New Challenges for

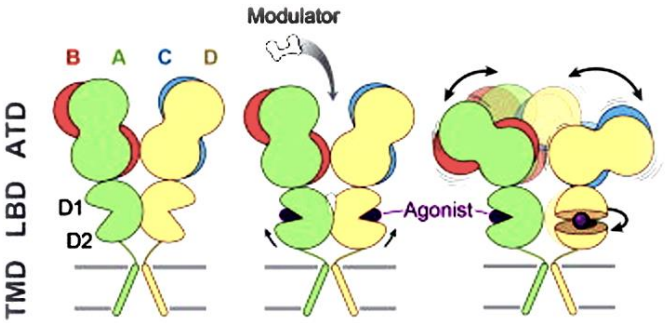
Processing

Heterogeneity

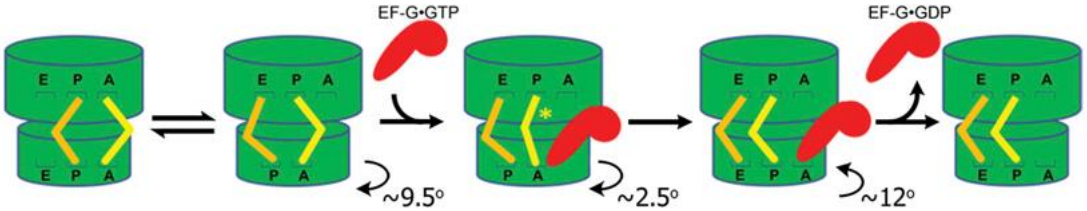
Nikolaus Grigorieff

hhmi | janelia Research Campus

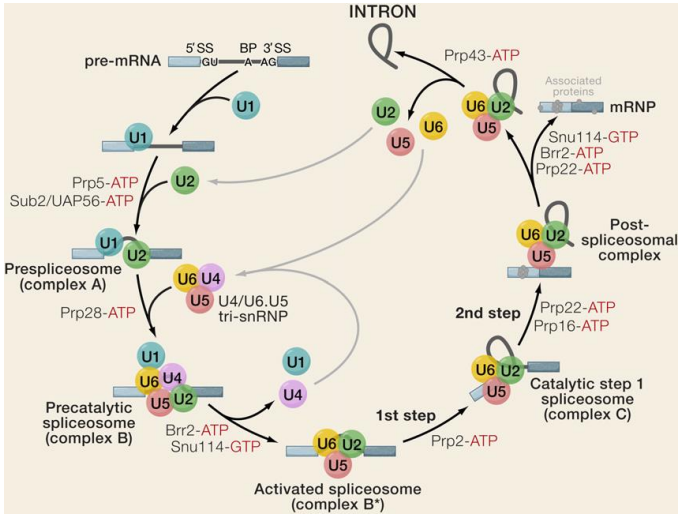
Heterogeneity and Biology



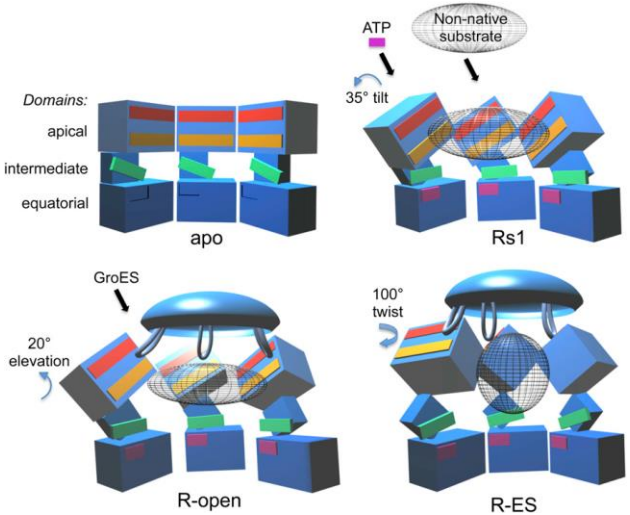
Glutamate receptor, Dürr et al 2014



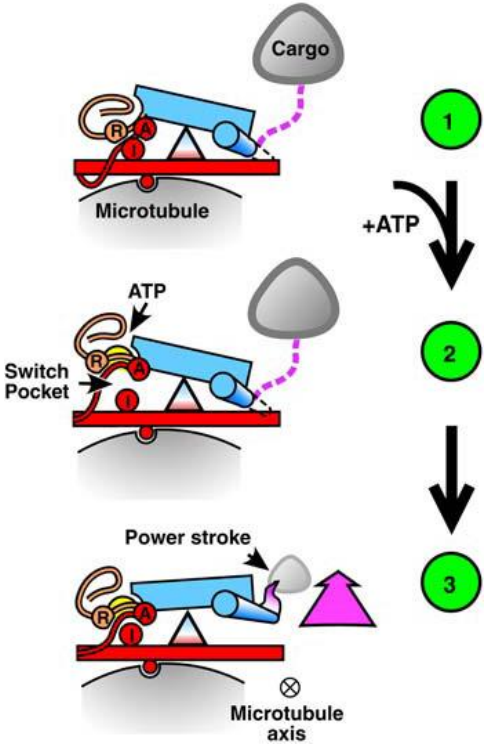
Translocation, Brilot et al 2013



Spliceosome, Wahl et al 2009



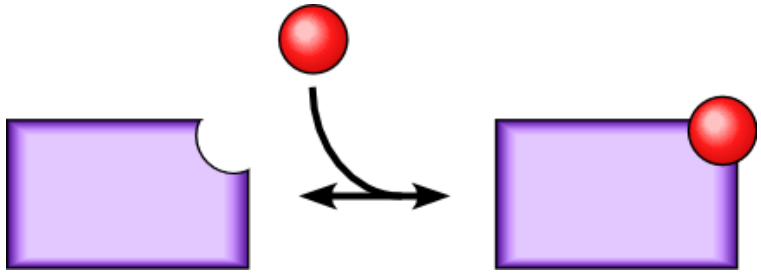
GroEL/GroES ATP cycle
Clare et al 2012



Kinesin power stroke
Sindelar & Downing 2010

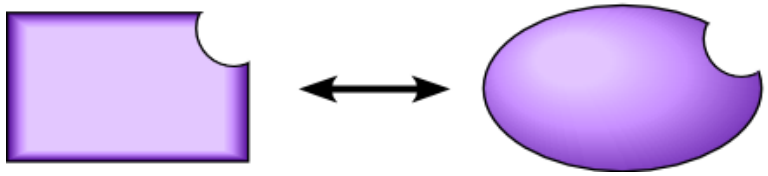
Types of Heterogeneity

Compositional

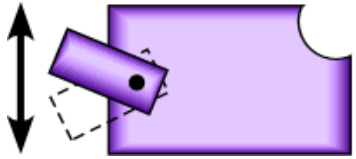


Conformational

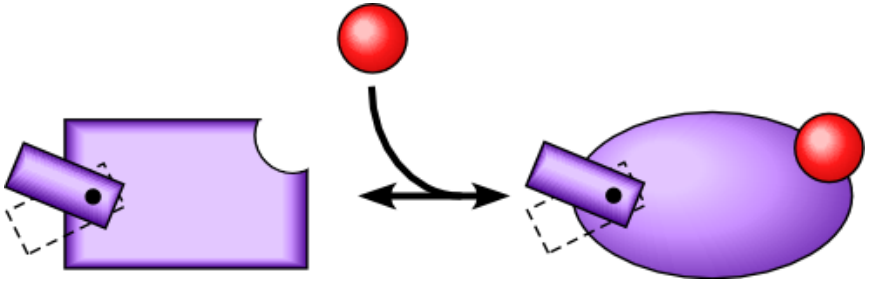
discrete



continuous

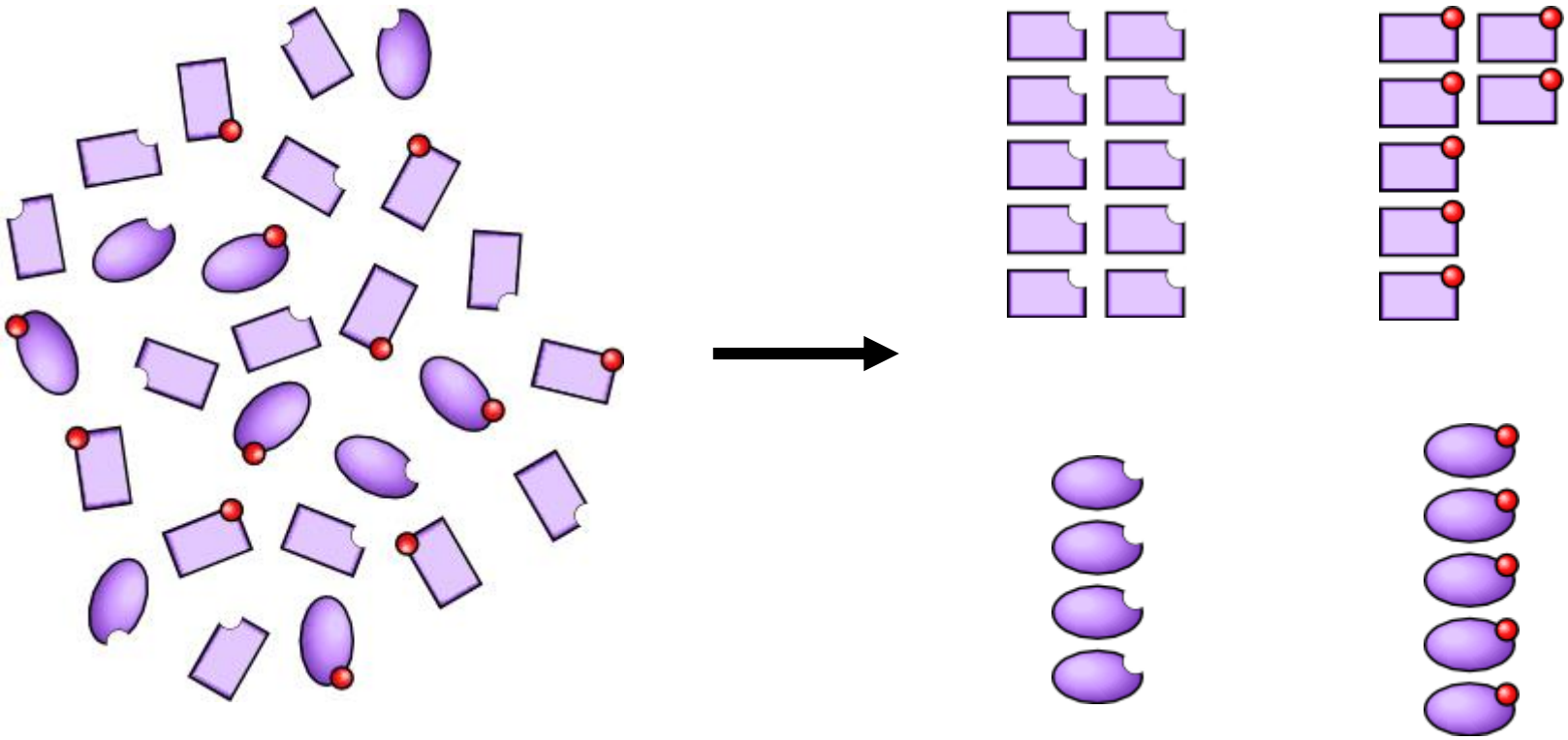


General

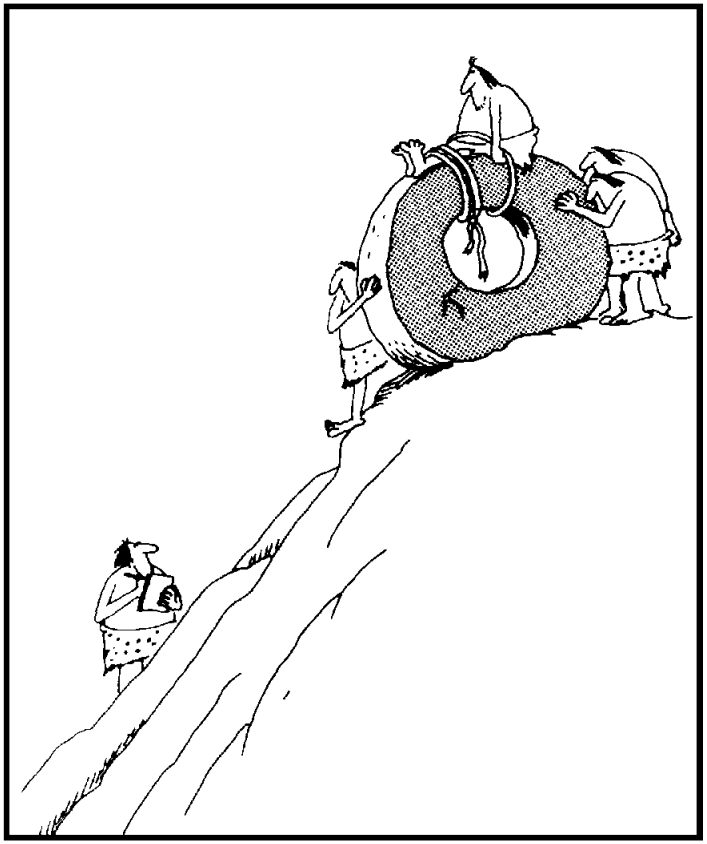


Classification Goal

Group images based on their similarity.



A Hypothetical Experiment

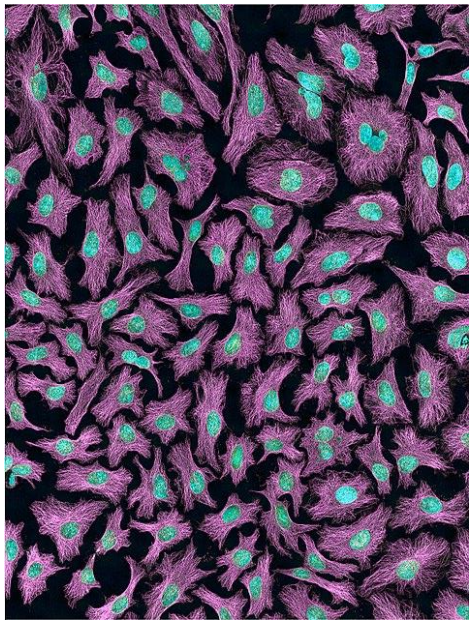


Early experiments in transportation

Larson, The Far Side

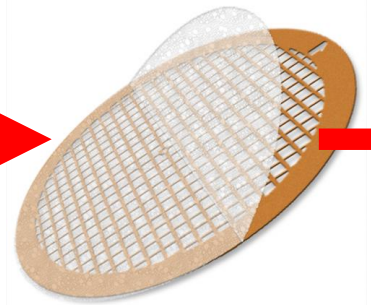
Wishful Thinking

HeLa cells

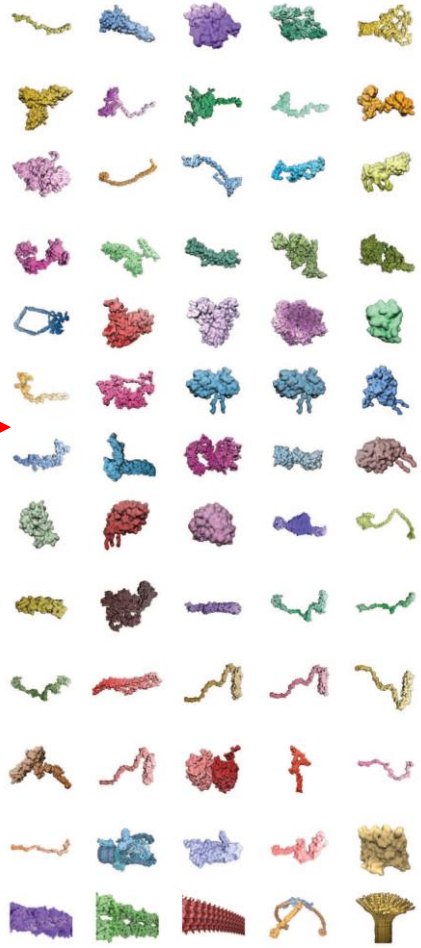


Blender

EM grid



3D structures



What are the challenges?

commons.wikimedia.org

amazon.com

emresolutions.com

Wilhelm et al. 2014

Challenge: Size of Dataset

- Assume 1000 different molecular species with M_w > 100 kDa
- Assume linear histogram with maximum concentration difference of 100-fold
- Require minimum of 30,000 particles per species

➢ Required dataset: 1000 x 100/2 * 30,000

= **1.5 billion particles**

Challenge: Processing Time

- Assume 1.5 billion particles
- Assume $n \log n$ dependence on particle number (fast sorting), 8h/7h for 2D/3D classification of 130,000 particles

➢ 2D classification: **19 years**

➢ 3D classification: **17 years**

Challenge: Small Classes

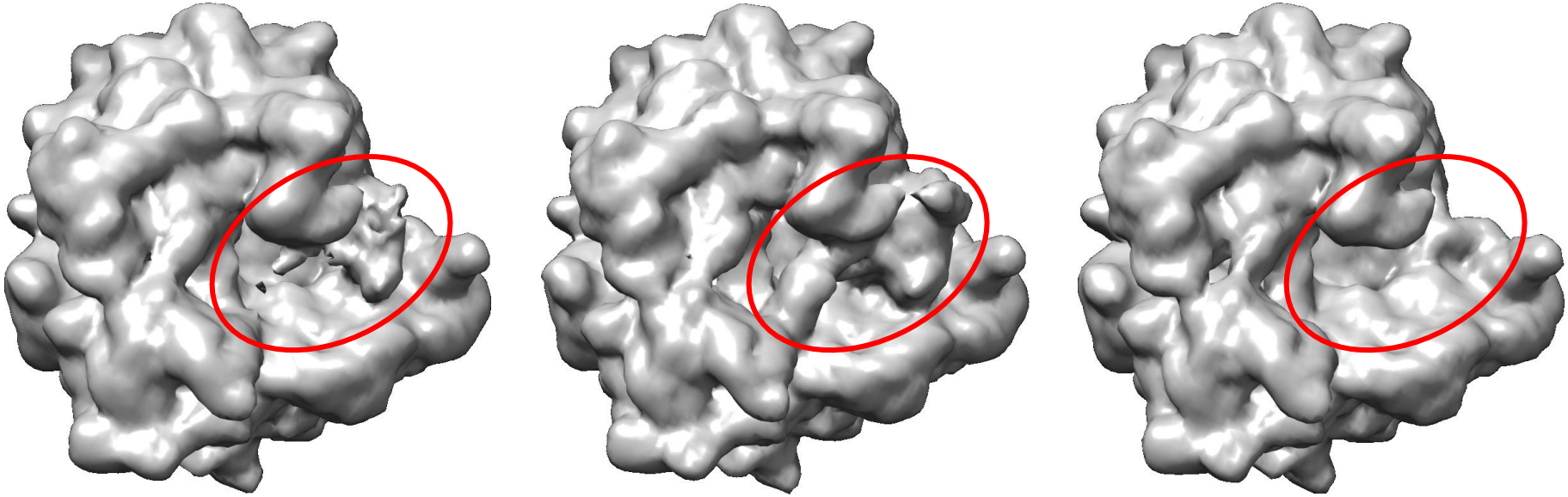
- Assume that smallest population is 100x smaller than largest population
- Larger classes tend to ‘attract’ particles from smaller classes (Yang et al. 2012, ISAC)

➢ Detectability will depend on size & shape of molecule/complex

➢ Particles may be discarded in 2D classification that might be assignable in 3D

Challenge: Convergence

- Incomplete separation of classes



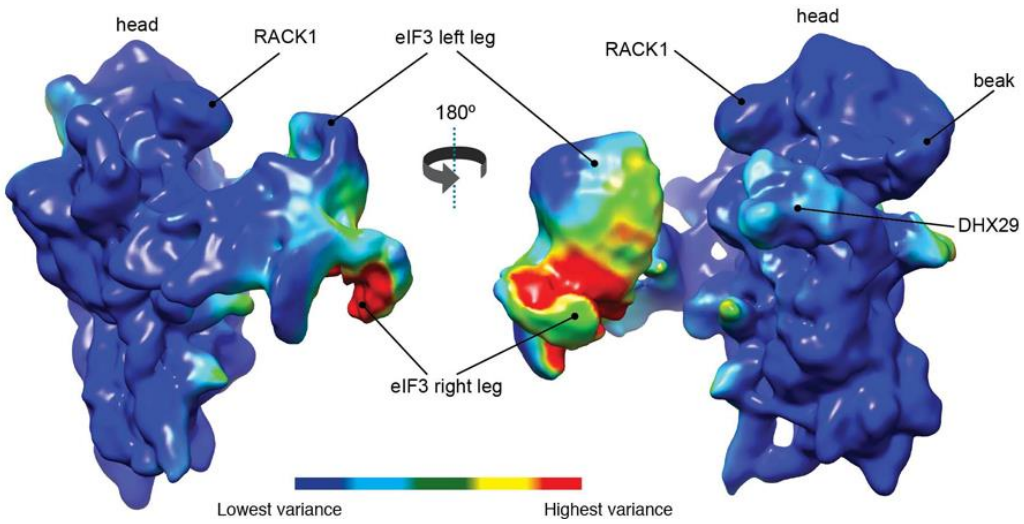
6.4% 2.4% 3.3%

70S ribosome + EF-G

Brilot et al. 2013

Challenge: Detection

40S ribosomal subunit bound to CSFV-IRES, DHX29 and eIF3



- Computationally expensive
- Very sensitive to particle misalignments
- Noisy/low resolution

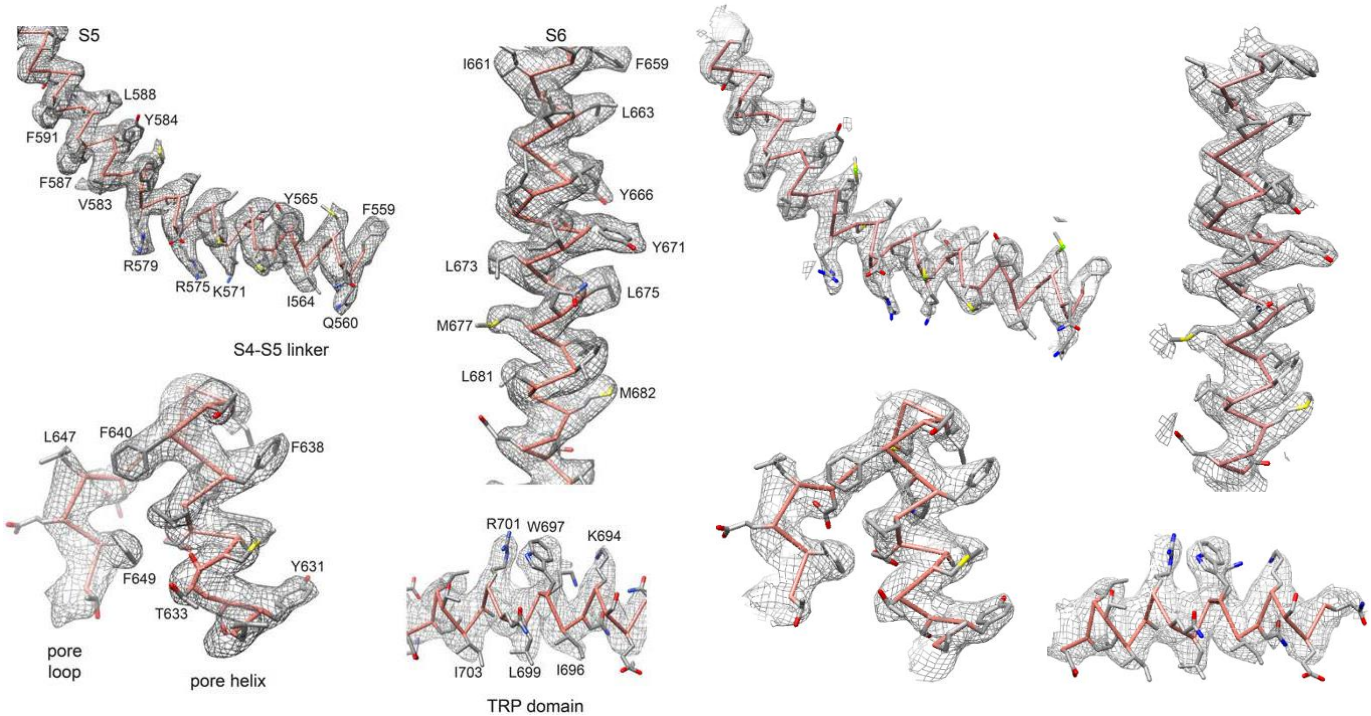
26317 particles (one class out of 630k particles)
40k bootstrap volumes

Hashem et al. 2013

Challenge: Reproducibility

TRPV1 channel

Dataset:
88915 particles
(300 kV, K2)



<u>Relion</u>
Refinement & classification
35645 particles (40%)

<u>Frealign</u>
Refinement & classification
38326 particles (44%)

Overlap: 23230 particles (~60%)

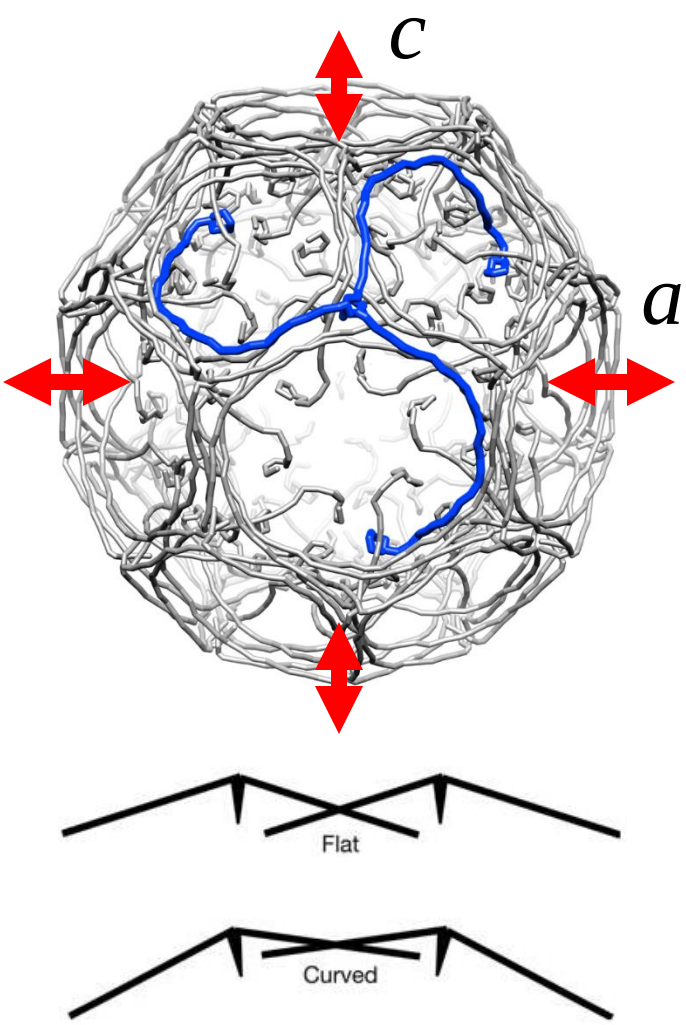
Liao et al. 2013

Challenge: Interpretation

- Current techniques classify pixels, not features
- Classes may still be mixtures
- States may be missing
- Results are irreproducible

➢ Structural interpretation may be difficult

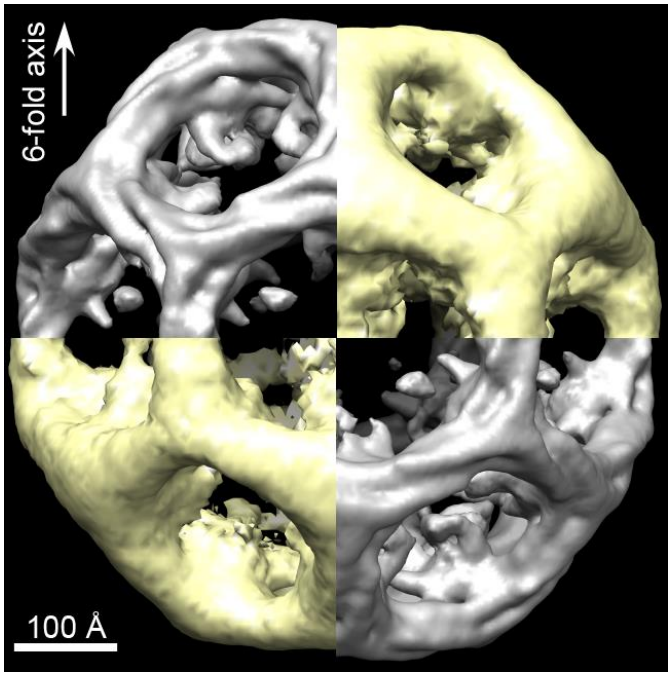
Challenge: Continuous States



$$\mathbf{Q} = \begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & c \end{pmatrix}$$

Clathrin cage

bound to auxilin and Hsc70

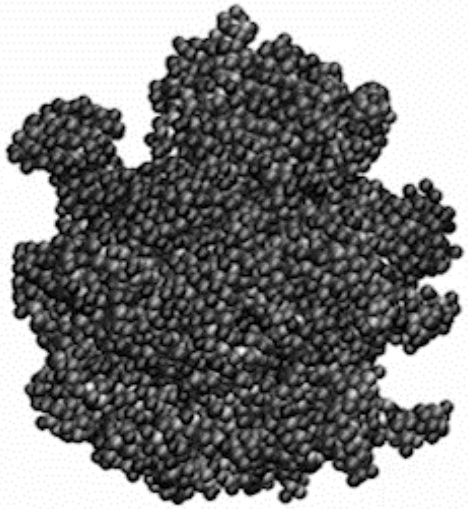


Model	FSC at 22 Å (σ = 0.016)
$\Delta a = -\Delta c$ const. surface	0.157
$\Delta a = -\sqrt{\Delta c}$ const. volume	0.145
No deformation	0.107
$\Delta a = 5\Delta c$	0.108

Fotin et al. 2004, Xing et al. 2010

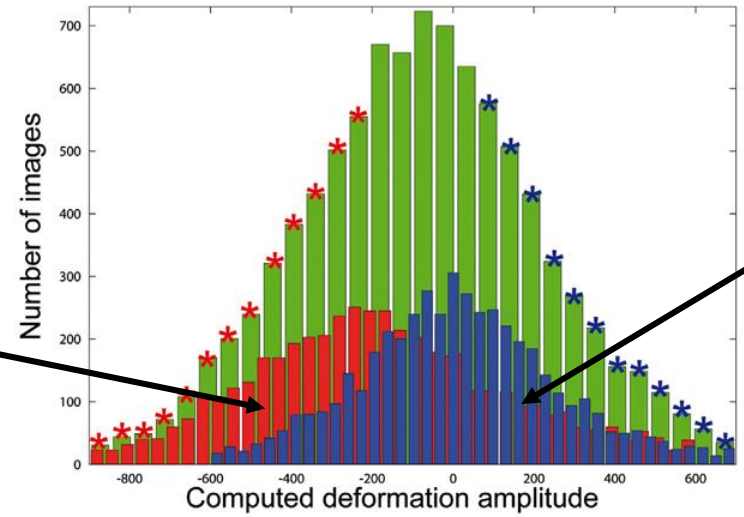
Normal Modes

70S ribosome + EF-G

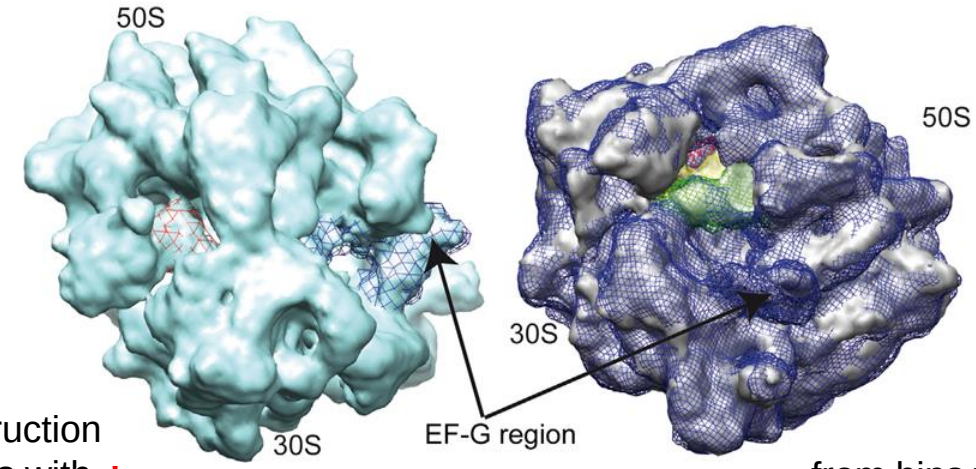


Normal mode corresponding to ratcheting

70S ribosome + EF-G (rotated)



70S ribosome (non-rotated)



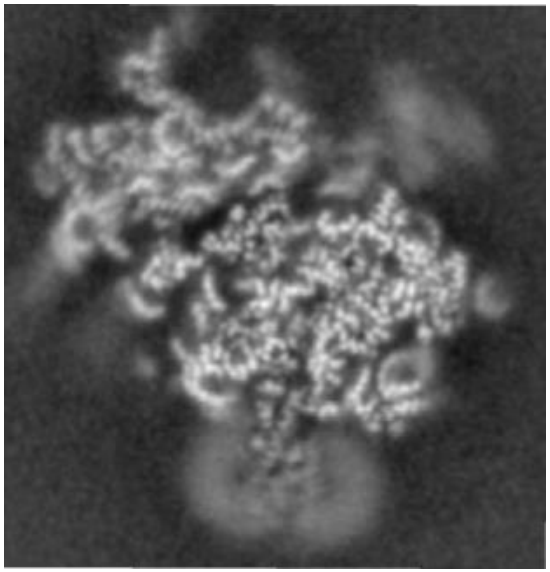
Reconstruction from bins with *

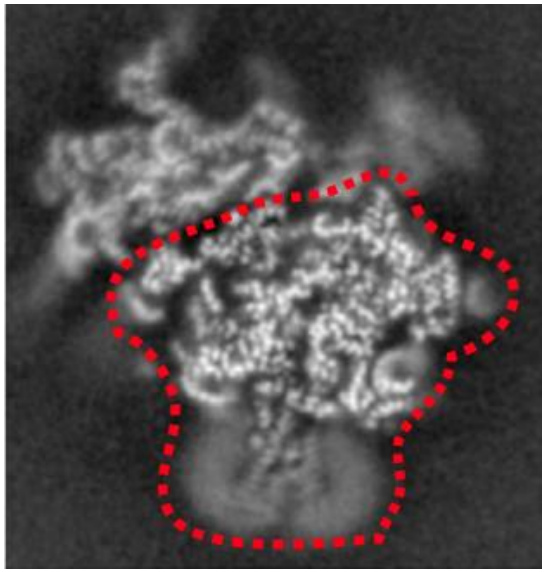
from bins with *

Jin et al. 2014

Alignment With Masks

80S ribosome + Sec61



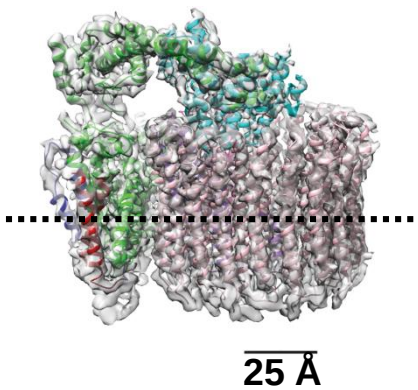


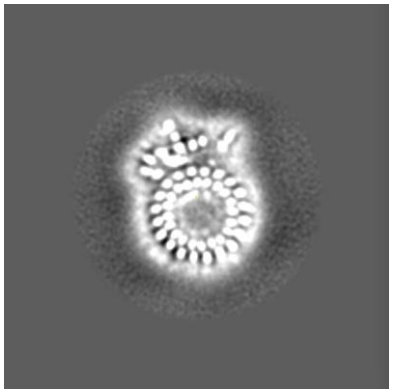
60S ribosome + Sec61

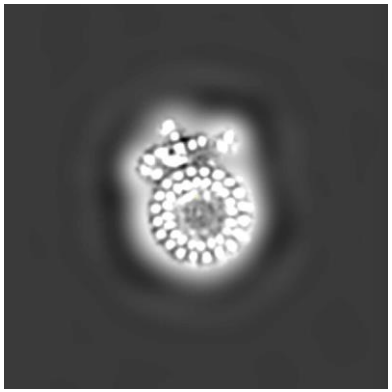
Voorhees et al. 2014

Masking And Filtering

V_O motor of a eukaryotic V-ATPase



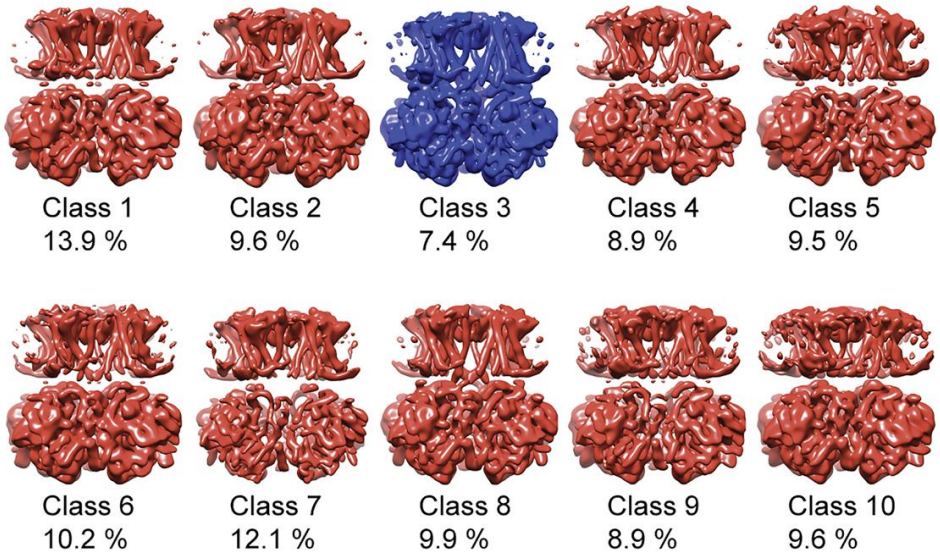


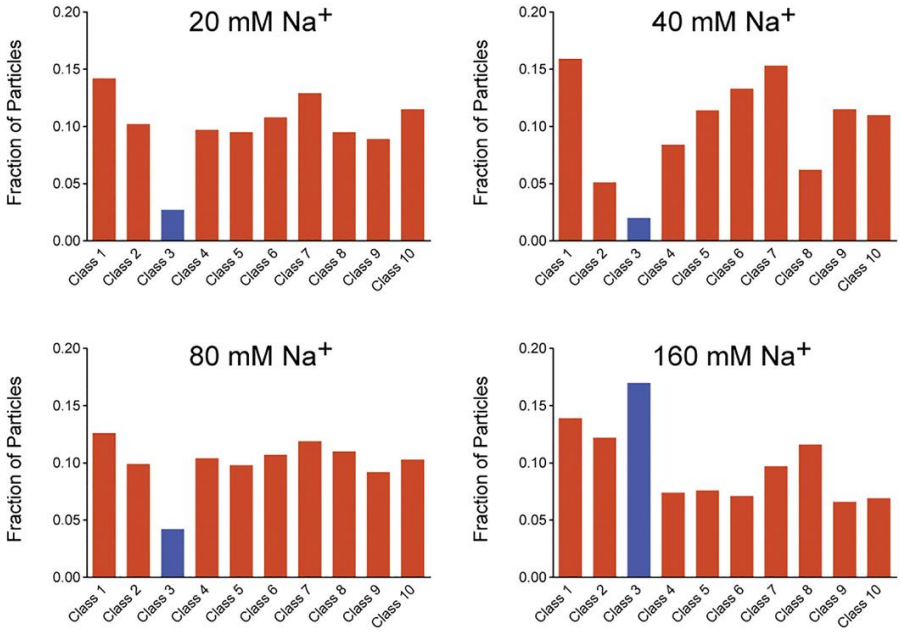


Mazhab-Jafari et al 2016

Structural Dynamics

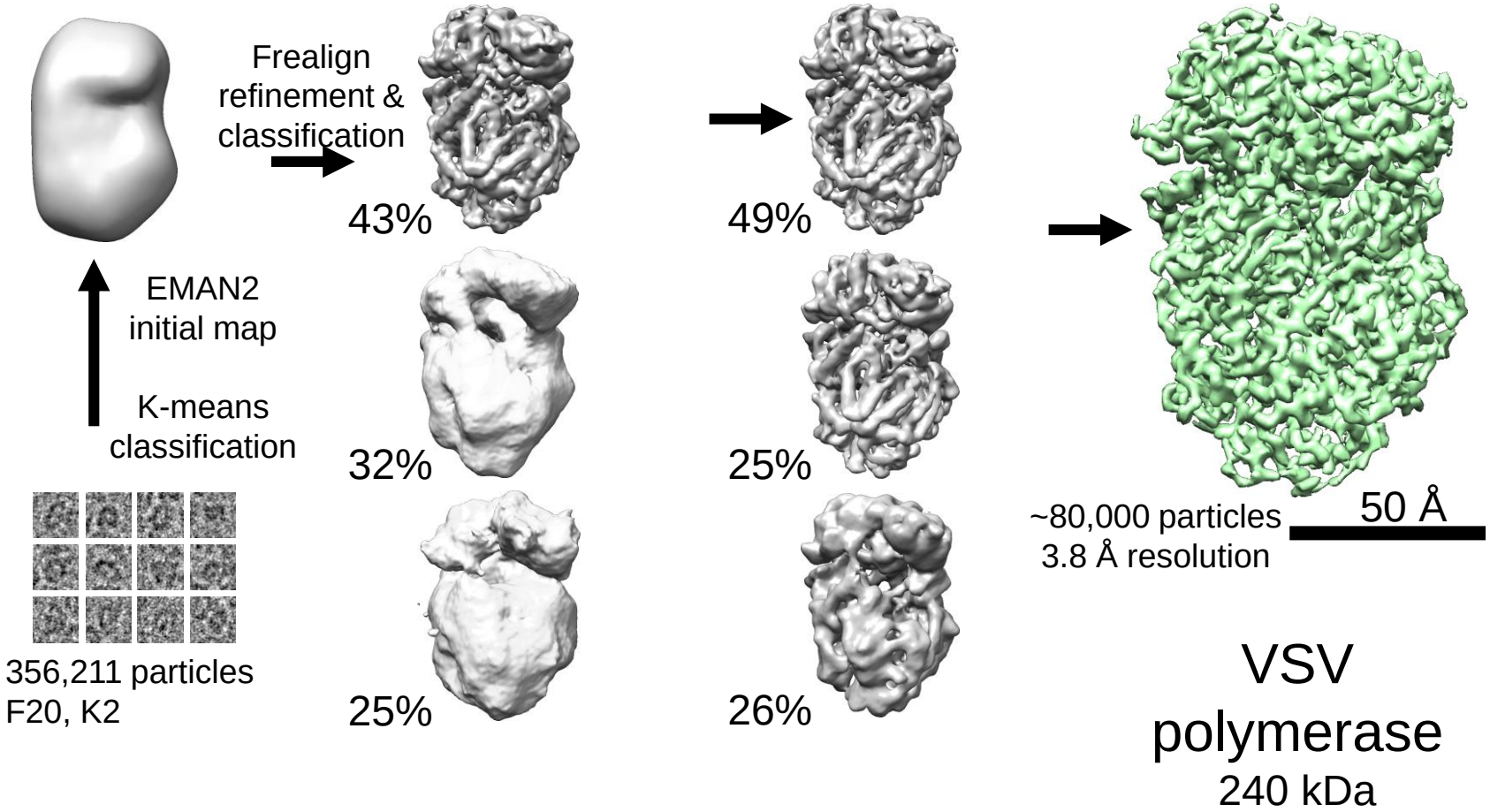
Slo2.2, a Na^+-dependent K^+ channel





Hite & MacKinnon 2017

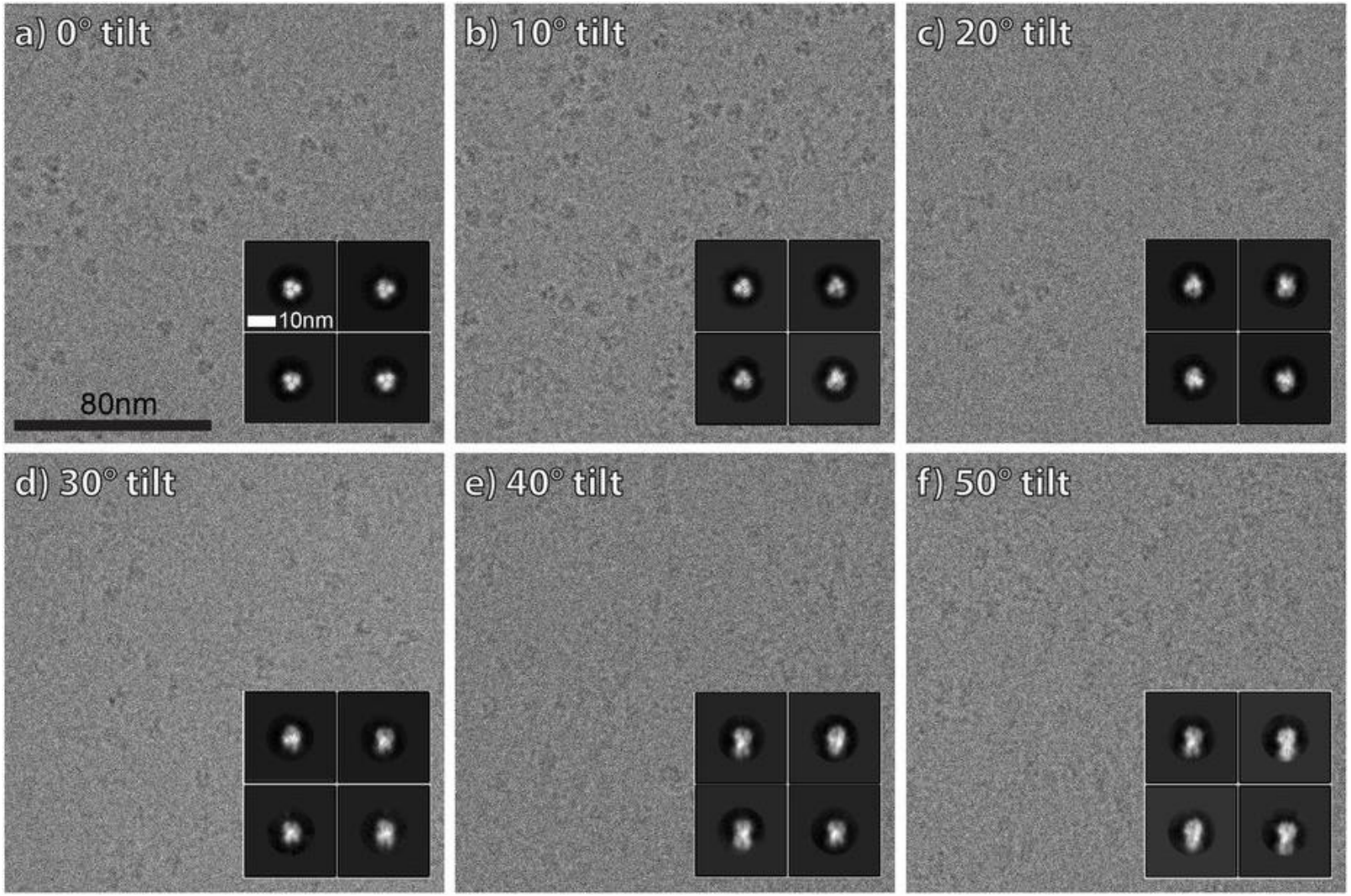
Challenge: Junk Classes



- Junk may not affect all classes equally

Liang et al. 2015

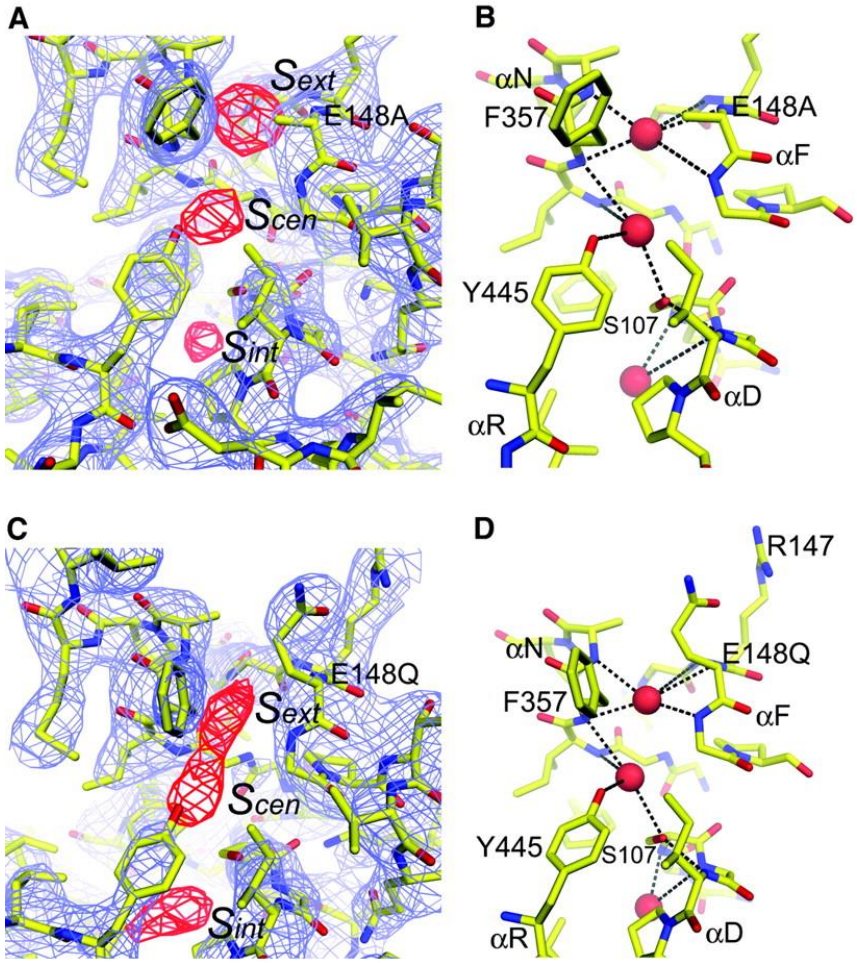
Challenge: Preferred Views

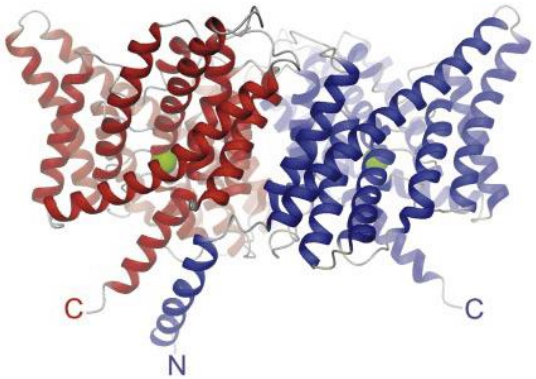


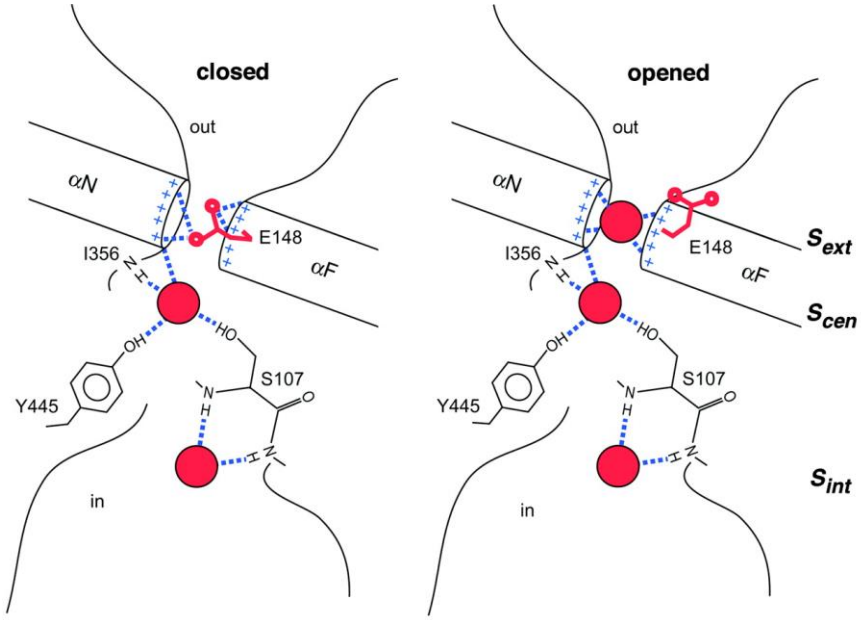
Tan et al. 2017

Challenge: Small Changes

Prokaryotic ClC Cl^- channel

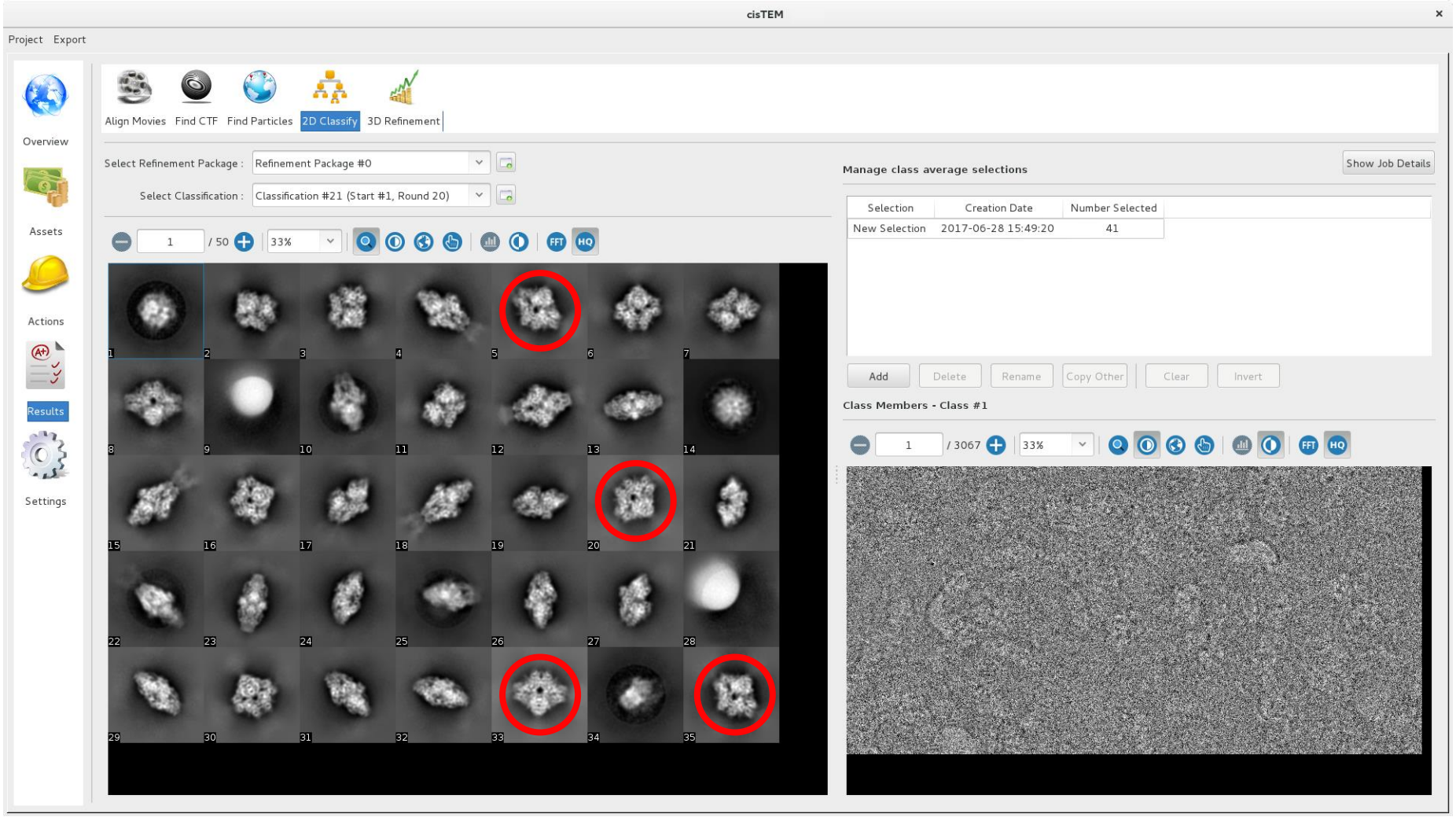






Dutzler et al. 2002/2003

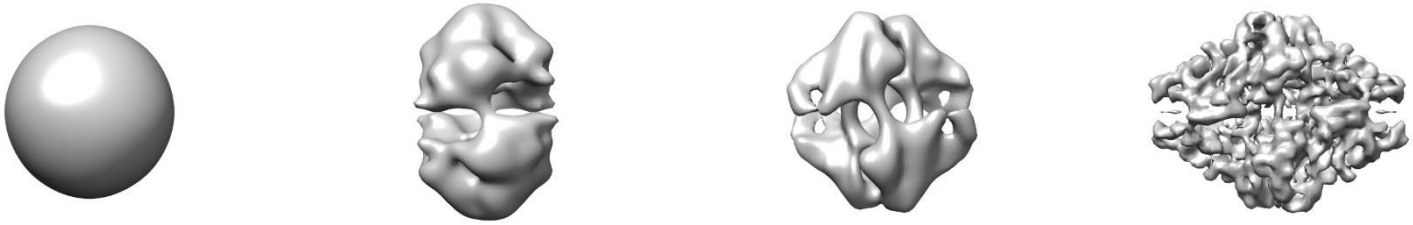
Challenge: Number of Classes



Grant, Rohou & Grigorieff

Challenge: Ab-Initio 3D

D2
460
kDa



Start | Cycle 9 | Cycle 17 | Cycle 40 | **0.7 h**

C1
240
kDa

Start | Cycle 9 | Cycle 27 | Cycle 40 | **4.2 h**

O
440
kDa

Start | Cycle 9 | Cycle 25 | Cycle 40 | **0.3 h**

Grant, Rohou & Grigorieff

Computational Resources

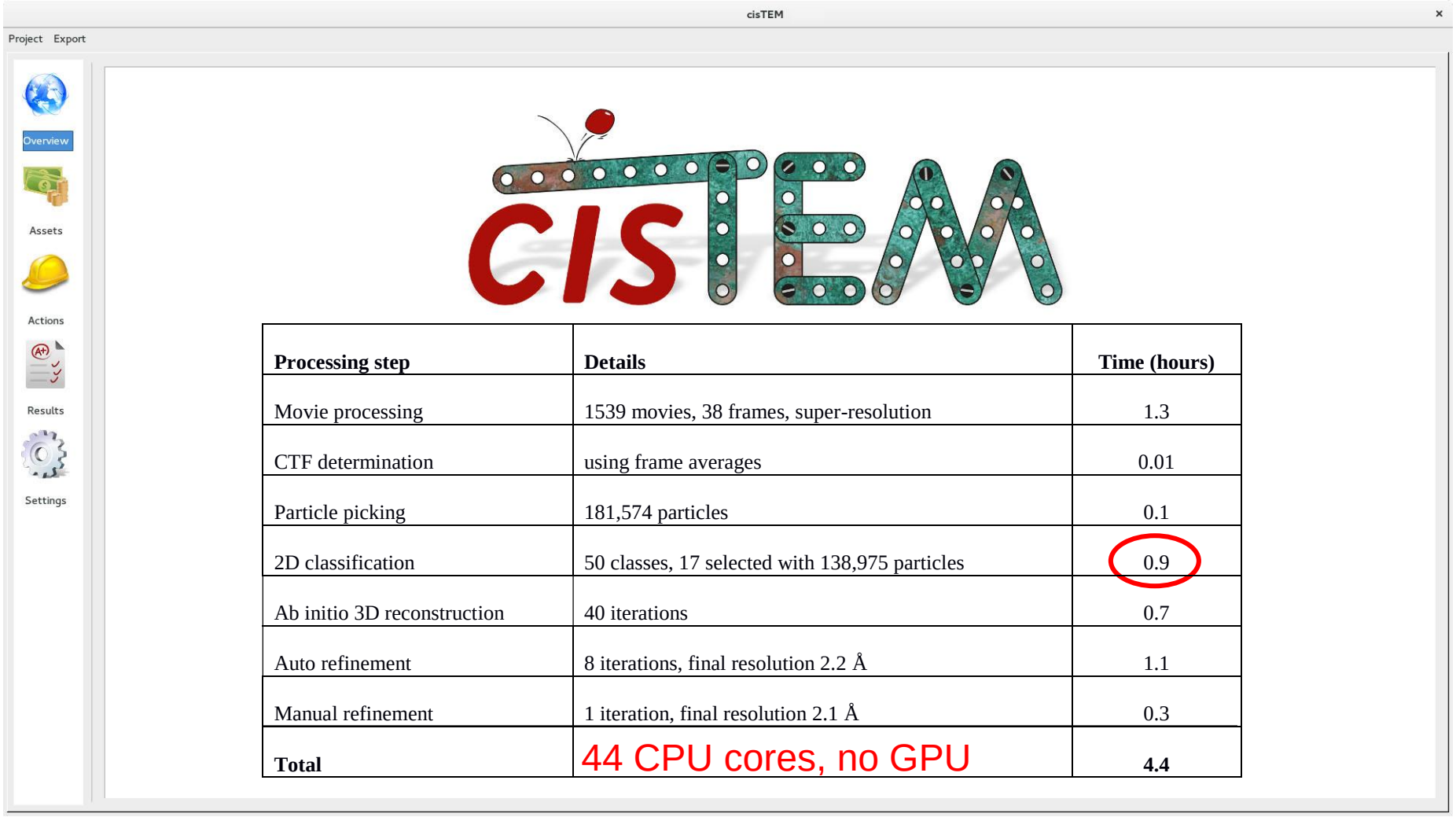
CISTEM

Computational Imaging System for Transmission Electron Microscopy

Tim Grant

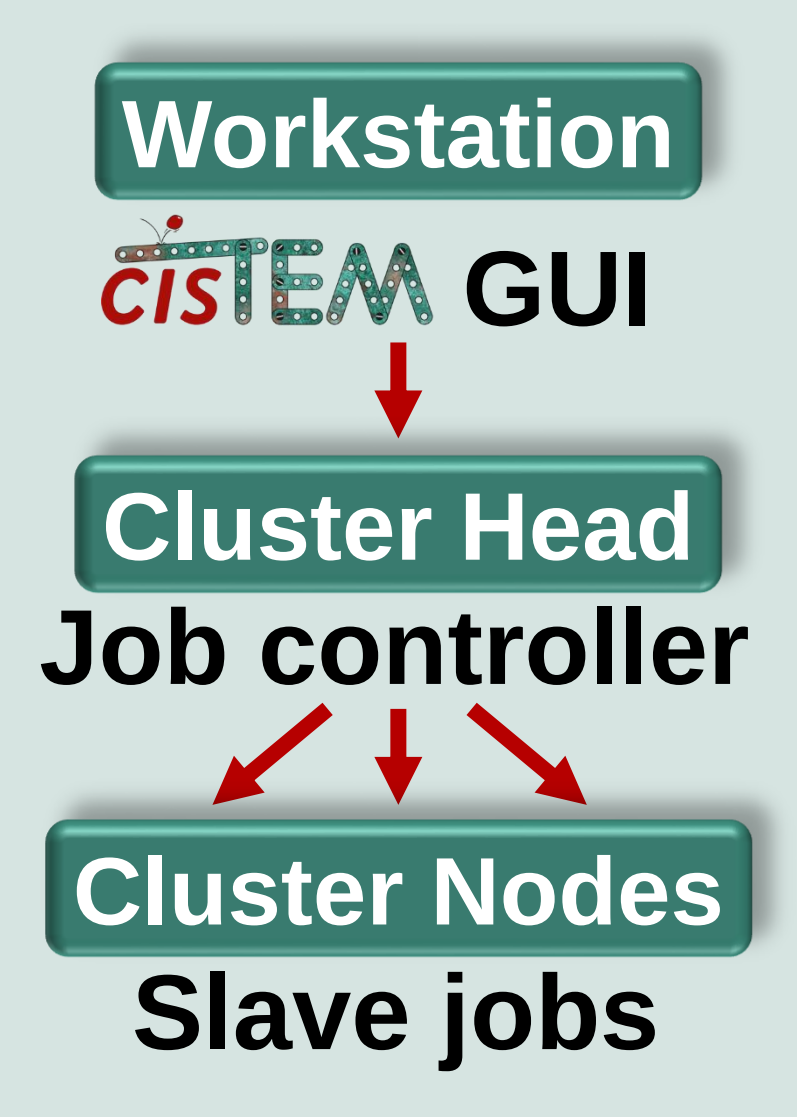
Alexis Rohou

*cis*TEM GUI



Processing step	Details	Time (hours)
Movie processing	1539 movies, 38 frames, super-resolution	1.3
CTF determination	using frame averages	0.01
Particle picking	181,574 particles	0.1
2D classification	50 classes, 17 selected with 138,975 particles	0.9
Ab initio 3D reconstruction	40 iterations	0.7
Auto refinement	8 iterations, final resolution 2.2 Å	1.1
Manual refinement	1 iteration, final resolution 2.1 Å	0.3
Total	44 CPU cores, no GPU	**4.4**

Flexible Architecture



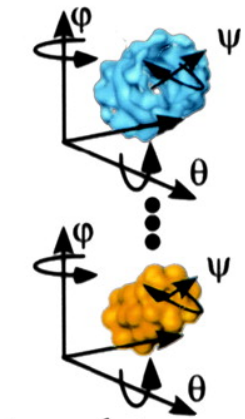
Challenge: Processing Time

- Assume 1.5 billion particles
- Assume $n \log n$ dependence on particle number, 0.9h for 2D classification of 180,000 particles on 44 CPU cores

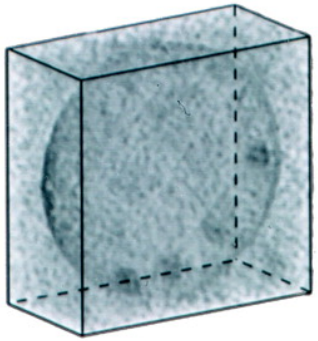
➢ 2D classification: **5 h** on 5000 CPU cores

Finding Molecules in a Heterogeneous Mess

3D Template Matching

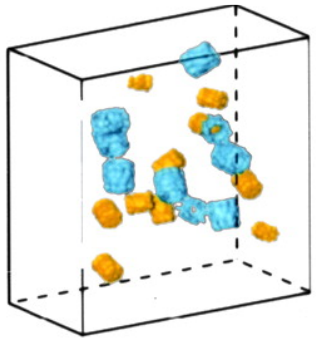


templates



tomogram

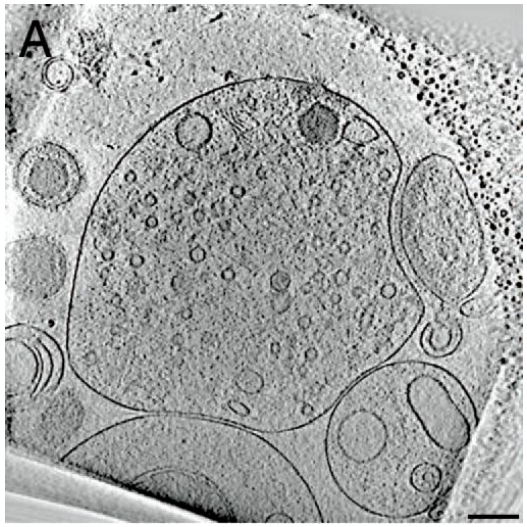
Templates match visible features

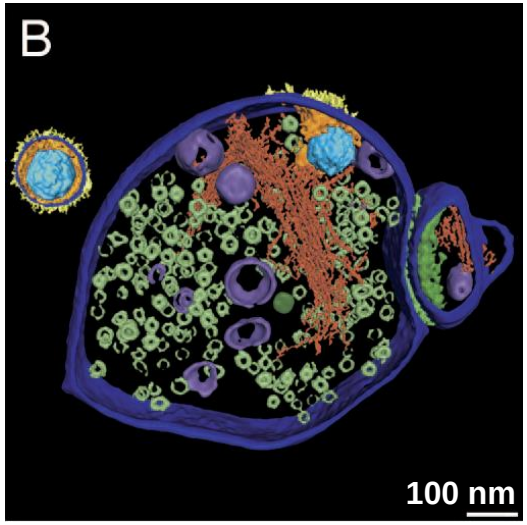


identified molecules

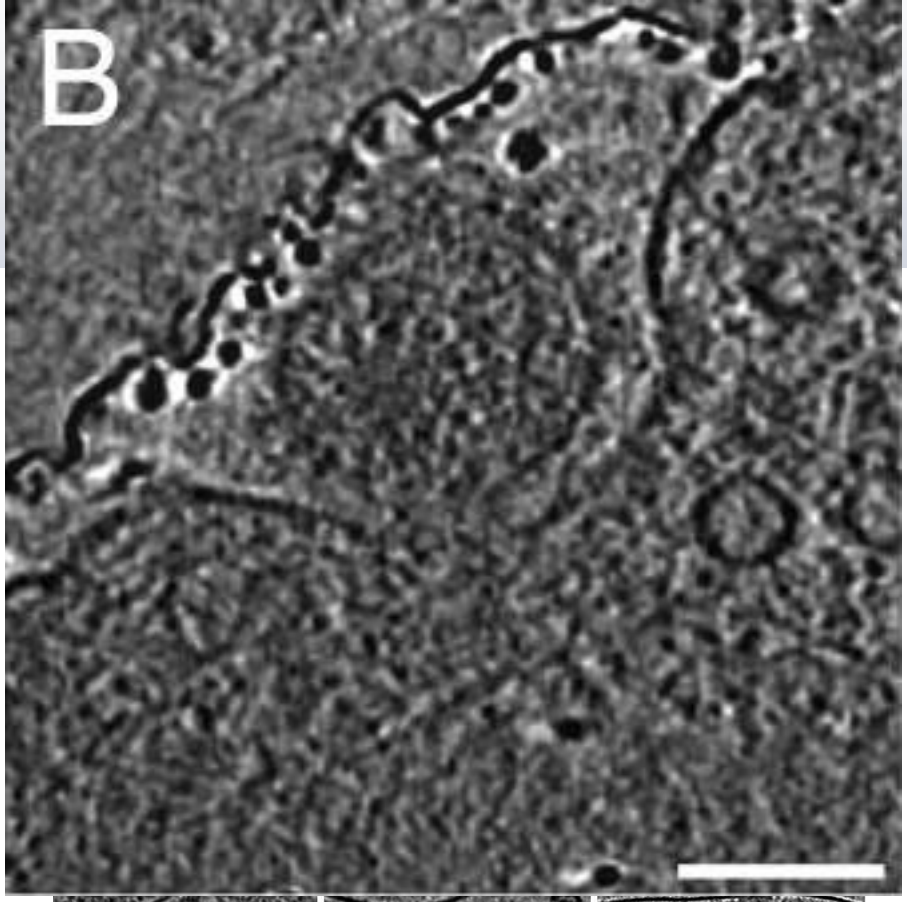
Frangakis et al. 2002

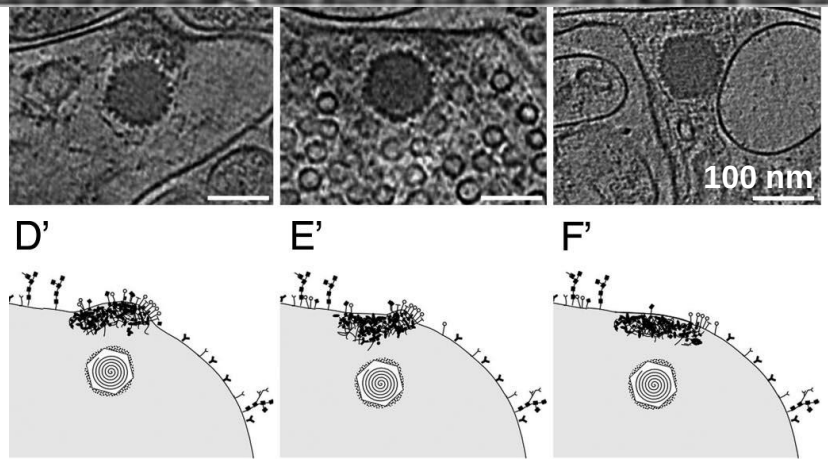
Dense





Virus
Viral tegument
Glycoproteins
Actin filaments
Synaptic vesicles
Vesicles
Synaptic cleft
Membrane





Herpes virus entering a synaptosome

Maurer et al. 2008

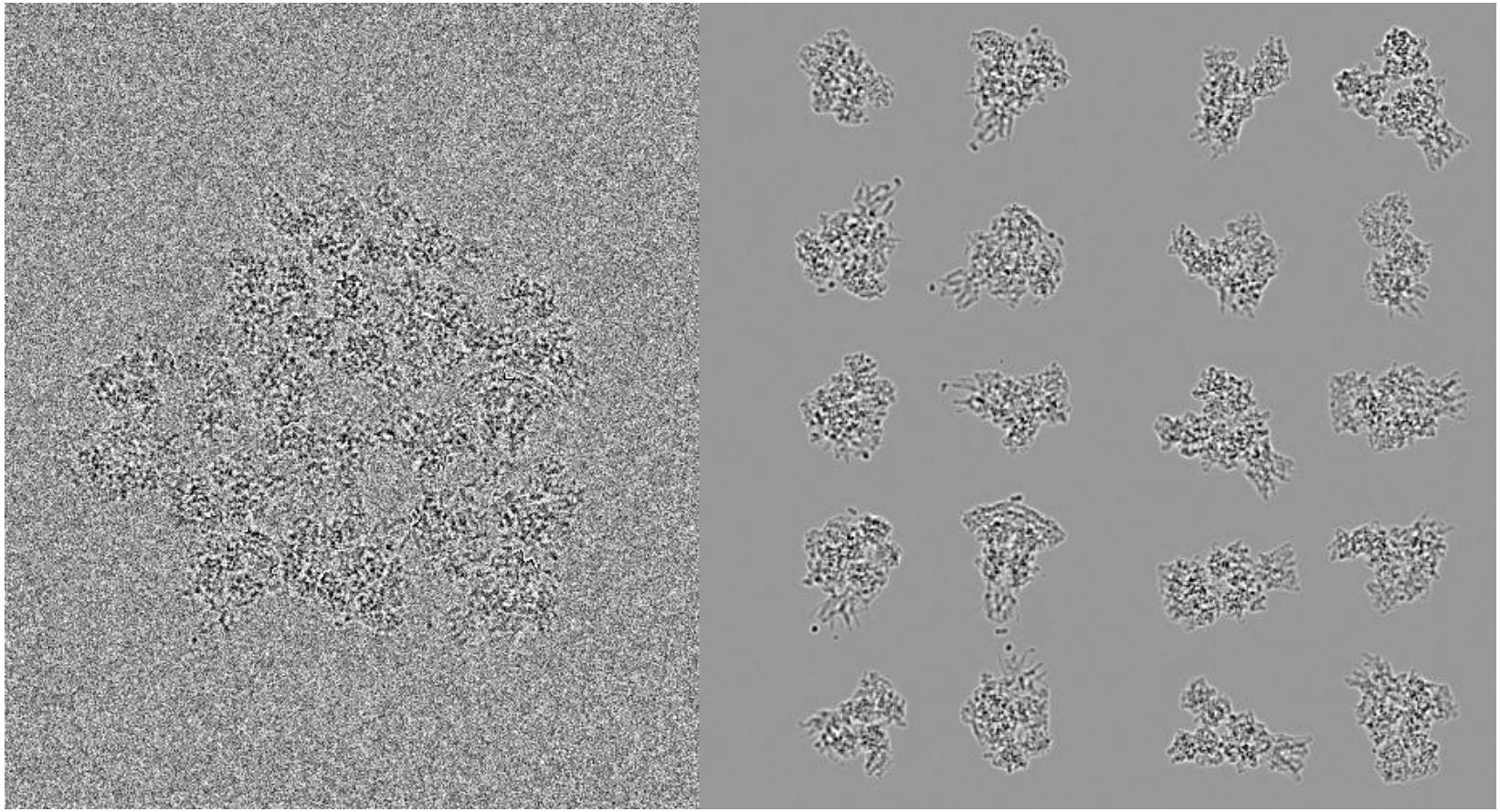
High Resolution Fingerprints

Close-to-focus cryo-EM image

High resolution EM image

NMDA receptor

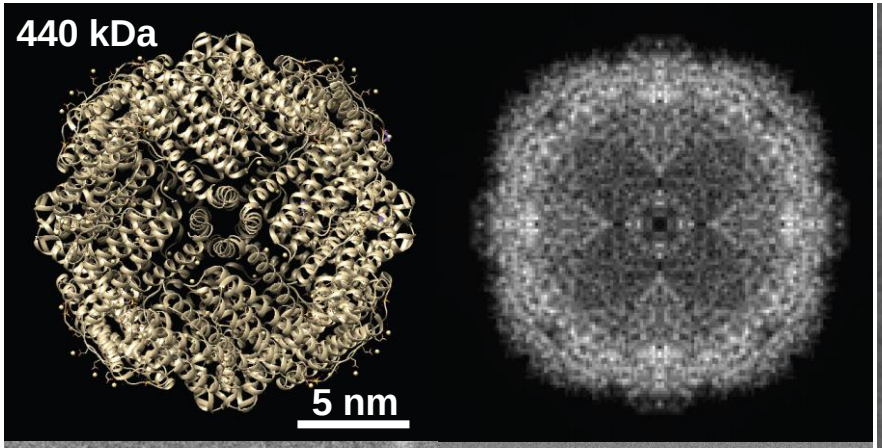
AMPA receptor

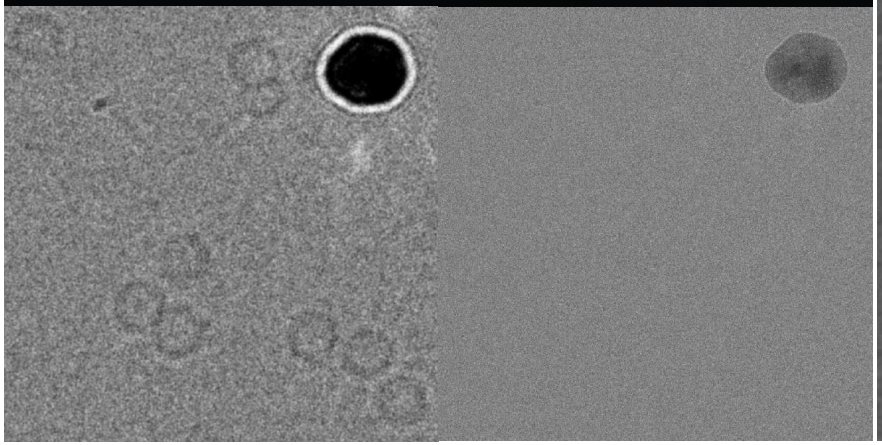


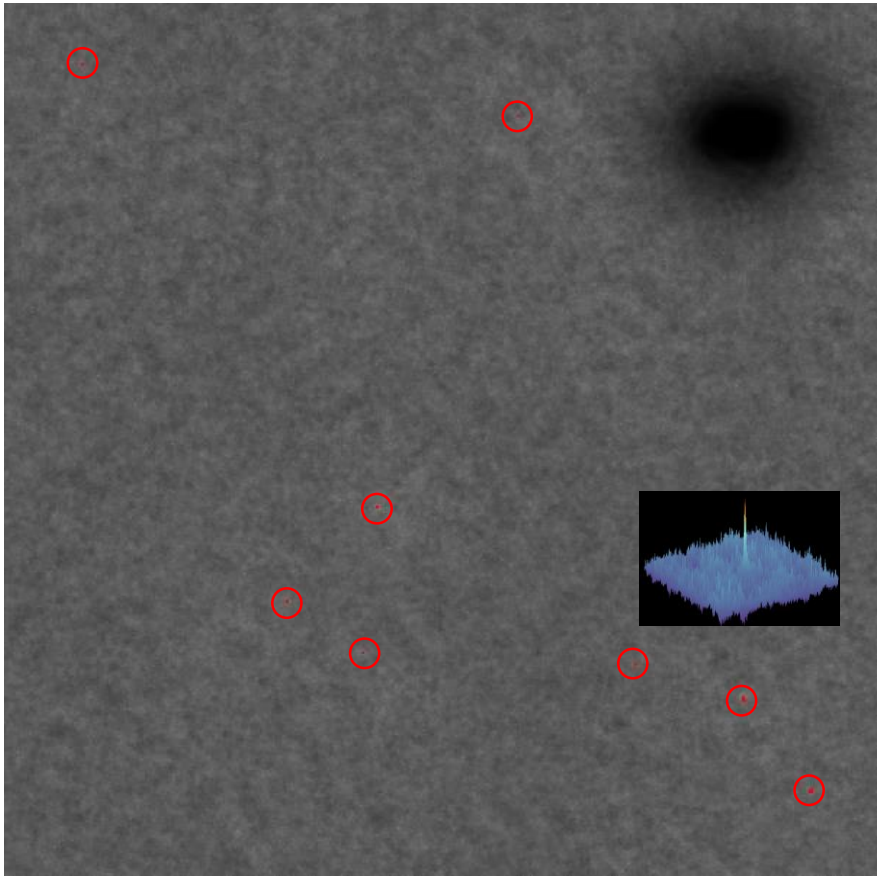
Finding Molecules

Apoferritin

Projection







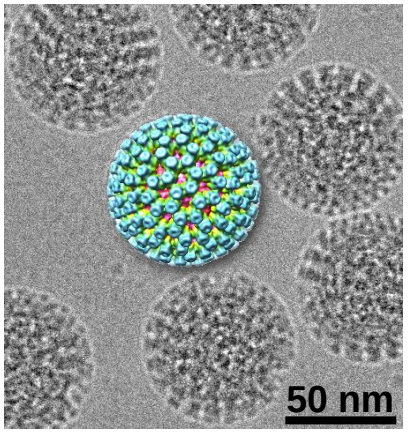
Cryo-EM image

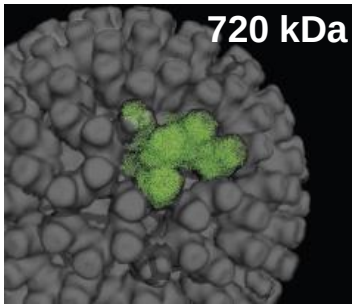
Close to focus

Correlation map

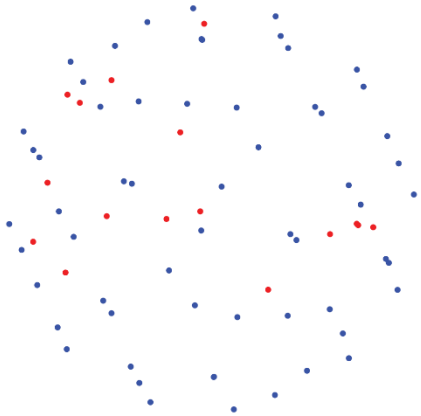
Rickgauer et al. 2017

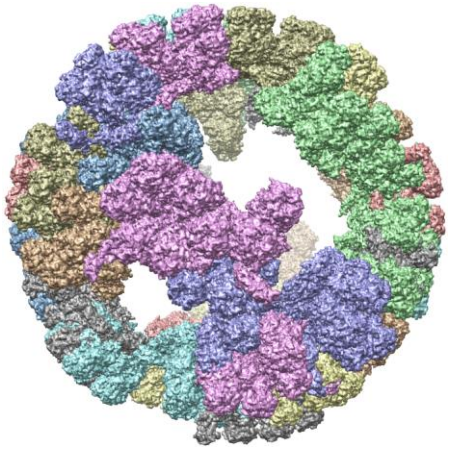
Finding Asymmetric Units



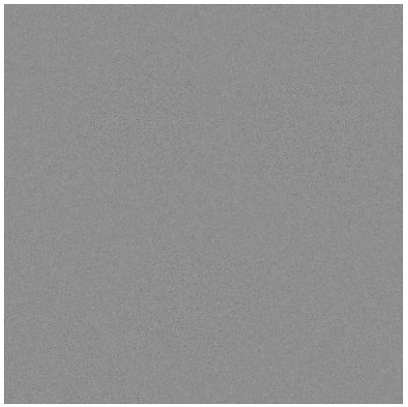


60 asymmetric units:
13 VP6 + 2 VP2

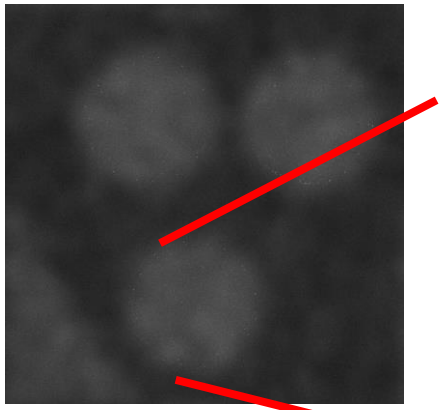




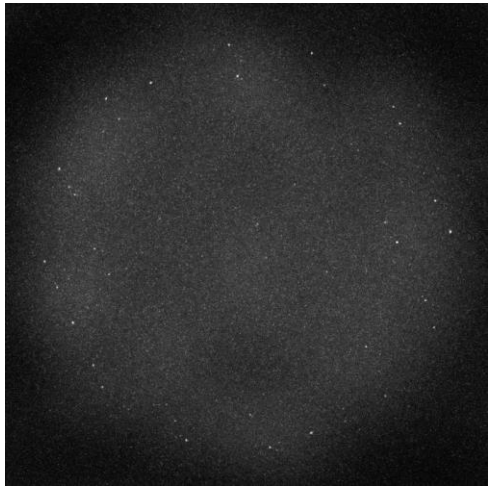
+ defocus search

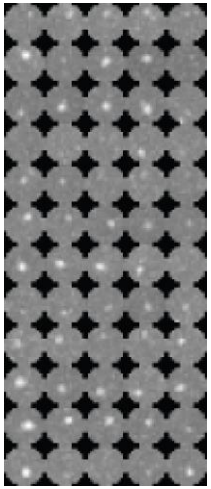


0.3 µm underfocus



Correlation map



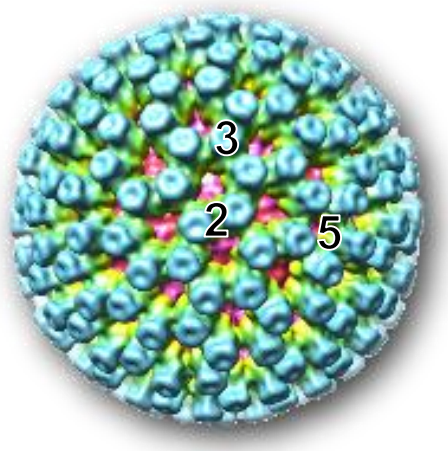


75% of expected positions found

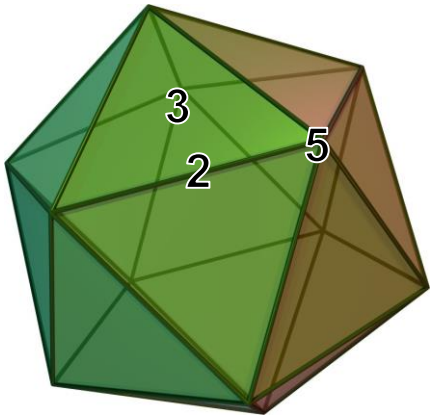
Rickgauer et al. 2017

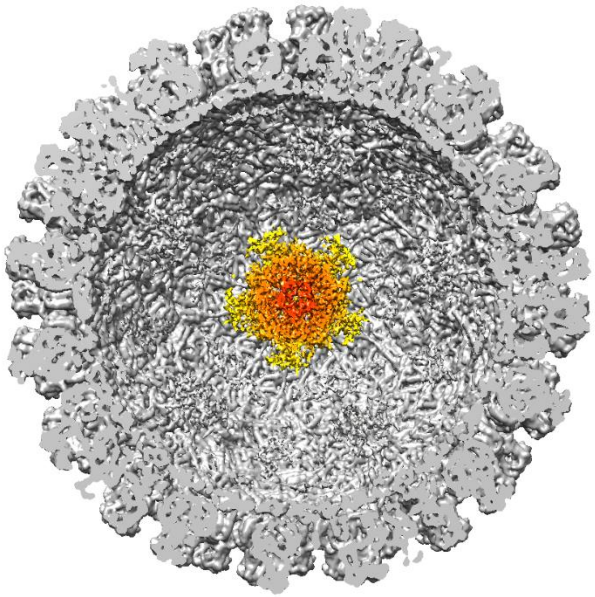
Finding RNA Polymerase

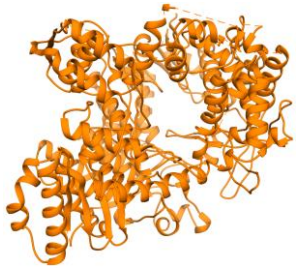
DLP



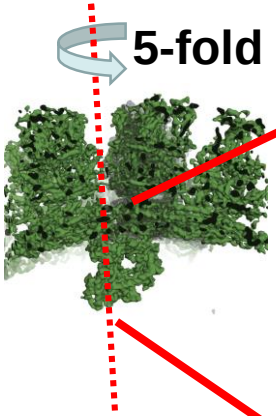
Icosahedron

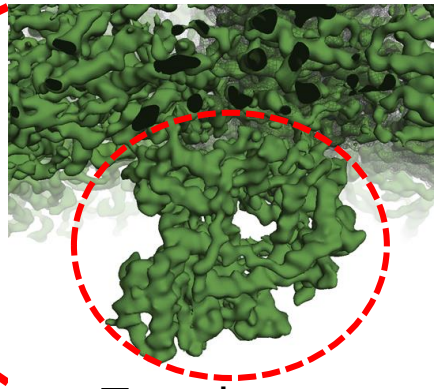




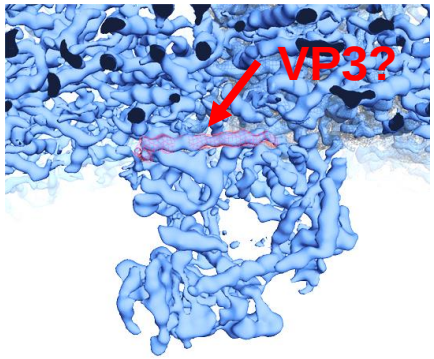


RNA polymerase
(VP1, 115 kDa)





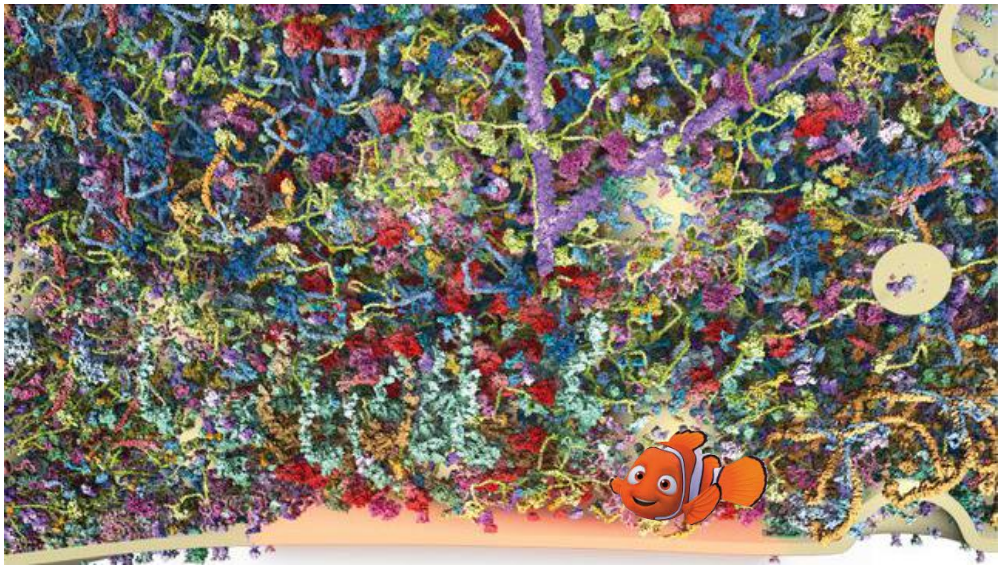
Template



Experimental density
15,265 vertices averaged

Rickgauer et al. 2017

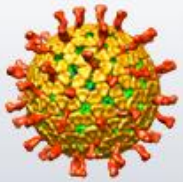
Finding Nemo



Synaptic bouton

- Current molecular weight limit:
  - **~300 kDa** when orientations are not constrained
  - **~100 kDa** with constraints (e.g. membrane)
- If images are perfect: limit lowered to **30 kDa**.
- Positional accuracy:
  - **1 Å** horizontally
  - **~20 Å** vertically

Wilhelm et al. 2014



Summary and Questions

- How do we detect heterogeneity?
  - Search for weak/blurred density, calculate variance maps.
- How do we make sure it does not lead us to the incorrect result?
  - Carful biochemistry, repeat analysis with different starting conditions, check that the results make structural/biological sense.
- How to distinguish conformational vs. compositional variability?
  - Biochemistry, classification, modeling, possibly 3D MSA of bootstrap volumes.
- What are the prospects for getting to atomic resolution for a small and heterogeneous particle?
  - Guess: 50 kDa particle with 10-20 kDa heterogeneity should be possible.
- Are there some samples that will never be amenable to high resolution reconstruction?
  - Very likely, for example if a particle contains large unstructured domains.

<u>Bottom line</u>

Better biochemistry**, bigger** datasets**, bigger** computers**, better** algorithms

Acknowledgements

Template matching

Peter Rickgauer

Winfried Denk

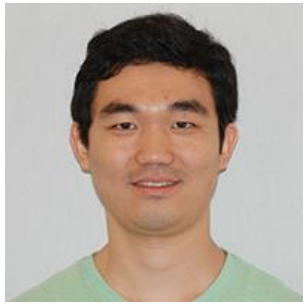
*cis*TEM

Tim Grant

Alexis Rohou

Janelia cryo-EM

Zhiheng Yu



Chuan Hong

Rick Huang