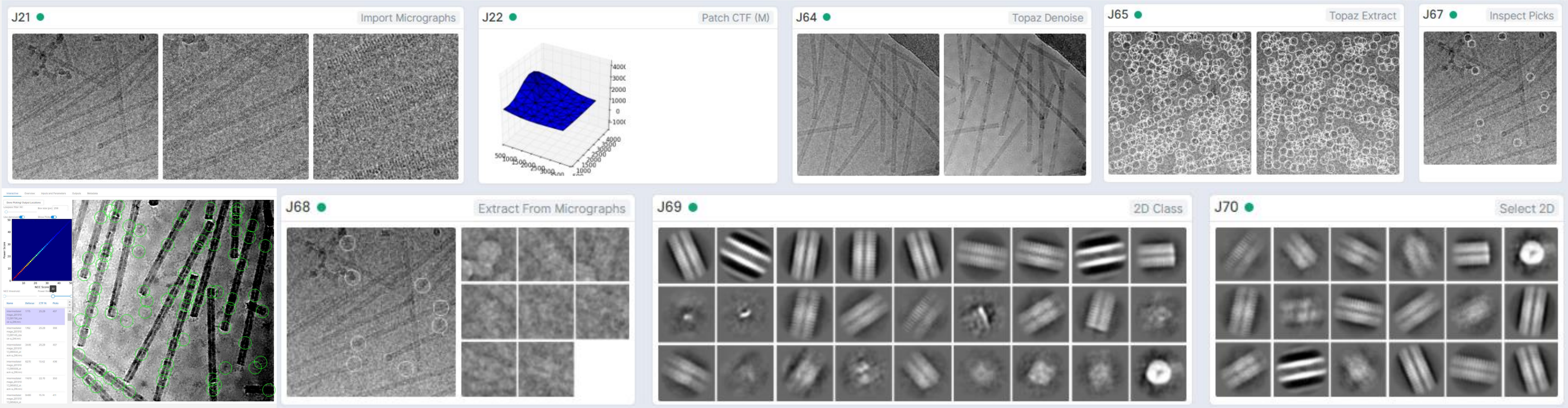
General Recommended Workflow
shown in CryoSPARC



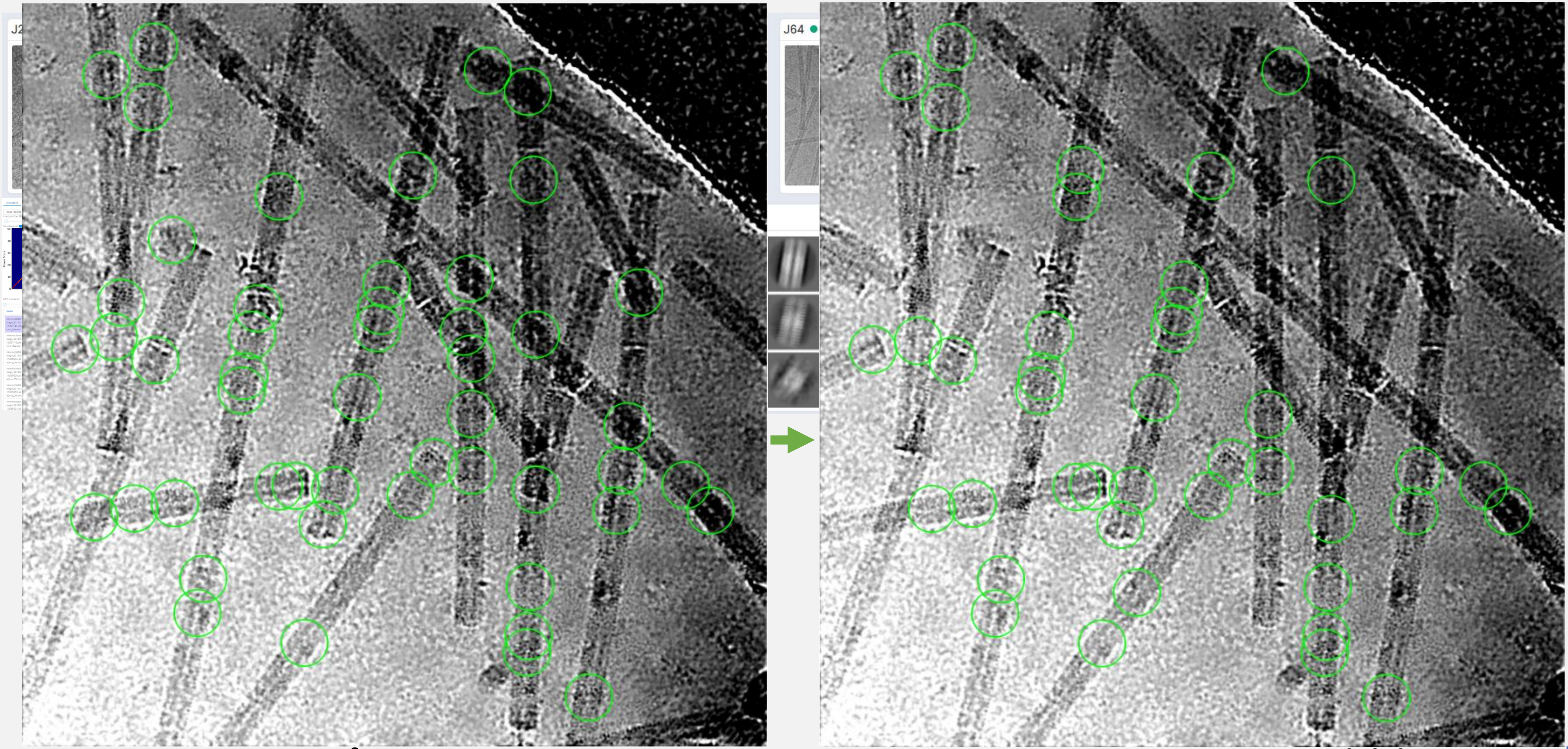
General Recommended Workflow
shown in CryoSPARC



General Recommended Workflow
shown in CryoSPARC

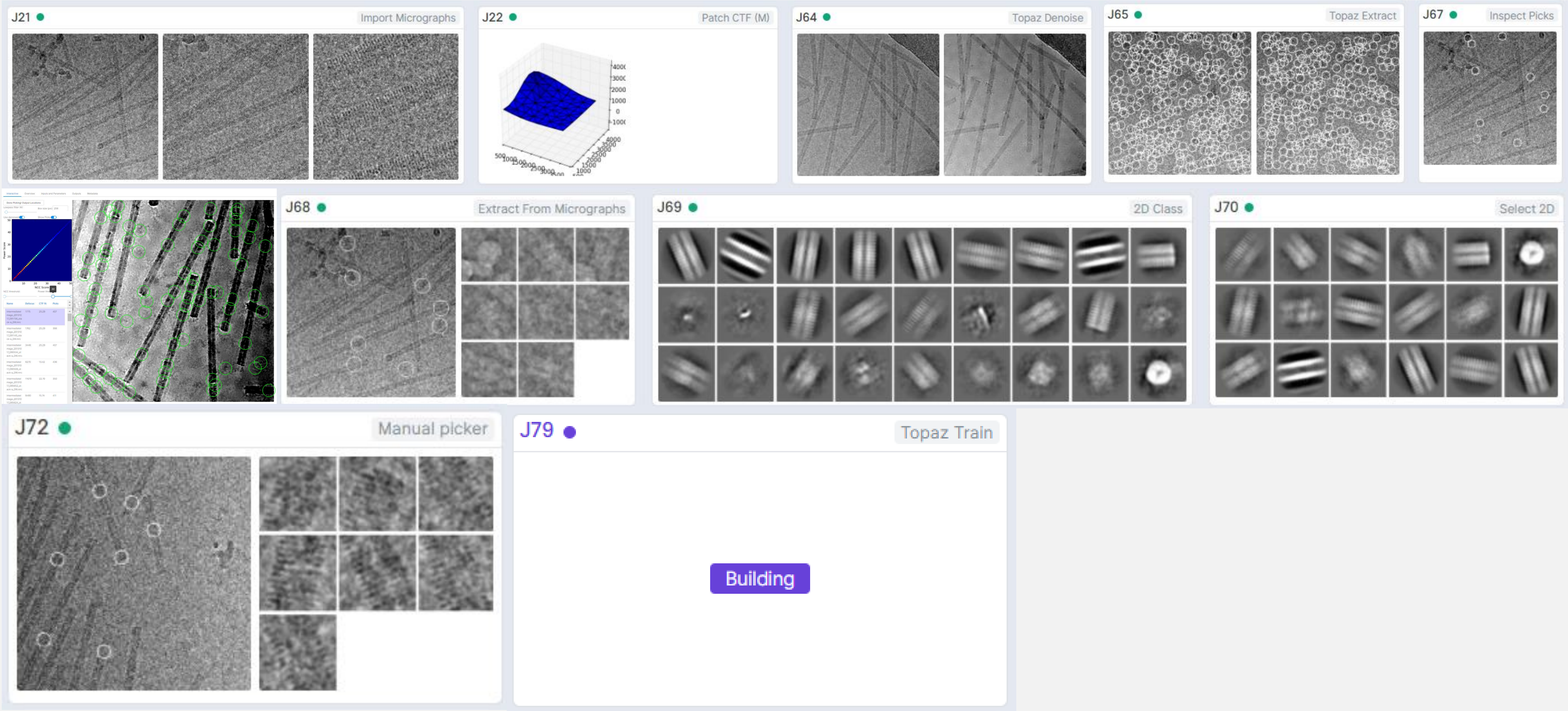


General Recommended Workflow
shown in CryoSPARC



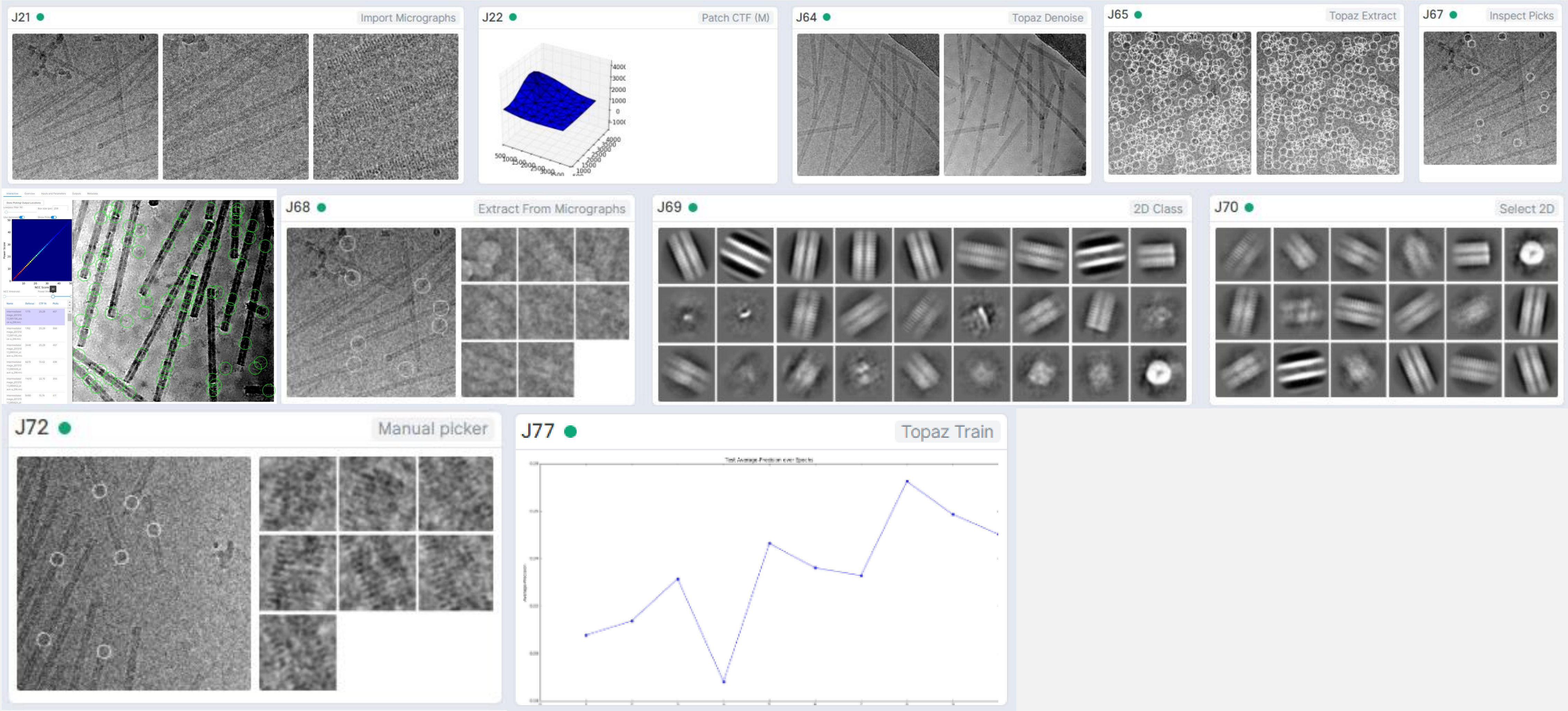
Manual curation – very important step!!!

General Recommended Workflow
shown in CryoSPARC

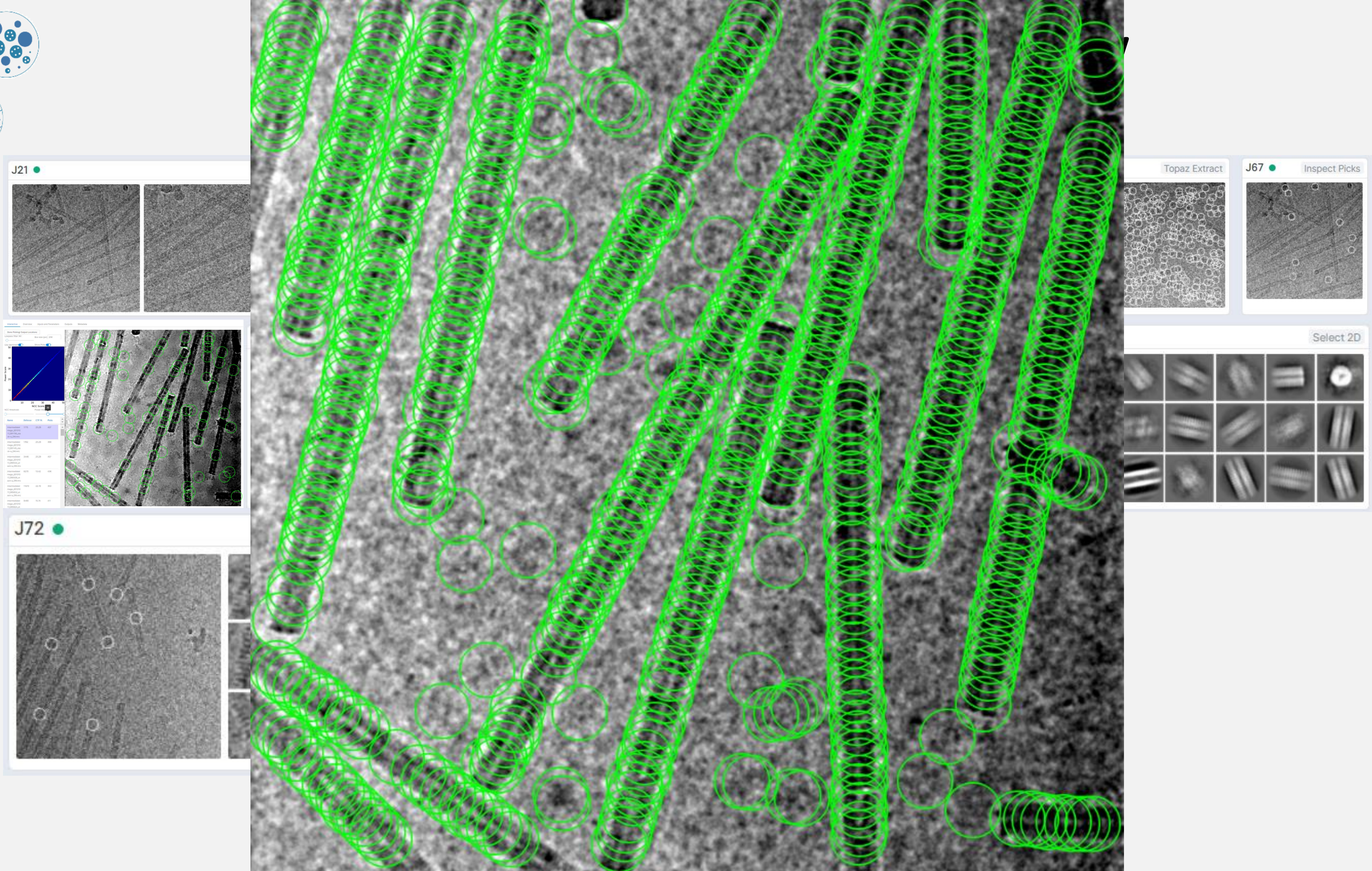


Trained with 3.3k manually curated particles

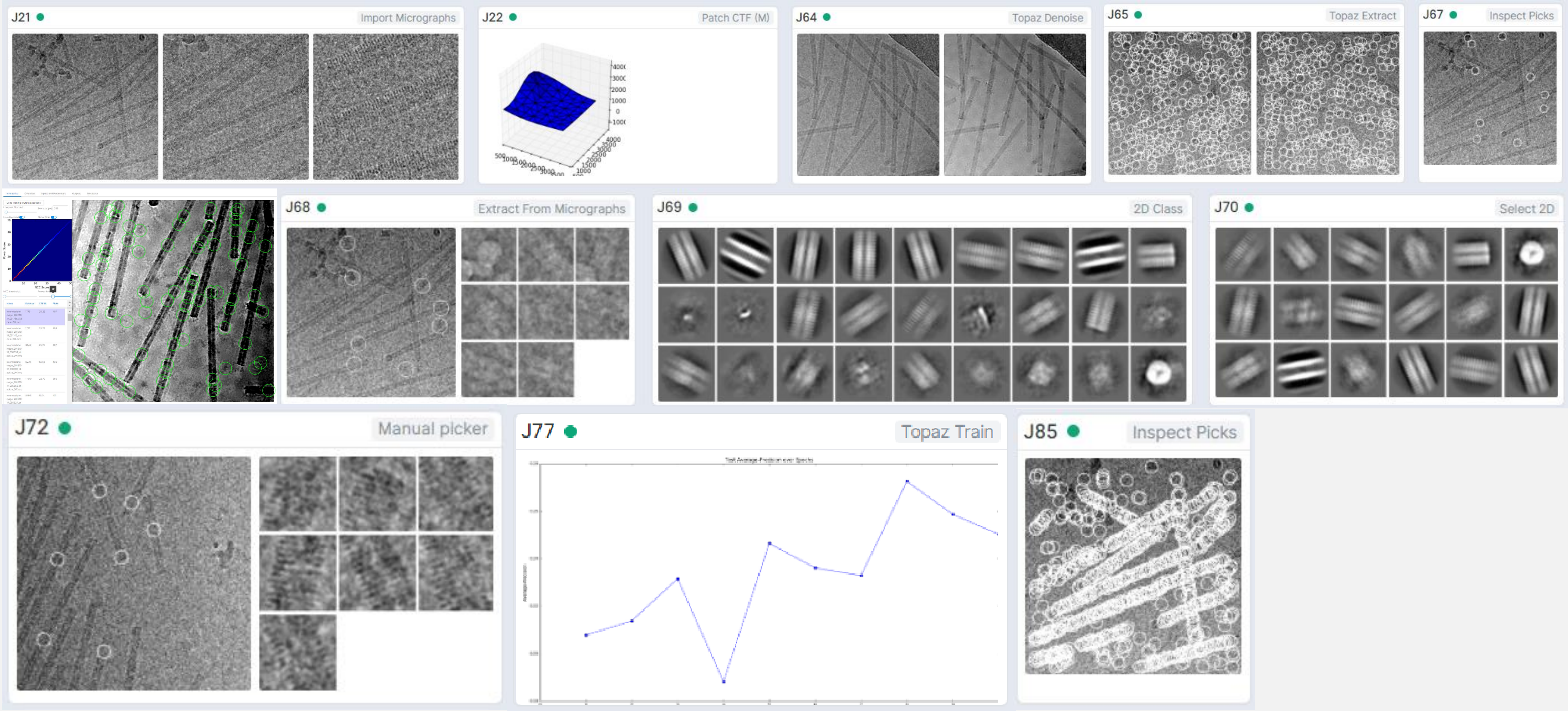
General Recommended Workflow
shown in CryoSPARC



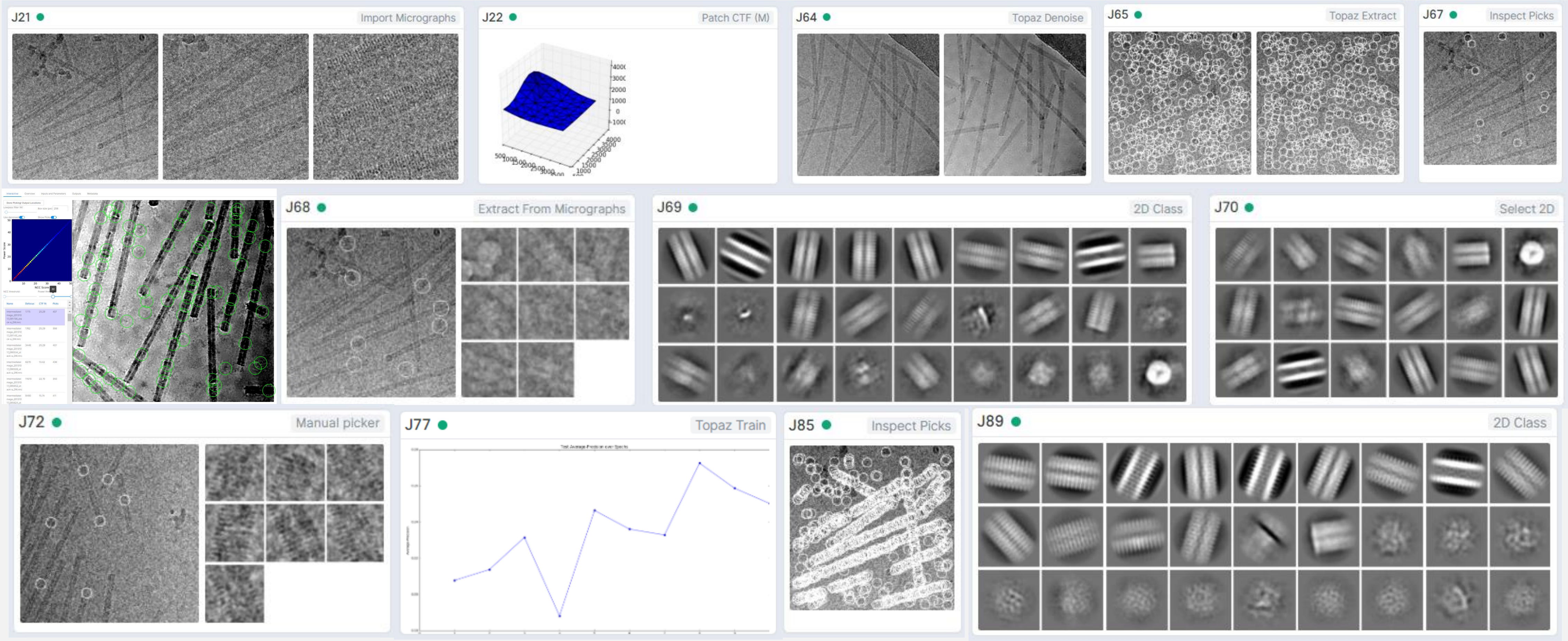
Trained with 3.3k manually curated particles


J21
J72
Topaz Extract
J67
Inspect Picks
Select 2D

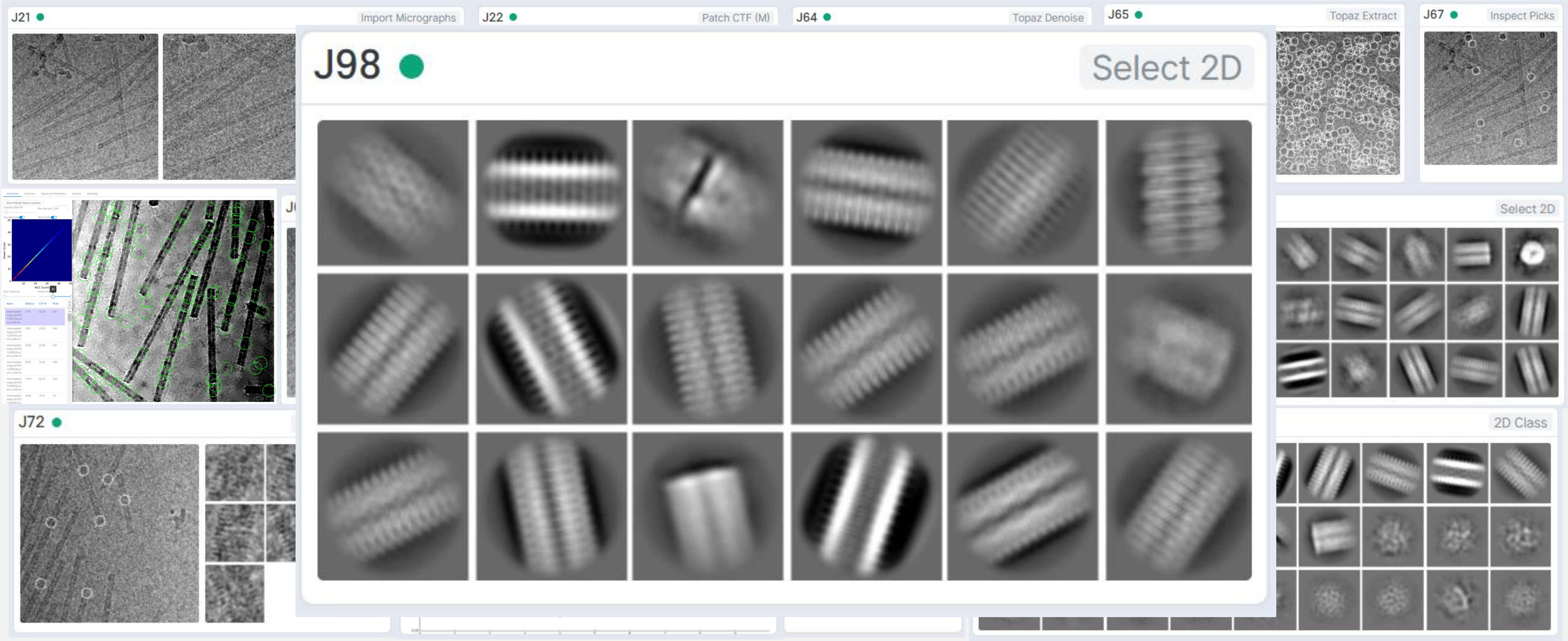
General Recommended Workflow
shown in CryoSPARC



General Recommended Workflow
shown in CryoSPARC



General Recommended Workflow
shown in CryoSPARC



91% of particles picked were kept

General Recommended Workflow
shown in CryoSPARC