Squeezing Blood From a Stone: Challenges in Single Particle Cryo-EM Data Processing

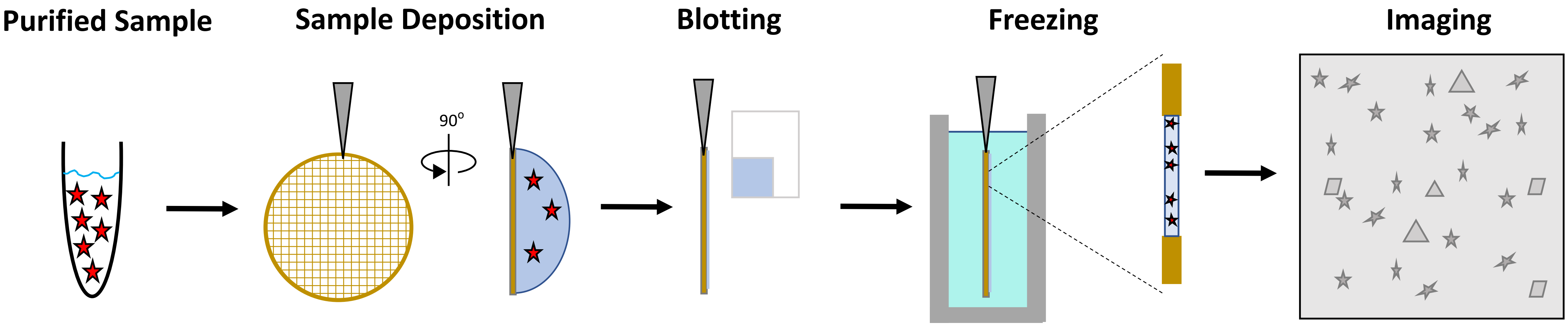
Philip T. McGilvray

Nanoimaging Services

03/09/22

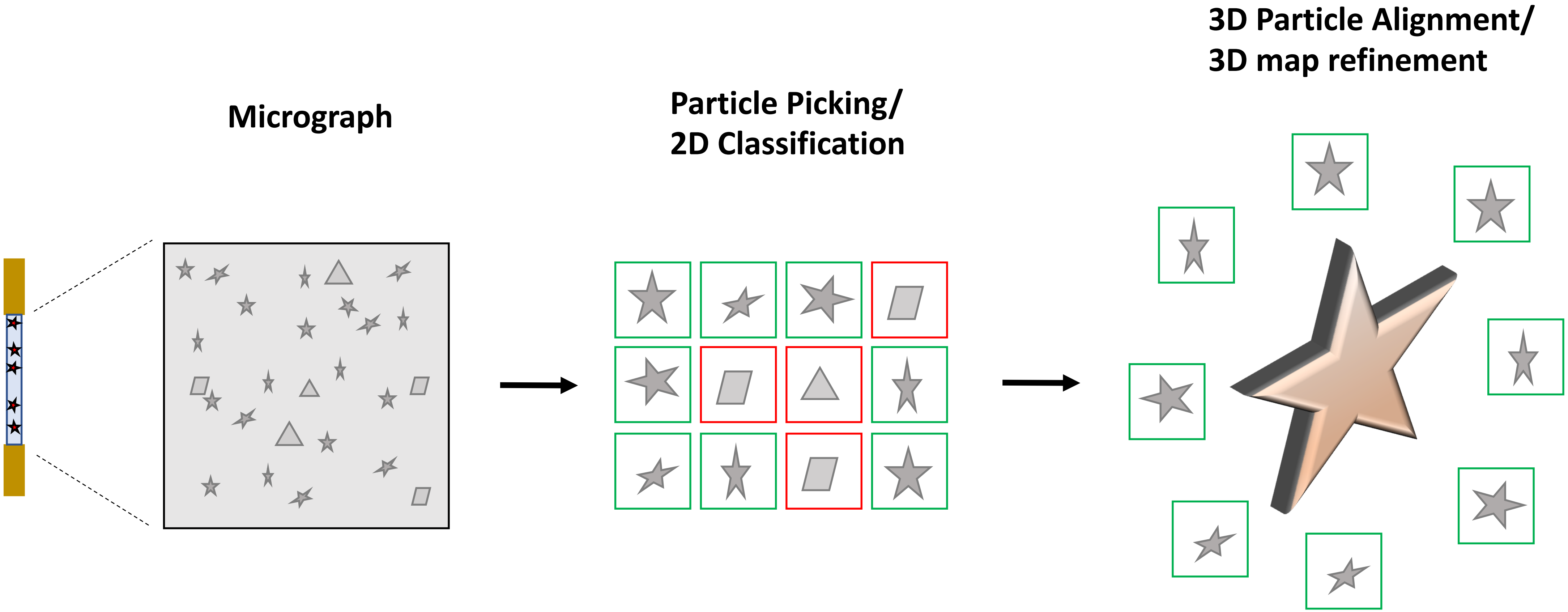
Sample preparation for single particle Cryo-EM

- The goal of single particle Cryo-EM is to generate micrographs of well dispersed, hydrated, frozen biomolecules



Single Particle Analysis

- The goal of single particle analysis is to align images of a homogenous particles in order to generate high resolution 3D maps



Micrograph

Particle Picking/ 2D Classification

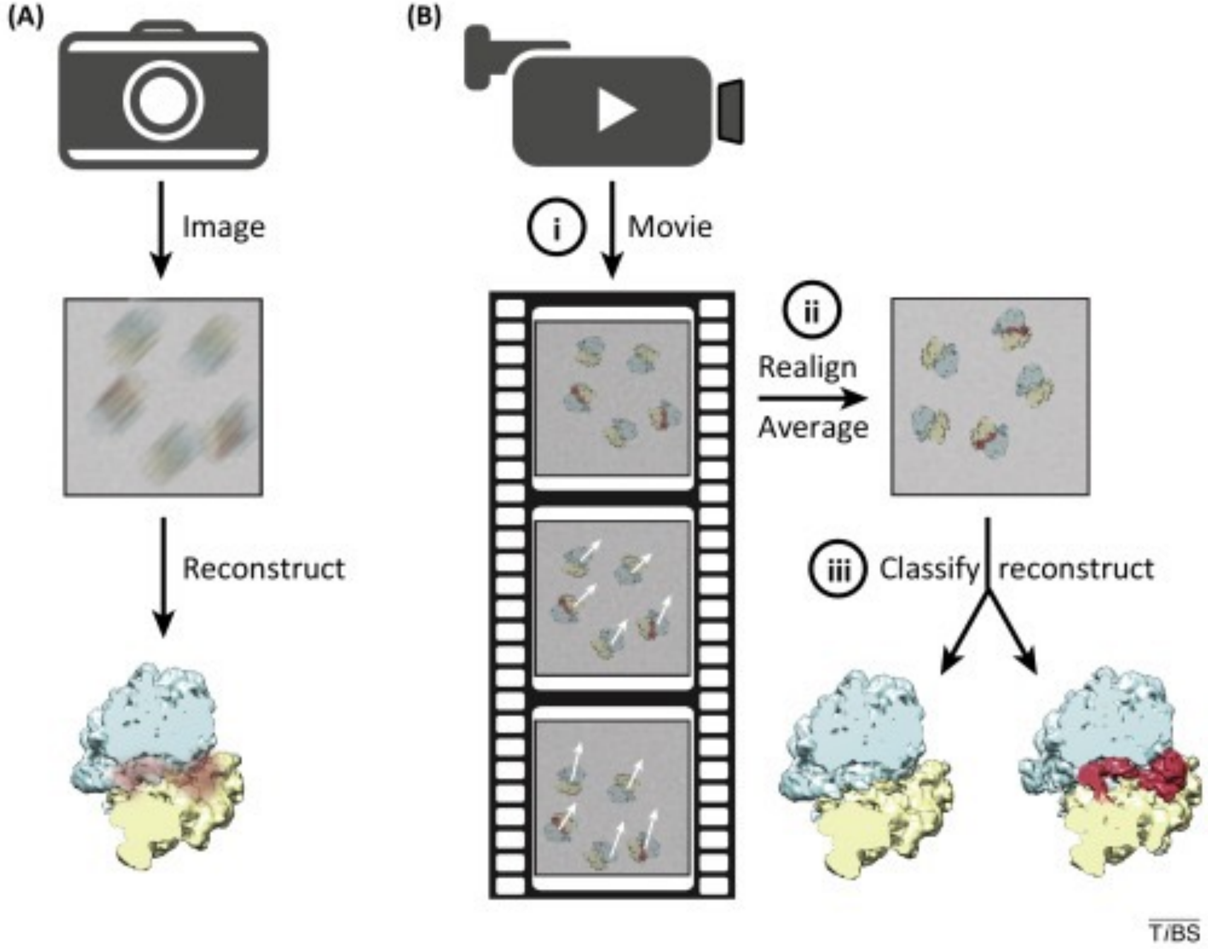
3D Particle Alignment/ 3D map refinement

Recent technical advances made Cryo-EM a frontline technique for protein stricture determination

- It has only been possible to routinely solve high resolution structures using Cryo-EM since ~2014

- Facilitated by several technical developments
  - Fast Imaging Direct Electron Detectors
  - Motion Correction
  - Improved analysis software

- Structural analysis of complex biomolecules (somewhat) routine
  - Membrane proteins
  - Large complexes
  - RNA/DNA



(Bai et al, TiBS 2015)

Single particle analysis deals with a range of challenges from the samples to the computing requirements

- Samples tend to have problems!
  - Can we work through these problems computationally?
  - What methods exist to work through them?
  - Why are they necessary?

- How much is all this going to cost me?
  - Scopes, cameras
  - Hardware/software
  - STORAGE!!!!
  - Connectivity
  - Facilities and maintenance

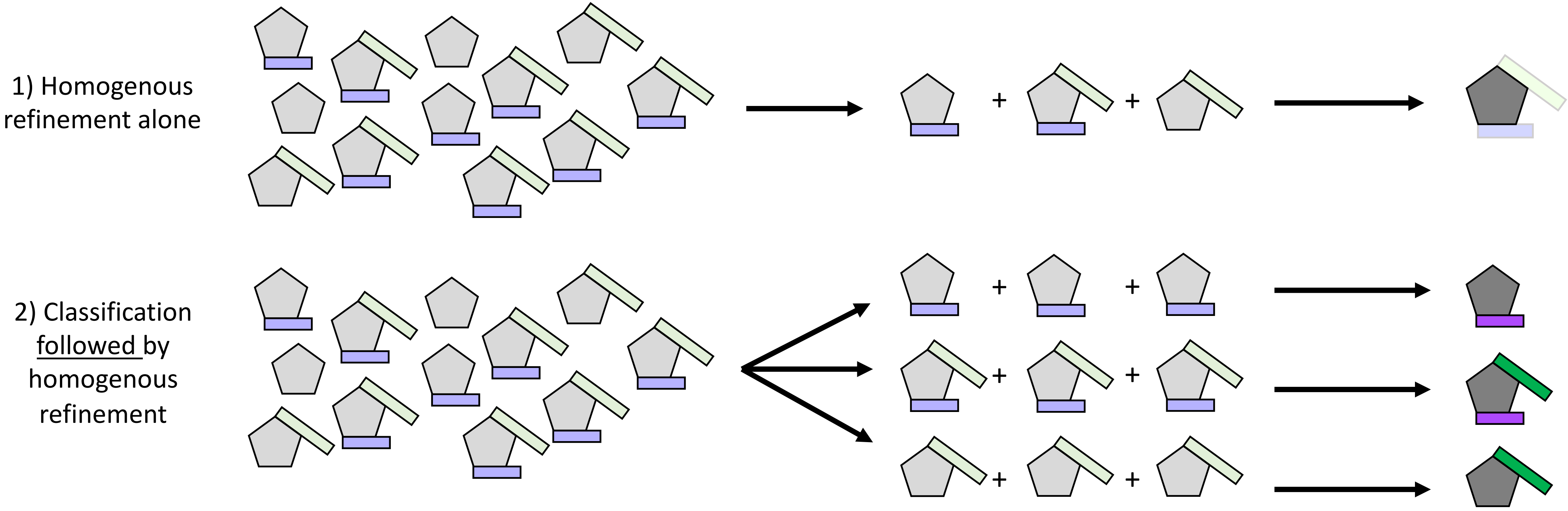
Single particle analysis deals with a range of challenges from the samples to the computing requirements

- Samples tend to have problems
  - Can we work through these problems computationally?
  - What methods exist to work through them?
  - Why are they necessary?

- How much is all this going to cost me?
  - Scopes, cameras
  - Hardware/software
  - STORAGE!!!!
  - Connectivity
  - Facilities and maintenance

Classification reduces heterogeneity by grouping like particles

- Single particle analysis assumes compositional and conformational homogeneity
- Biomolecules are (importantly!) neither

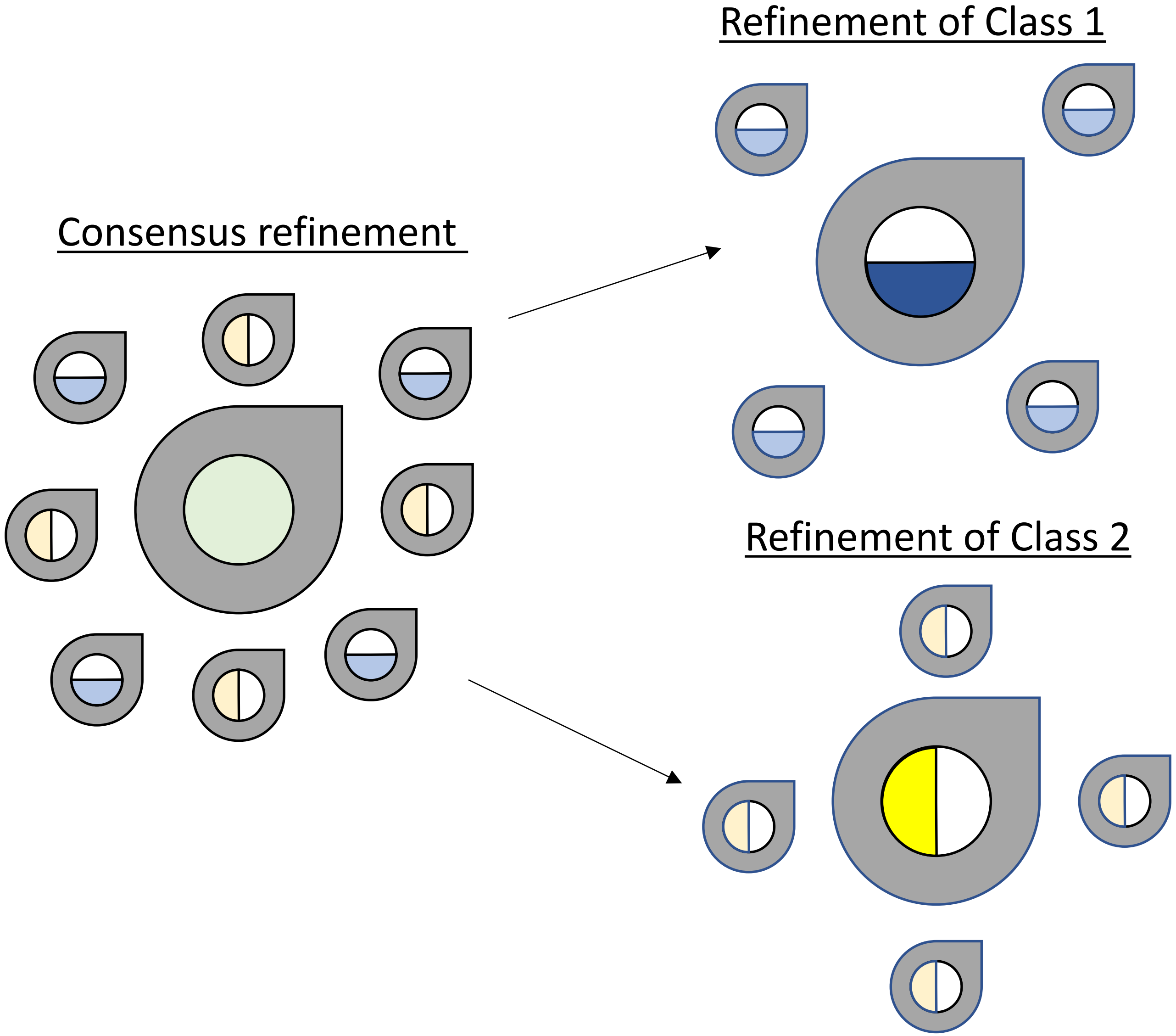

1) Homogenous refinement alone
2) Classification followed by homogenous refinement
+
+
+
+
+
+
+
+

Classification of particles without alignment can identify subtle heterogeneity

- Alignment driven by homogenous features at the expense of heterogenous features

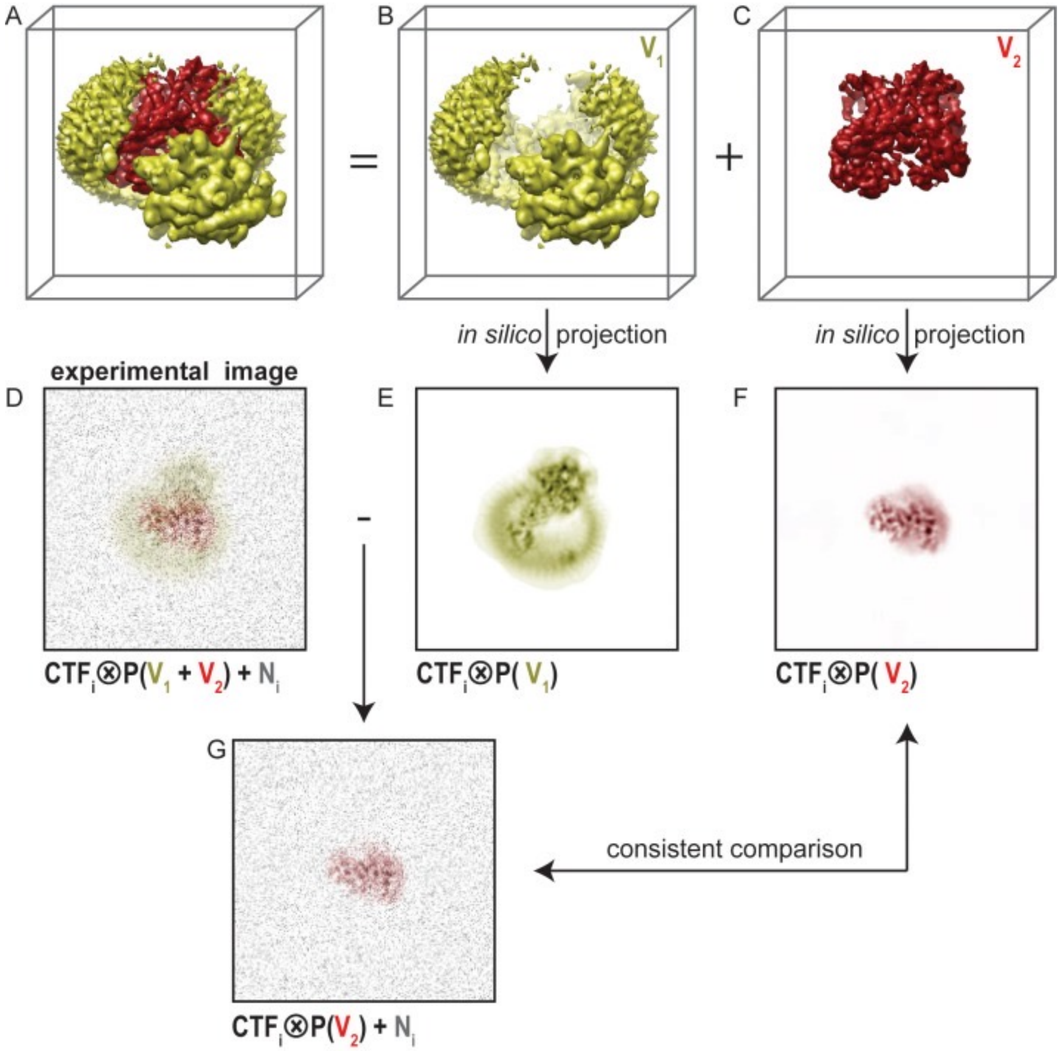
- Classifying previously aligned particles sorts heterogenous features and preserves high signal alignments

- Frequently done with masking to focus classification on a particular area of the molecule



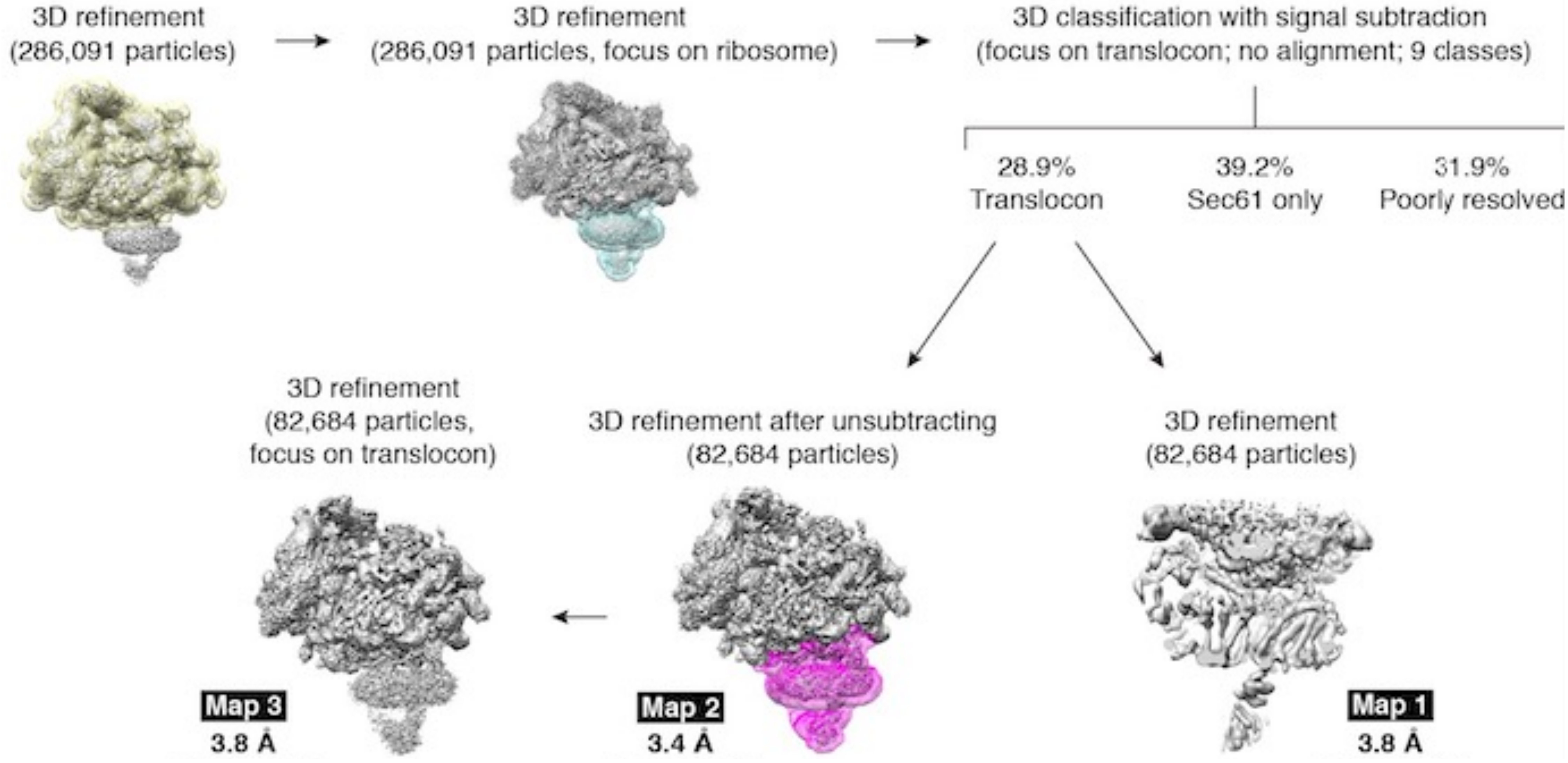
Removing high signal regions from the particles can improve alignment of low signal regions

- Masked classification contains alignment to a certain region of a reference
- High signal outside of the mask can prevent good alignment
  - Not enough signal in mask to secure strong alignment
- Can remove the high signal noise from the particles in the data set
  - Makes a same-to-same comparison reference to data, less noise to promote misalignment
- Allows for alignment of very low signal regions



(Bai et al, eLife 2015)

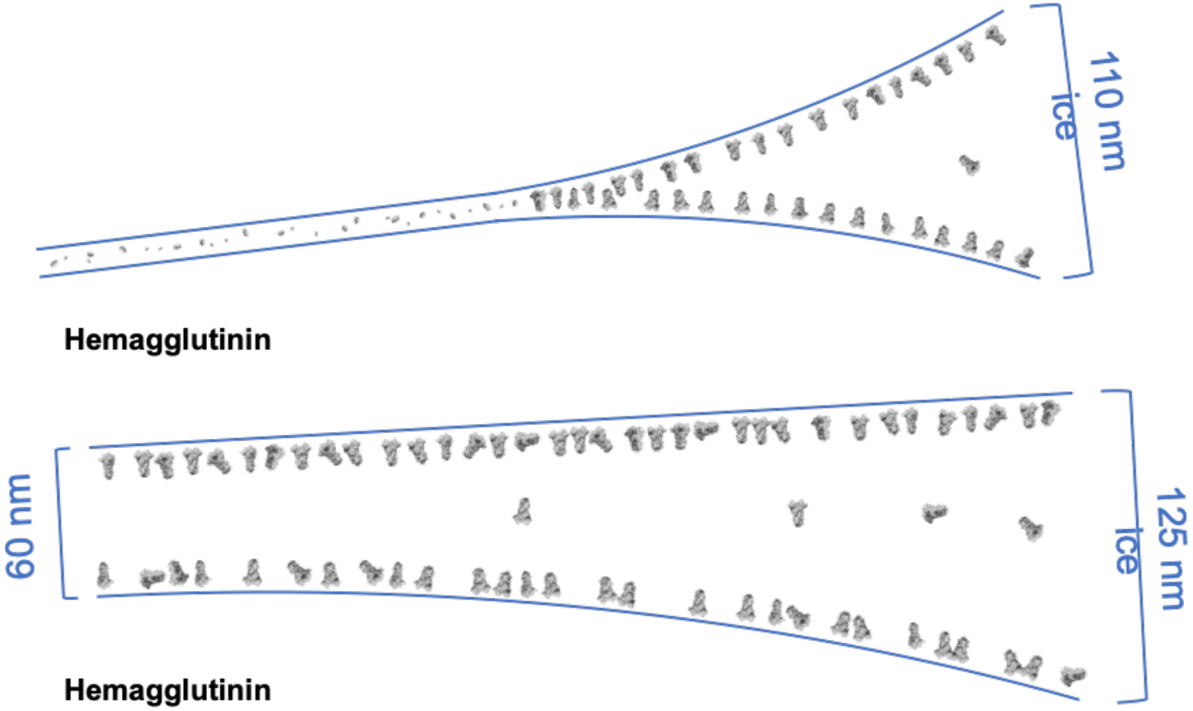
Iterative approaches – masked classification and refinement of signal subtracted particles

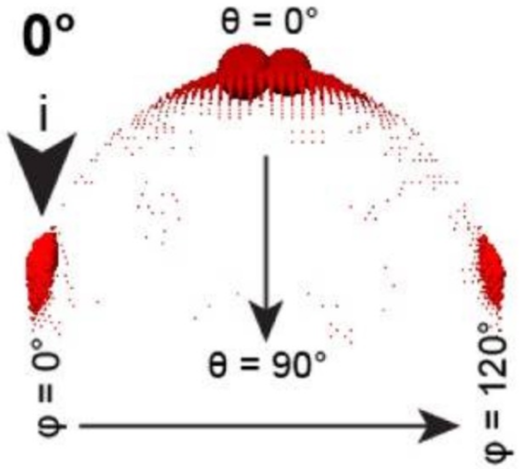


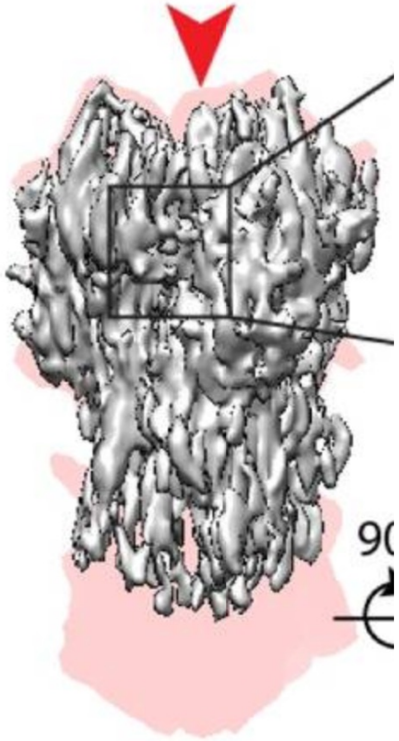
(McGilvray et al, eLife 2019)

Strong preferential orientation reduces signal and biases particle alignment

- Sometimes particles interact strongly with the air-water interface preventing them from tumbling freely
- This reduces the number of particle views available; strong signal in particular orientations
- leads to anisotropic 3D reconstruction and “stretching”



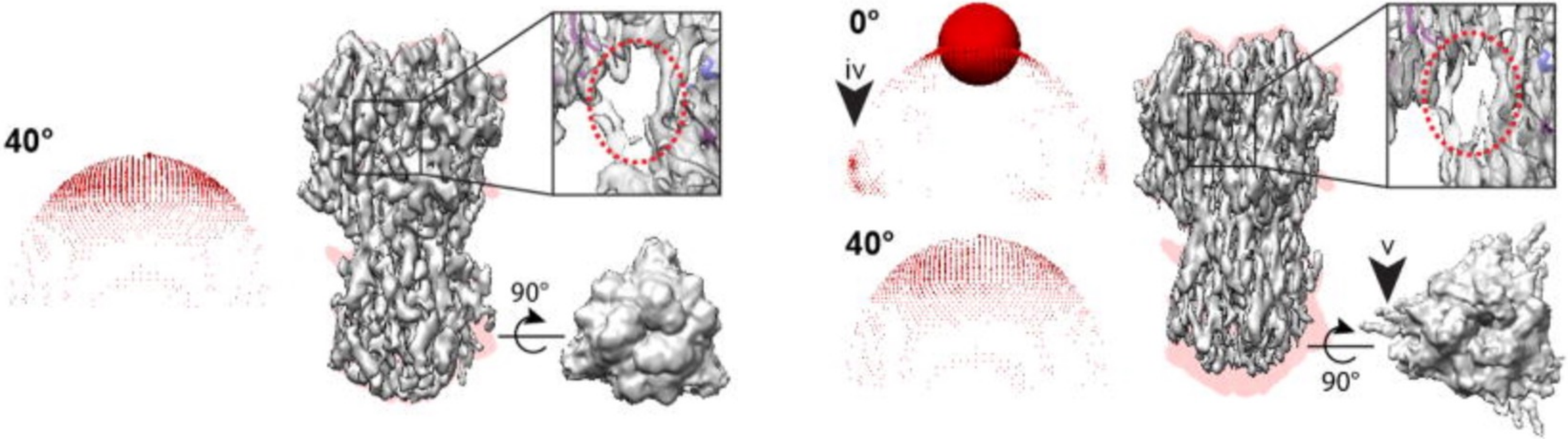




(Noble et al, eLife 2015; Tan et al, Nat Methods 2017)

Mild preferential orientation can be overcome computationally

- With mild orientation bias most views are biased but there's a significant number of other views as well
- It is possible to overcome this by:
  - Being generous with the 2D classes you keep
  - Using extensive 3D classification/heterogenous refinement to generate a batch of well aligned particles which show little bias



Tan et al, Nat Methods 2017)

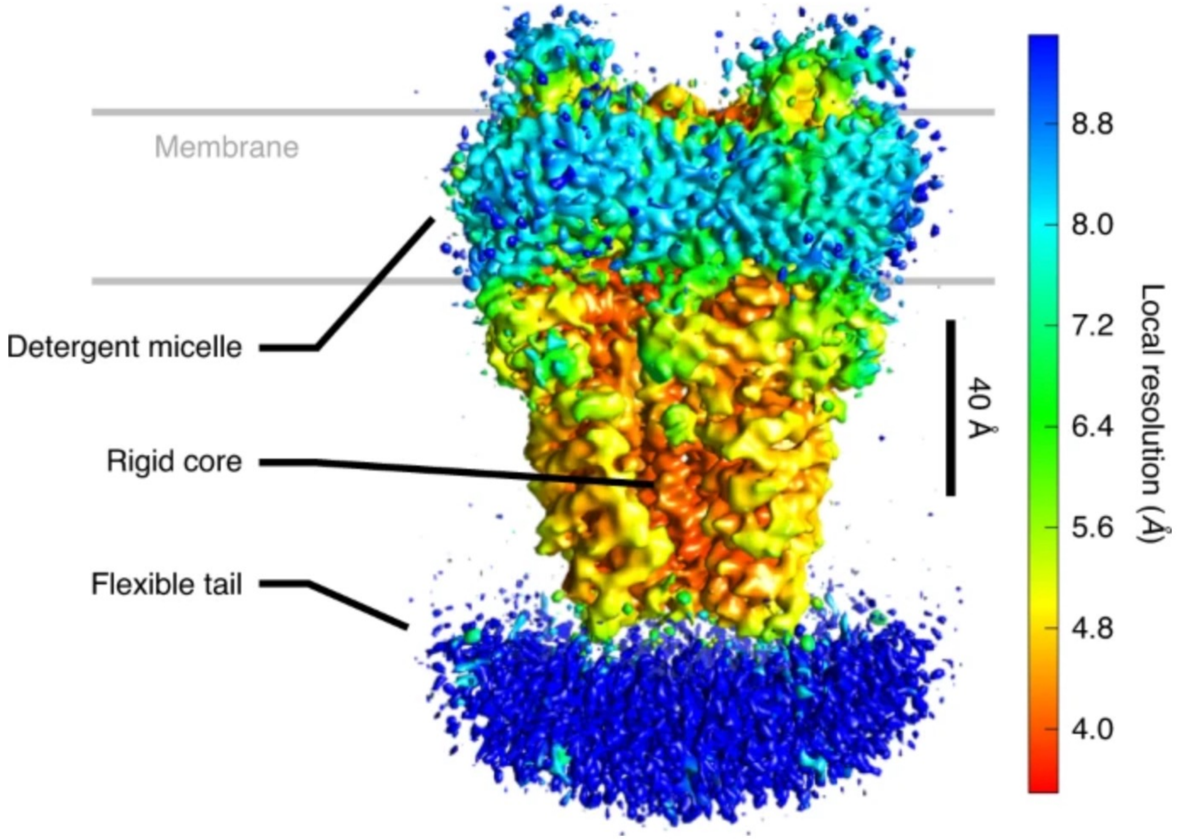
Refinement of homogenous yet dynamic particles - linear vs adaptive regularization

- After projection alignment and 3D density estimation, filter signal based on global resolution estimations to prevent overfitting to noise
  - Based on FSC

- Biomolecules are frequently dynamic and can’t be accurately described by a single resolution
  - Especially membrane proteins

- Single filters degrade potentially high-quality density to prevent over fitting poor quality density

TRPA1



(Punjani et al, Nature 2020)

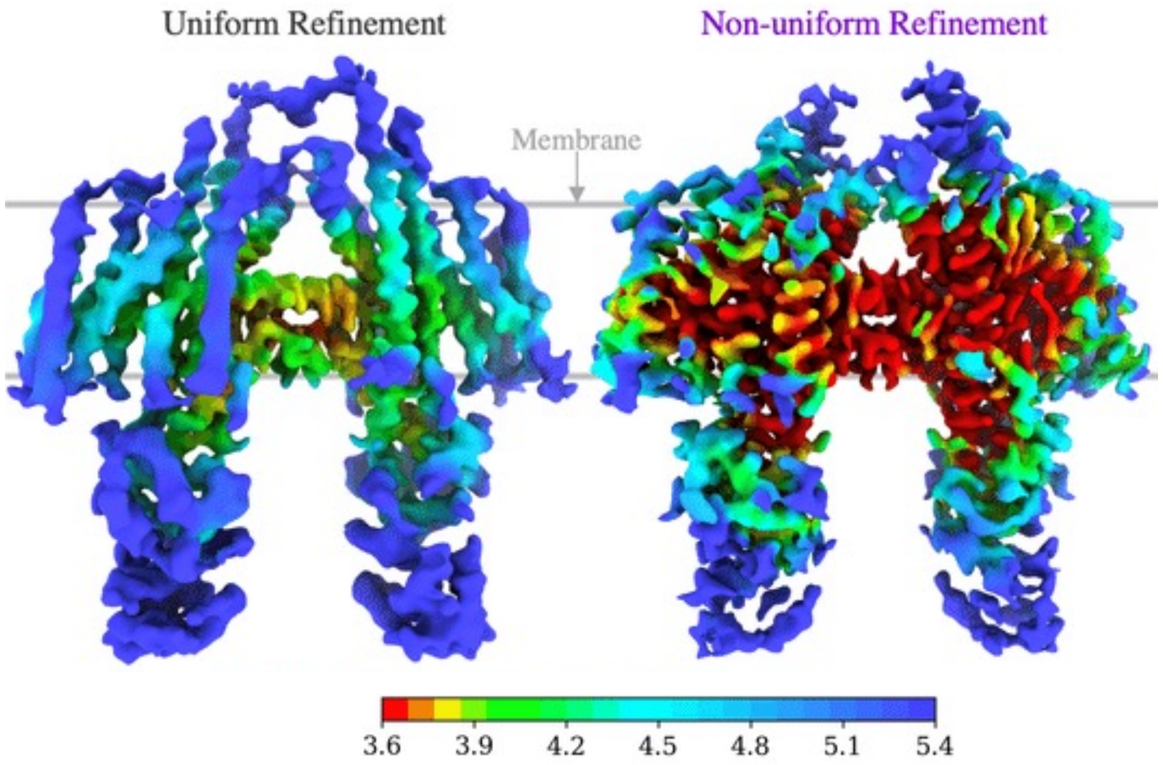
Refinement of homogenous yet dynamic particles - linear vs adaptive regularization

- Signal quality differs at different regions of a projection

- By recalculating new regularization parameters for different regions in a map during refinement we can promote high quality alignment in all regions

- Downsides
  - SLOW (2-4x time)
    - Throw more computers at it
  - Frequently unnecessary if your sample is high quality

STRA6-CaM



(Punjani et al, Nature 2020)

Single particle analysis deals with a range of challenges from the samples to the computing requirements

- Samples tend to have problems
  - Can we work through these problems computationally?
  - What methods exist to work through them?
  - Why are they necessary?

- How much is all this going to cost me?
  - Scopes, cameras
  - Hardware/software
  - STORAGE!!!!
  - Connectivity
  - Facilities and maintenance

High Performance Computing in CryoEM

- Near real-time image and data processing to support rapid target enablement and med-chem cycles
  - ~2-4 GPUs & ~10-50 CPUs per project
  - Fast local storage and high bandwidth to main storage

- Support for multiple simultaneous users at multiple sites
  - Personal computers and peripherals for each user

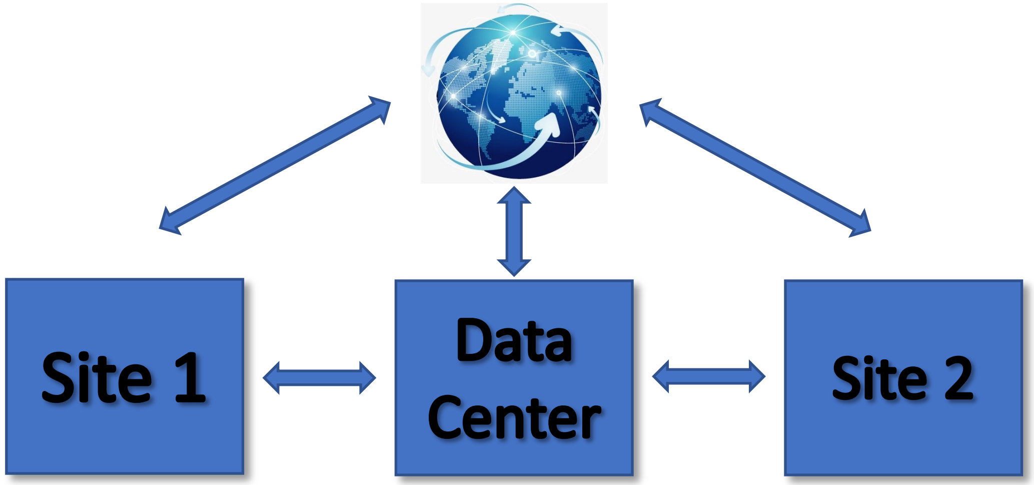
- $100,000’s for equipment
  - Servers
  - Licenses

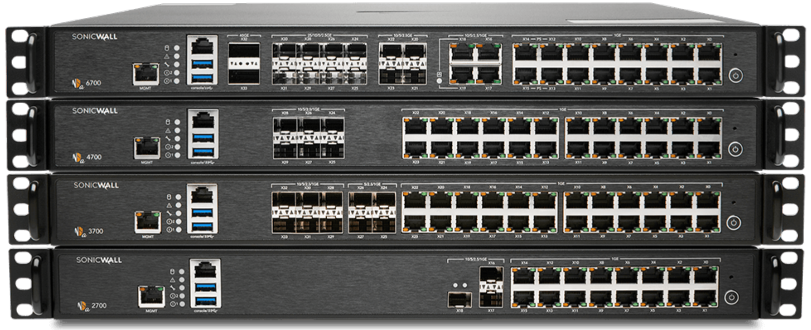
Data Storage in CryoEM

- Combined image storage and data processing
  - ~1.5 TB/day/microscope
    - ~45 TB/mo
    - ~500 TB/year
  - Average for different types of microscopes running 24/7
  - 1 PB of storage will last ~2 years/microscope
    - Ex: 2 scopes, 1 PB is enough for 1 year
    - NIS has 7 microscopes
  - Real time processing roughly doubles the amount of storage taken up by the collection alone

Networking Infrastructure in CryoEM

- High speed and high capacity bandwidth required to support large data transfers
- Up to several TB/day/microscope
- Multiple hardware/software firewalls
- Multiple network switches
- Public internet vs private fiber connections
- Can exceed $100,000/yr for internet access + private fiber depending on configuration





Continuing Computing Challenges in CryoEM

- Forecasting required capacity for storage and processing
- Response time on complex issues while running 24/7
- Maintaining uptime – continuity of services
- Redundancy – reducing single points of failure in cost effective manner
- Archives & disaster recovery for PB’s of data
- Faster microscopes
- Faster cameras
- Faster software

The best way to deal with your Cryo-EM problems … give NIS a call