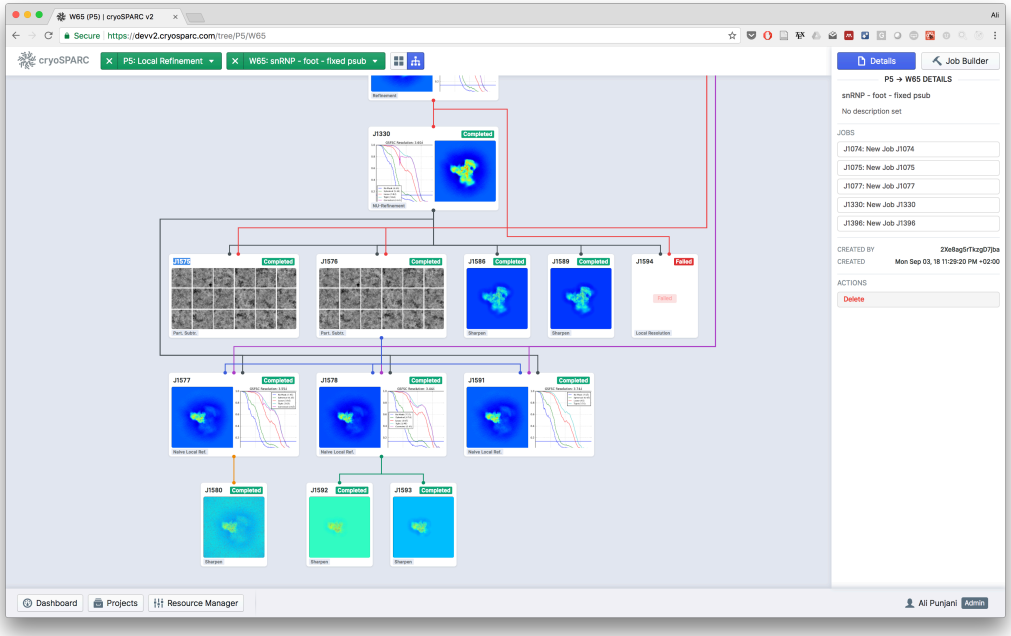
cryoSPARC

Ali Punjani
University of Toronto
Structura Biotechnology Inc.

EMBO Birkbeck
Sept 2019

CryoSPARC

Algorithm Development, Software, Performance

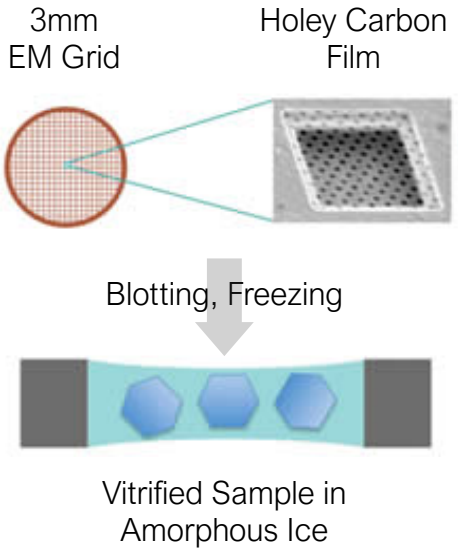


- Complete single-particle processing workflow
- Freely available for academia, supported by industry
- Under heavy development, regularly updated
- Methods and algorithms
  - Computational statistics, optimization theory, machine learning
- Professionally designed and built UI/software platform
- Ultra high-performance

Computational considerations for SPA

Computational considerations for SPA

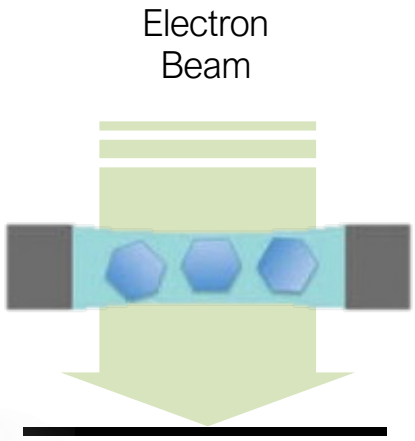
- Thin film of liquid sample on grid
- Rapid plunge-freezing to vitrify and preserve sample
- Results in protein particles frozen in random orientations in thin film of ice



- Ice thickness/quality
- Particle concentration
- Sample heterogeneity
- Orientation distributions
- Denaturation/aggregation

Computational considerations for SPA

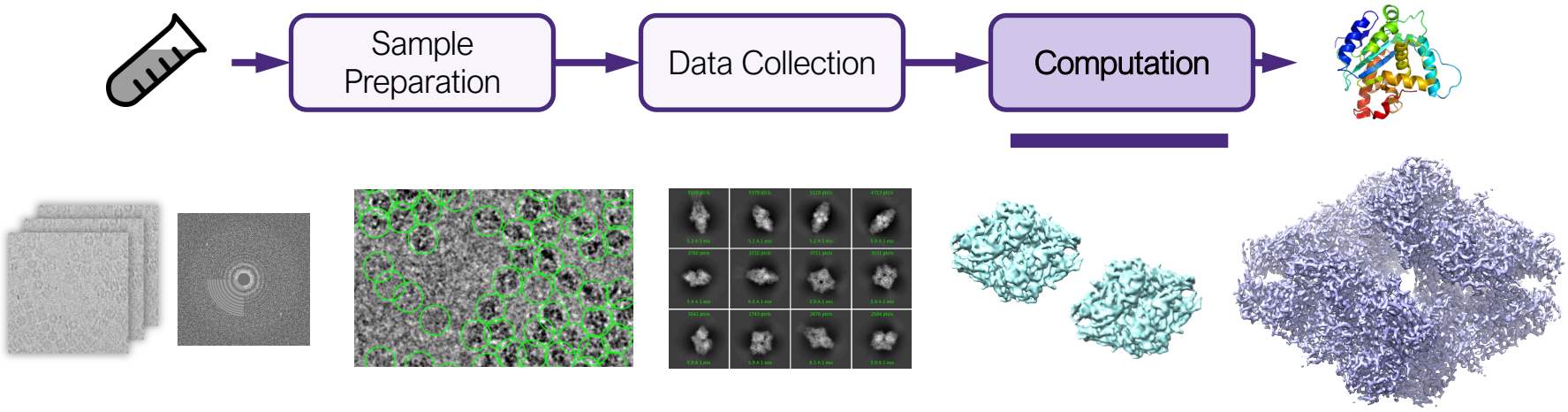
- Imaging of sample in Transmission Electron Microscope
- Avoiding radiation damage: very low electron dose (high noise)



Direct Detector captures 2D projections

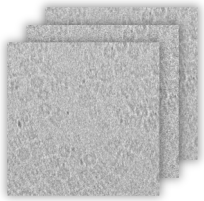
- Phase contrast
- Microscope aberrations
- Stage drift
- Sample deformation
- Radiation damage
- Very high noise levels

Computational considerations for SPA

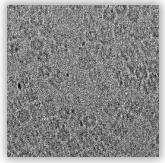


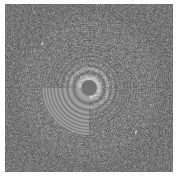
- Pre-processing
- Locate particles in ice
- Exclude junk/broken particles
- 2D-to-3D reconstruction
- Resolving heterogeneity
- High-resolution refinement

Single-particle workflow

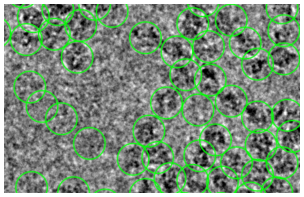


2000 movies
2TB

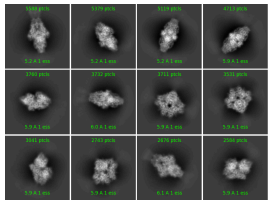




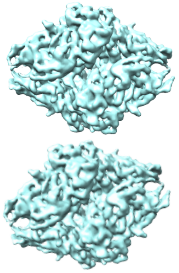
2000 micrographs
200GB



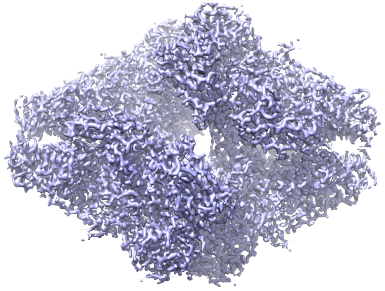
1M particles
400GB



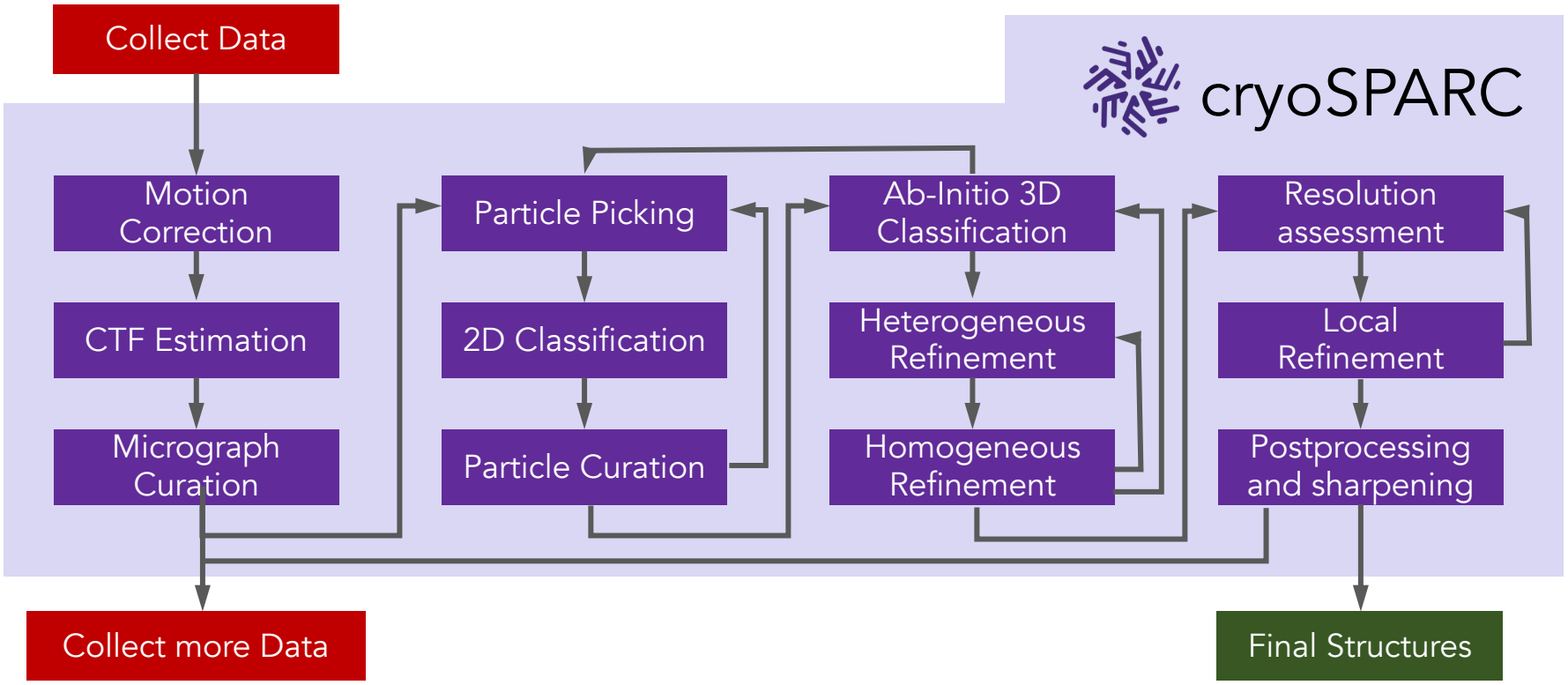
500K particles
200 2D-classes



200K particles
5 3D classes



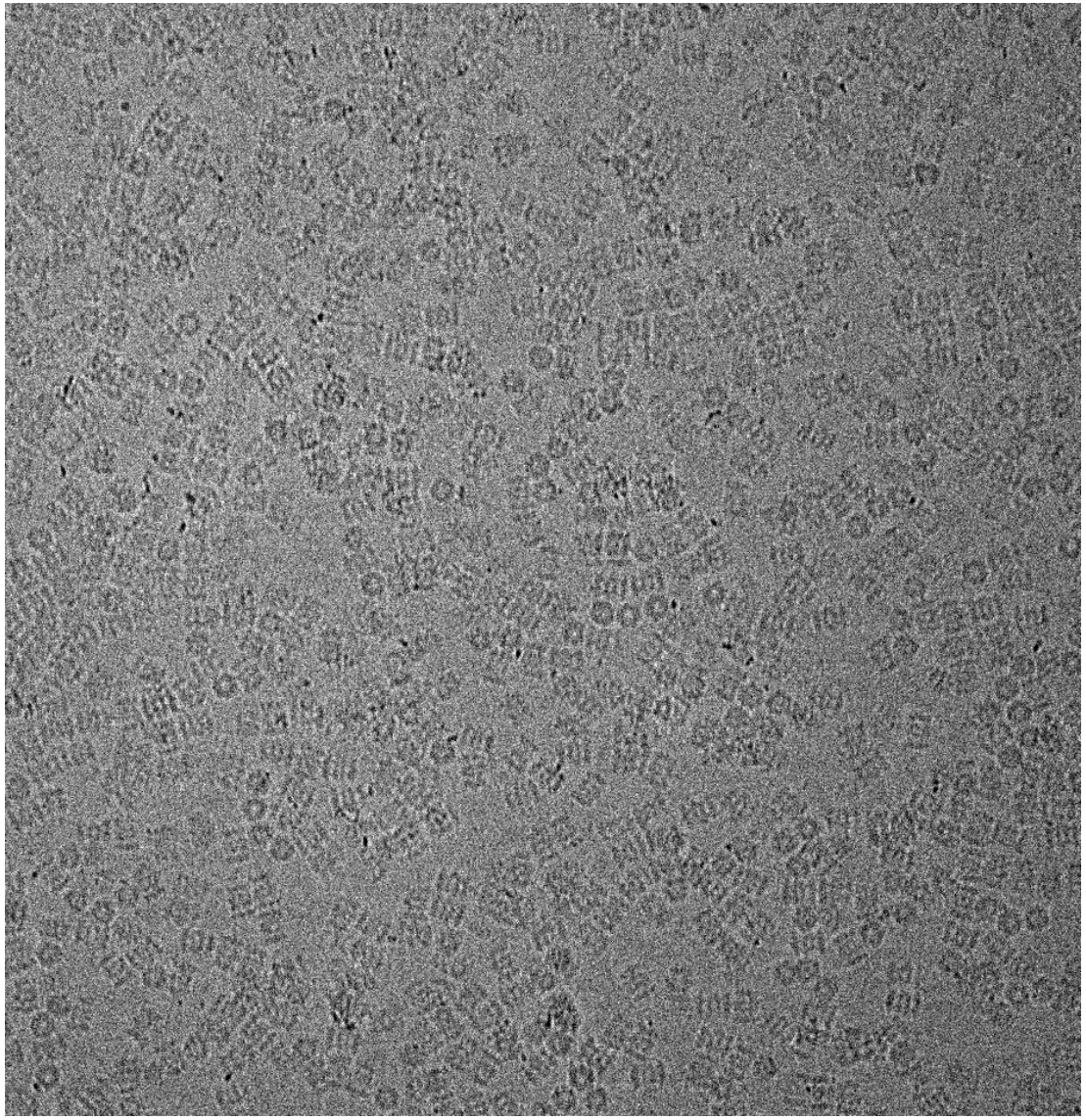
100K particles
3Å resolution



Tutorial Dataset

T20S Proteasome

- Data collected at NRAMM (Titan Krios)
- Publically depositited at EMPIAR (10025)
- 0.6575 A pixel size (super resolution)
- 8k x 8k, 38 frames
- 53 electrons/$A^{2`}$
- Large, stable, no heterogeneity, no flexibility
- Subset of 20 movies
- TIFF format compression



Start Motion Correction

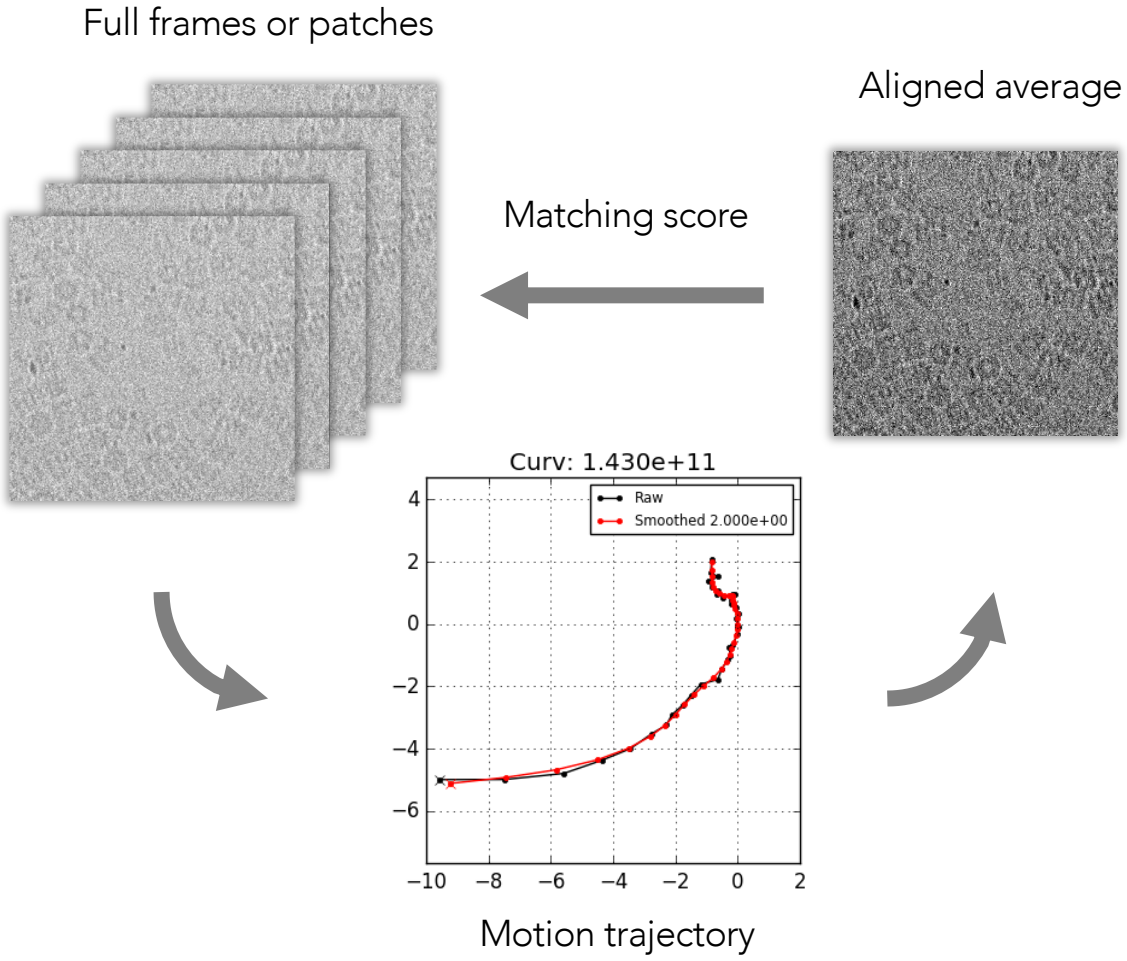
T20S Proteasome

- Full Frame Motion Correction (Multi) job
- Use output from import job
- Default params

Motion correction

General Principle

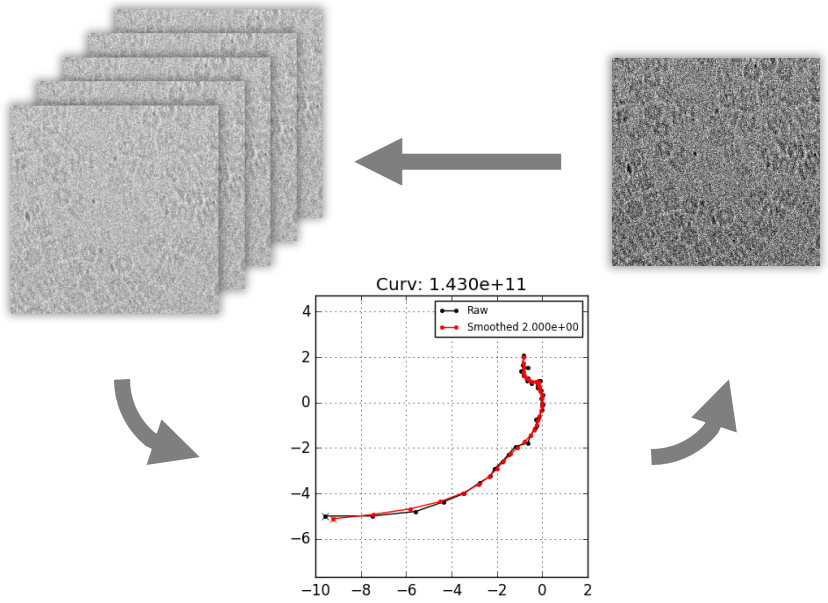
- Maximize matching score over trajectories
- Global motion and local motion



Motion correction

Variant of *alignparts_lmbfgs* [Rubinstein & Brubaker 2015]

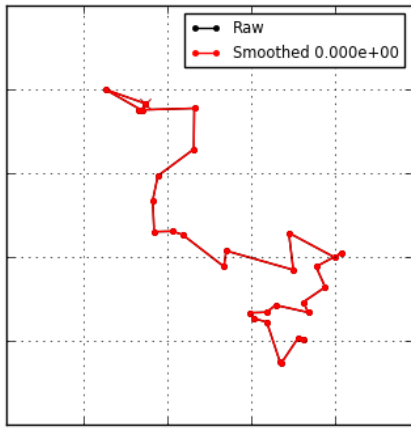
- Maximize cross-correlation between aligned average and frames
- Optimize over trajectories with gradient based quasi-Newton optimizer (LMBFGS)
- Apply a smoothness penalty to trajectory

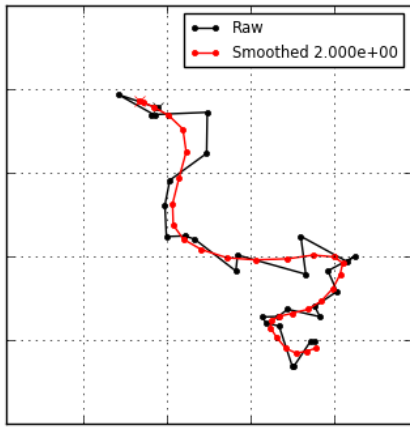


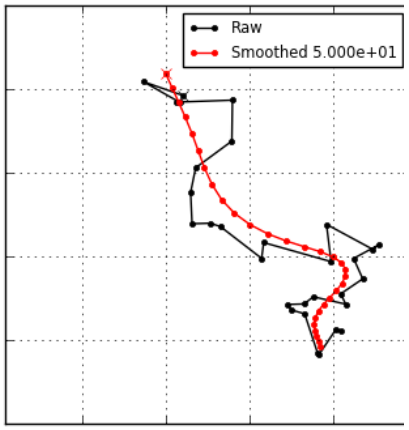
Motion correction

Variant of *alignparts*

- Maximize cross-correlation between aligned average and frames
- Optimize over trajectories with gradient based quasi-Newton optimizer (LMBFGS)
- Apply a smoothness penalty to trajectory



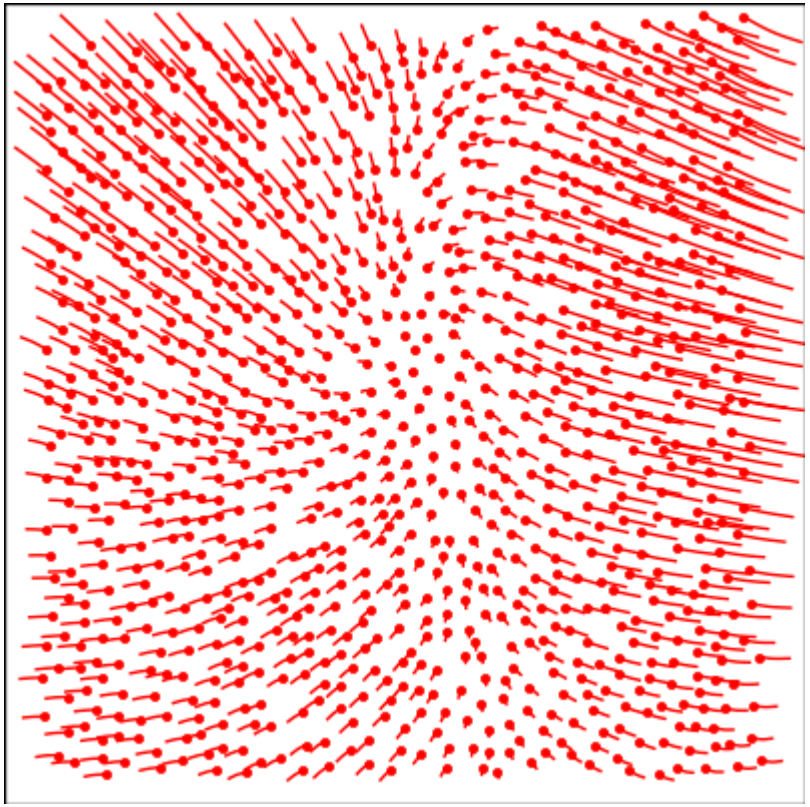




Motion correction

Variant of *alignparts*

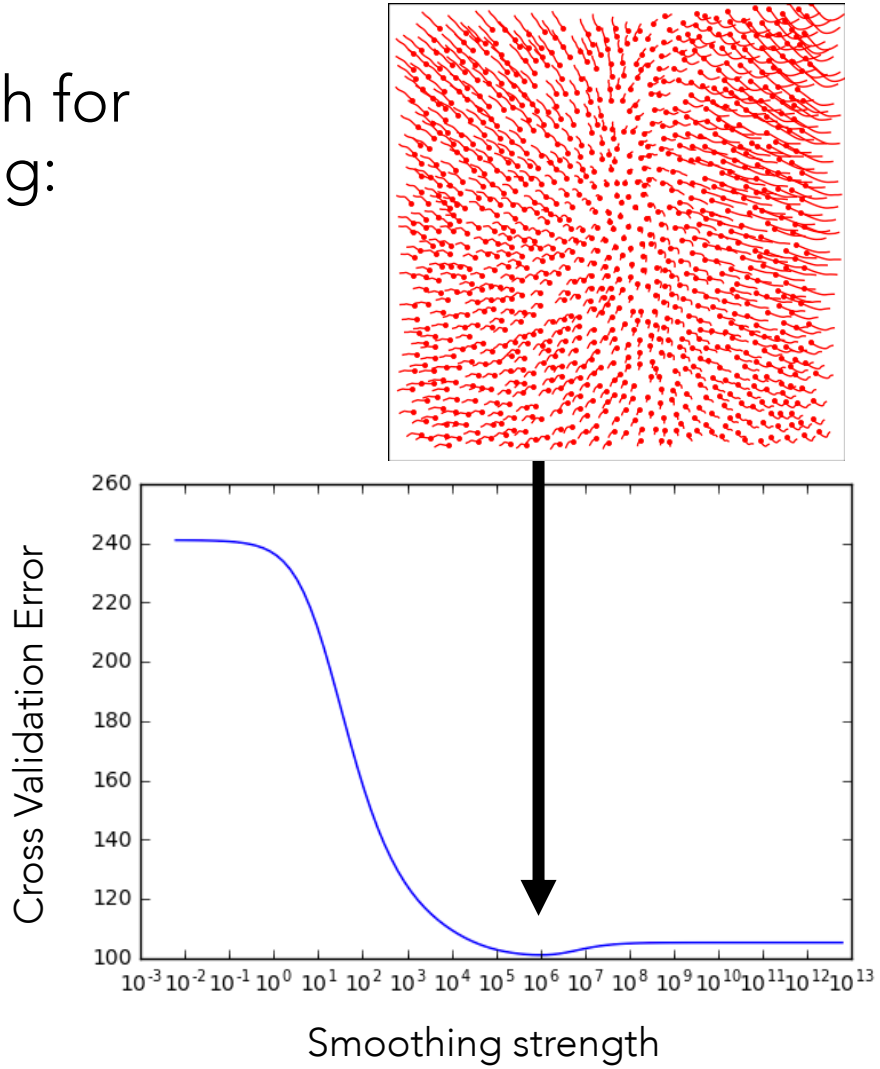
- Maximize cross-correlation between aligned average and frames
- Optimize over trajectories with gradient based quasi-Newton optimizer (LMBFGS)
- Apply a smoothness penalty to trajectory



Motion correction

How much smoothing?

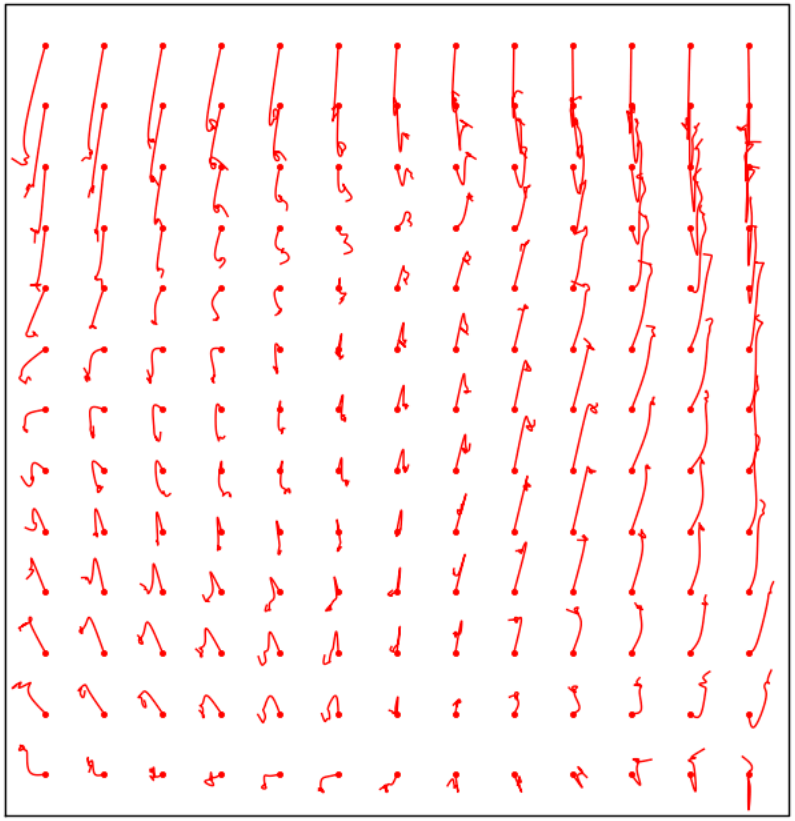
- New cross-validation approach for data-driven optimal smoothing:
- Optimal Smoothing
  - No tuning parameters
  - 10 seconds per movie



Motion correction

New patch-based Local motion correction

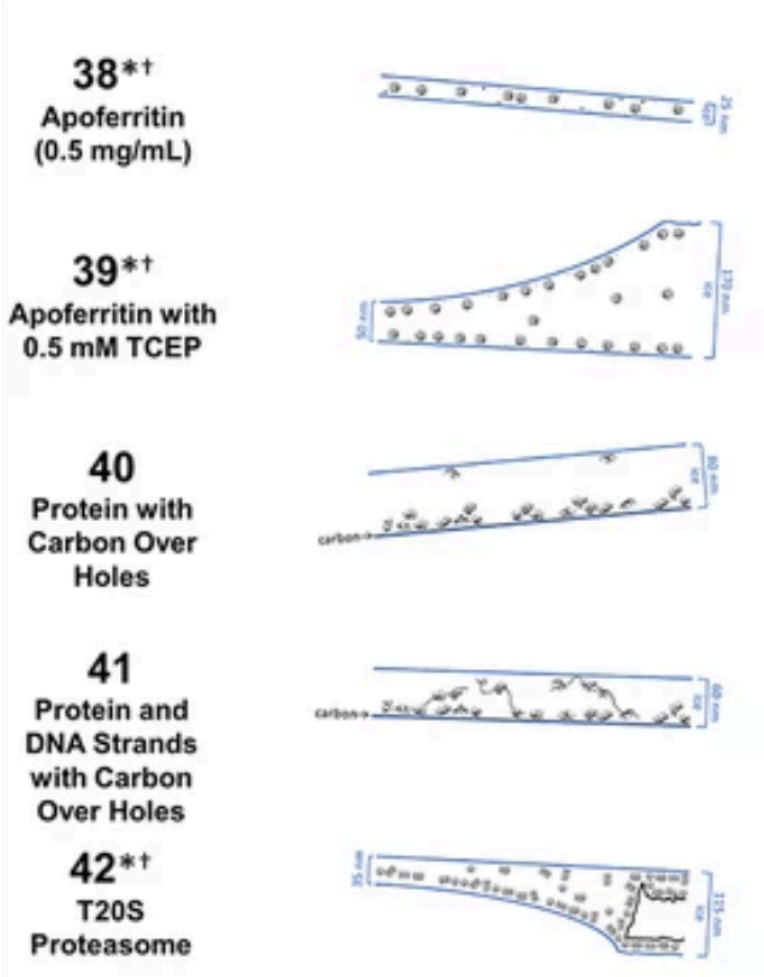
- Global rigid motion and local sample deformation together
- No need for particle positions
- Performs interpolation to create motion corrected micrographs
- Dose weighting



CTF Estimation

New Patch-based Local CTF Estimation

- Fast GPU Implementation
- Automatically estimates defocus variation for tilted, bent, deformed samples
- Accurate for all particle sizes and type including flexible and membrane proteins

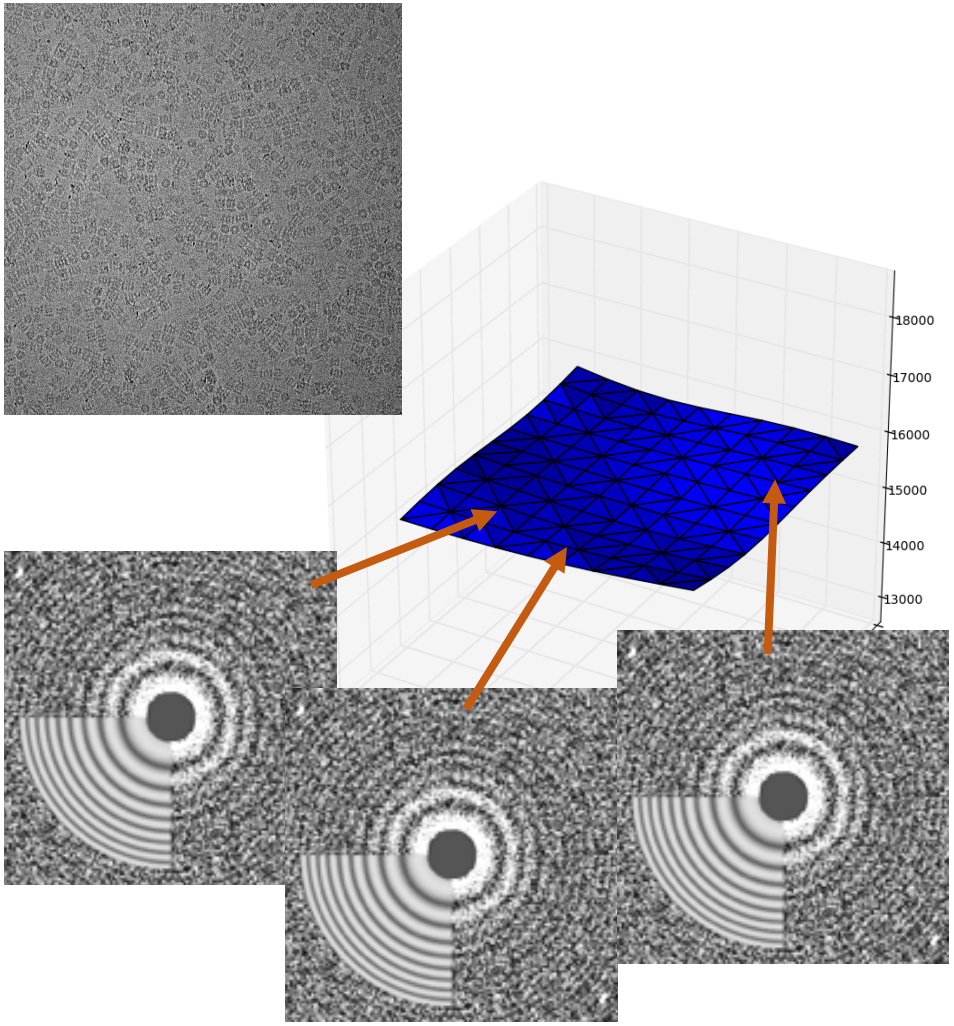


Noble et al eLife 2018

CTF Estimation

New Patch-based Local CTF Estimation

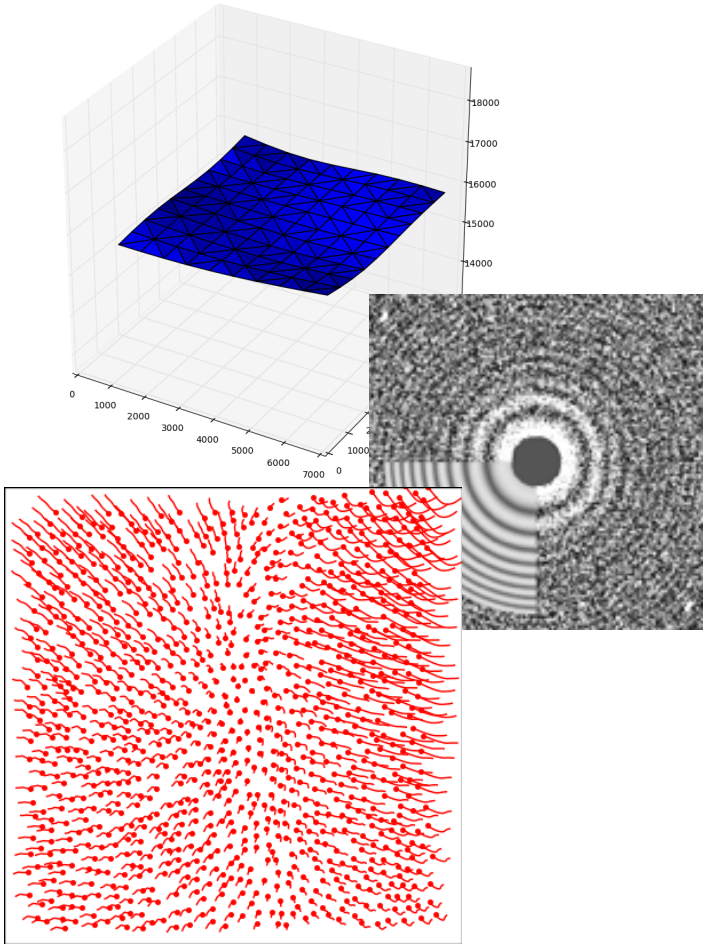
- Fast GPU Implementation
- Automatically estimates defocus variation for tilted, bent, deformed samples
- Accurate for all particle sizes and type including flexible and membrane proteins



Exposure curation

Often tedious

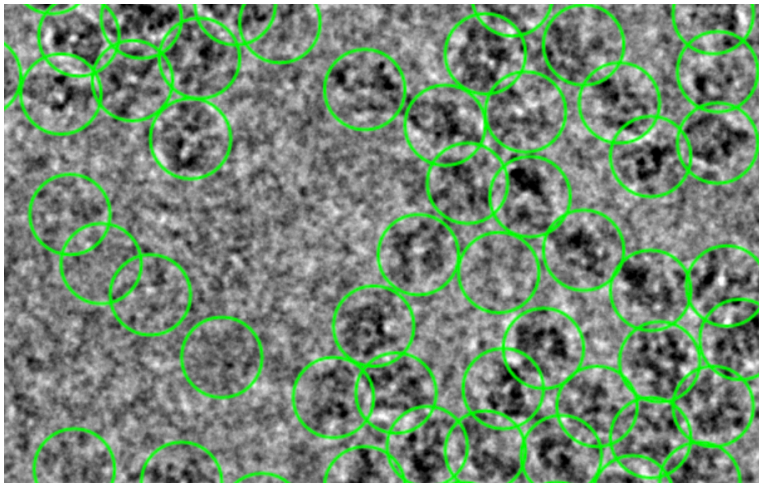
- Based primarily on motion trajectories and CTF fit quality scores
- Looking for thinnest ice, least motion
- Exposure curation tool

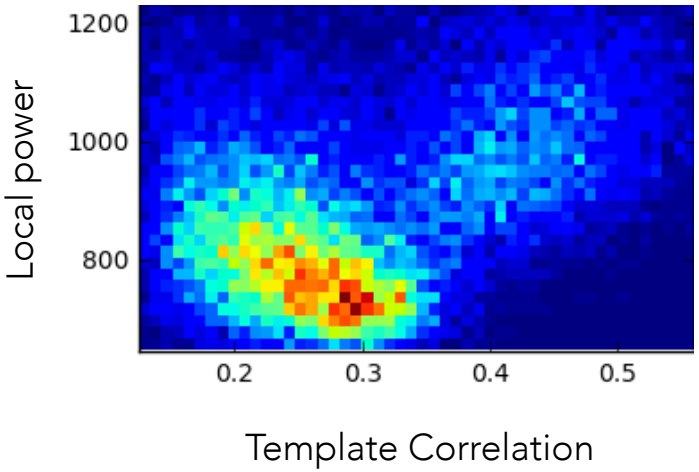


Particle Picking

Template or blob based picking

- Correlation-based scoring
  - Template bias (Einstein-from noise)
- Local power score
  - Unbiased measure of signal content
- GPU Implementation:
  - < 1 second per micrograph
- Unbiased blob-based particle picker
  - Circular and elliptical blobs

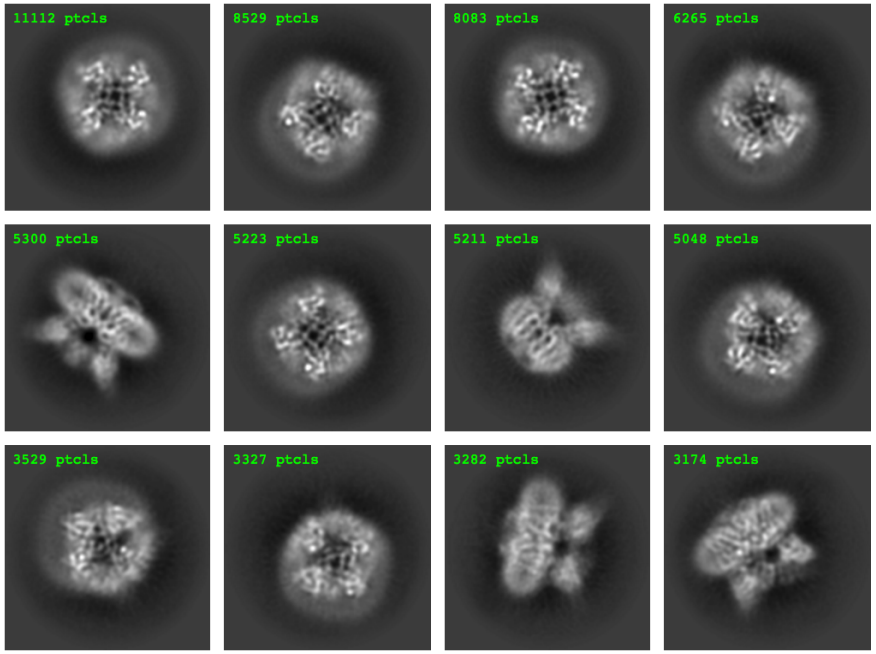




Fast 2D Classification

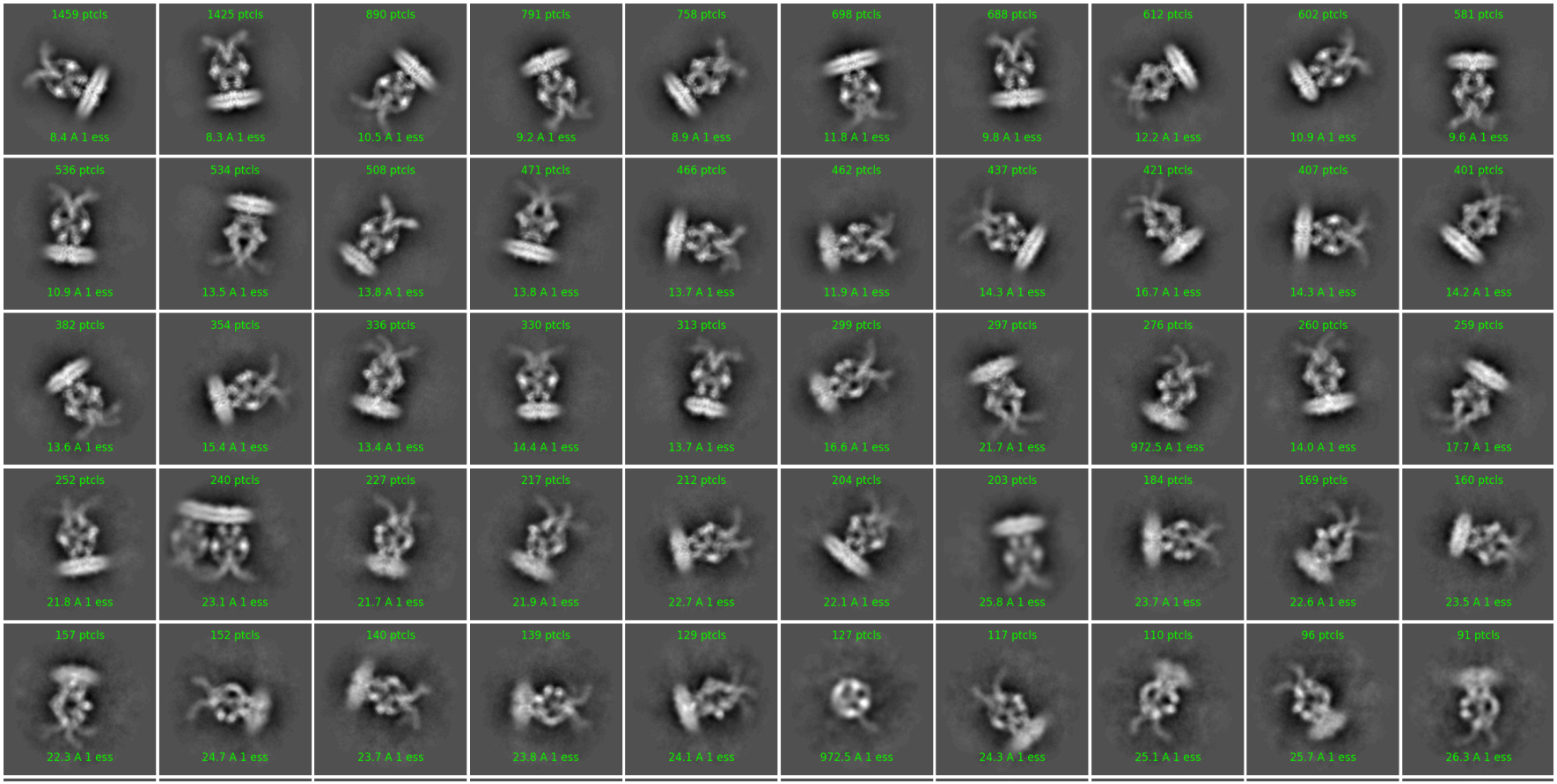
Inspect and curate massive datasets

- Remove junk particles and improve homogeneity of particle stacks
- Scales well with dataset size
- Online-EM algorithm



Fast 2D Classification

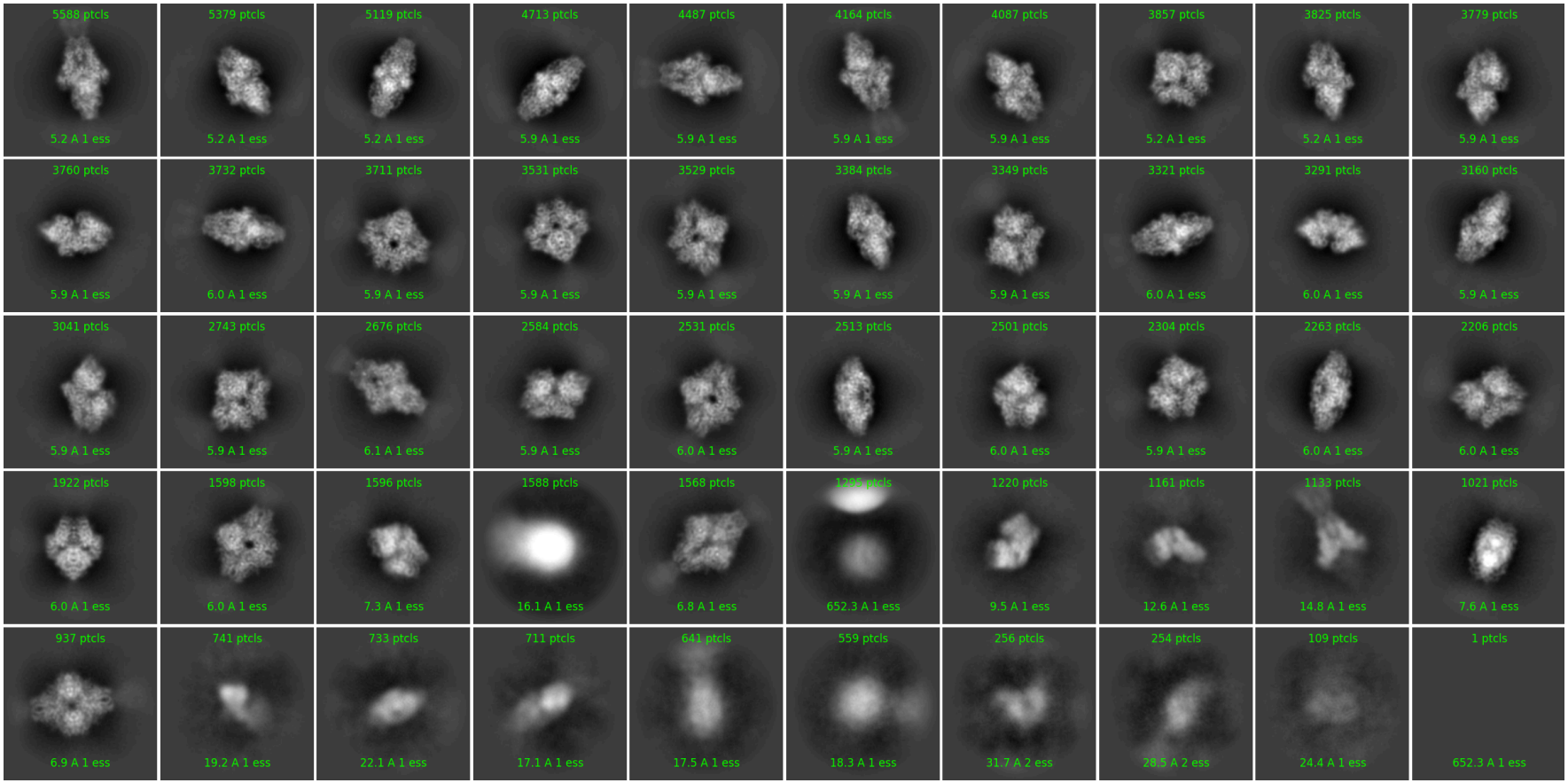
125,000 particles with 50 classes in 12 minutes on 1x GPU



Fast 2D Classification

125,000 particles with 50 classes in 12 minutes on 1x GPU

1,000,000 particles with 50 classes in 2 hours on 1x GPU

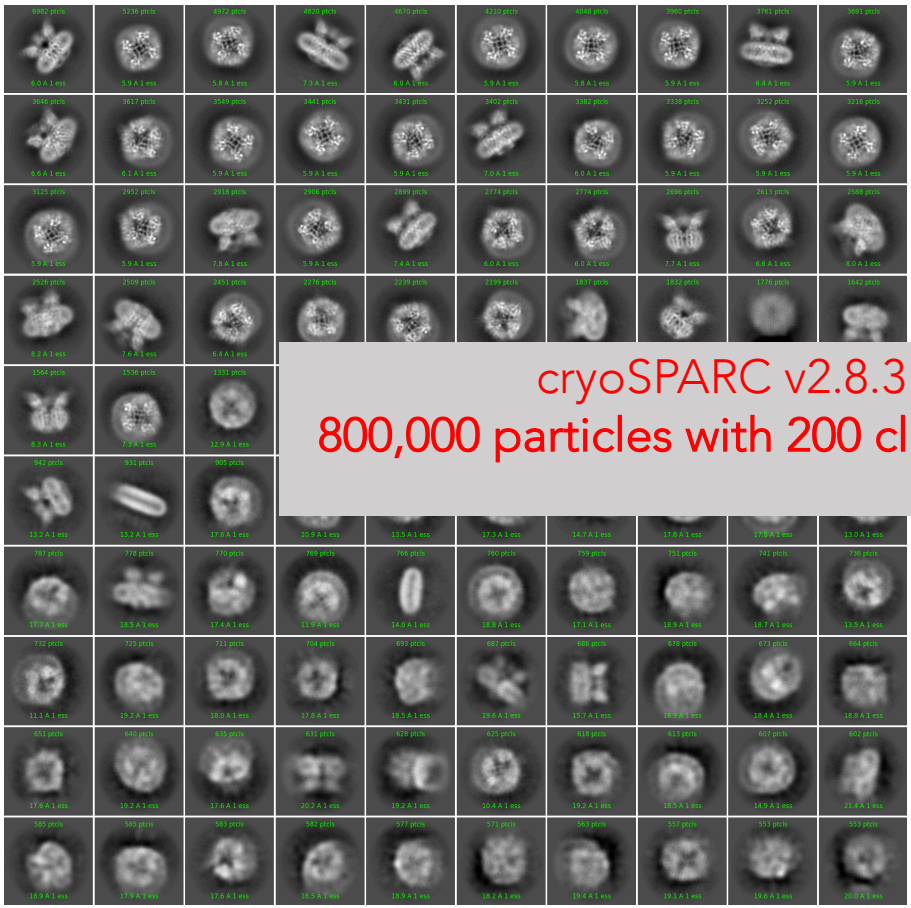


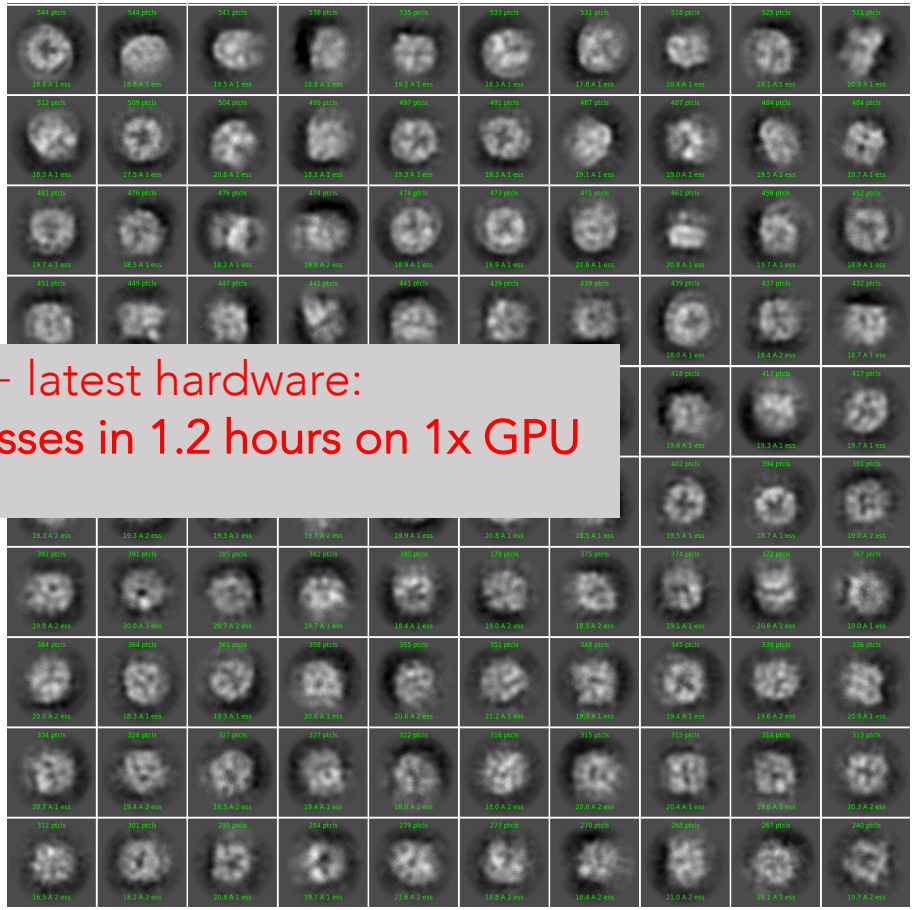
Fast 2D Classification

125,000 particles with 50 classes in 12 minutes on 1x GPU

1,000,000 particles with 50 classes in 2 hours on 1x GPU

200,000 particles with 200 classes in 2 hours on 1x GPU





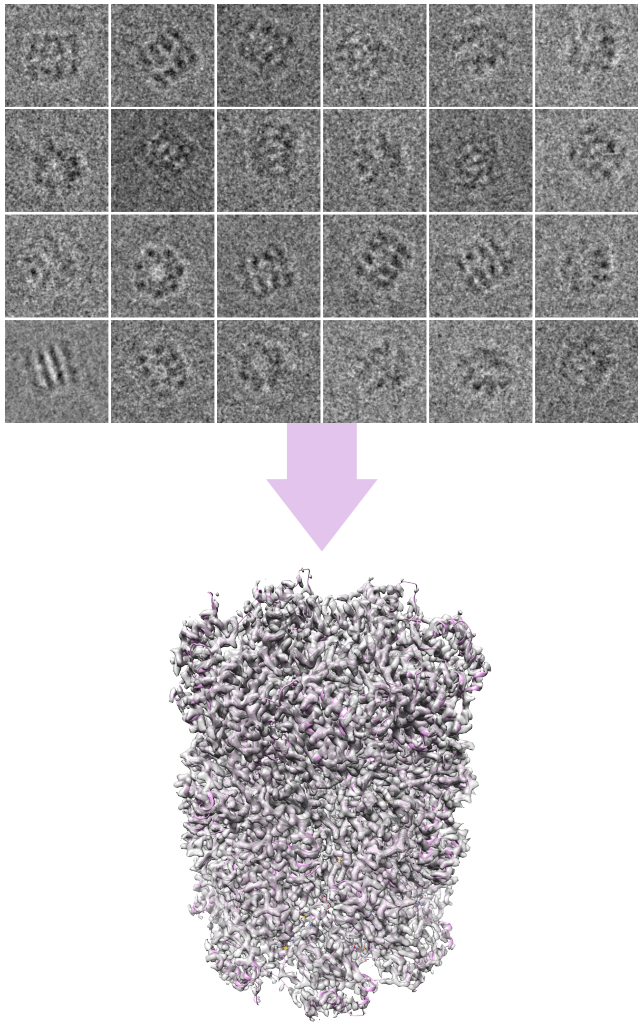
cryoSPARC v2.8.3 + latest hardware:

800,000 particles with 200 classes in 1.2 hours on 1x GPU

SPA 3D Reconstruction

What makes the problem difficult

- Unknown pose of each particle
  - 3D Orientation + 2D Shift
- High noise level
  - Irreducible due to beam damage
- Many particles (100,000+)
- Corruption by microscope contrast transfer function
- Multiple conformational states or distinct particles
- Flexibility and disorder



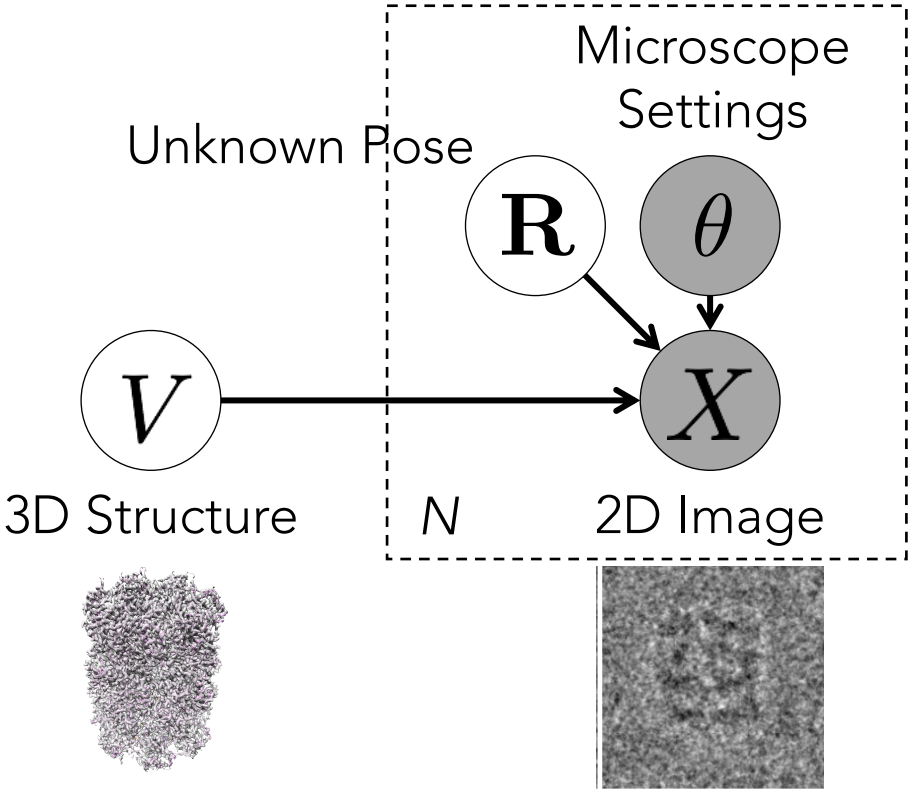
Probabilistic foundations

Probabilistic graphical model

- Fully "Bayesian" treatment:

$$p(V|X_1, X_2, ...X_N)$$

- Meaning we want to know all the possible 3D structures that could explain the images
- Arrow from *V to X* is the image formation model



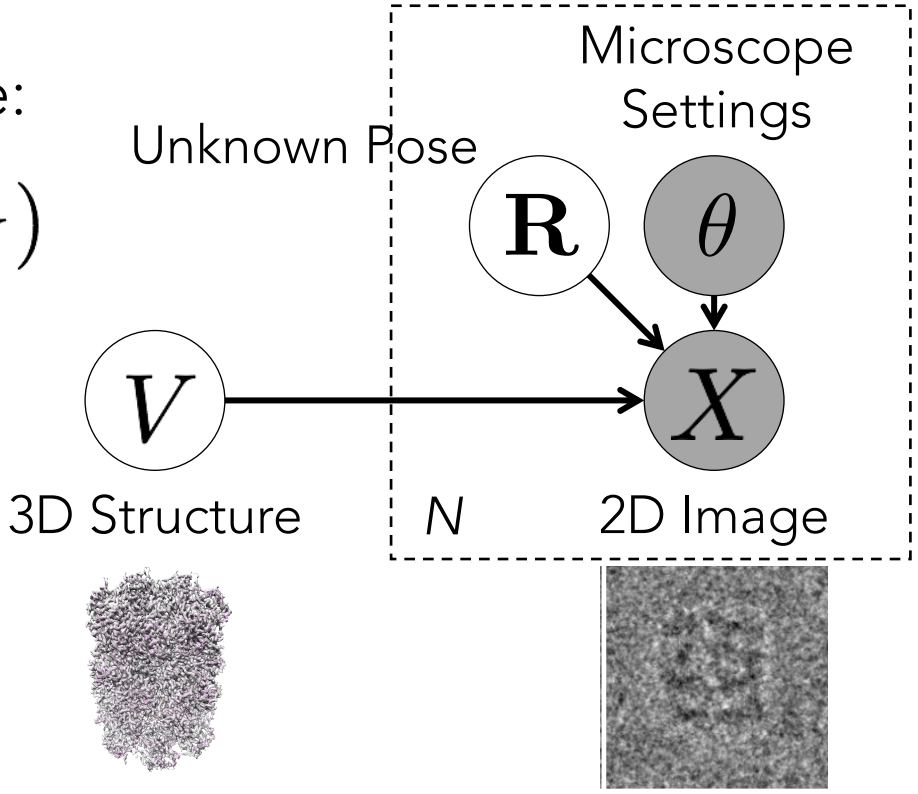
Probabilistic foundations

Making inference computationally feasible

- Maximum probability estimate:

$$\max_{V} p(V|X_1, X_2, ...X_N)$$

- Now only looking for the single estimate of V that best explains the data
- Different choices of V yield all SPA techniques



Probabilistic foundations

Various choices of *V*

- Homogeneous refinement:

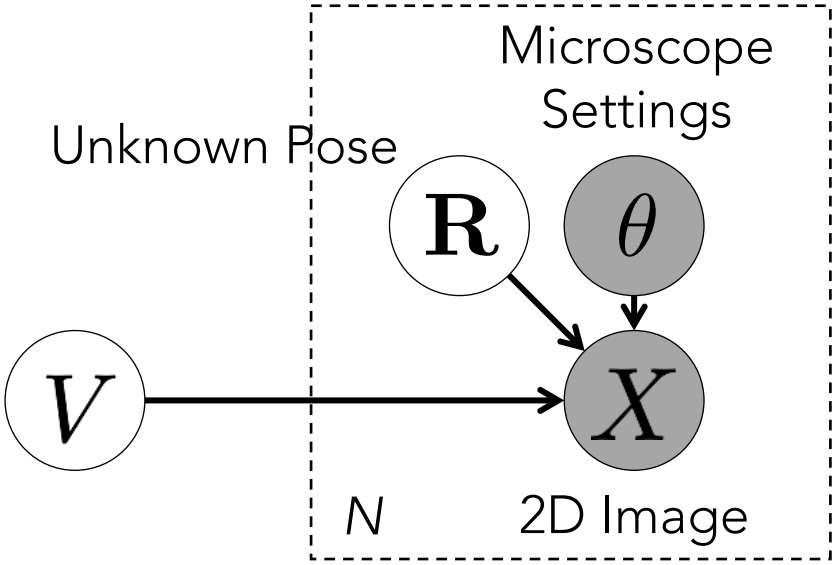
$$V = V$$

- Discrete Heterogeneity:

$$V = V_k | k = C_i$$

- Local refinement/multiple rigid bodies:

$$V = V_1 + R_{rel}(V_2)$$



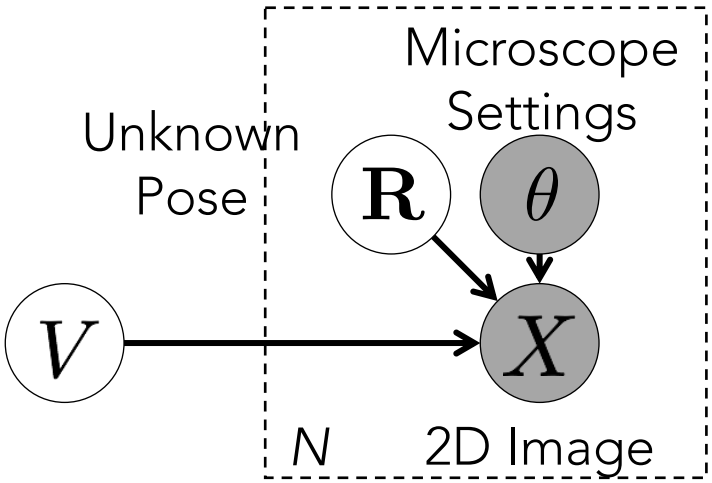
- In all cases, need to solve an <u>optimization problem:</u>

$$\max_V p(V|X_1, X_2, ...X_N)$$

Probabilistic foundations

Likelihood function

- Likelihood of observing an image given *V* depends on noise, CTF, etc
- Marginalization or maximization over poses *R*



$$\max_V p(V|X_{i=1..N}) = \max_V \prod_i^N p(X_i|V) \cdot p(V)$$

$$= \max_V \left(\prod_i^N \int_R p(X_i|V,R) dR \right) \cdot p(V)$$

$$\approx \max_V \max_{R_i} \prod_i^N p(X_i|V,R_i) \cdot p(V)$$

Optimization problems

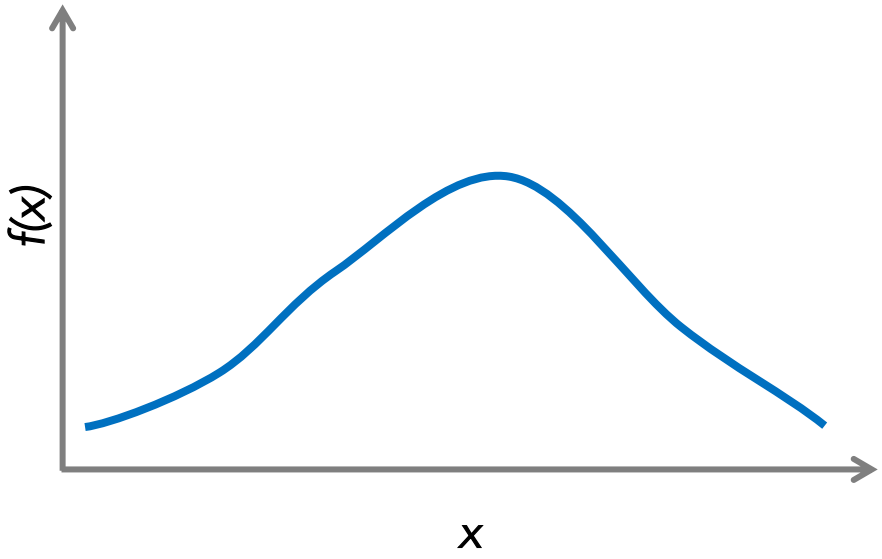
Mathematical properties

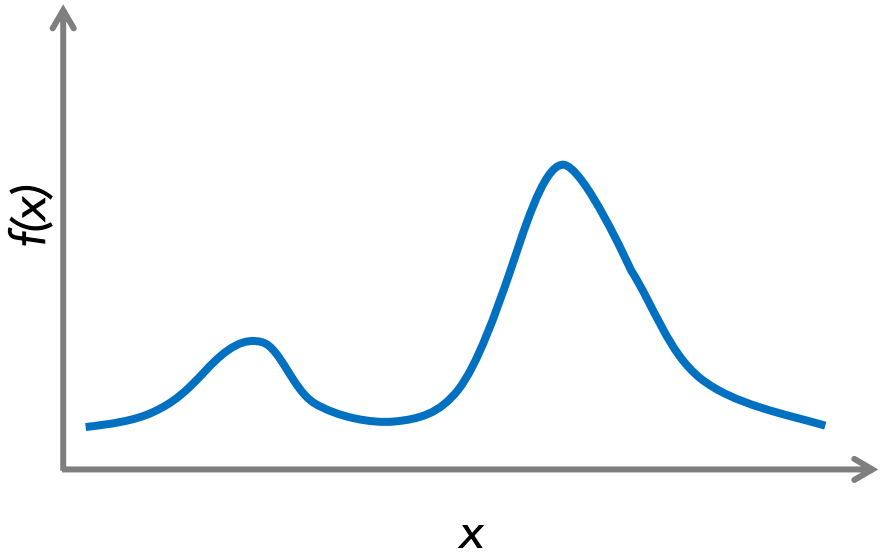
$$\max_x f(x)$$

- Optimization is a very well studied topic in mathematics
- Almost all other computational problems in EM are optimization problems:
  - Motion correction
  - CTF estimation
  - Model building
- Major characteristics of optimization problems are key

Optimization problems

Property: Convexity



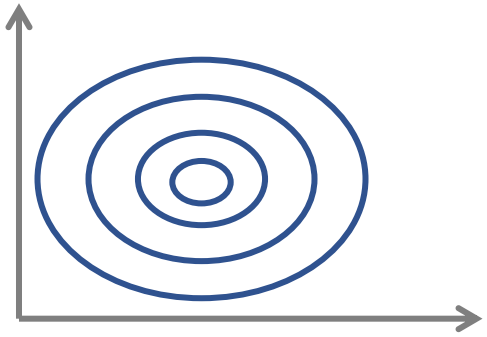


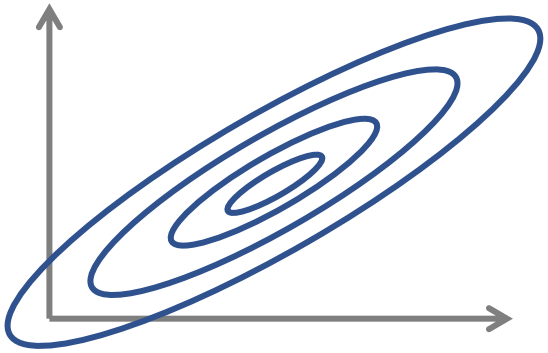
- Single global optimum
- Guaranteed unique solution
- Generally inexpensive to solve

- Multiple local optima
- Need to “explore”
- Generally exponential cost

Optimization problems

Property: Conditioning





- Least coupling of variables
- Simple to solve

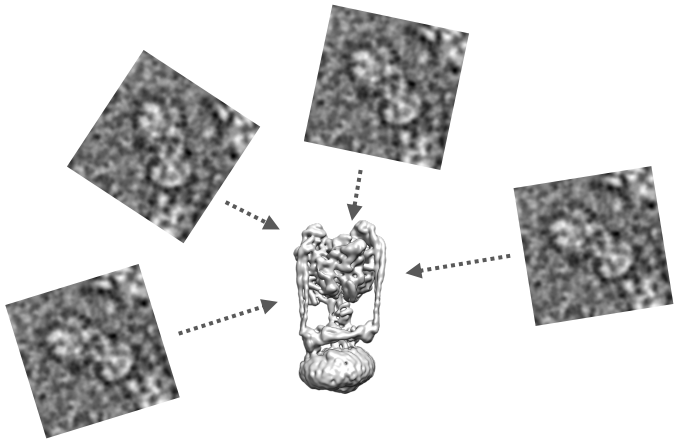
- Highly coupled variables
- Difficult to solve

Optimization problem for SPA

Homogeneous case

$$\max_{V} \max_{R_i} \prod_{i}^{N} p(X|V, R_i) \cdot p(V)$$

- Globally Non-convex:
  - multiple local optima
- Poor-conditioning:
  - small change in angle changes the structure, and change in structure changes angles
- Locally convex and well-conditioned
  - Once near enough to a local optimum



Optimization algorithms

Coordinate Ascent

$$\max_{V} \max_{R_i} \prod_{i}^{N} p(X|V, R_i) \cdot p(V)$$

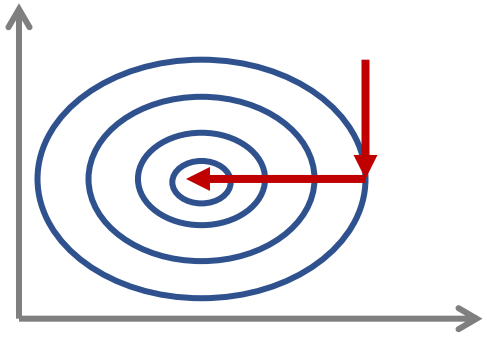
- Also known as iterative refinement, expectation-maximization
  - Hold first variable fixed, optimize the second
  - Hold second variable fixed, optimize the first
- Provably guaranteed to converge:
  - Globally for a convex problem
  - To the nearest local optimum for a non-convex problem
- Convergence rate:
  - Very fast for well conditioned problems
  - Very, very slow for poor condition

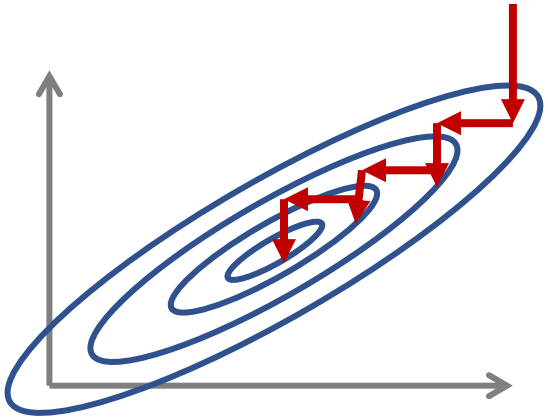
Optimization algorithms

Coordinate Ascent

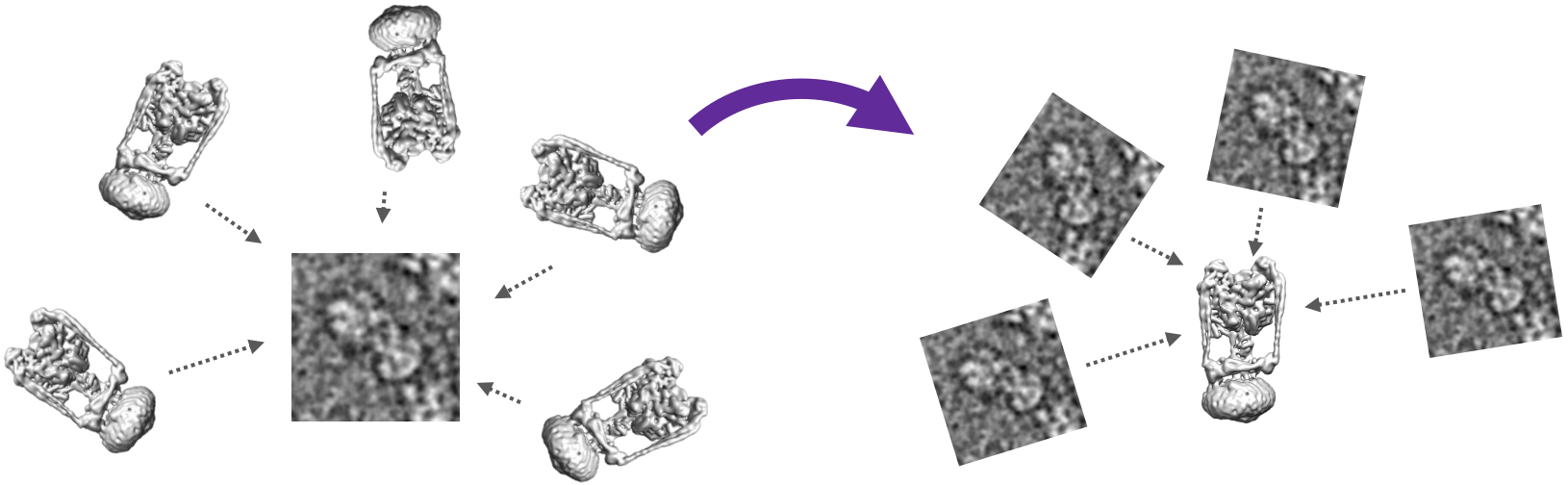
$$\max_{V} \max_{R_i} \prod_{i}^{N} p(X|V, R_i) \cdot p(V)$$

- Hold first variable fixed, optimize the second
- Hold second variable fixed, optimize the first
- Each subproblem becomes simple





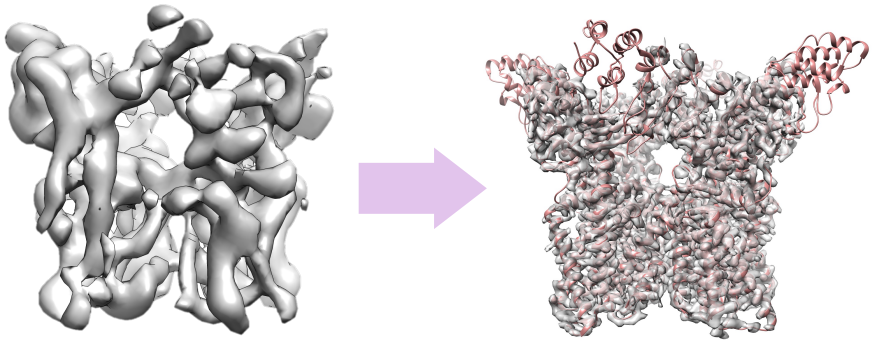
Iterative refinement (Expectation Maximization)



Alignment
(Structure fixed)

Reconstruction
(Alignments fixed)

$$\max_{V} \max_{R_i} \prod_{i}^{N} p(X|V, R_i) \cdot p(V)$$



Optimization algorithms

Coordinate Ascent

- Well suited for optimization of our SPA problem once near the optimal solution
- How to solve without "knowing the answer"?
- How does this change in the more complex case of heterogeneity?

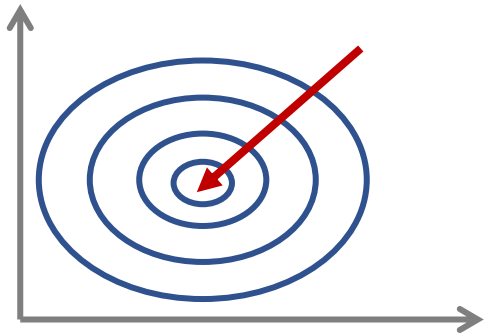
- Consider a different optimization algorithm:
  - Stochastic gradient descent

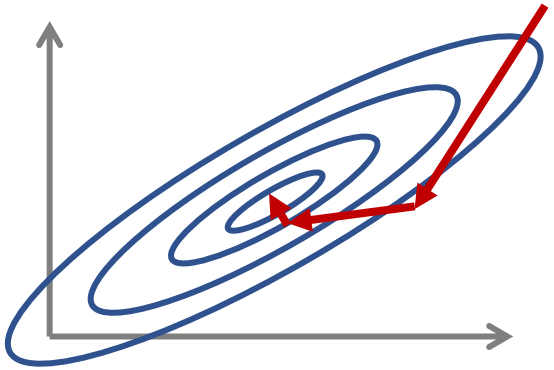
Optimization algorithms

Gradient Descent

- Can take large steps, in multiple dimensions at once
- Gradient is direction of steepest descent:

$$G = \frac{d}{dV}\left(\prod_{i}^{N} p(X|V, R_i) \cdot p(V)\right)$$





Optimization algorithms

Stochastic gradient descent

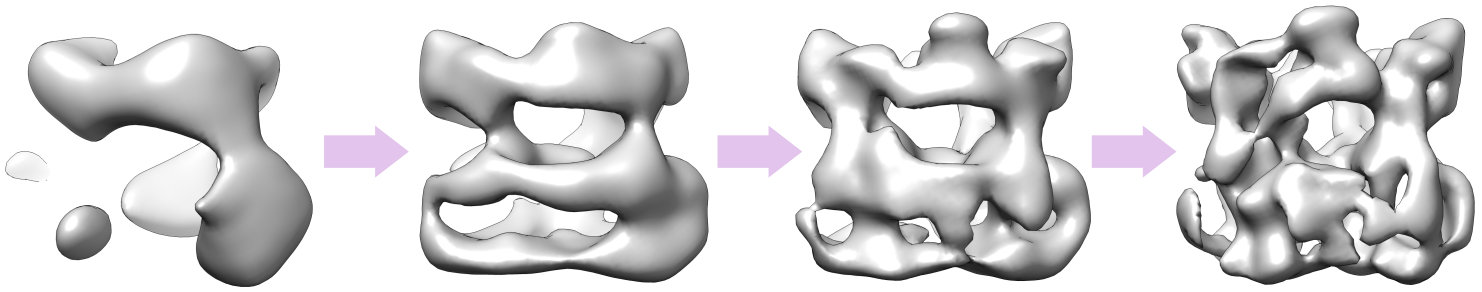
$$\max_V p(V|X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, ...X_N)$$

Iter 1 $\max_V p(V|\quad X_2,\quad X_4, X_5,\quad)$

Iter 2 $\max_V p(V|X_1,\quad X_3,\quad X_7,\quad)$

Iter 3 $\max_V p(V|\quad X_6,\quad X_8,\quad X_N)$

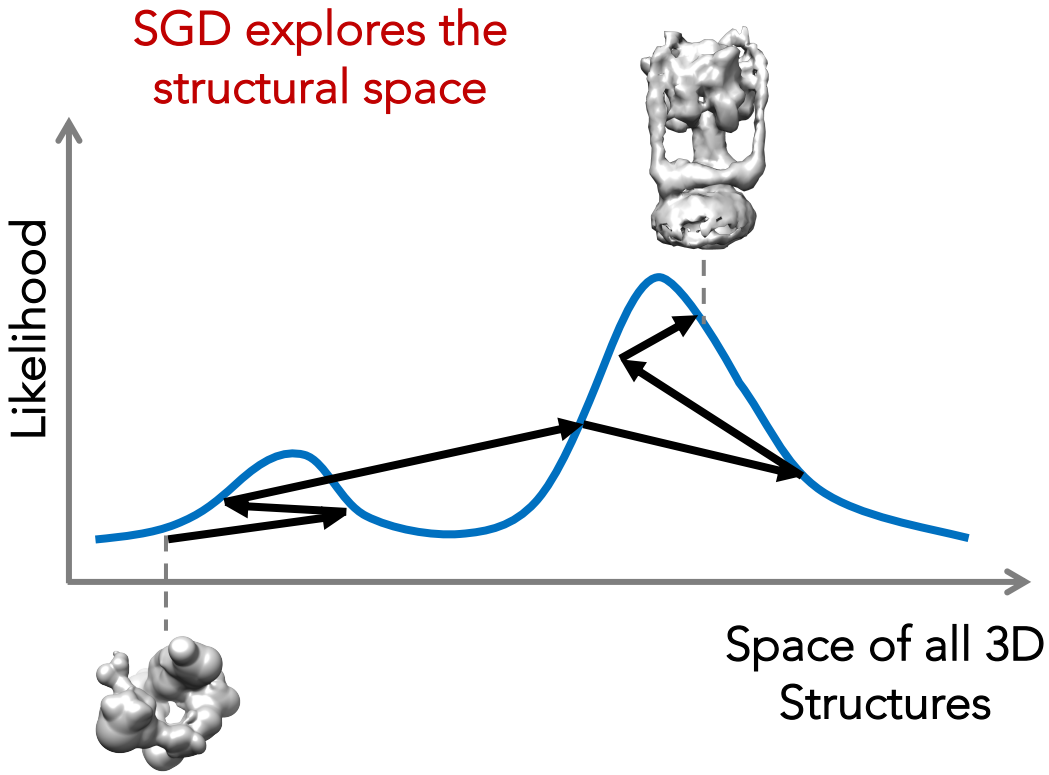
Randomly select small subsets of images at each iteration



Many **noisy** incremental changes

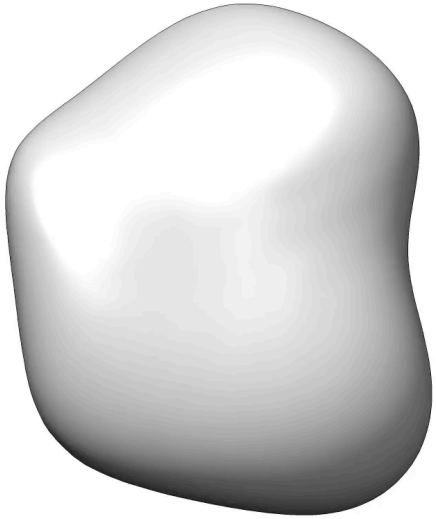
Ab-initio reconstruction: SGD

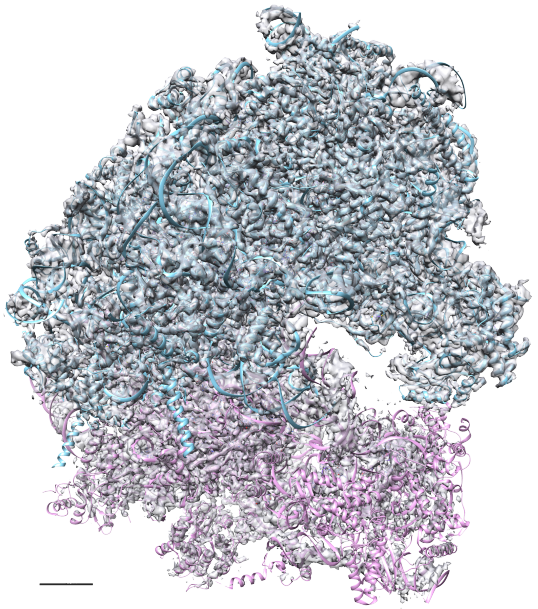
- Stochastic optimization
  - Class of modern statistical methods
- Stochastic Gradient Descent (SGD)
  - Very successful variant
  - Powers modern deep learning



Punjani, Rubinstein, Fleet & Brubaker *Nature Methods* 2017

Ab-initio reconstruction: 80S Ribosome





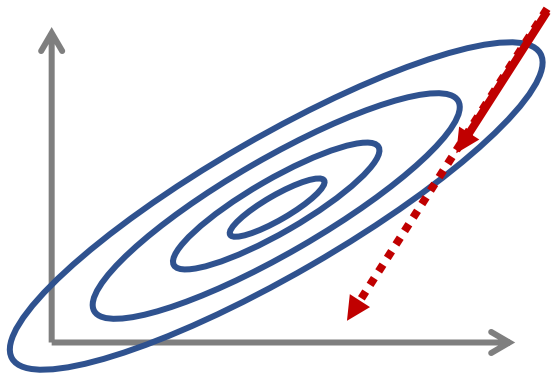
Data from Wong et al. *eLife* 2014

Ab-initio reconstruction: SGD

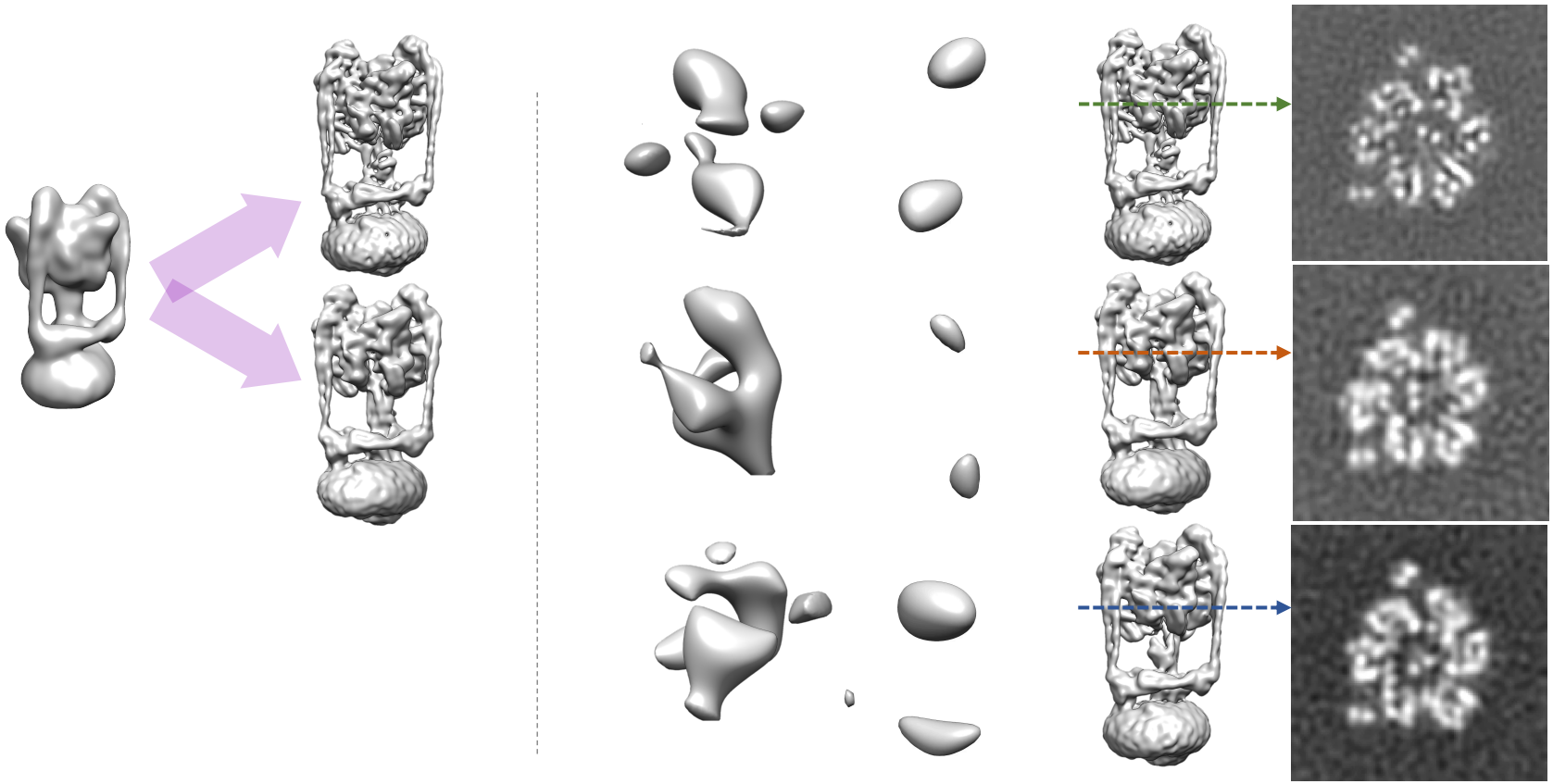
Heterogeneity

Explore vs. Exploit

- Additional classes add more variables and more non-convexity (local optima)
- Plain SGD alone is no longer effective
- Exploration vs. Exploitation
- In SGD controlled by “step size” and “minibatch size”
- In cryoSPARC: “class similarity score”

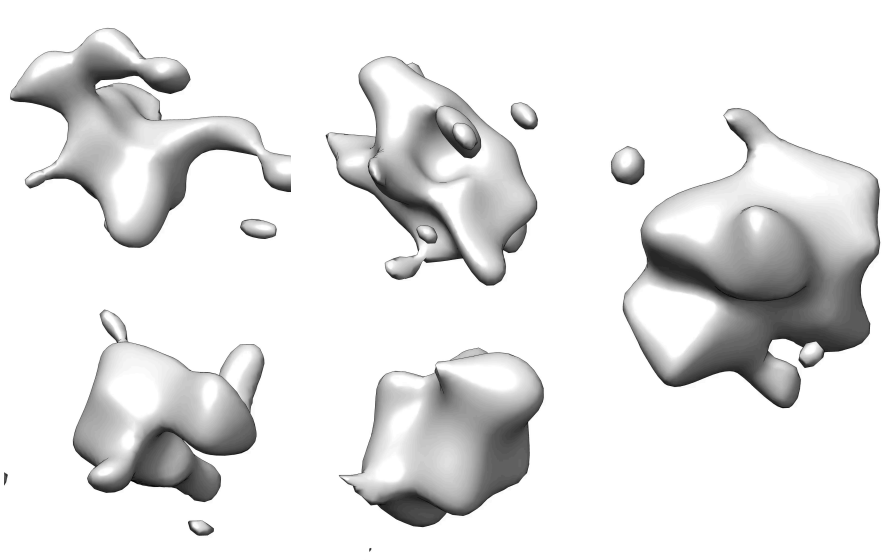


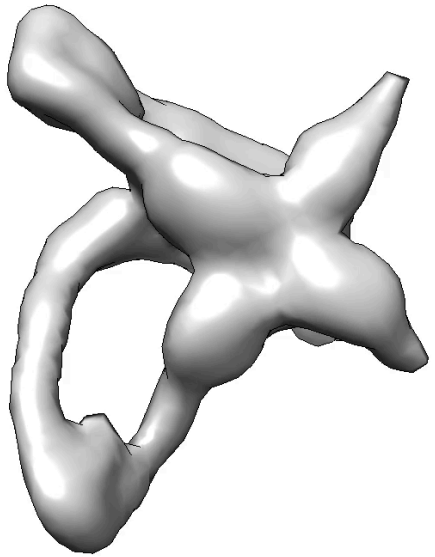
Heterogeneous samples: ATPase



Data from Schep et al. *PNAS* 2016

Heterogeneous samples: Holiday Junction



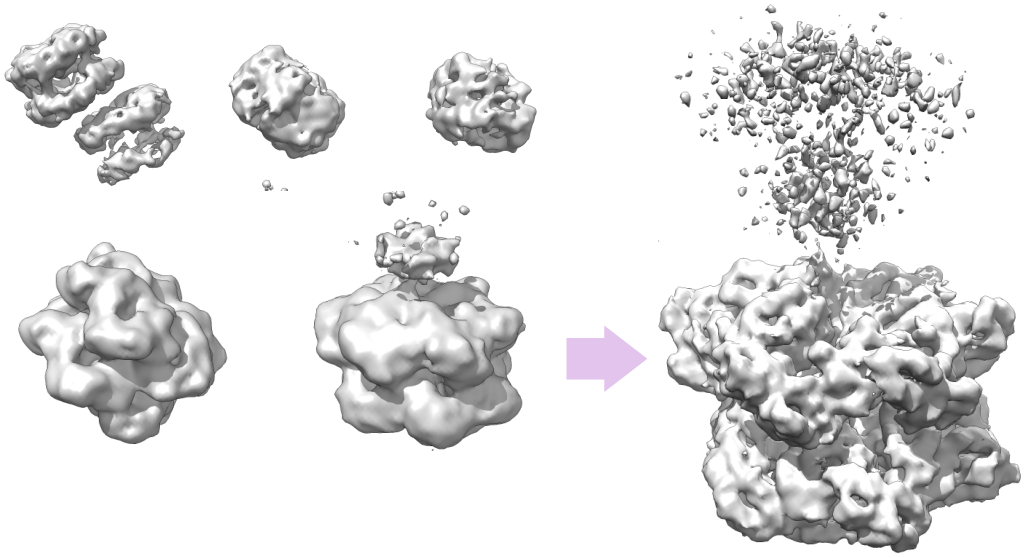


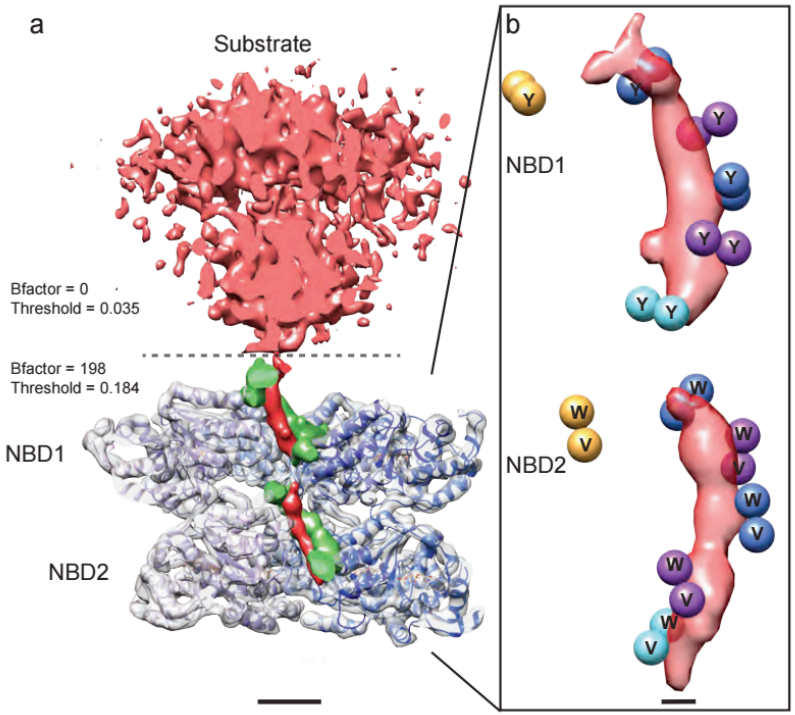
Data from Laxmikanthan et al. *eLife* 2016

Heterogeneous samples: AAA+ Unfoldase

170K particles
5 classes ab-initio

30K particles
4.6A refinement





Data and figures from Ripstein et. al *eLife* 2017

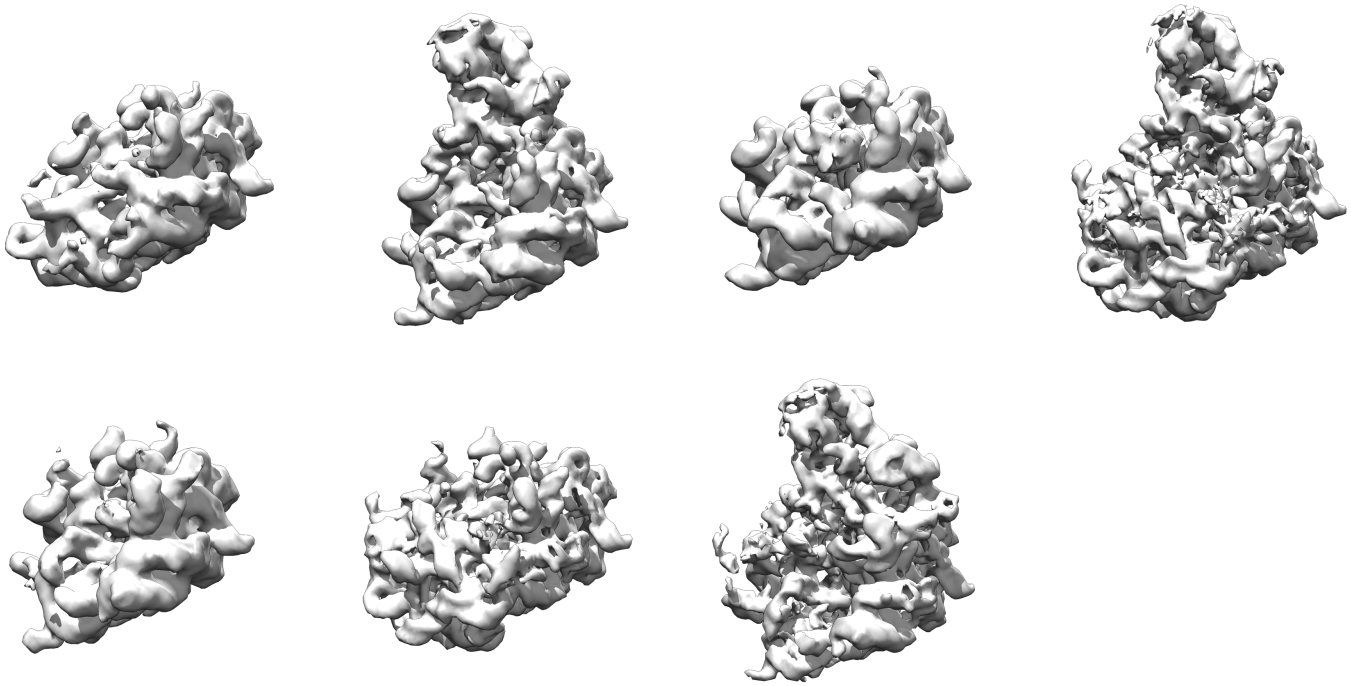
Ab-initio Heterogeneity

Important considerations

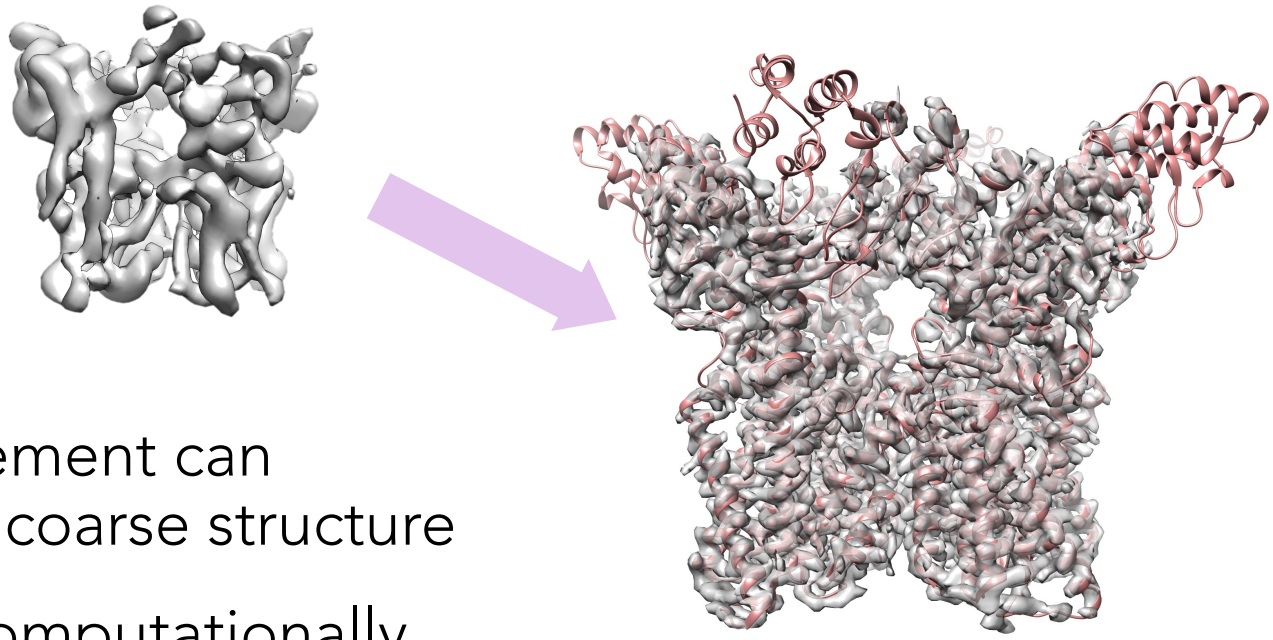
- Cannot fix missing views
- Too few classes: average of structures
- Too many classes: 3D classes start to just become 2D and have only a single view
- Reproducibility: can run multiple times each time will use a different random seed
- Symmetry can cause issues if enforced

Multi-reference Refinement (3D classification)

- One or more starting models (from ab-initio reconstruction)

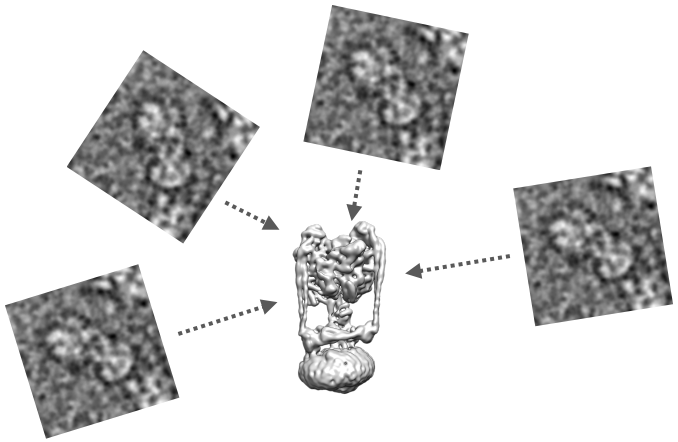


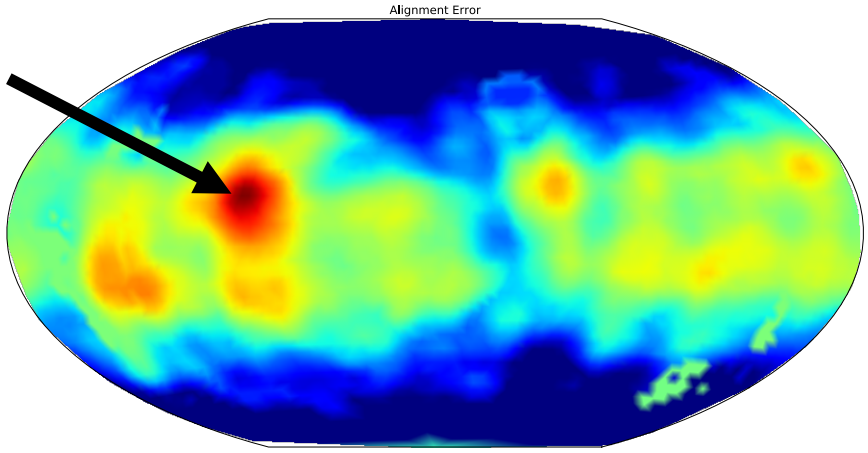
High resolution refinement



- Iterative refinement can proceed from coarse structure
- Usually very computationally expensive
- Many refinements in a complex workflow

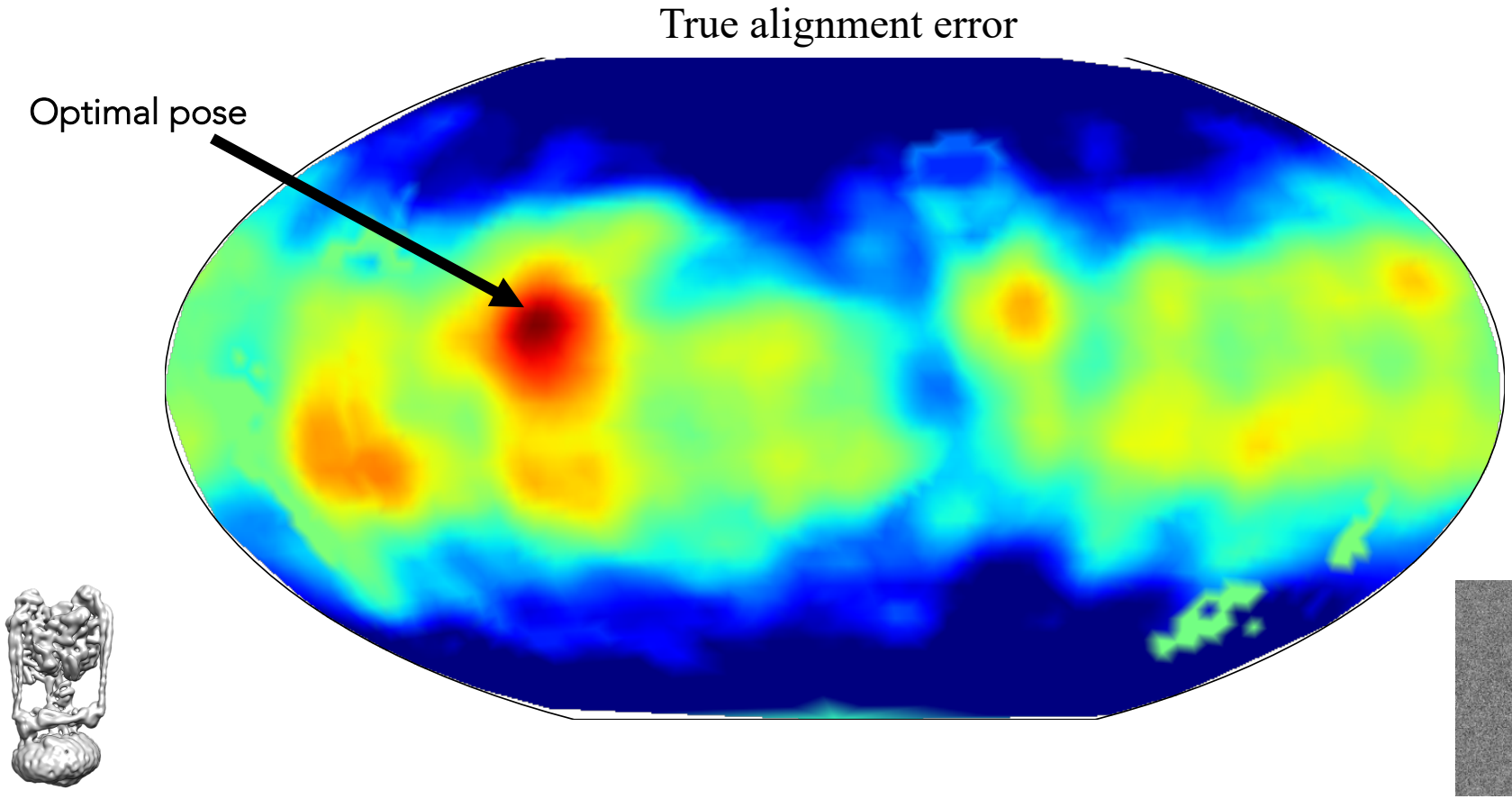
2D-to-3D image alignment

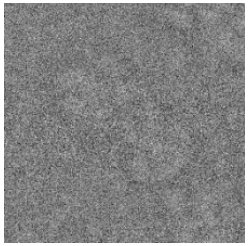




- 5D pose search is expensive, for every image
- Existing techniques search exhaustively or locally
- Branch-and-bound reduces computational expense

Branch and bound method

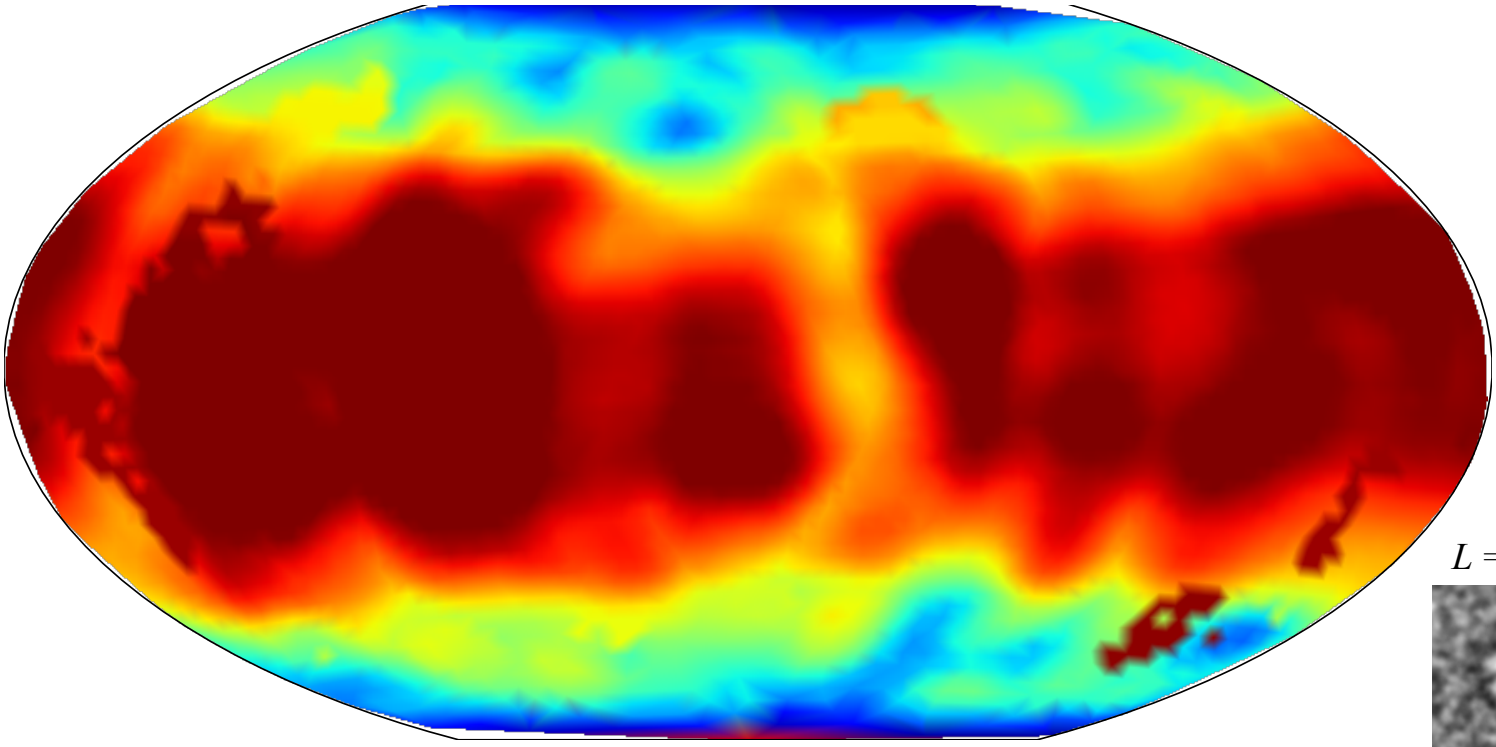


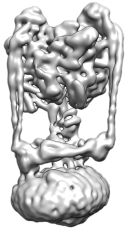


Punjani, Rubinstein, Fleet & Brubaker *Nature Methods* 2017

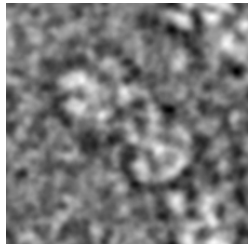
Branch and bound method

Bound on error





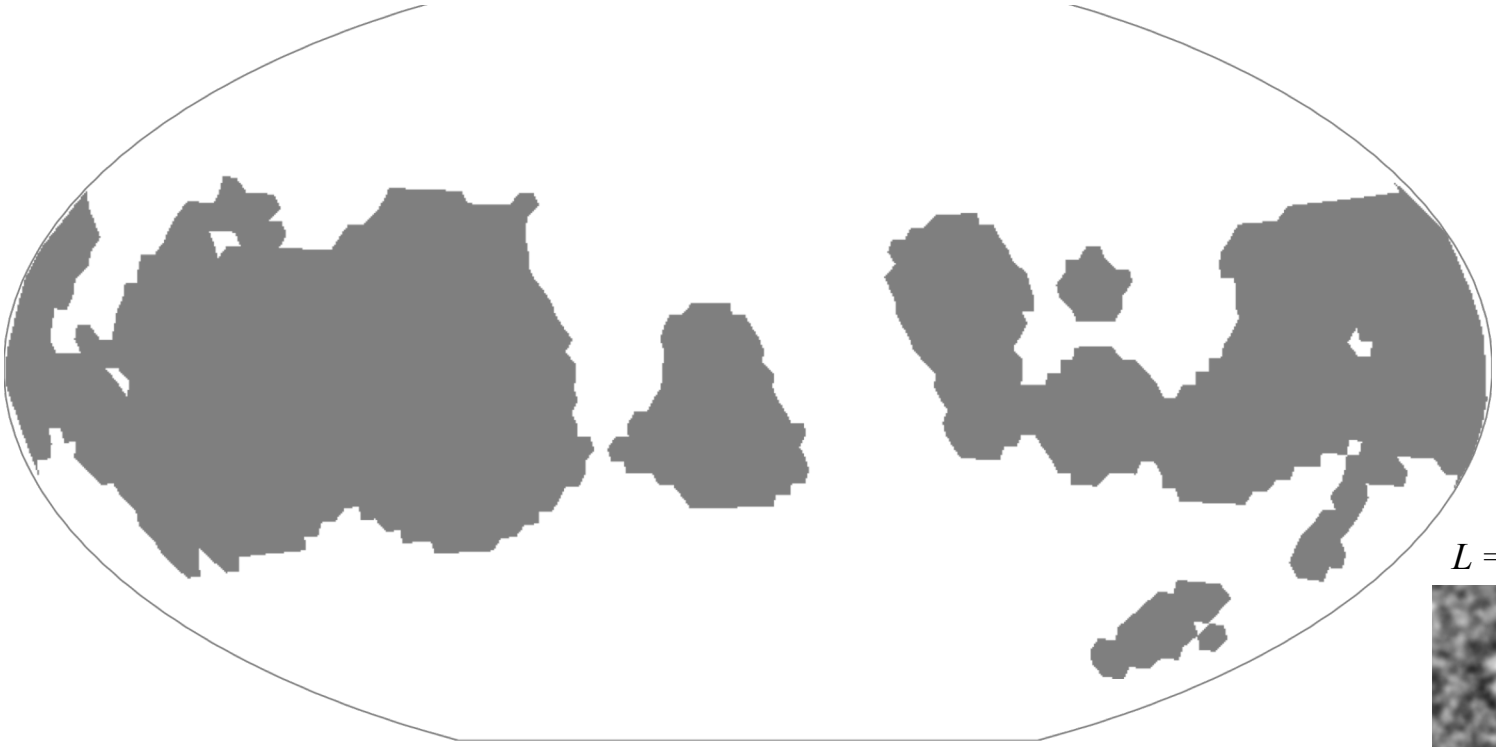
$L = 15$ (24Å)

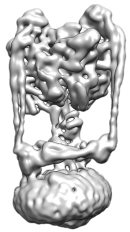


Punjani, Rubinstein, Fleet & Brubaker *Nature Methods* 2017

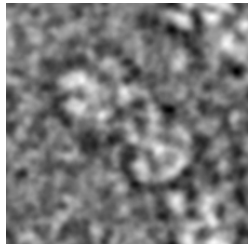
Branch and bound method

Remaining candidates at $L = 10$ (37Å)





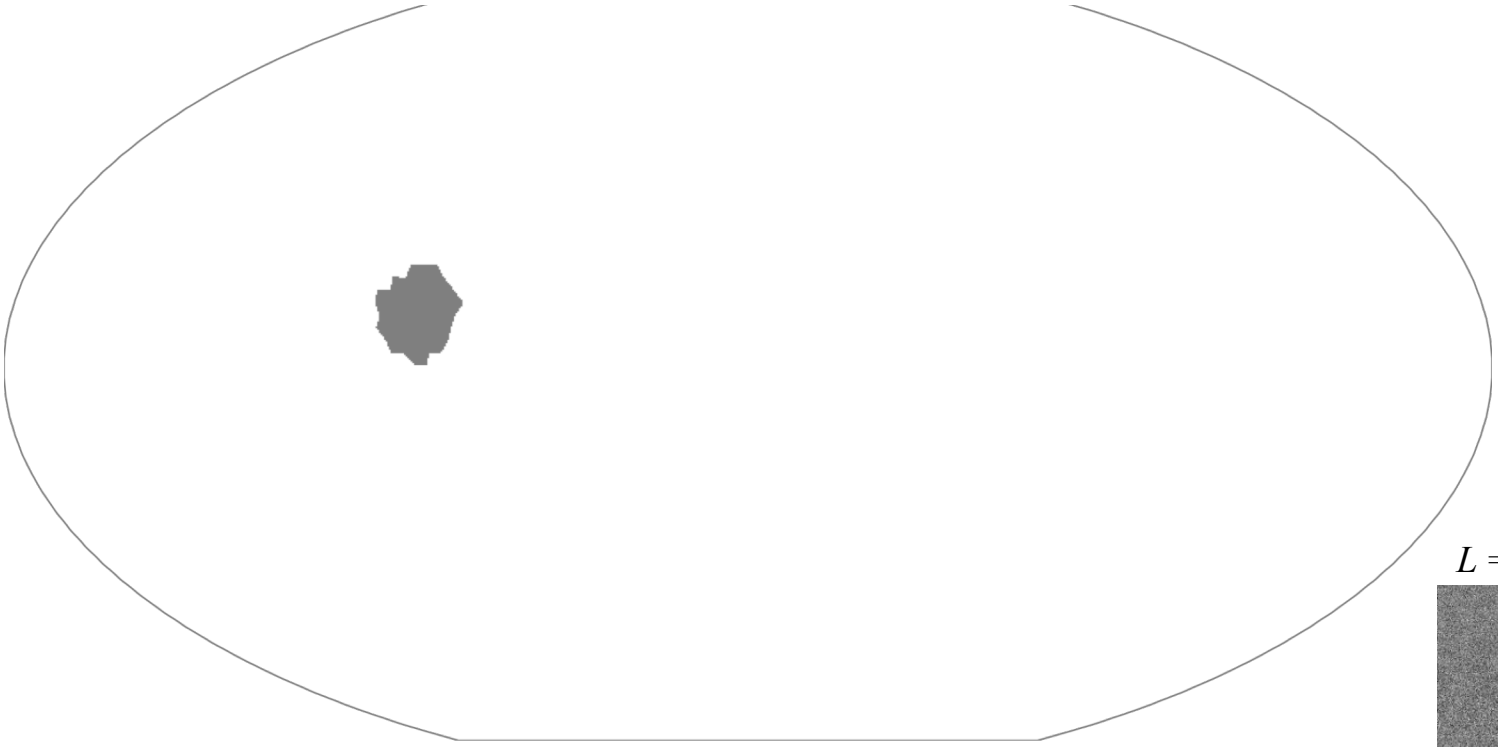
$L = 15$ (24Å)

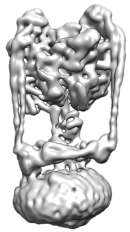


Punjani, Rubinstein, Fleet & Brubaker *Nature Methods* 2017

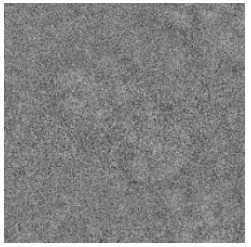
Branch and bound method

Remaining candidates at L = 15 (24Å)



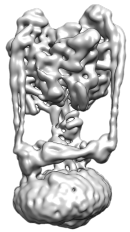


L = 20 (14Å)



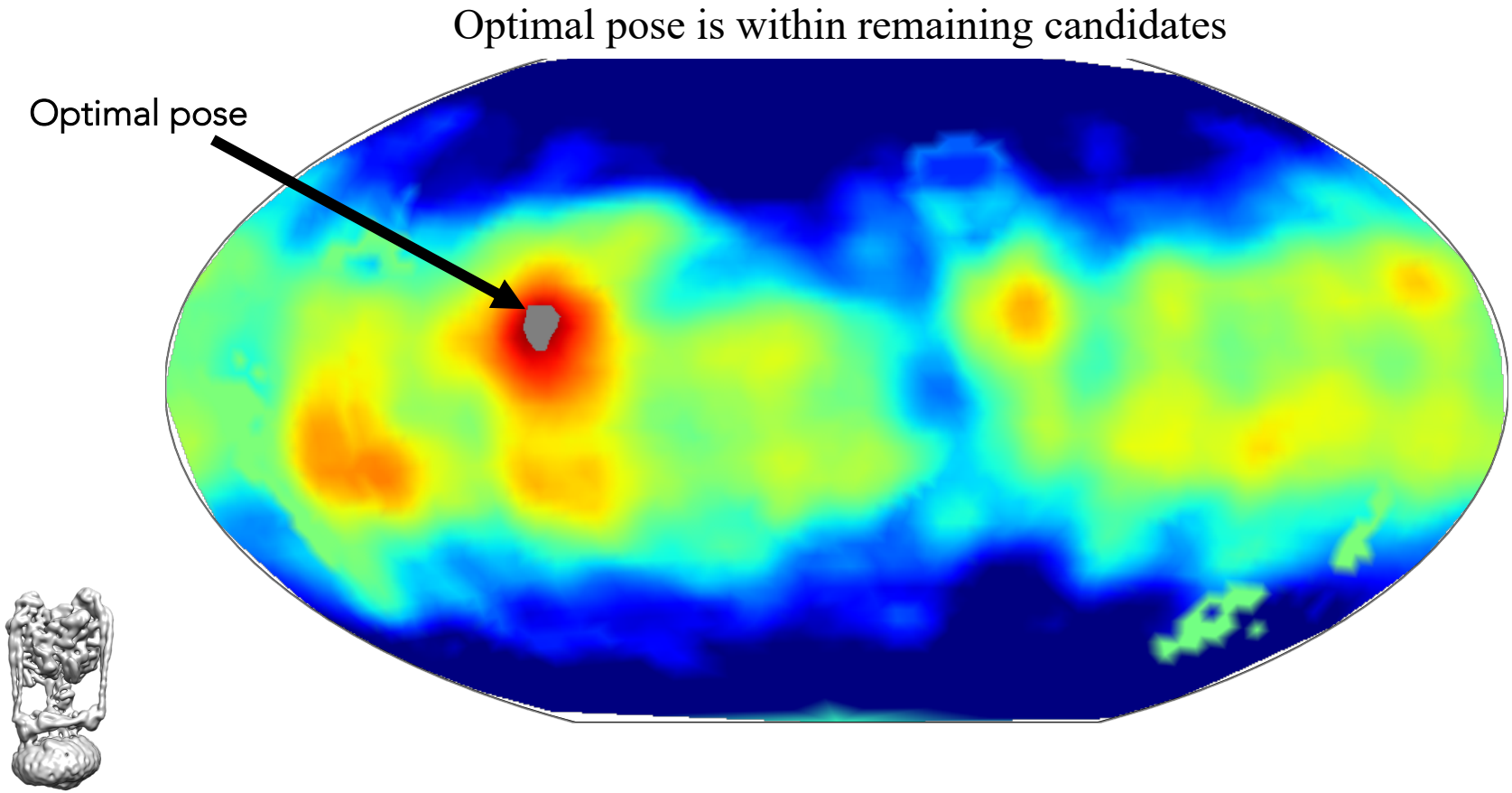
Punjani, Rubinstein, Fleet & Brubaker *Nature Methods* 2017

Branch and bound method



Punjani, Rubinstein, Fleet & Brubaker *Nature Methods* 2017

Branch and bound method



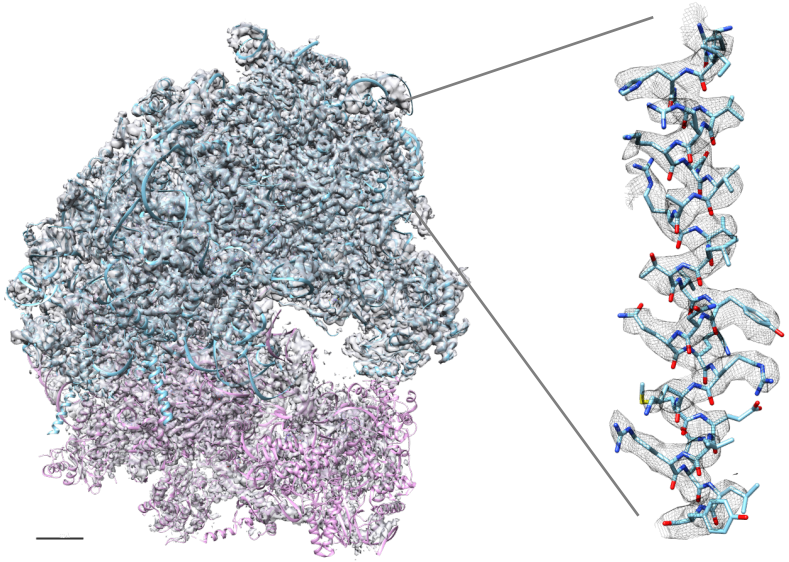
Punjani, Rubinstein, Fleet & Brubaker *Nature Methods* 2017

High-speed high-resolution refinement

- Iterative refinement is very expensive – multiple rounds of alignment and reconstruction required
- Branch and bound (BnB) algorithm drastically reduces computation required without any loss in quality
- **High-resolution structures in minutes**
- **Refinements are repeated dozens of times in a real cryoEM workflow**

80S ribosome dataset:
105,000 particles refined to 3.2Å

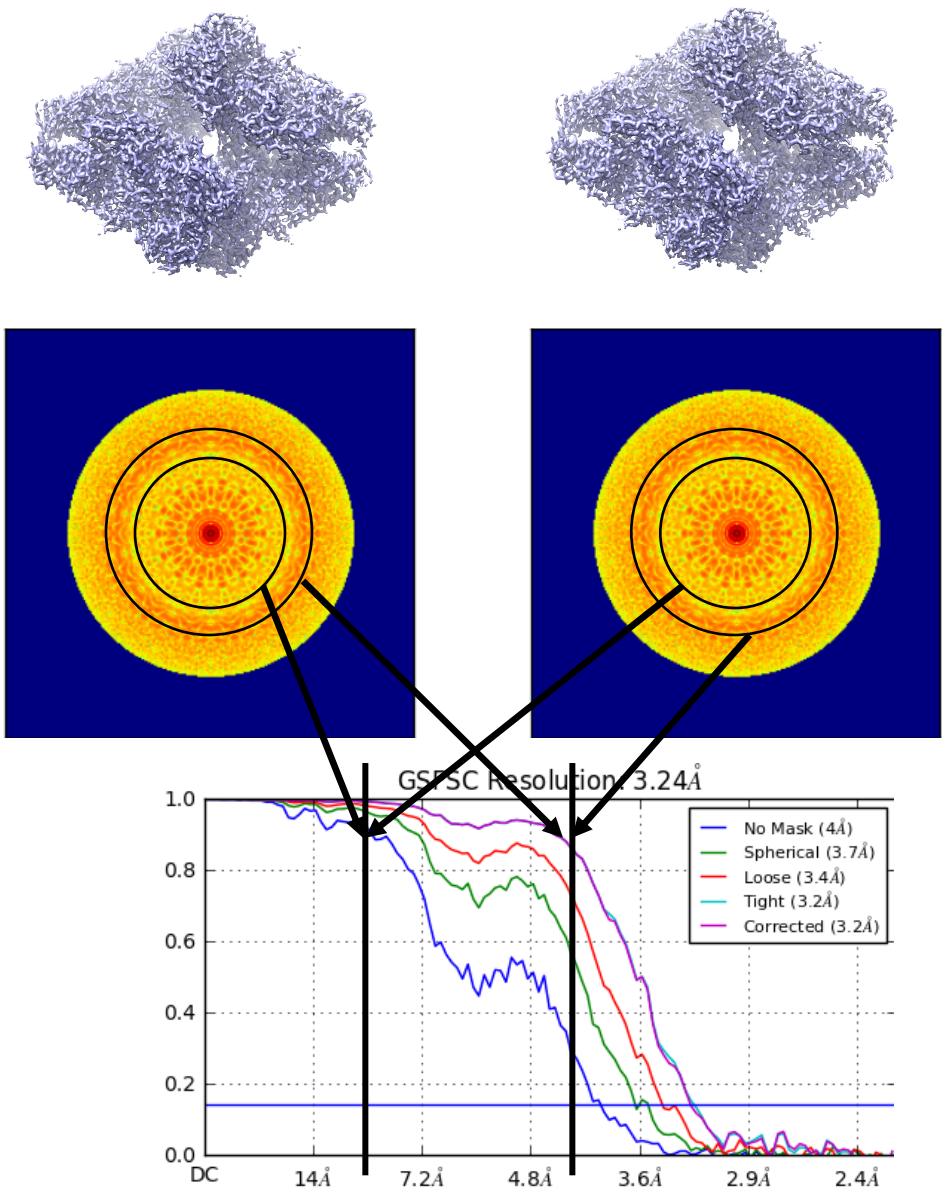
11 min on 1x NVIDIA V100



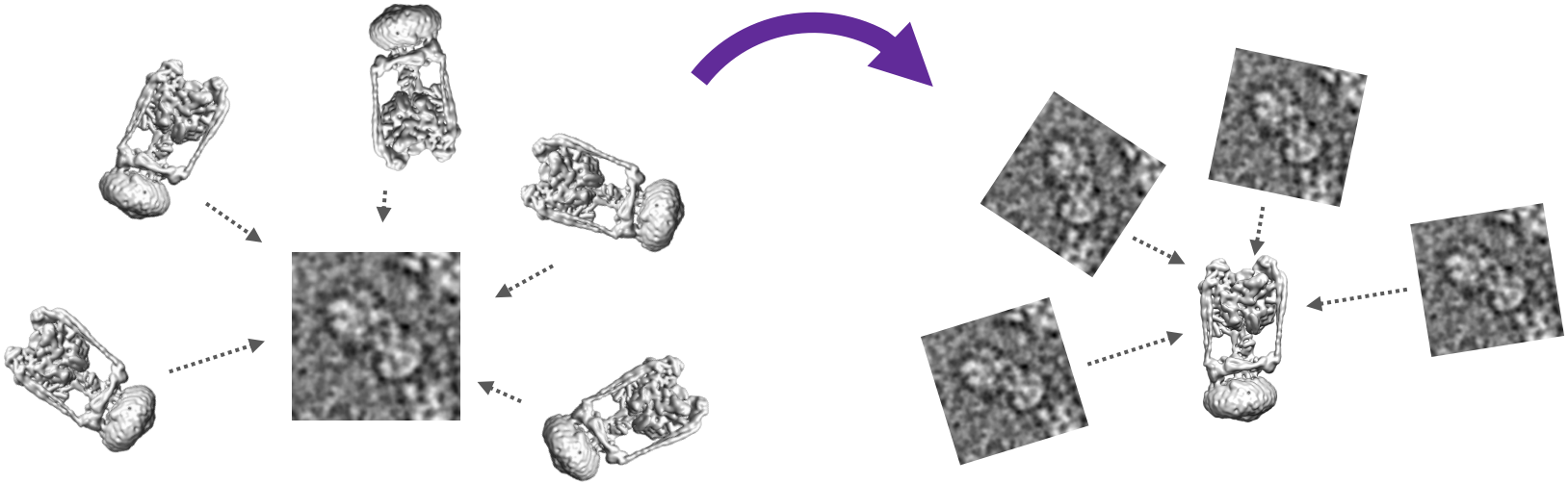
Punjani et al. Nature Methods 2017

Measuring Resolution

- Dataset split in half
- Maps reconstructed from each half
- Compare corresponding shells of Fourier components
- Check where Fourier Shell Correlation drops below a threshold
- Mask is critical!

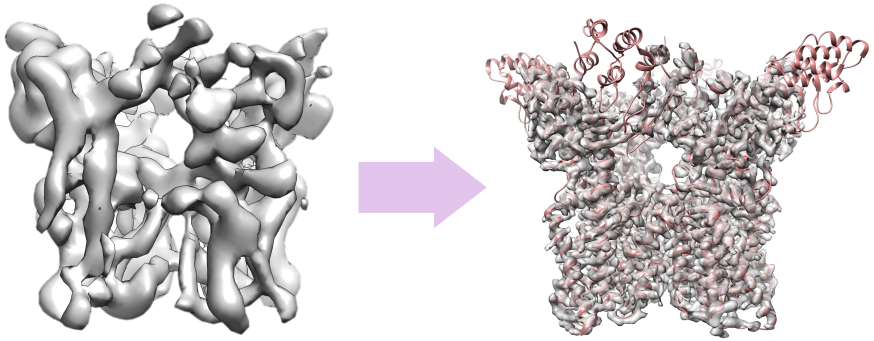


Iterative refinement (Expectation Maximization)



Alignment
(Structure fixed)

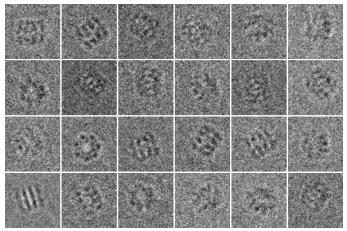
Reconstruction
(Alignments fixed)

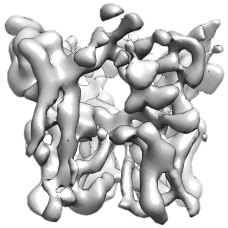


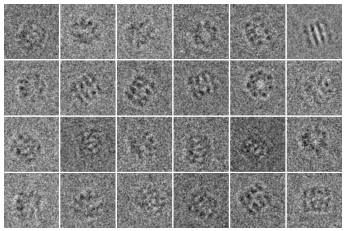
Overfitting!

Preventing overfitting: Gold-Standard FSC

[Scheres and Chen, 2012]

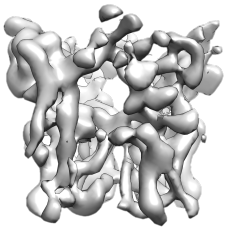


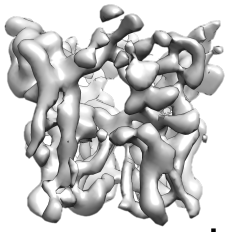




Half "A"

Half "B"





FSC

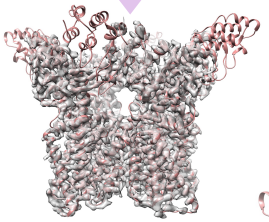
FSC

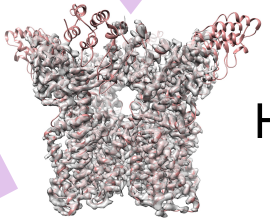
FSC

FSC

Independent Refinements

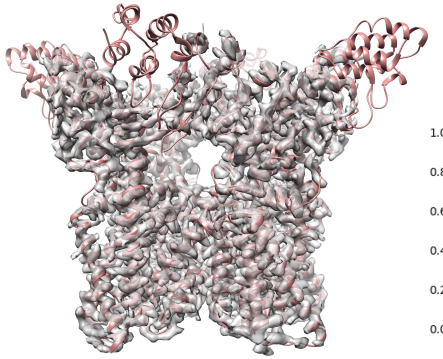
Independent Refinements

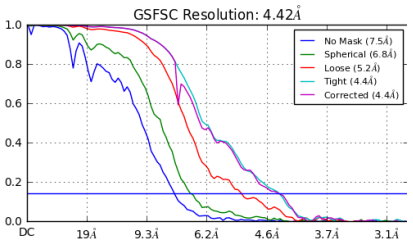




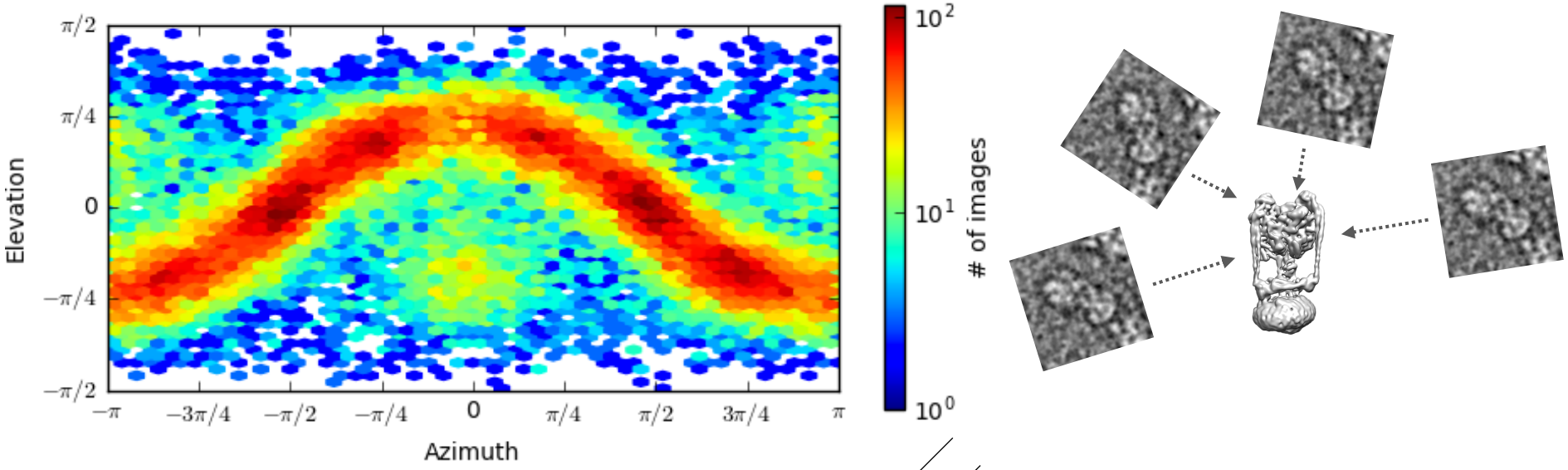
Half Map A

Half Map B

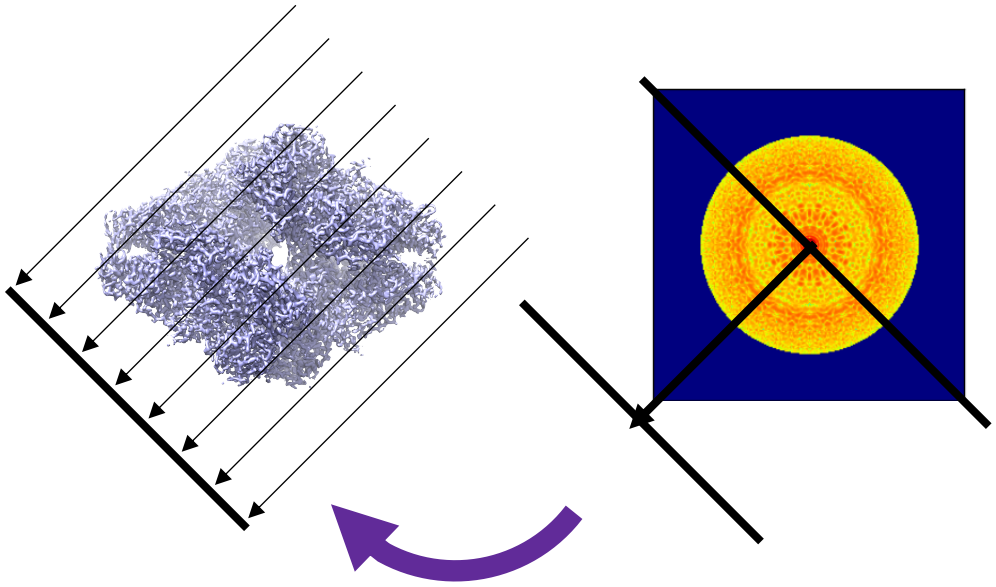




Orientation distribution

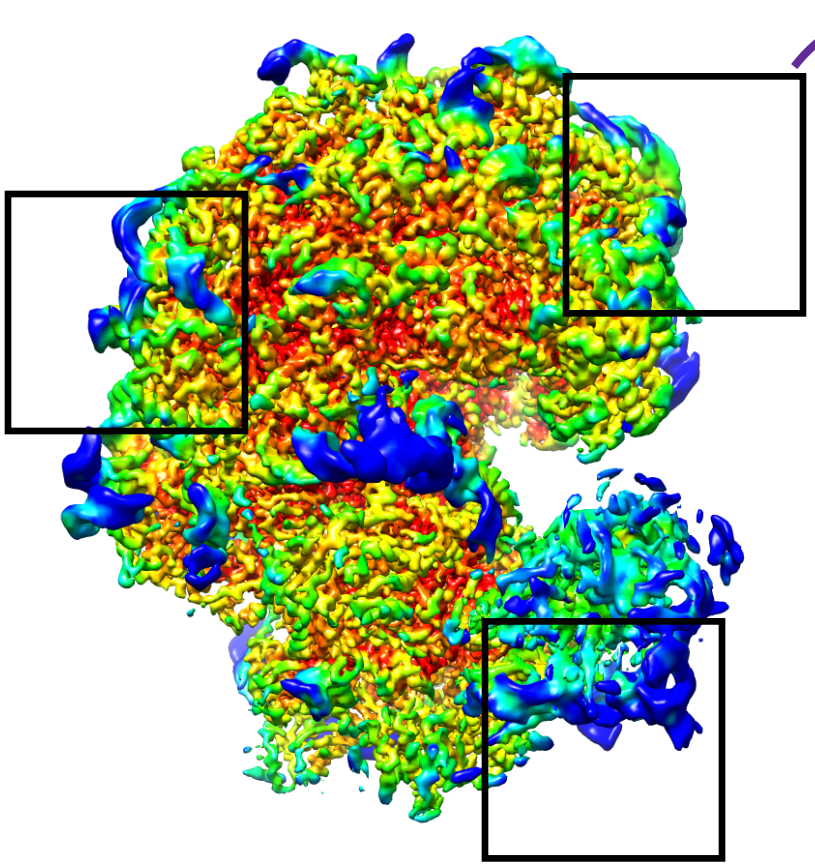


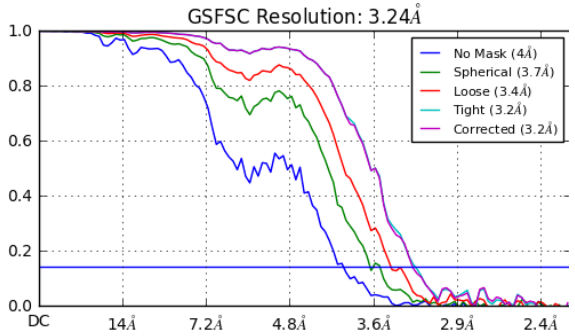
- Need coverage of viewing directions
- Not necessarily every direction, but enough to cover 3D Fourier space
- Very common resolution limitation



Local resolution and filtering

- Fast GPU implementations

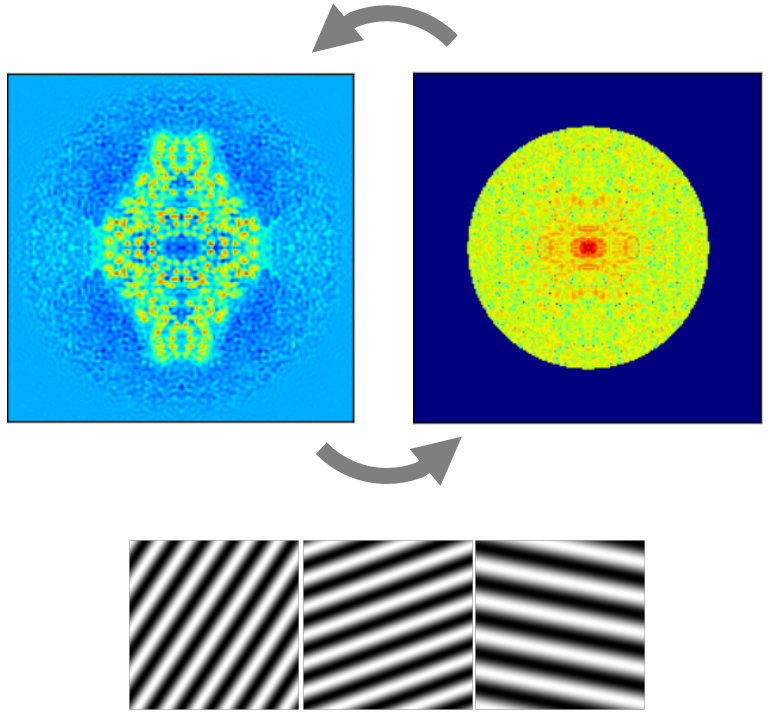




Non-uniform refinement

Conventional refinement:

- Grounded in Fourier basis
  - Fourier slice theorem
  - CTF correction
  - 2D-3D Backprojection
  - Fourier Shell Correlation

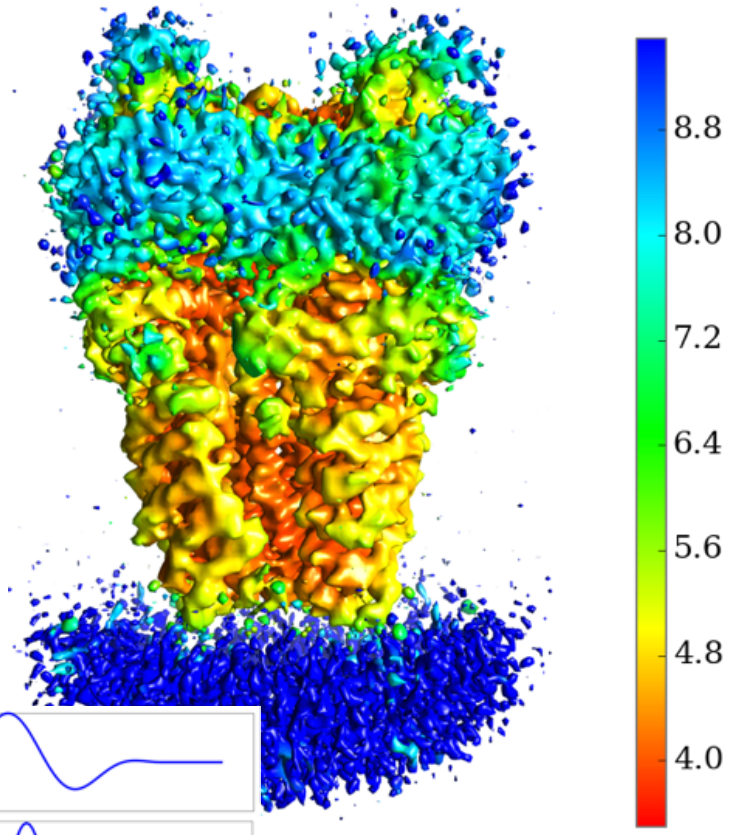


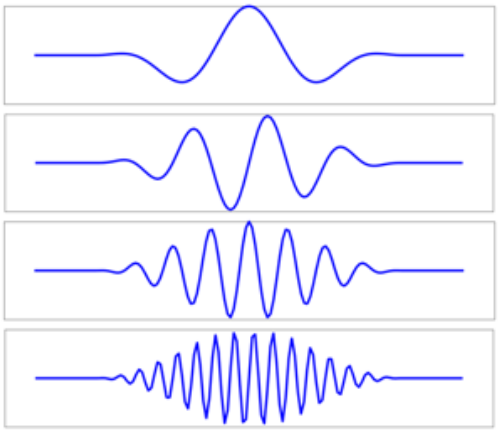
Non-uniform refinement

Key intuition:

- Regularization in Fourier basis inappropriate
  - Disordered regions
    - Detergent, lipid nanodisc, etc
  - Solvent-protein boundary
  - Fractional occupancy
  - Flexible or highly dynamic regions

TRPA1
EMPIAR 10024

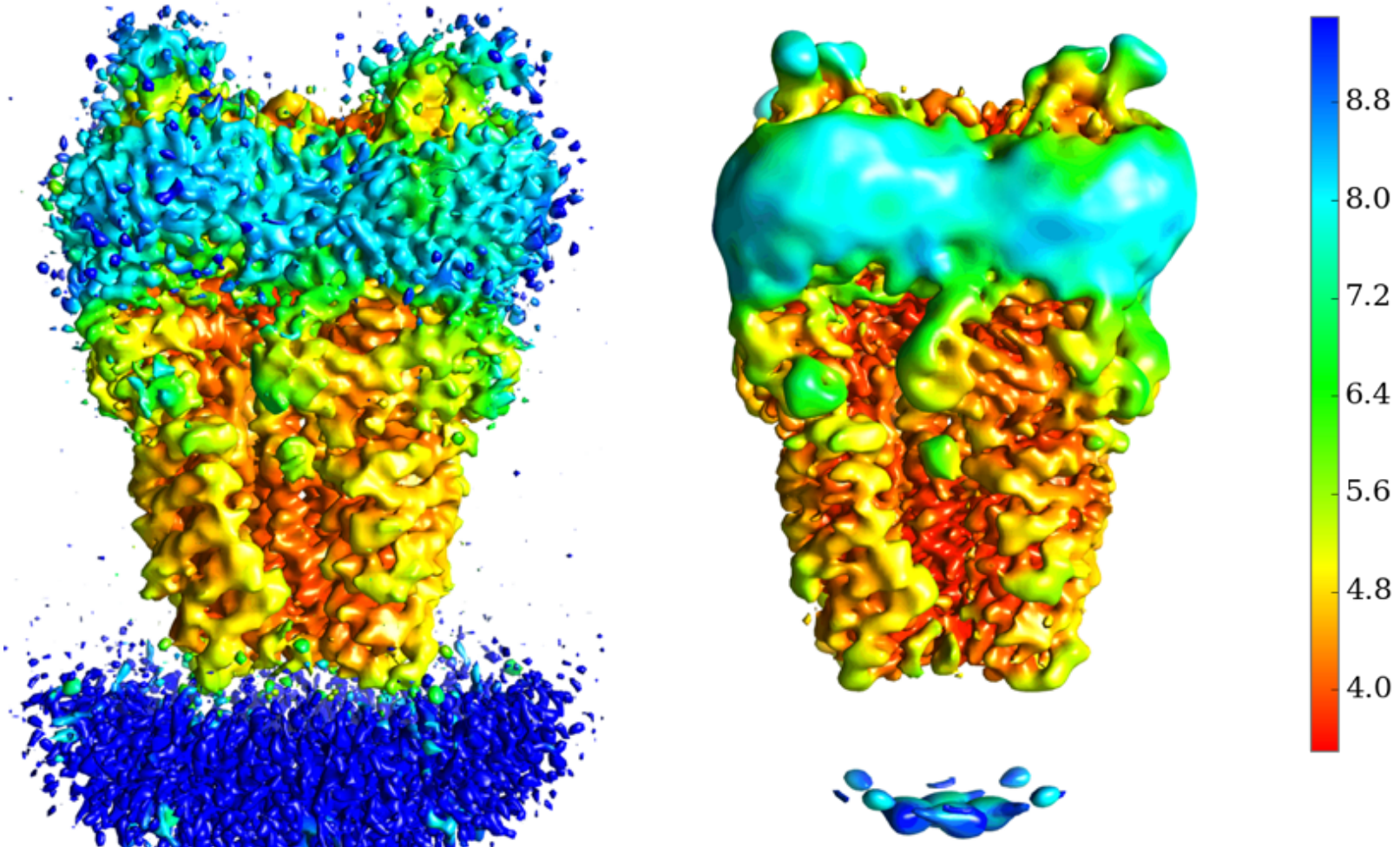




- Non-uniform refinement:

  Use a basis that is localized in space and frequency

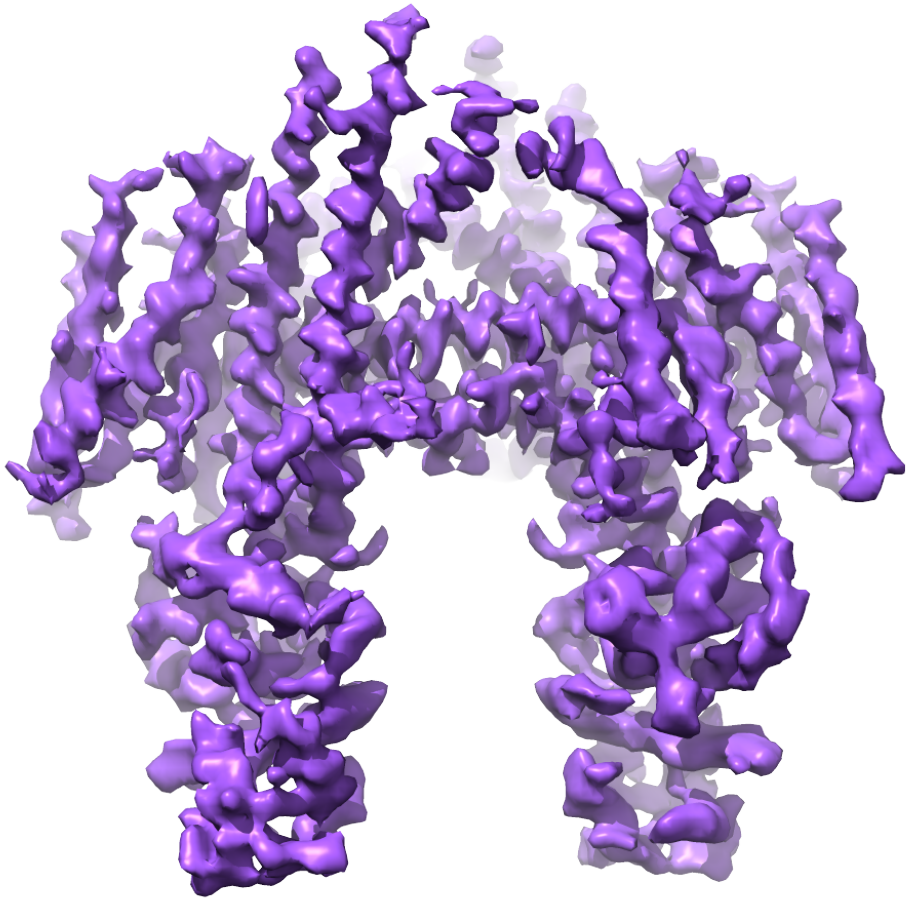
Non-uniform refinement



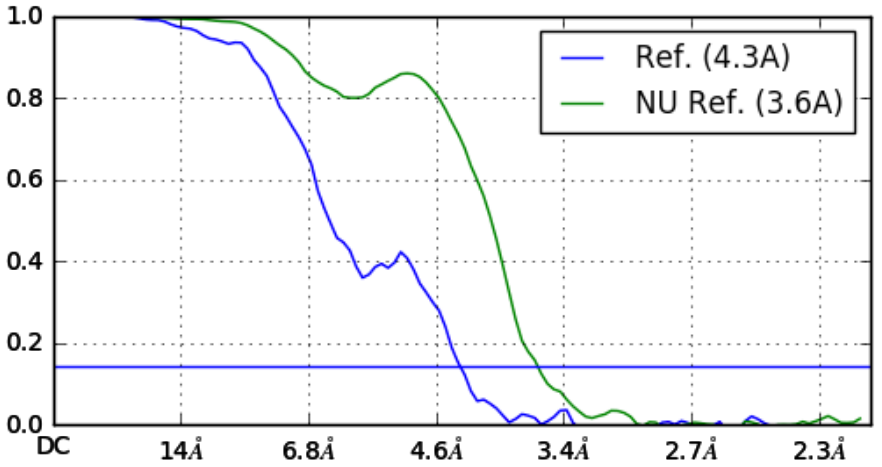
Non-uniform refinement

Oliver Clarke & Filippo Mancia, Columbia University

STRA6 receptor - 180 kDa membrane protein



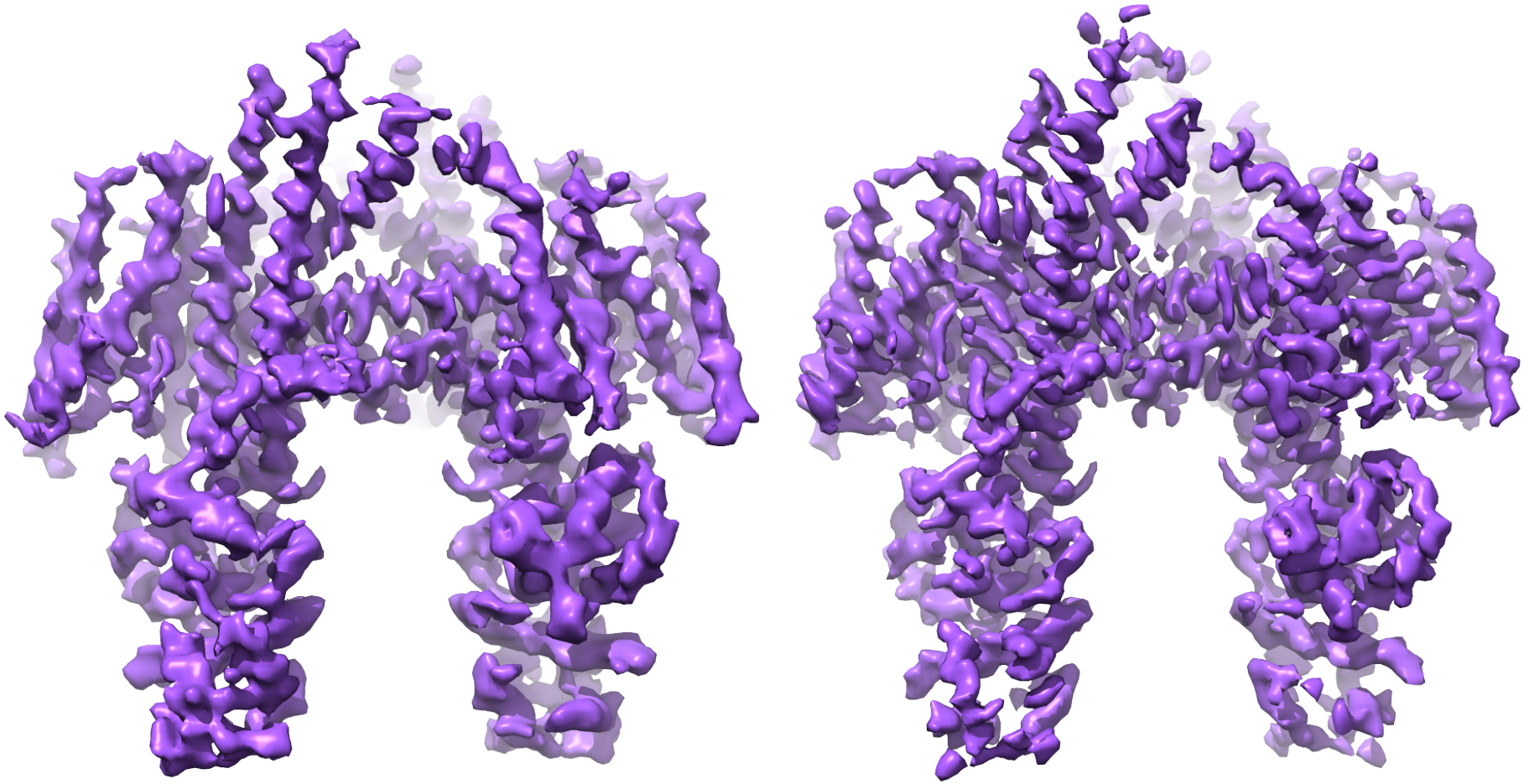
Conventional Refinement



Non-uniform refinement

Oliver Clarke & Filippo Mancia, Columbia University

STRA6 receptor - 180 kDa membrane protein

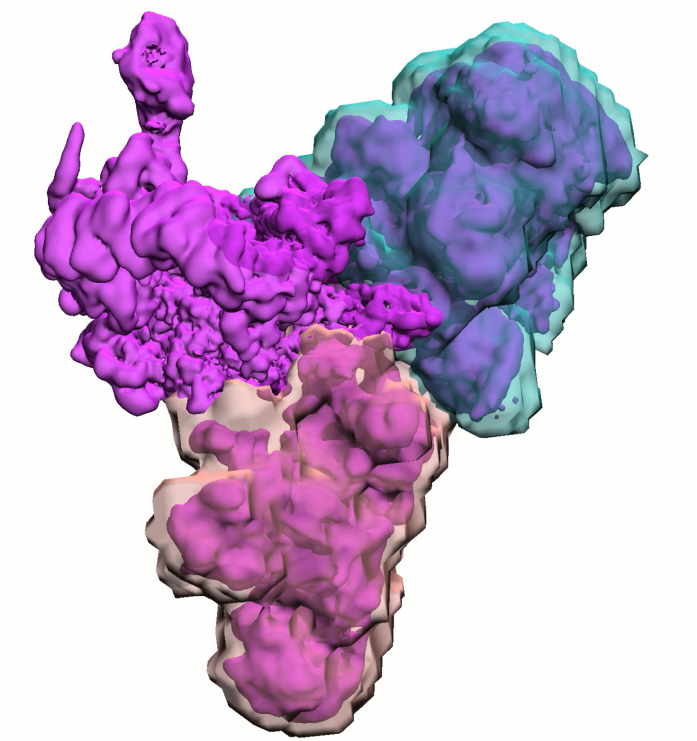


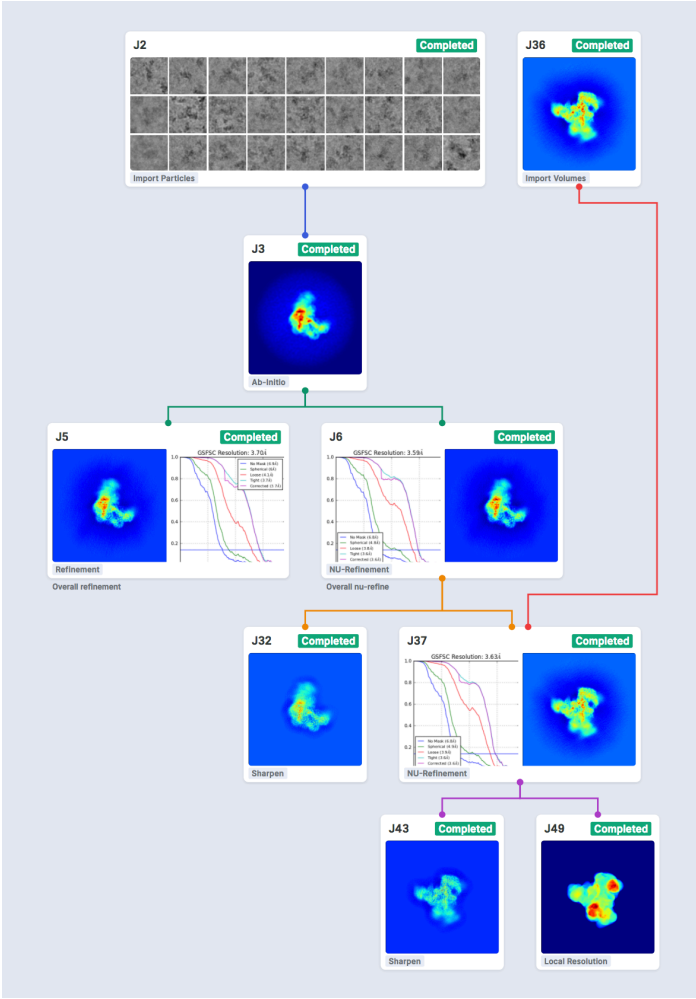
Conventional Refinement

Non-uniform Refinement

Local and focused refinement

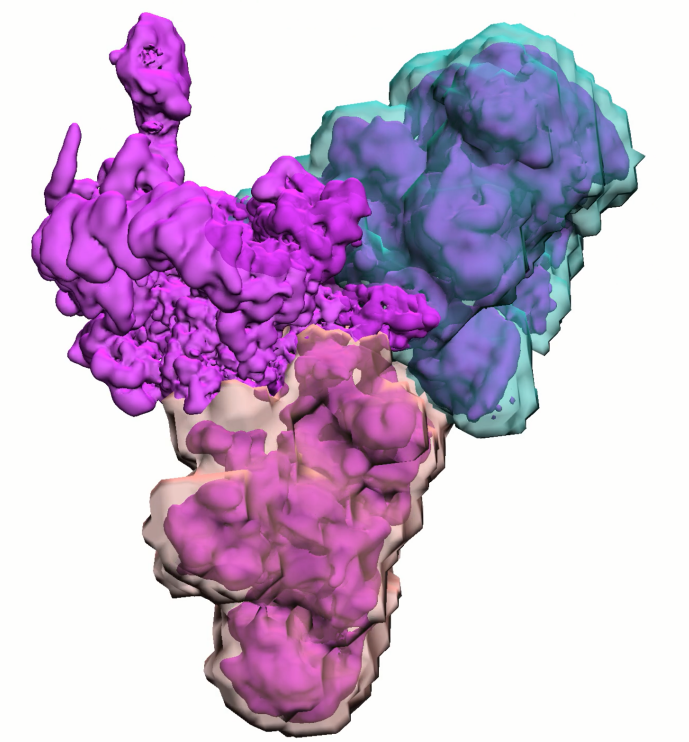
Masking, signal subtraction and non-uniform refinement



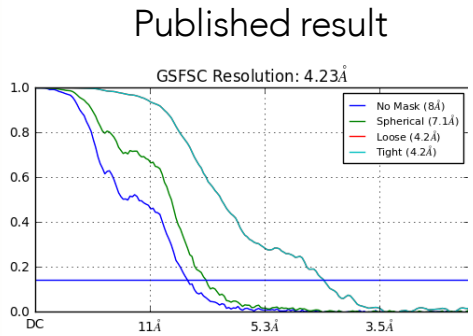


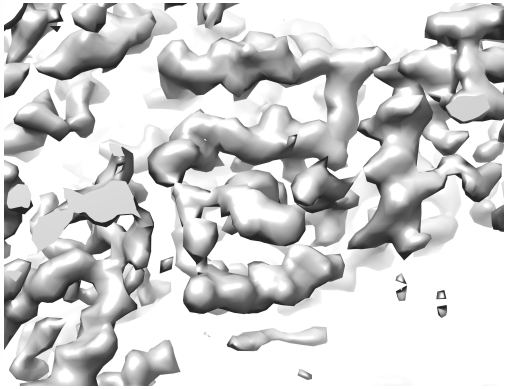
Local and focused refinement

Masking, signal subtraction and non-uniform refinement

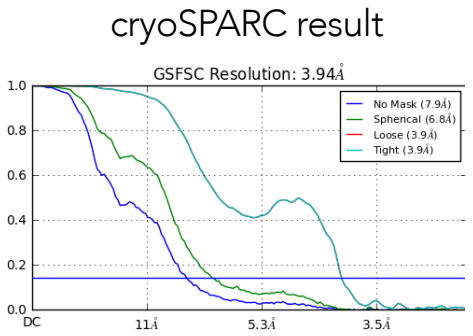


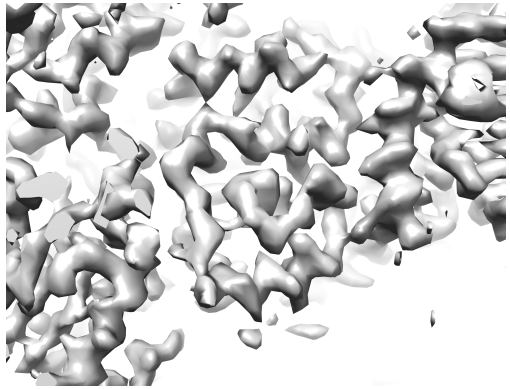
Published result





cryoSPARC result

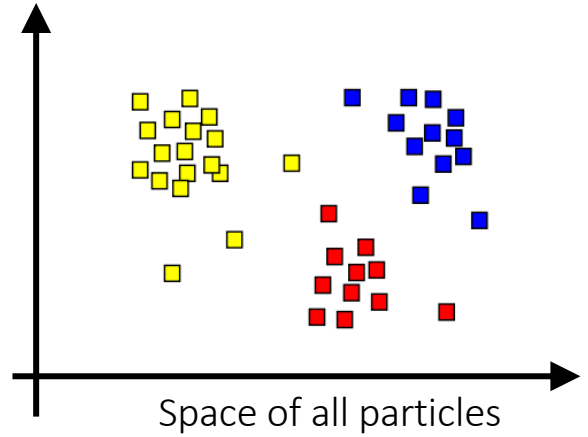




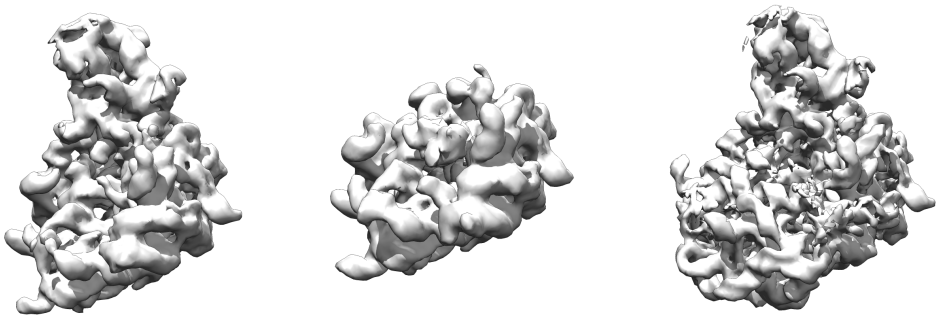
Conformational Heterogeneity

"3D classification" is clustering

- Significant biological insight in conformational landscapes
- Existing algorithms can deal with discrete heterogeneity



- How many classes?
- Continuous heterogeneity?



- New algorithm!

3D Variability Analysis (New! cryoSPARC v2.9)

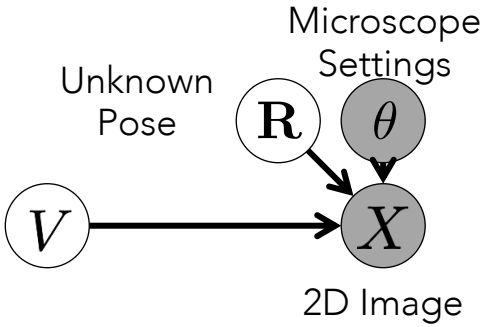
Directly determining 3D motion/conformational change

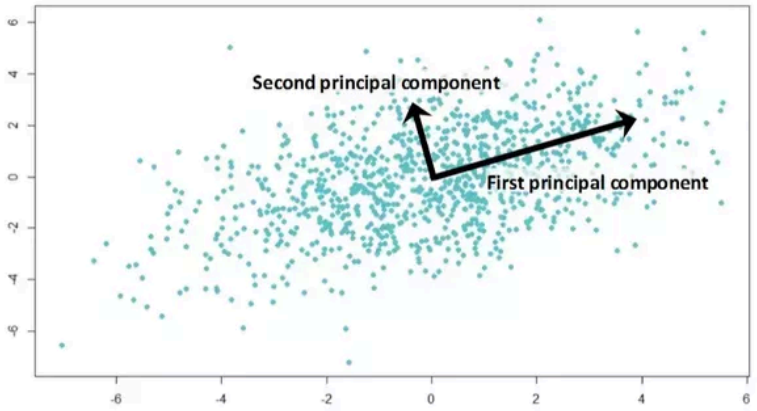
Consensus Volume 3D Covariance

3D Volumes $V_i \sim \mathcal{N}(\overline{V_0}, \overline{\Sigma})$

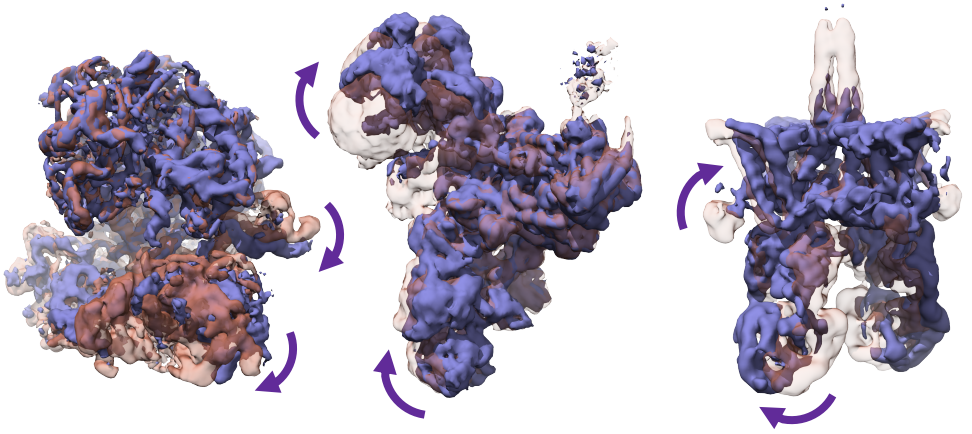
CTF Projection Image noise

Images $X_i \sim \mathcal{N}(\overline{C_i}\,\overline{P(\phi_i)}V_i, \overline{\sigma^2})$





- 3D Covariance matrix
- New algorithm to solve for top K eigenvectors
- Eigenvectors correspond to molecular motions
- Accounting for CTF, viewing directions, all particles simultaneously, high resolution



80S Ribosome
EMPIAR10028

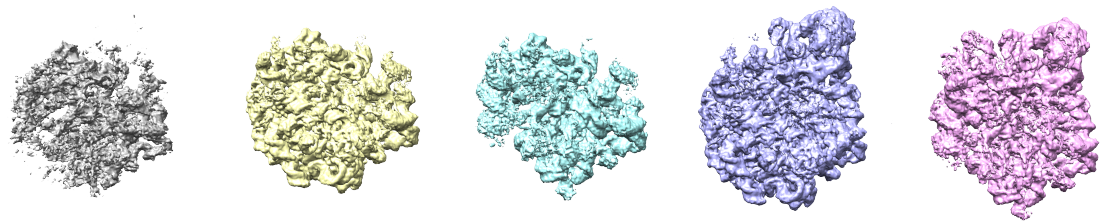
tri-snRNP
EMPIAR10073

Nav1.7
EMPIAR10261

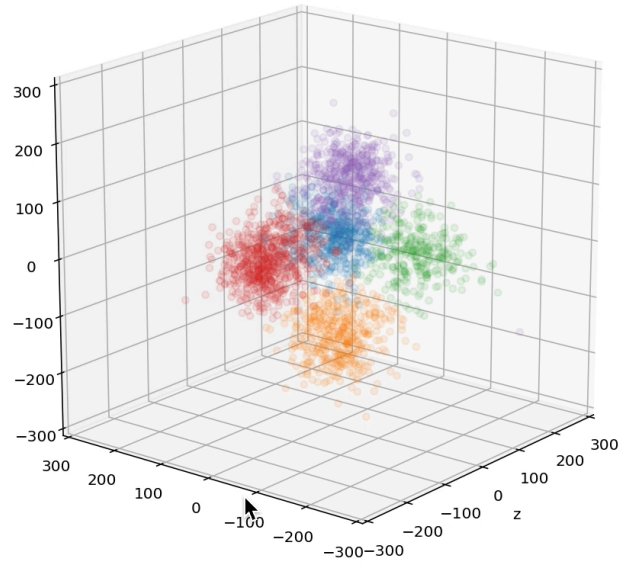
3D Variability Analysis – discrete heterogeneity

Directly determining 3D conformations

- Can measure "reaction coordinates" of individual particles
- Low-dimensional space directly shows clusters for discrete states
- Can run hierarchically

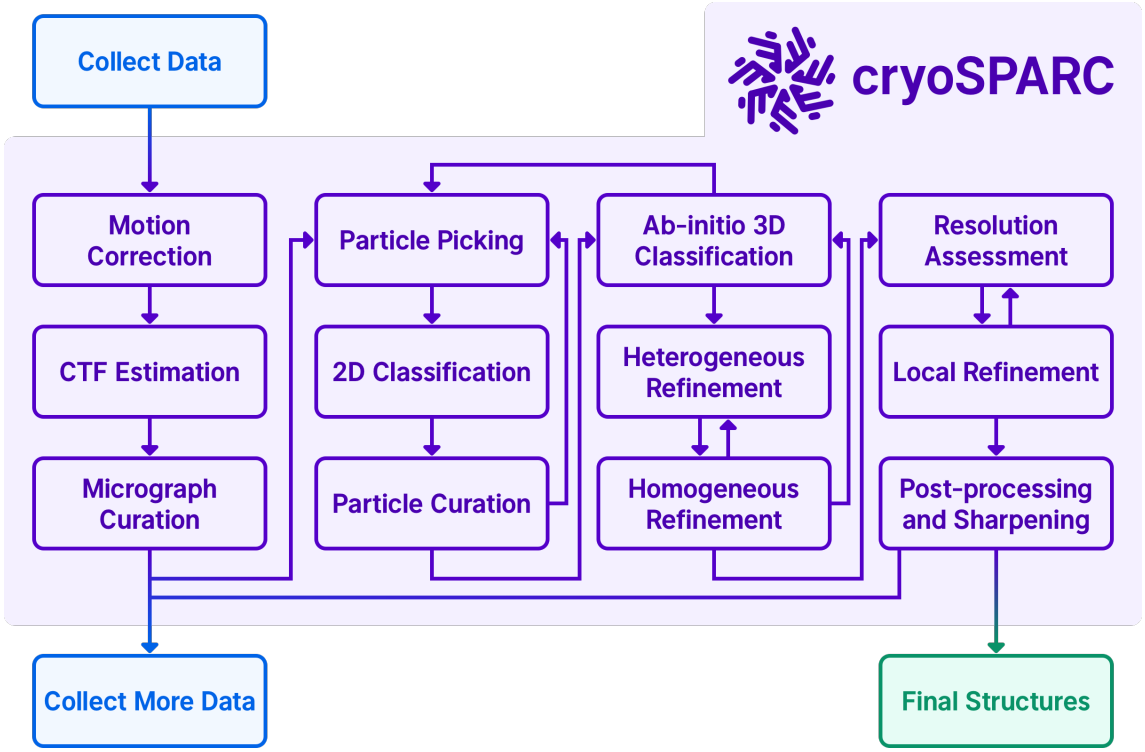


130K 50S Ribosome Particles
(EMPIAR 10076)
3 variability dimensions

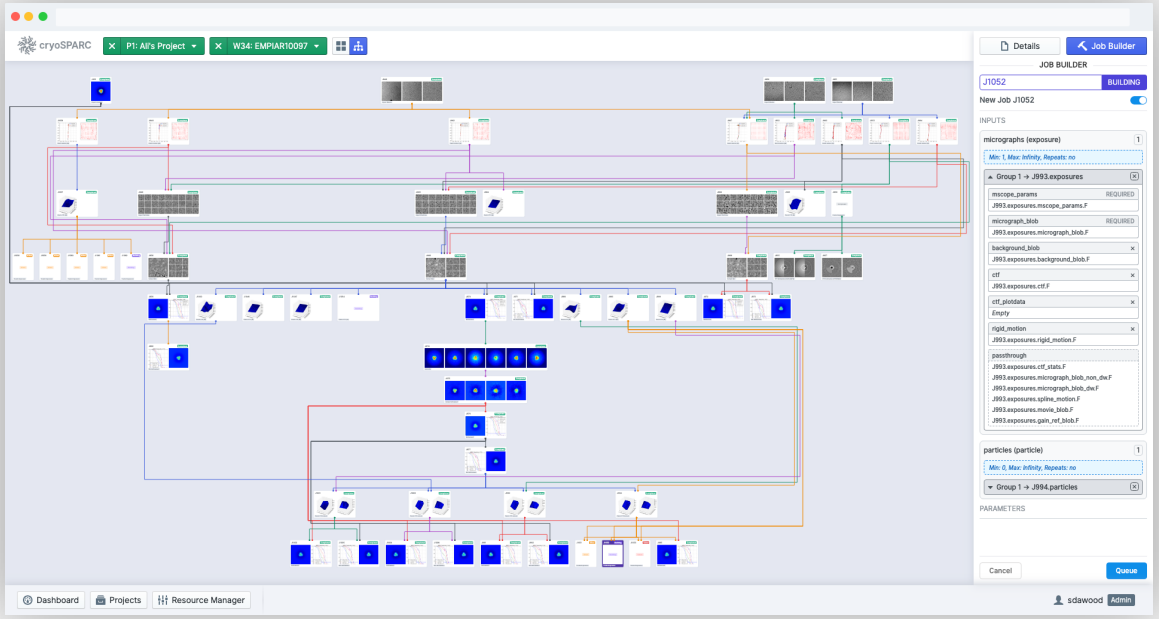


Non-trivial workflows

- Many tools needed for dealing with complex molecules
- Multiple iterations through processing pipeline required for near-atomic resolutions and identification of ligand binding sites
- Typical dataset sizes now multiple millions +

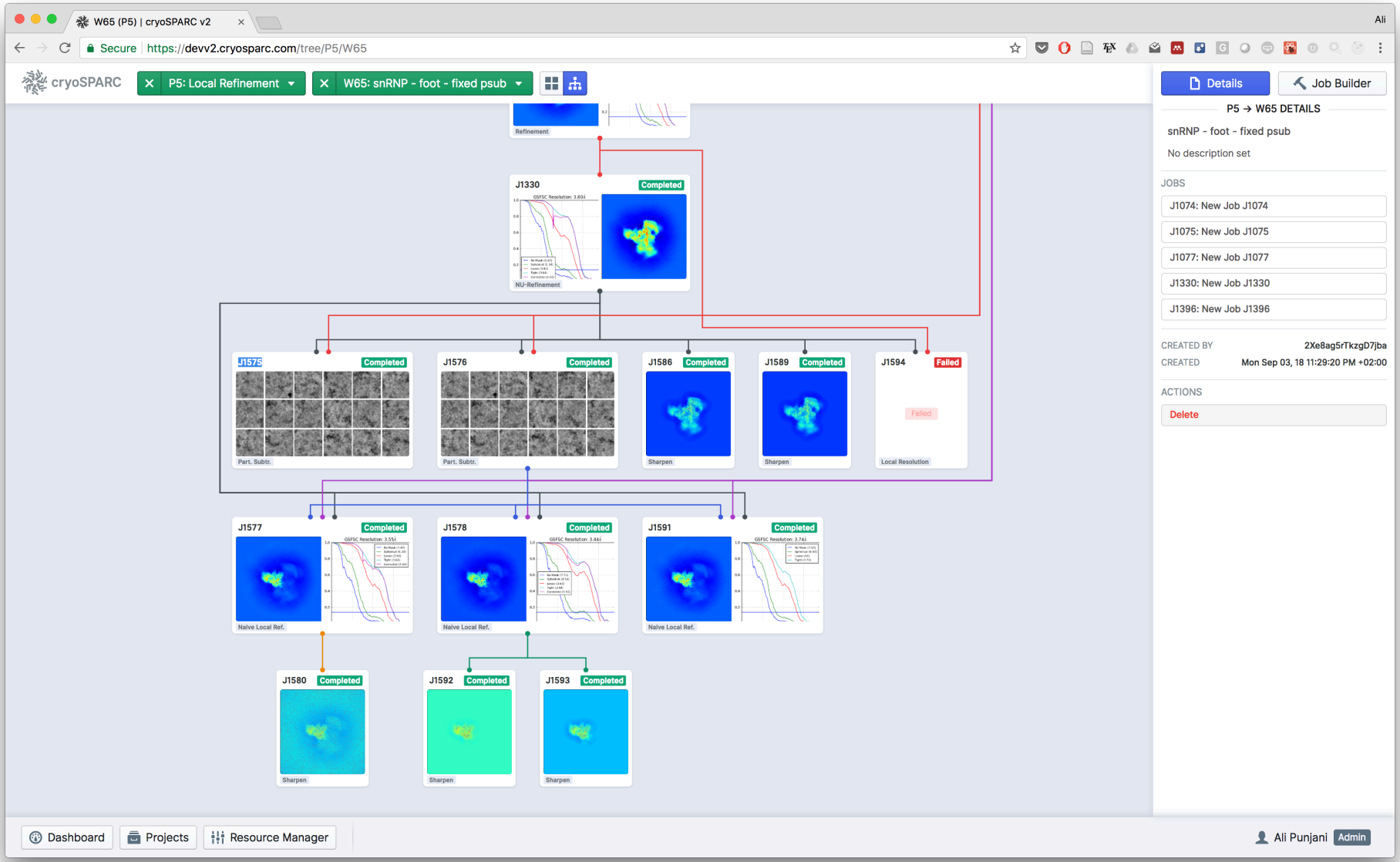


Robust software features and intuitive user experience



- ✓ Directly input and decompress raw data
- ✓ Project and workspace organization
- ✓ Tree view
- ✓ Drag and drop job builder
- ✓ Auto-tuning of hundreds of parameters
- ✓ Automated one-click workflows
- ✓ Interactive jobs
- ✓ Real-time experiment details and plots
- ✓ Direct downloads
- ✓ Computational resource manager
- ✓ Smart queuing/cluster integration
- ✓ User management
- ✓ Interoperability with other EM packages

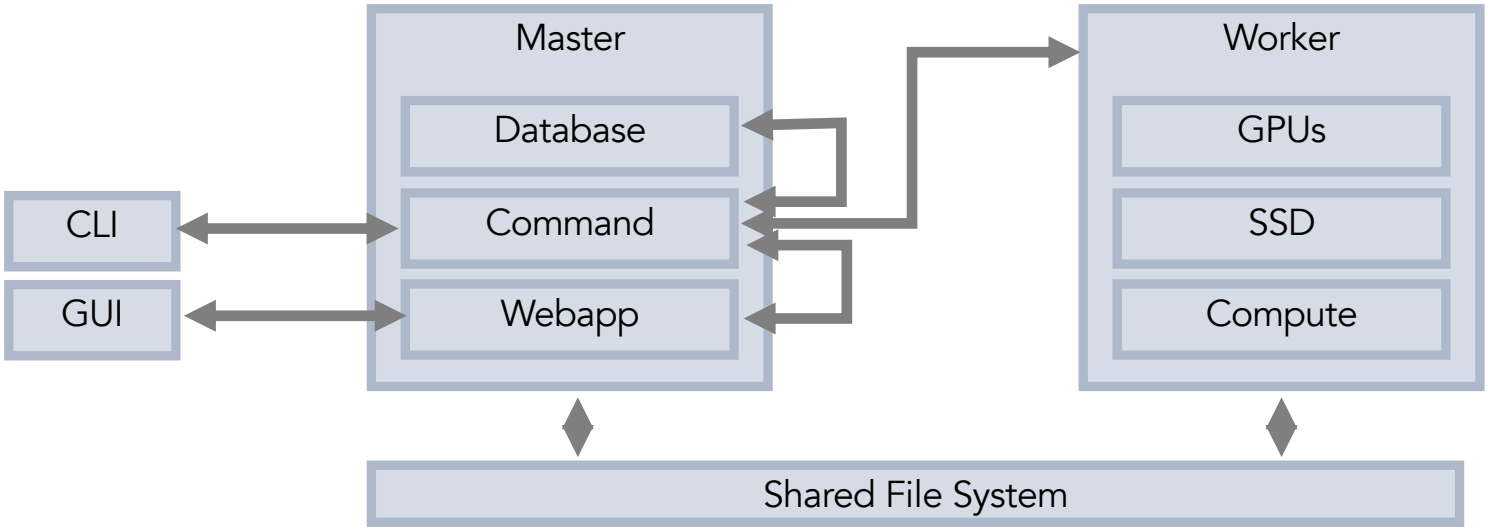
cryoSPARC v2 software system



Multi-node and cluster support

Modularized

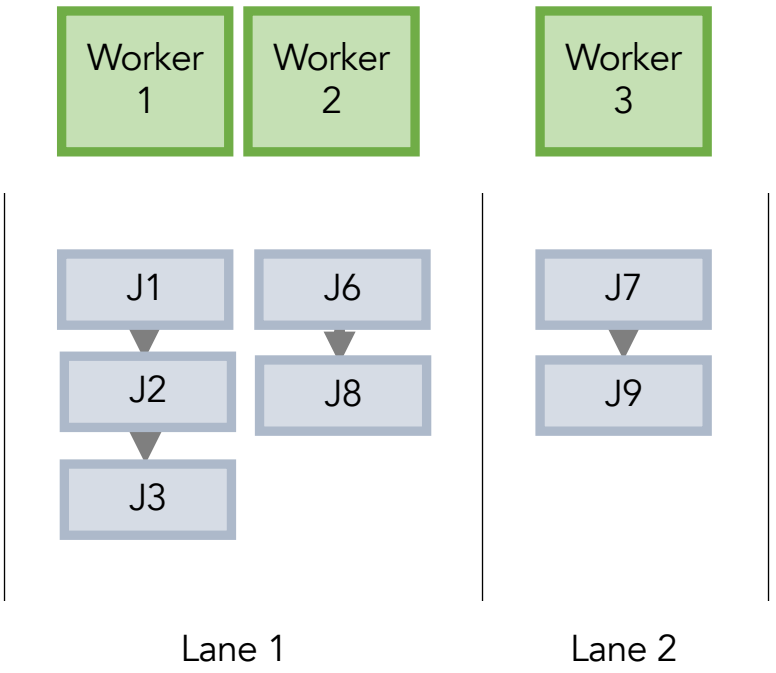
- Master and worker nodes



Scheduler

Smarter queueing

- Multiple independent lanes
- Each lane has workers on which jobs can be launched
- Within a lane, first in first out
- Jobs wait for their dependencies to complete before running
    - Easy to chain and queue jobs and let run autonomously



CryoSPARC Team

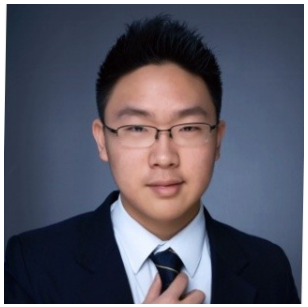
Ali Punjani

Suhail Dawood

Saara Punjani

Stephan Arulthasan

Ali Haydaroglu



Jay Yoo

David Fleet, PhD

University of Toronto

John Rubinstein, PhD

SickKids Hospital
University of Toronto

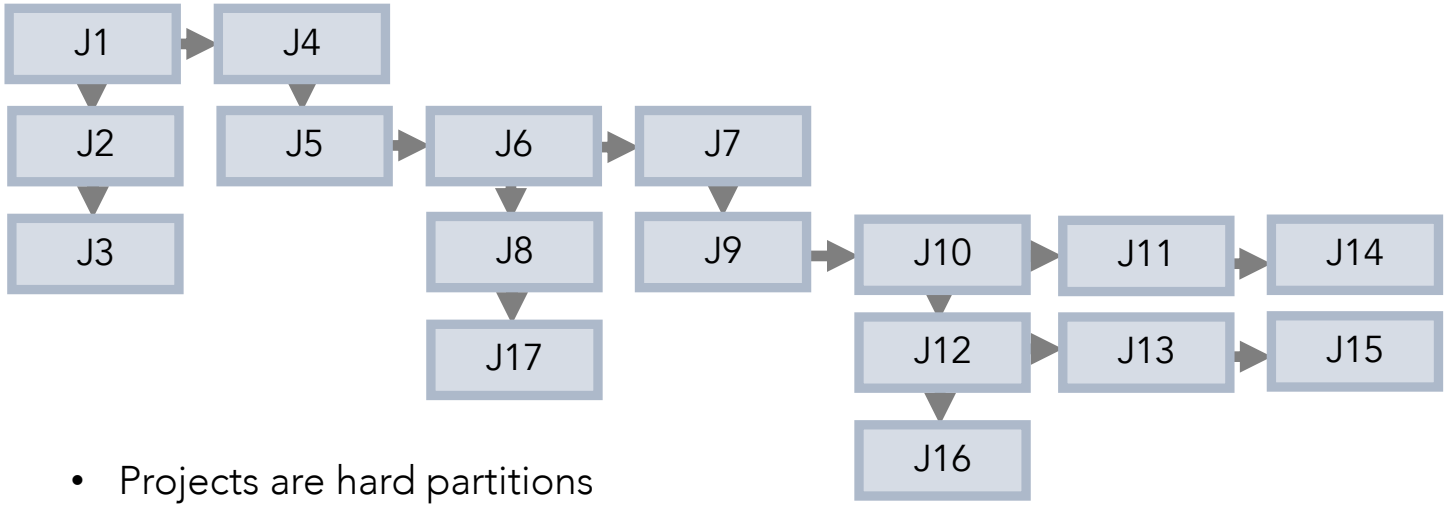
Marcus Brubaker, PhD

York University

Thanks!

Projects, Workspaces, Jobs

Tree structure with subtrees

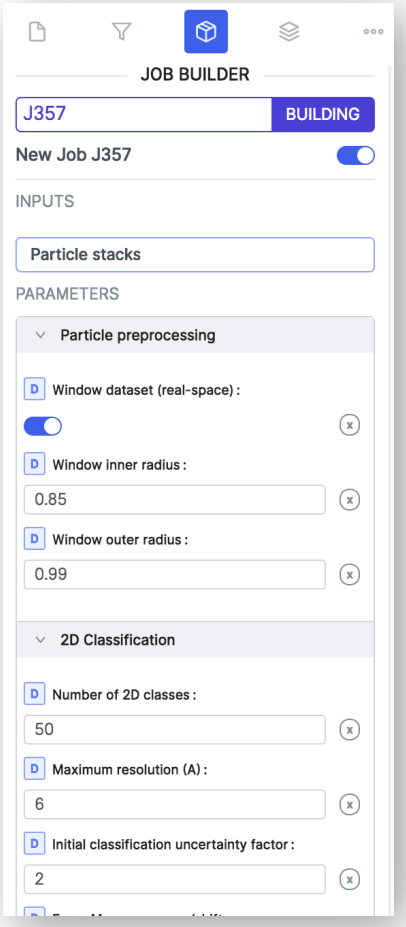


- Projects are hard partitions
- Jobs connect via results
- Complex workflows

Job builder

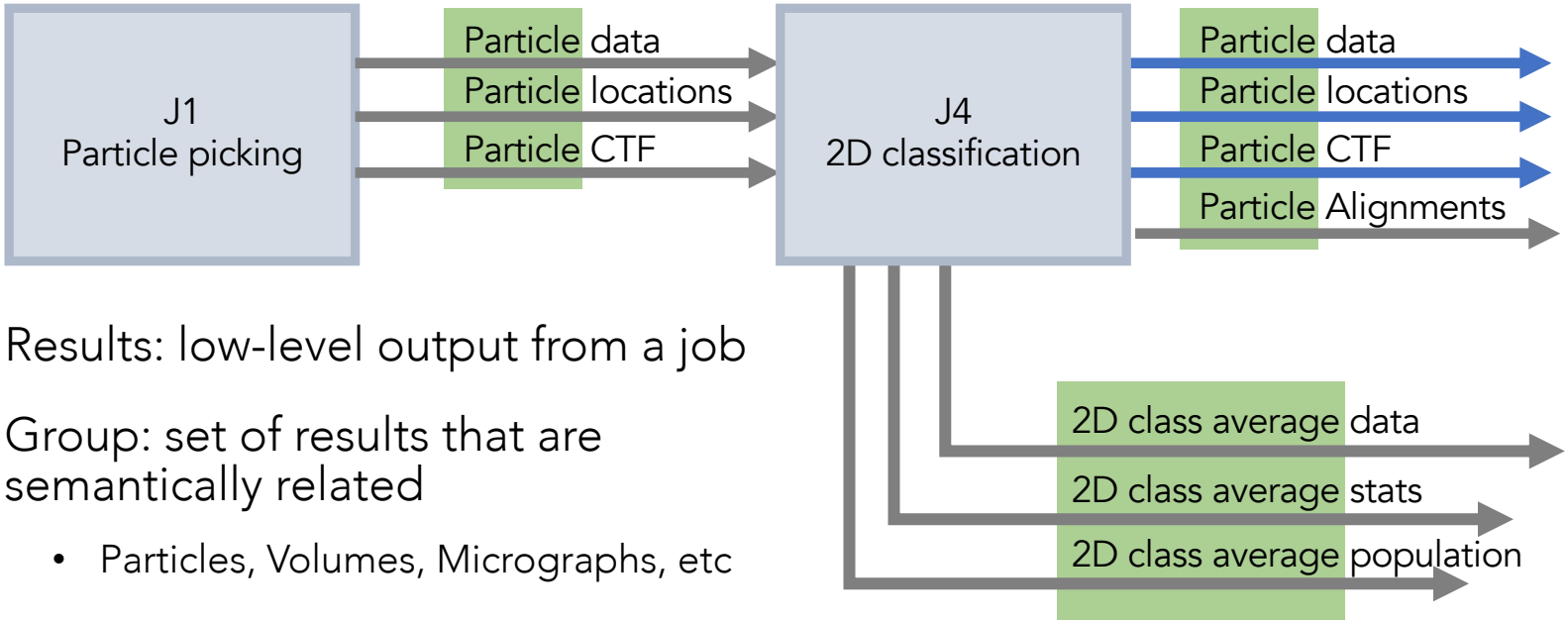
Drag and drop to create new jobs

- Connect inputs
- Set parameters
- Queue job



Results and Groups

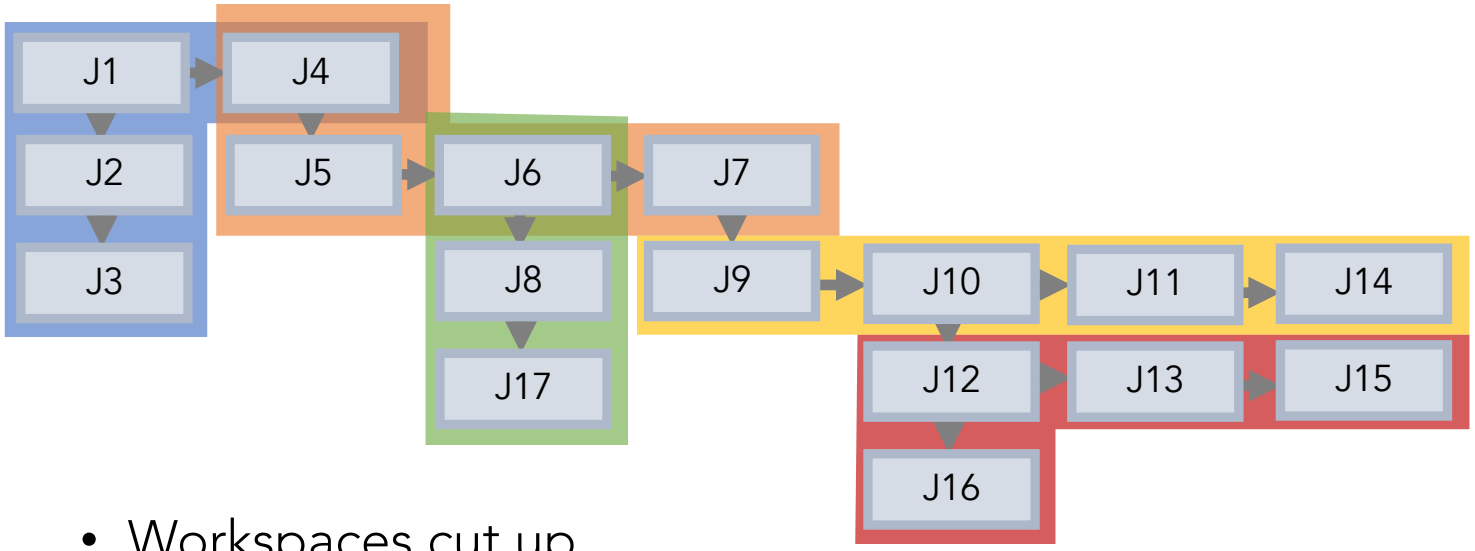
High level and low level containers



- Results: low-level output from a job
- Group: set of results that are semantically related
  - Particles, Volumes, Micrographs, etc
- Connect jobs using groups

Projects, Workspaces, Jobs

Tree structure with subtrees



- Workspaces cut up workflows into manageable chunks within a project