# New Challenges for

# Processing

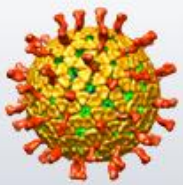
# Heterogeneity

Nikolaus Grigorieff

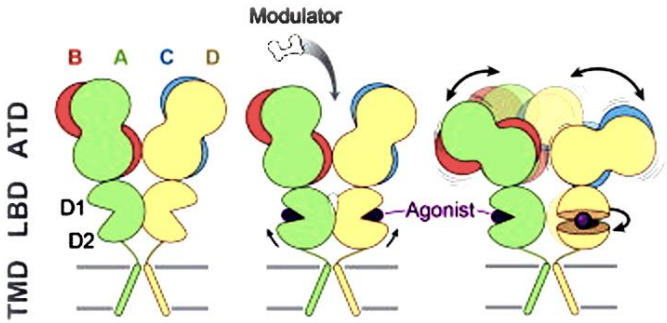

Larson, The Far Side

hhmi | janelia Research Campus

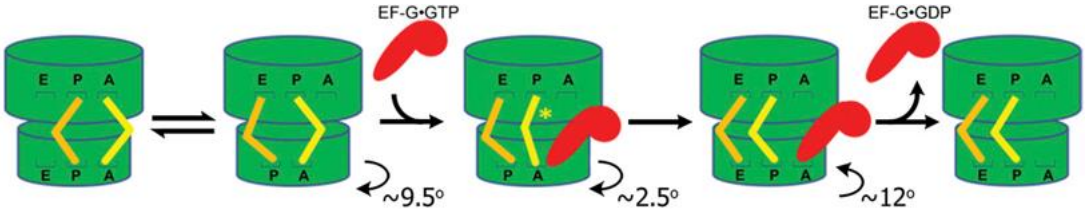
Brandeis



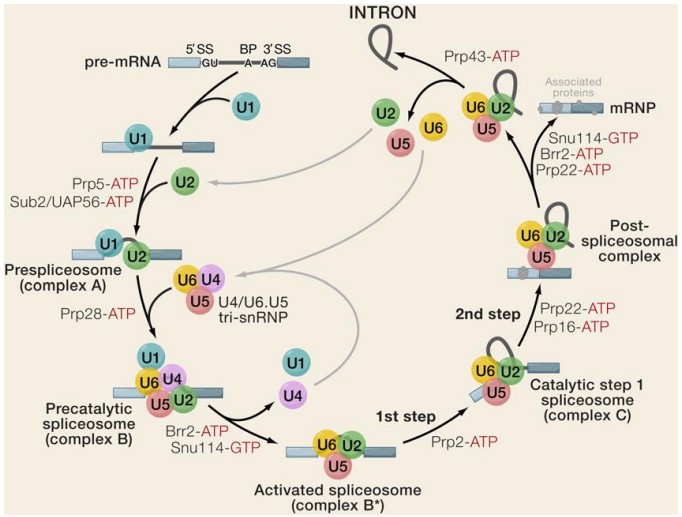
# Heterogeneity and Biology



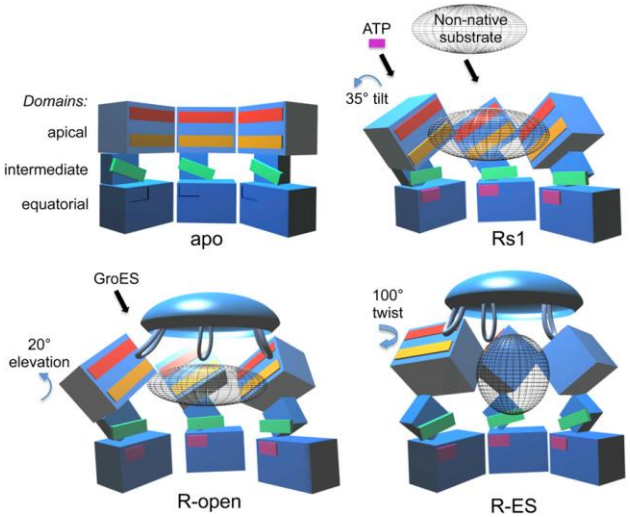

Glutamate receptor, Dürr et al 2014



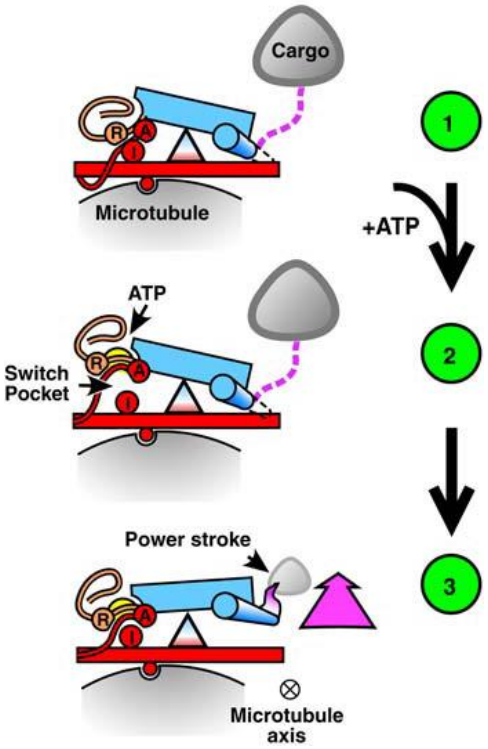

Translocation, Brilot et al 2013



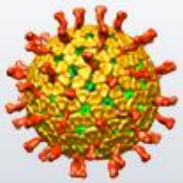

Spliceosome, Wahl et al 2009




GroEL/GroES ATP cycle
Clare et al 2012




Kinesin power stroke
Sindelar & Downing 2010



# Types of Heterogeneity

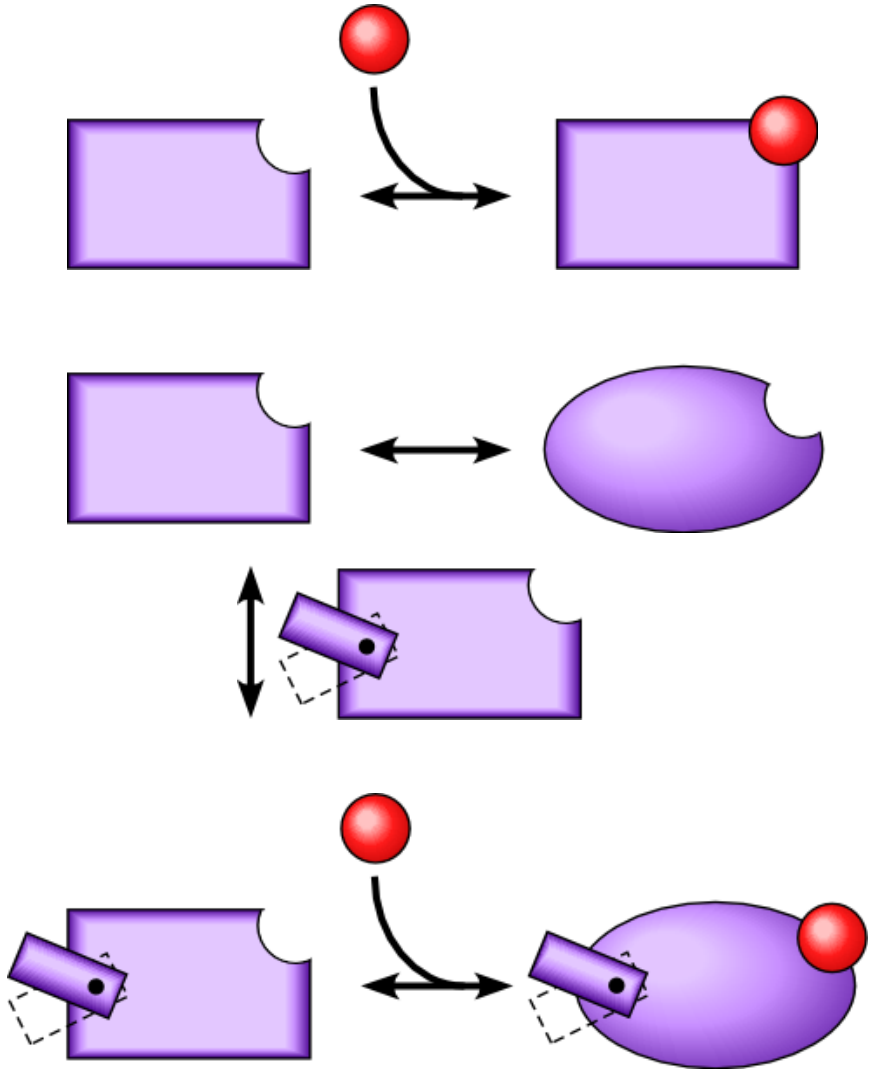
Compositional

Conformational

discrete

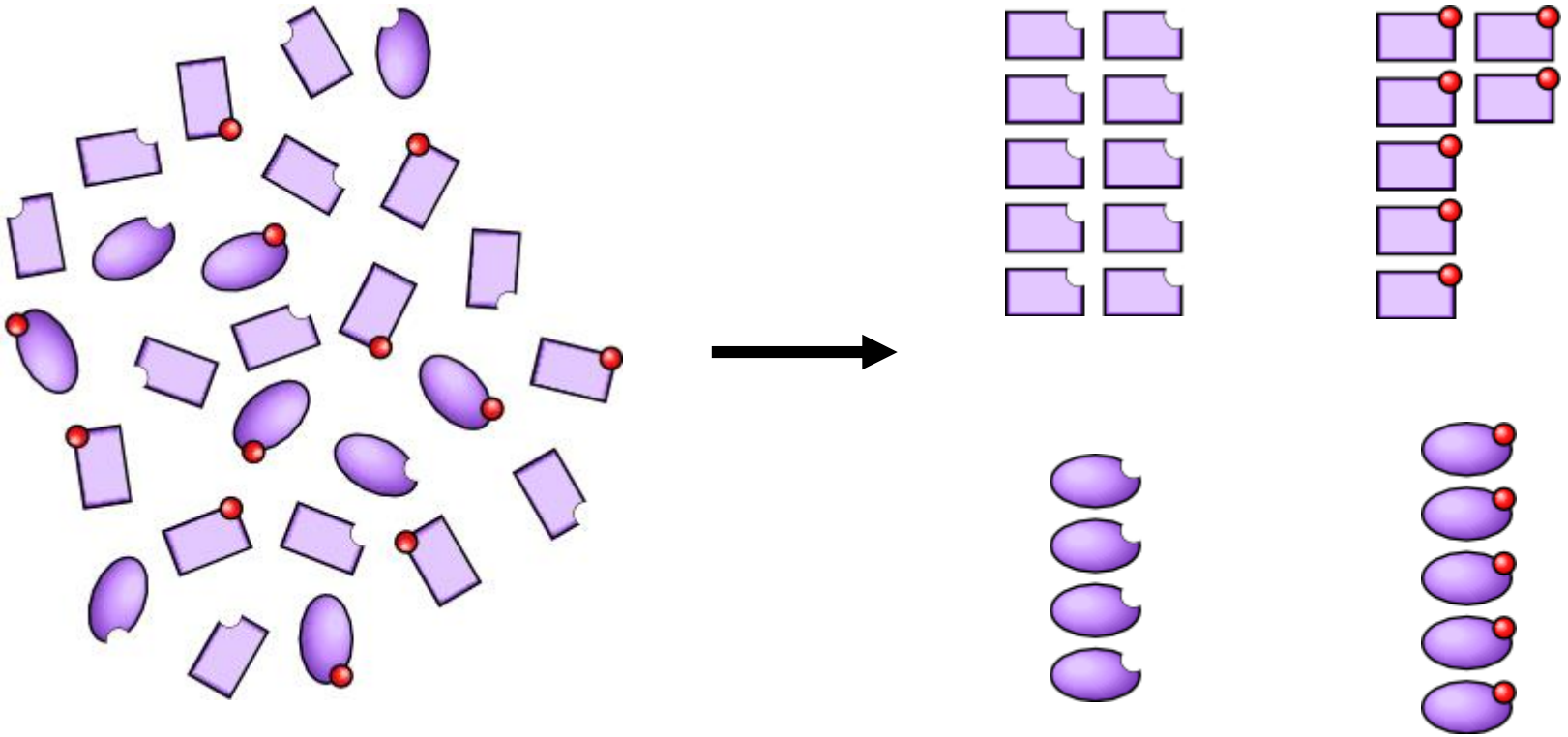
continuous

General



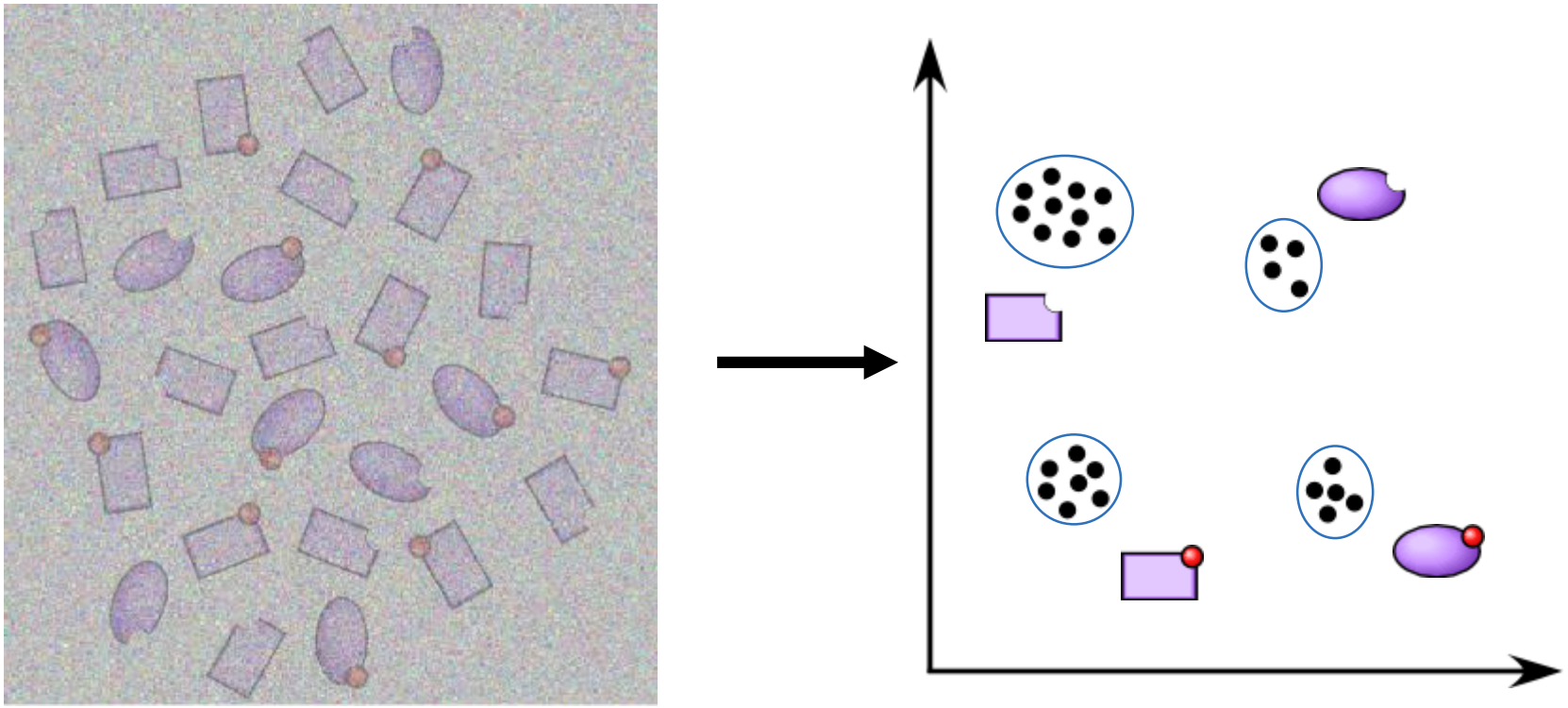
# Classification Goal

Group images based on their similarity.

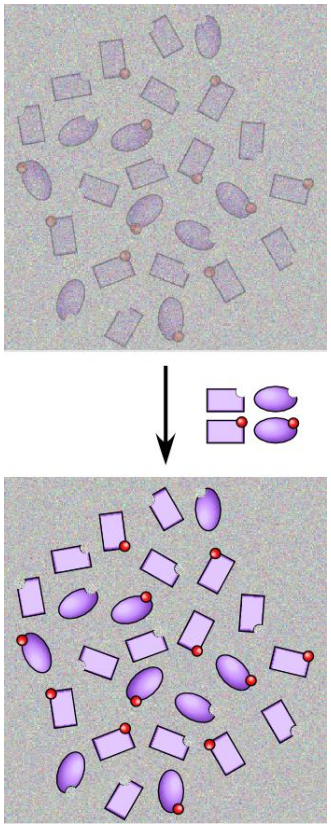


# Noisy Data

Group images based on their similarity so that **averaging** enhances common features (signal) and **reduces noise**.

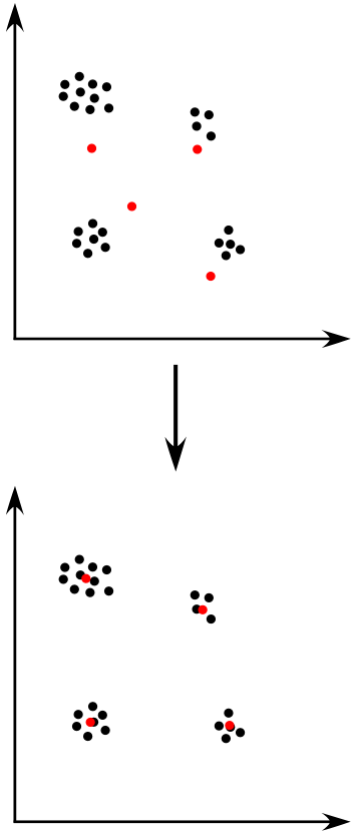


# Common Strategies



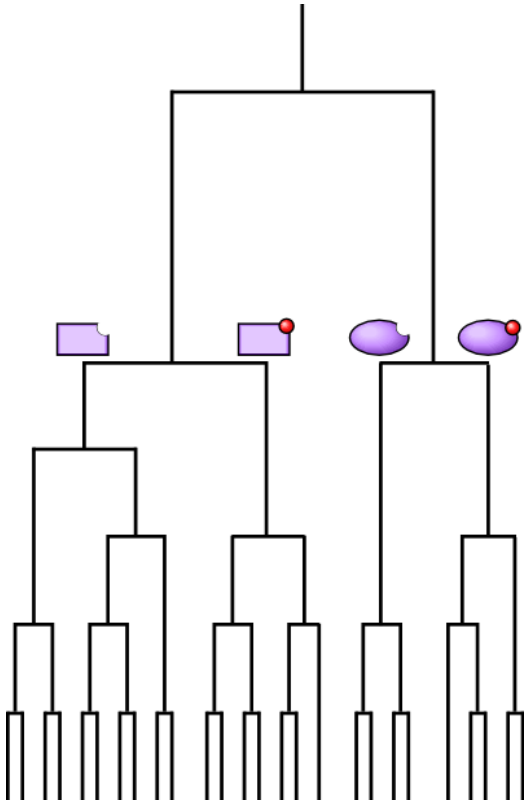
Supervised classification

(MRA)



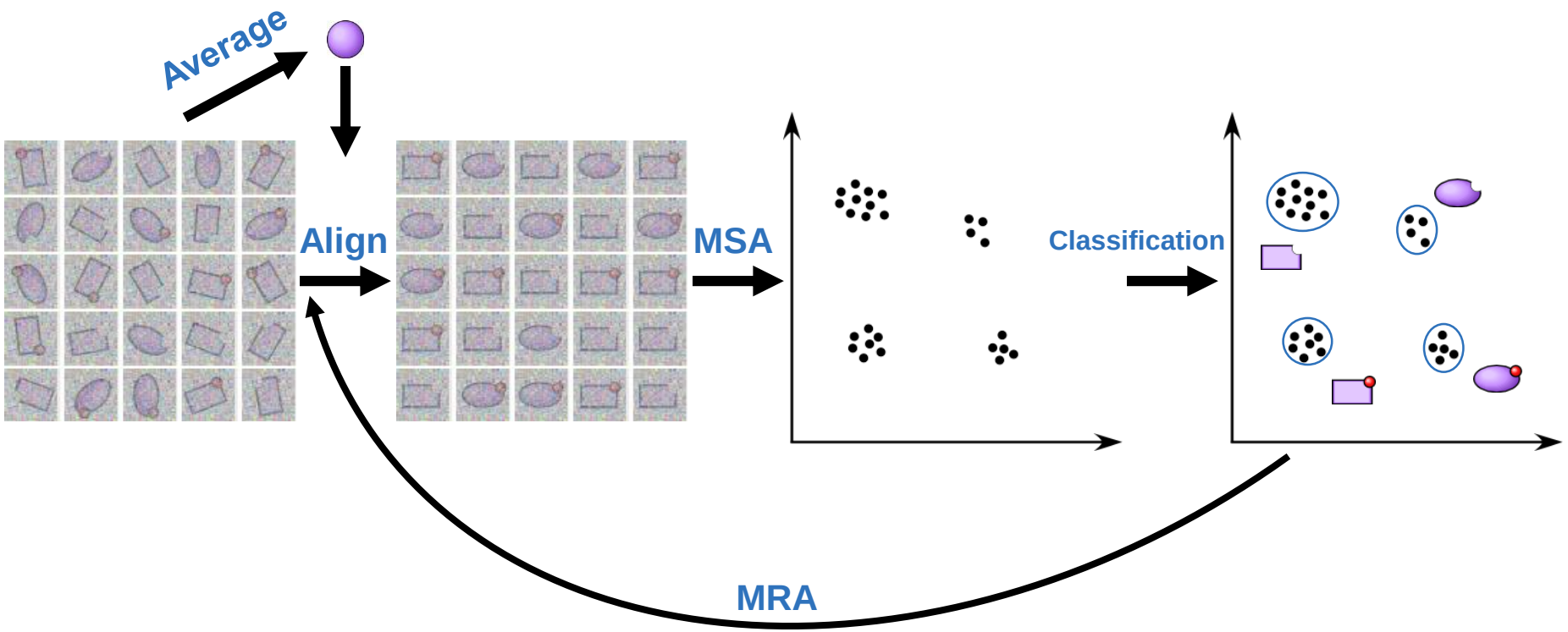
K-means

(MRA/multiparticle, ML2D/3D, ISAC)

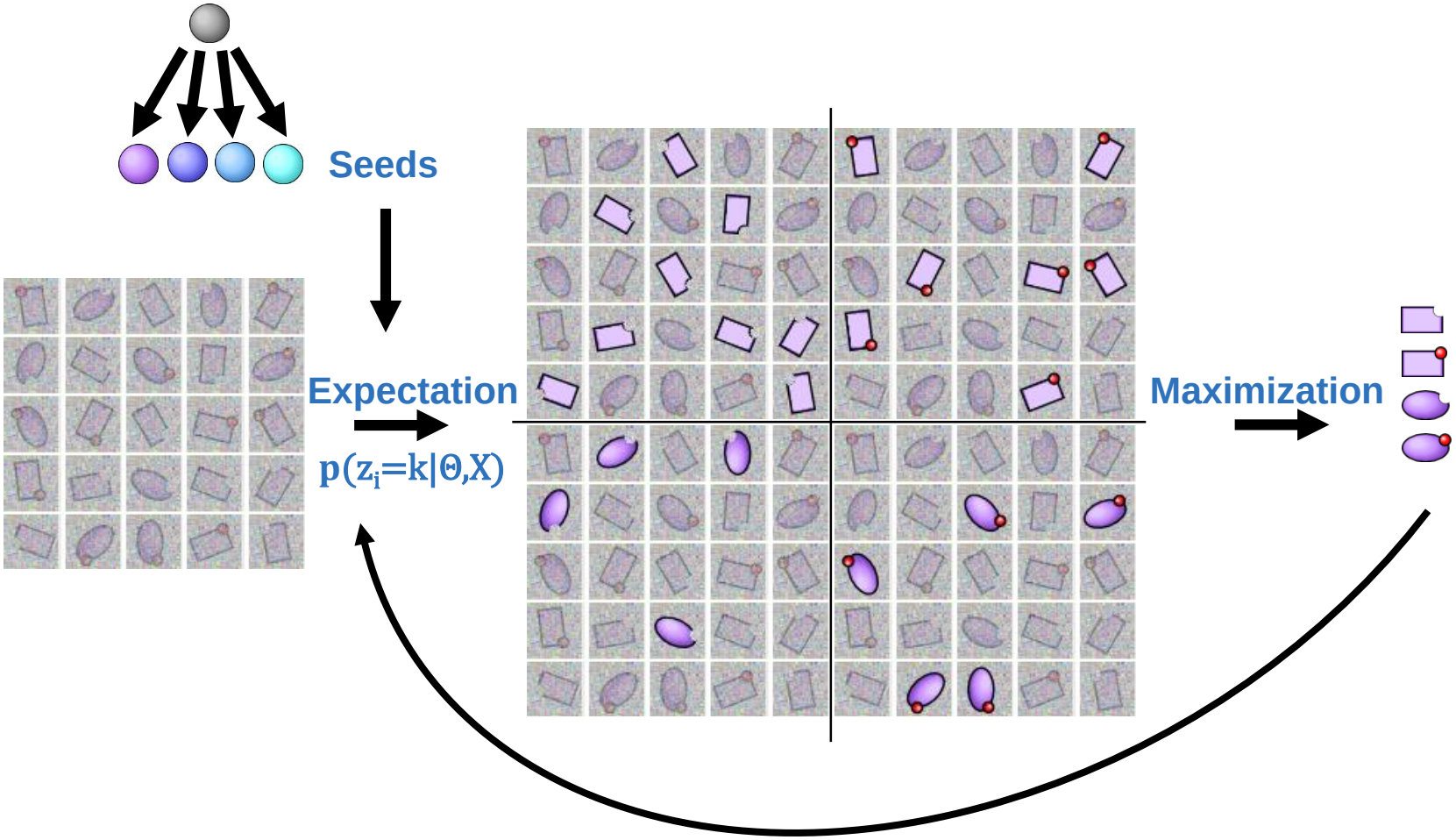


Hierarchical ascendant classification

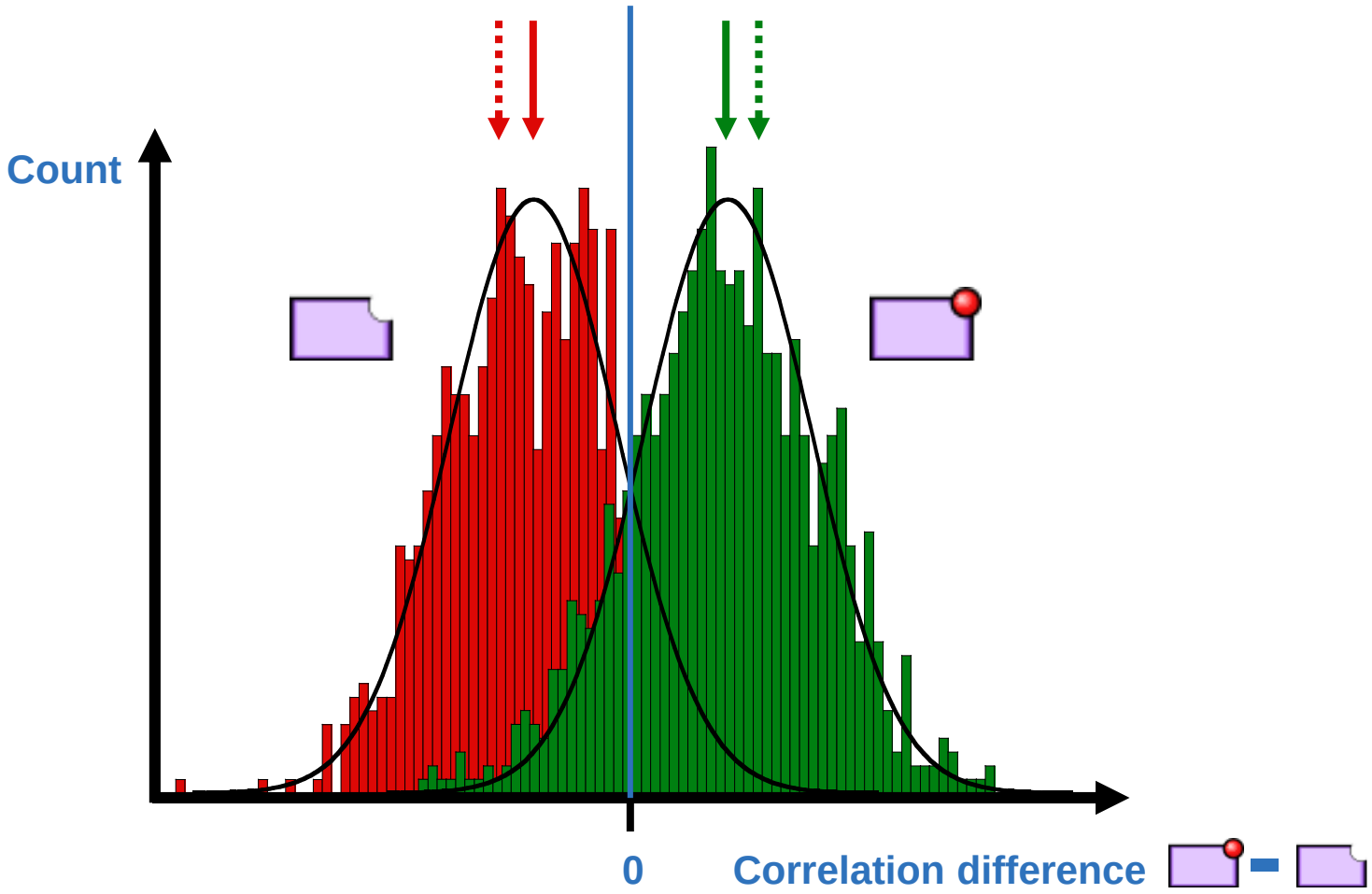
# Classification Procedure



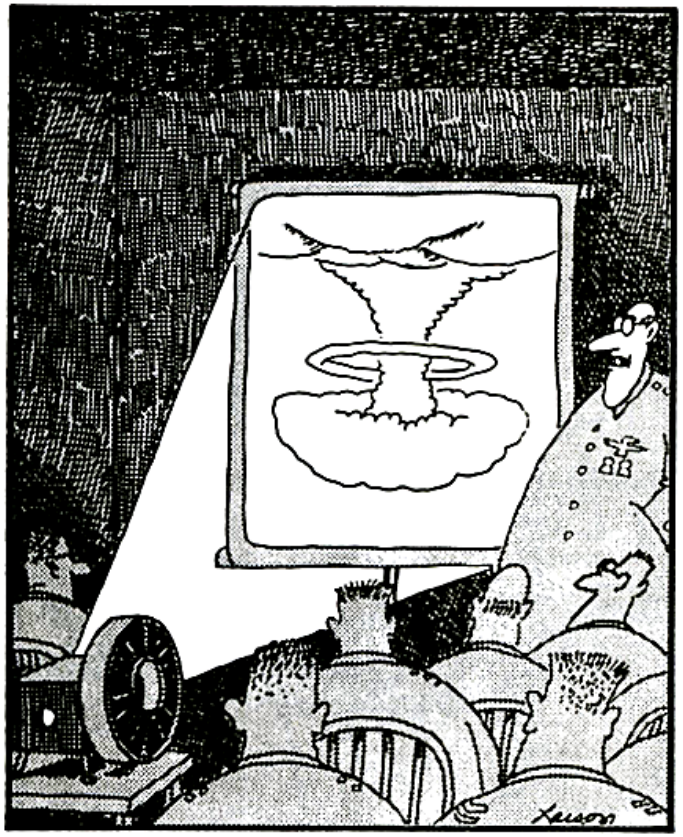
# ML Classification



# The Advantage of ML

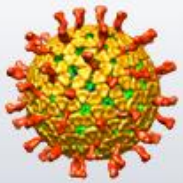


# Some Results

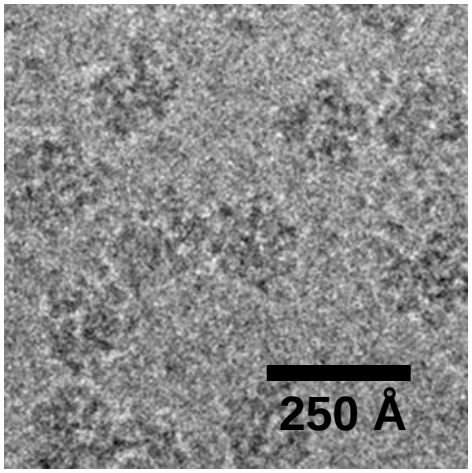


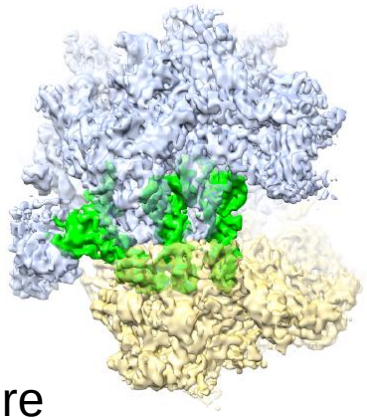
"Now this next slide, gentlemen, demonstrates the awesome power of our twenty megaton . . . For crying out loud! Not again!"

Larson, The Far Side



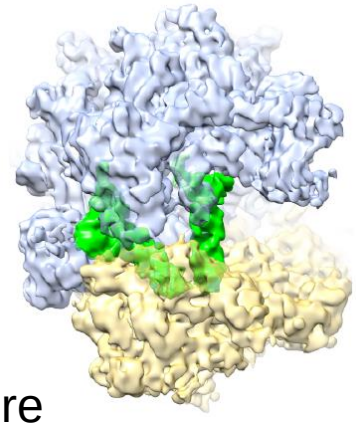
# The Beautiful Ribosome



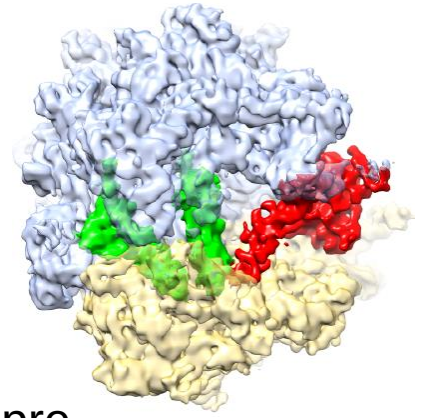
pre

27%



pre

3.5%



pre

2.4%

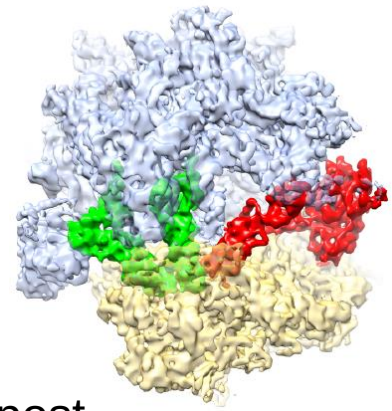
70S ribosome
+ EF-G

tRNA

EF-G

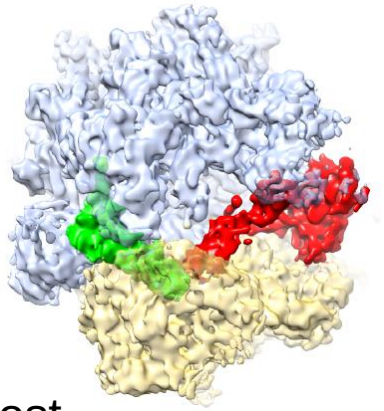
Dataset:
1.3 million particles
300 kV, Falcon I

100 Å



post

13%

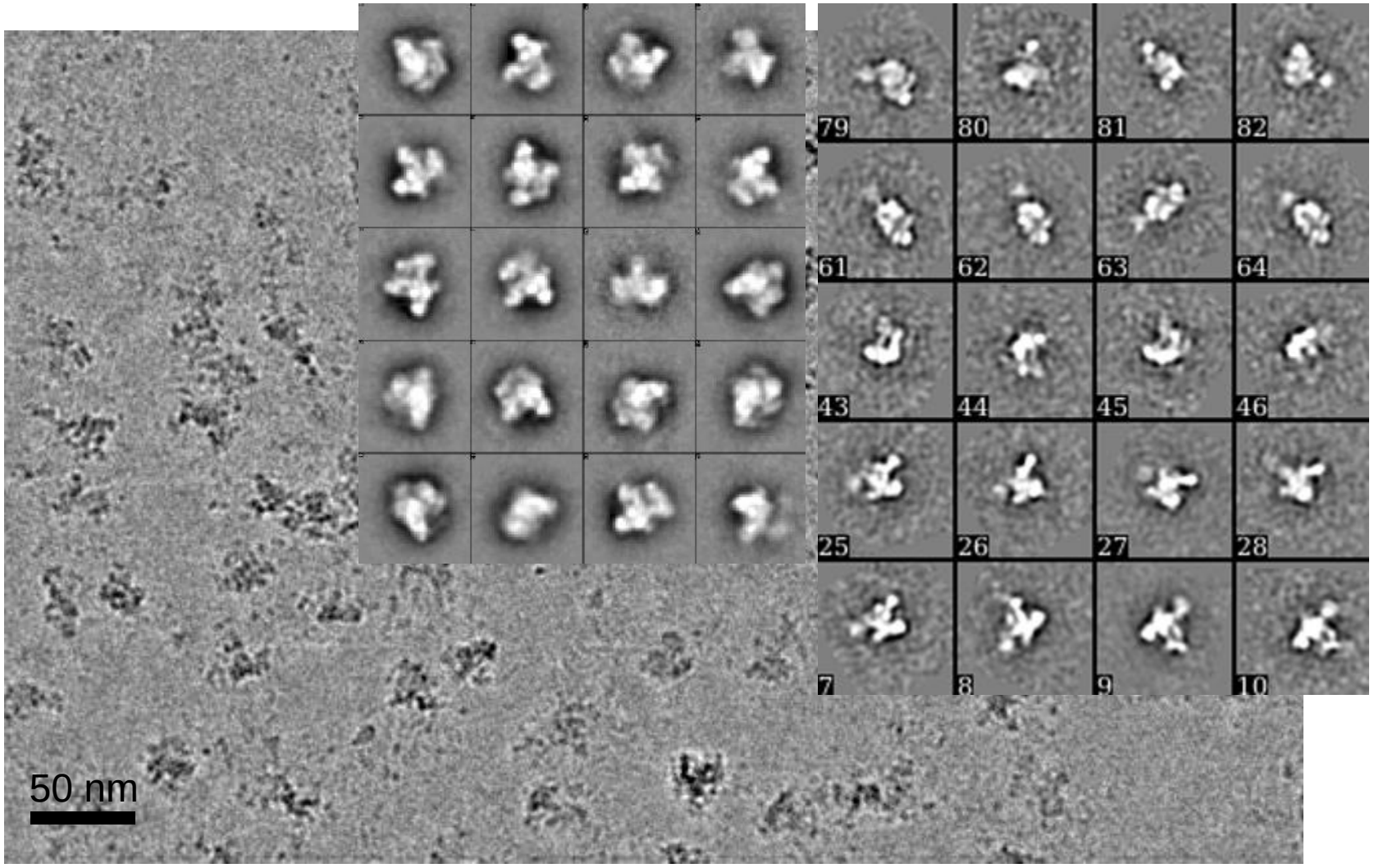


post

6.8%

**Brilot et al. 2013**

# A Dog's Breakfast




**Spliceosome**

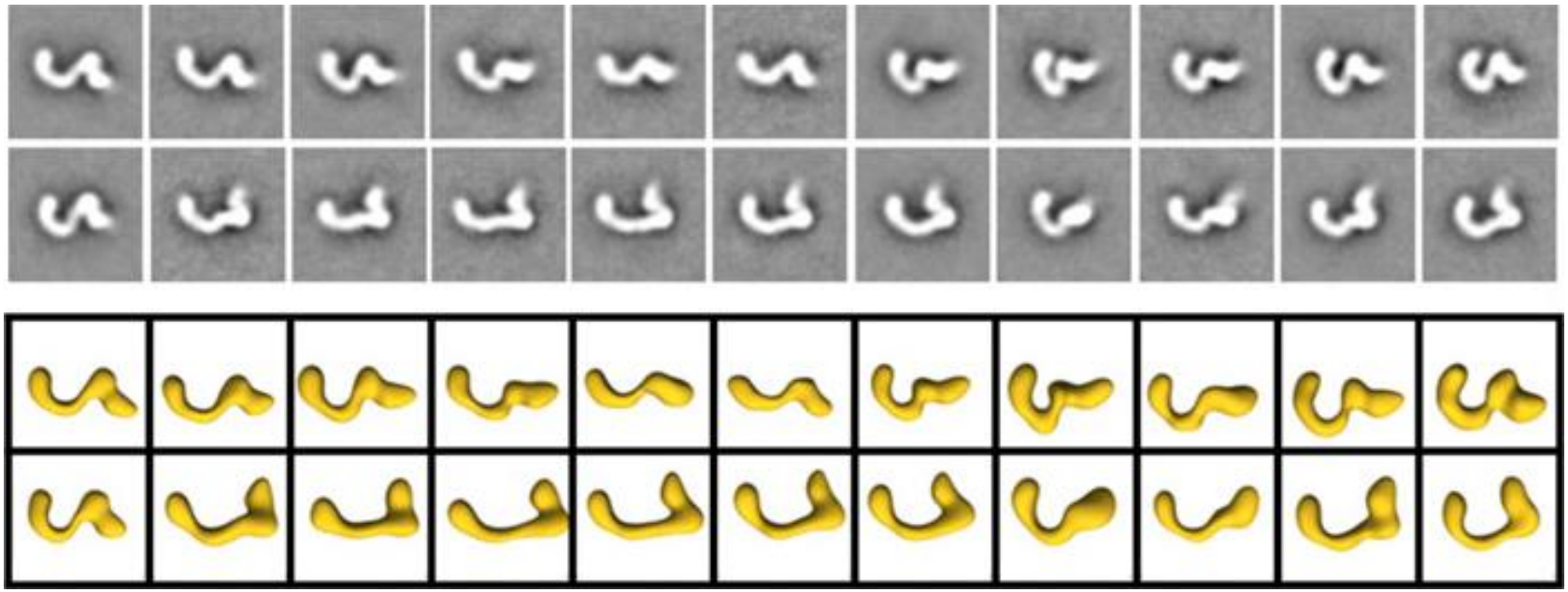
**Anna Loveland, unpublished**

# Classification and RCT/OTR

E3 ubiquitin ligase Ltn1

Negative stain data
180 kDa
Dataset: 68k particles, 12k final

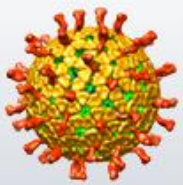
**ML2D, MRA, MSA, HAC**



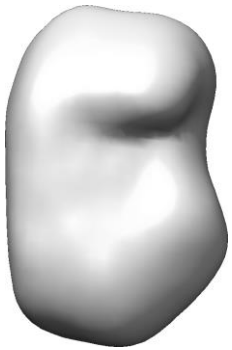
**Random conical tilt reconstructions**

200 Å

**Lyumkis et al. 2013**

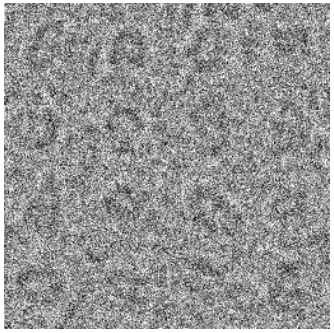


# Cleaning up Datasets



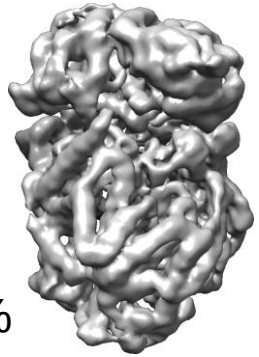
EMAN2
initial model

K-means
classification

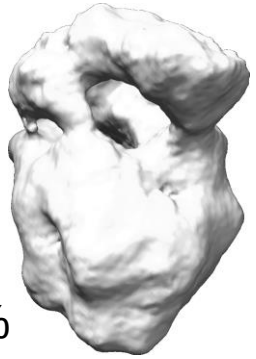


356211 particles
F20, K2

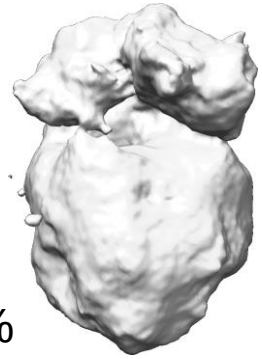
Frealign
refinement &
classification



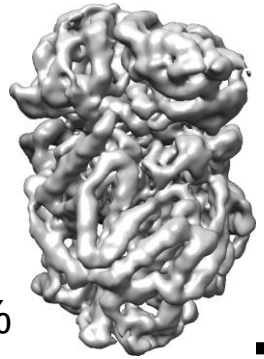
43%



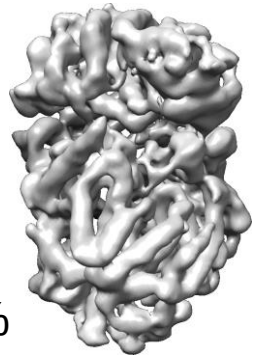
32%



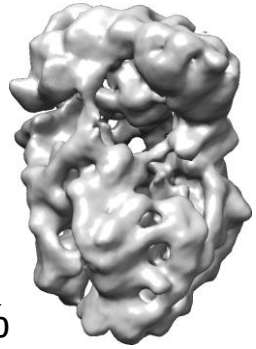
25%



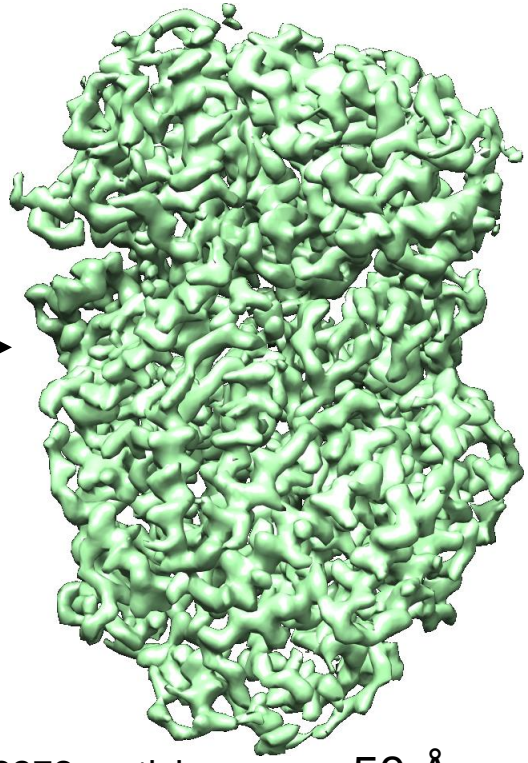
49%



25%



26%



82278 particles
3.9 Å resolution

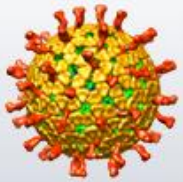
50 Å

VSV polymerase
250 kDa

**Bo Liang, Zongli Li, Simon Jenni, Tim Grant**
**Steve Harrison, Sean Whelan, Tom Walz**

# Problems & Limitations

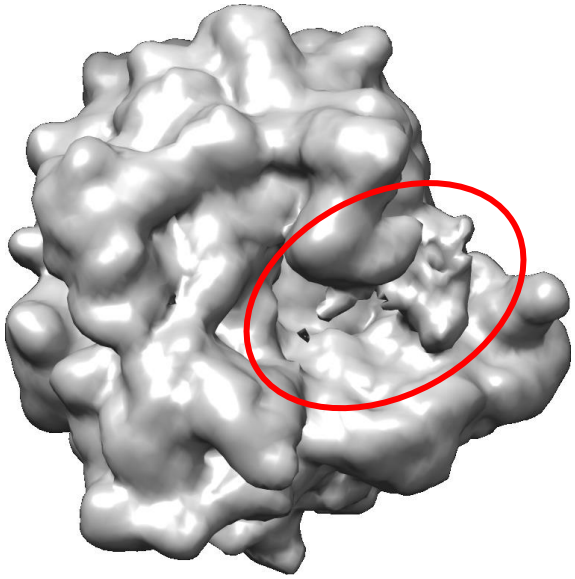

Larson, The Far Side



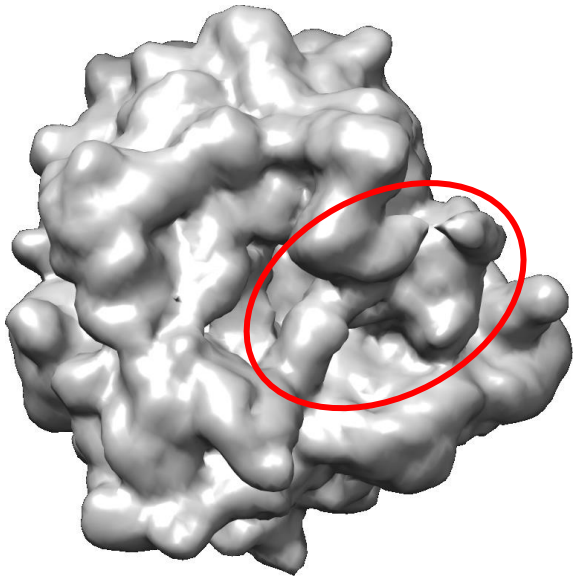
# Potential Pitfalls of K-Means

- Circular cluster shapes only
- Results sometimes strongly dependent on initial seeds
- May not converge to global optimum
- Incomplete separation of classes

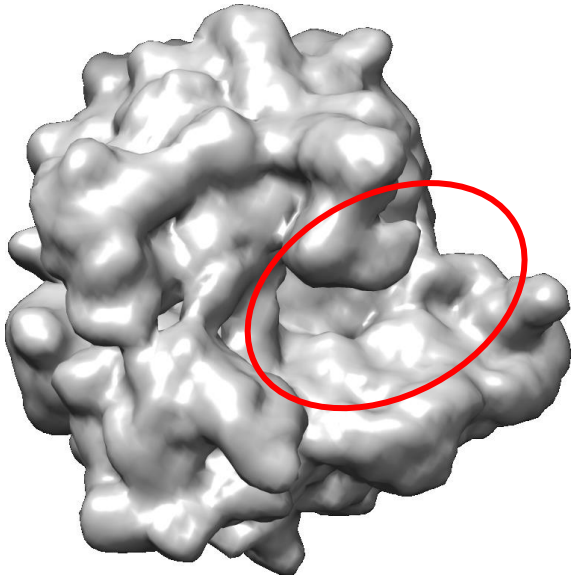
# Incomplete Separation



6.4%



2.4%



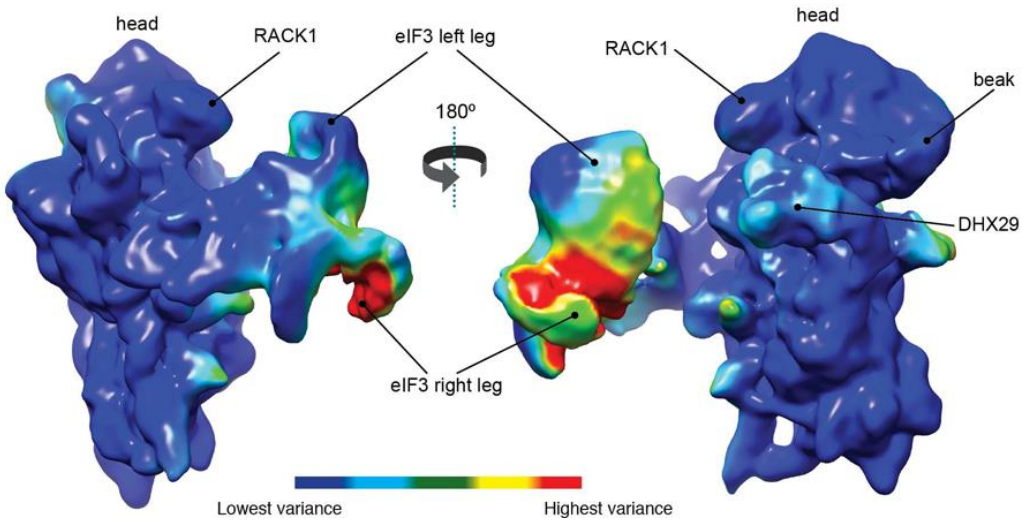
3.3%

70S ribosome + EF-G

**Brilot et al. 2013**

# Detecting Heterogeneity

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




- Computationally expensive
- Very sensitive to particle misalignments

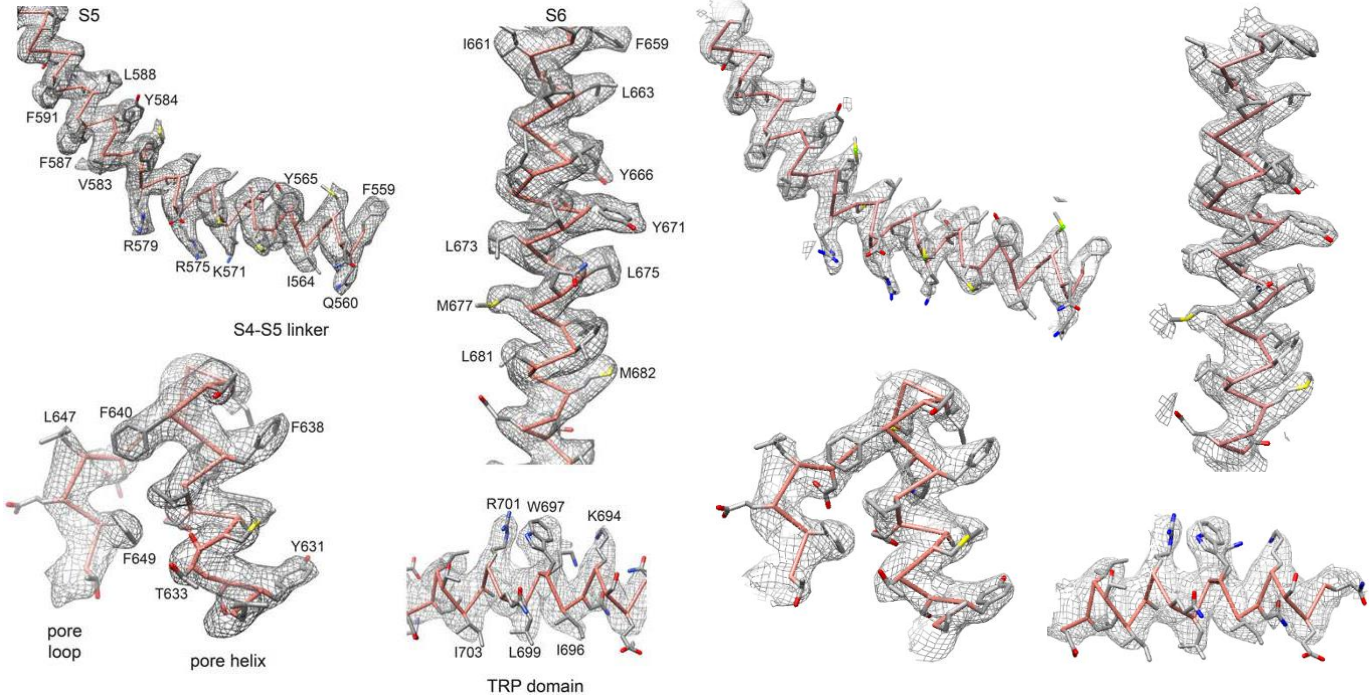
26317 particles (one class out of 630k particles)
40k bootstrap volumes

**Hashem et al. 2013**

# (Ir)reproducibility

TRPV1 channel

Dataset:
88915 particles
(300 kV, K2)

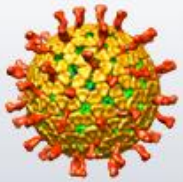



Relion
Refinement & classification
35645 particles (40%)

Frealign
Refinement & classification
38326 particles (44%)

Overlap: 23230 particles (~60%)

**Liao et al. 2013**

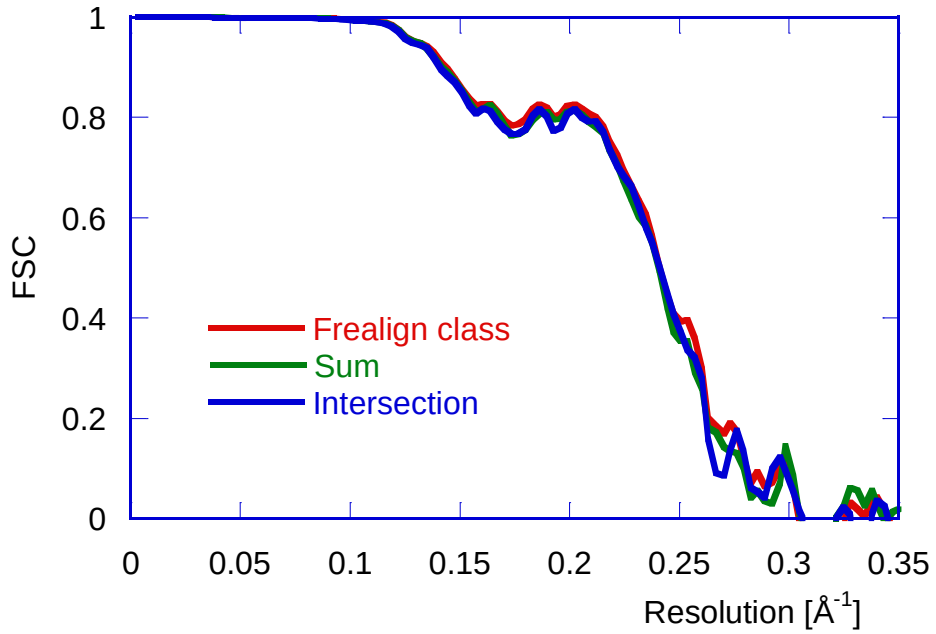


# Interesting Questions

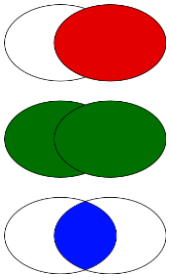
- What is wrong with the ~60% of particles that did not produce good reconstructions?

- Why is the overlap of classes not better?

# Resolution




- Adding or subtracting particles to the Frealign class slightly degrade resolution.
- Possible interpretation:
  - Additional particles from Relion class were misaligned by Frealign (and vice versa).
  - Classification can separate aligned form misaligned particles.

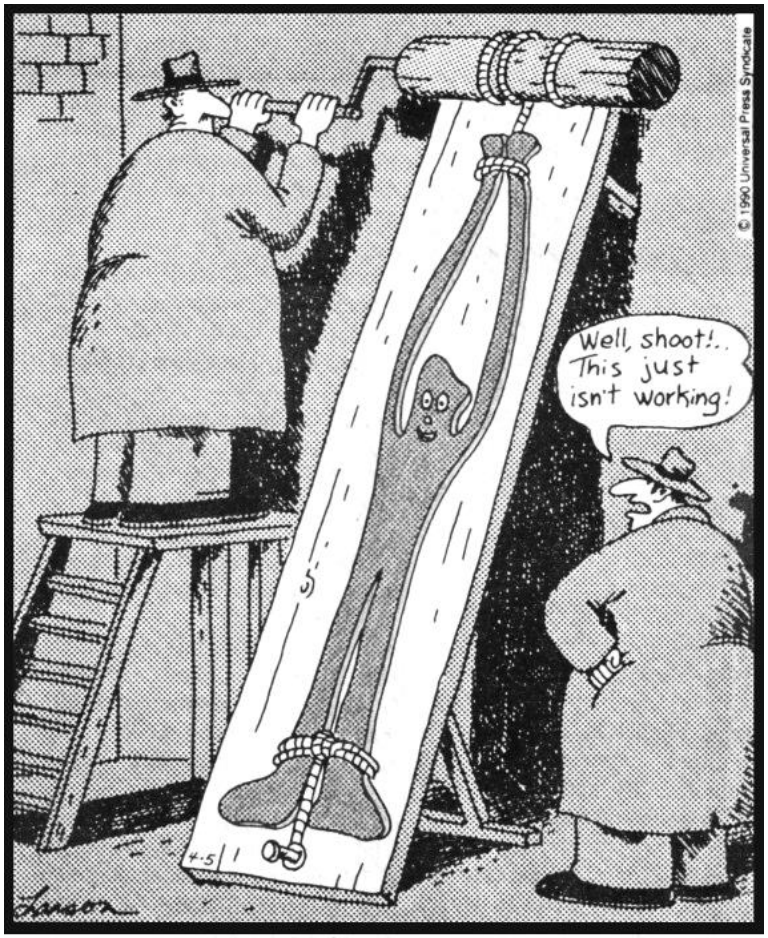


Frealign class (38326 particles)

Sum of Frealign and Relion classes (50755 particles)

Intersection of Frealign and Relion classes (23216 particles)

# Continuous Heterogeneity

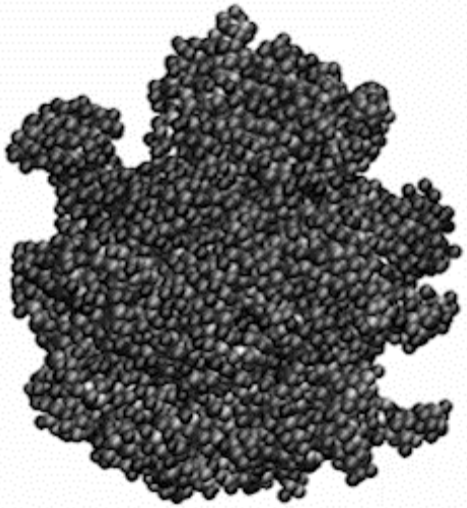



Special Agent Gumby falls into the frustrated hands of the enemy.

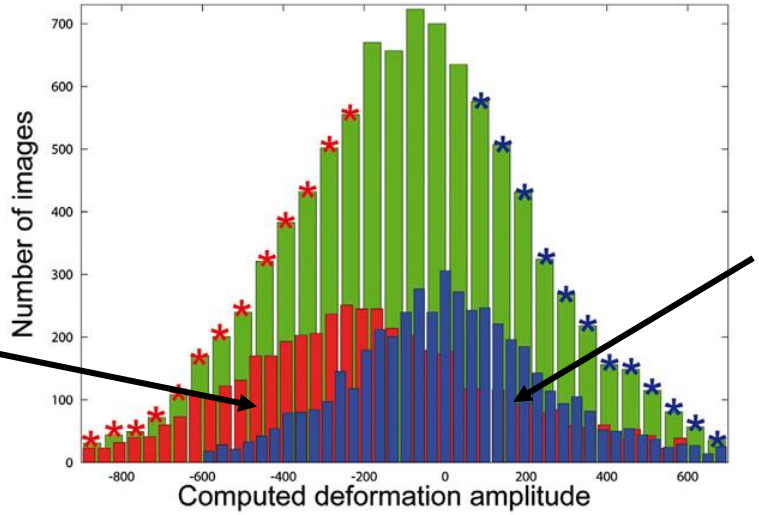
Larson, The Far Side

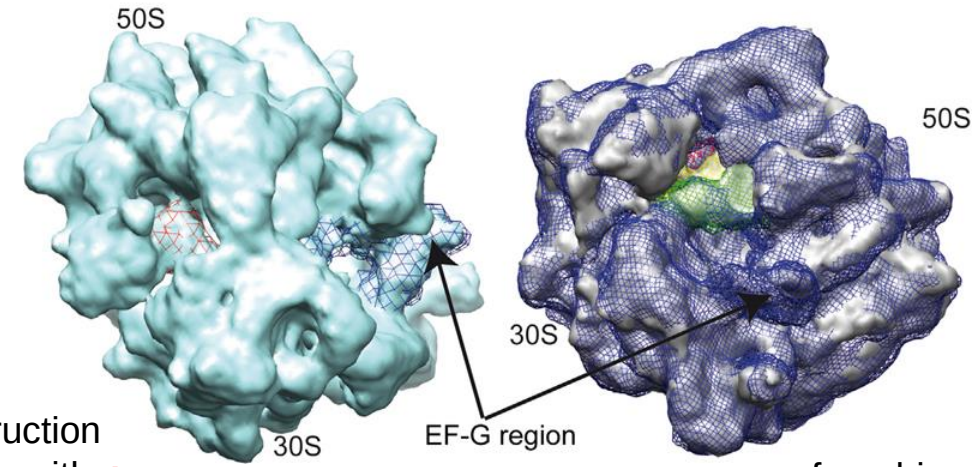
# Normal Modes

70S ribosome + EF-G



Normal mode corresponding to ratcheting



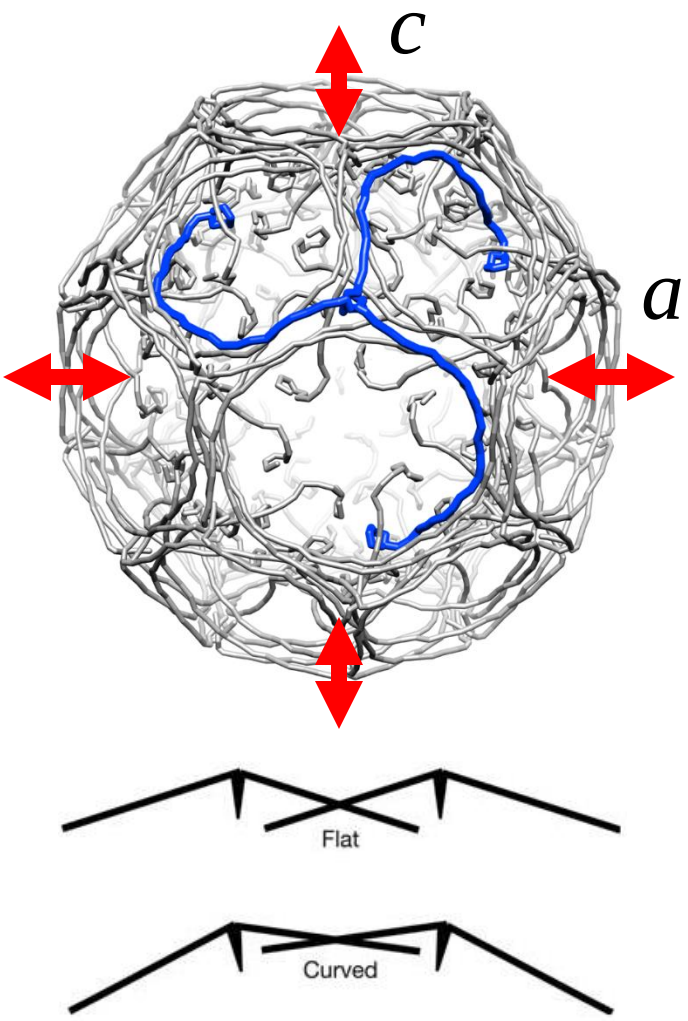




Reconstruction from bins with *

from bins with *

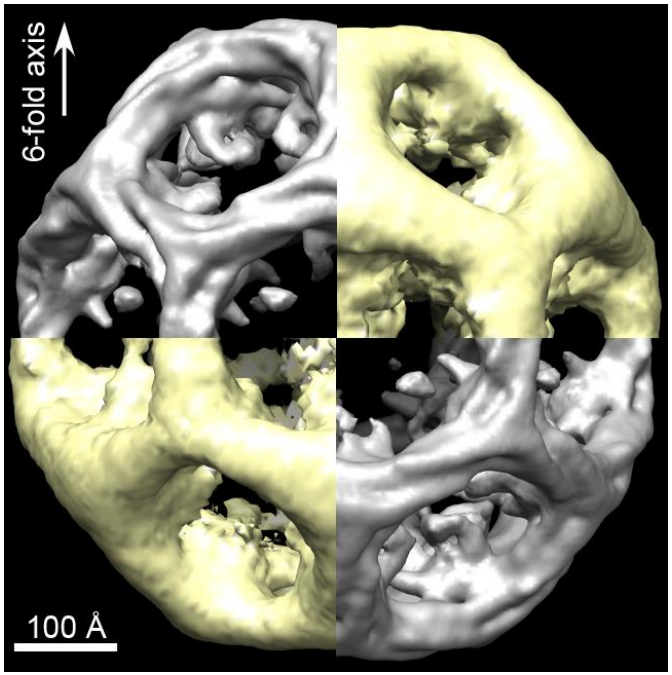
**Jin et al. 2014**

# Deformable Particles




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

Clathrin cage
bound to auxilin and Hsc70

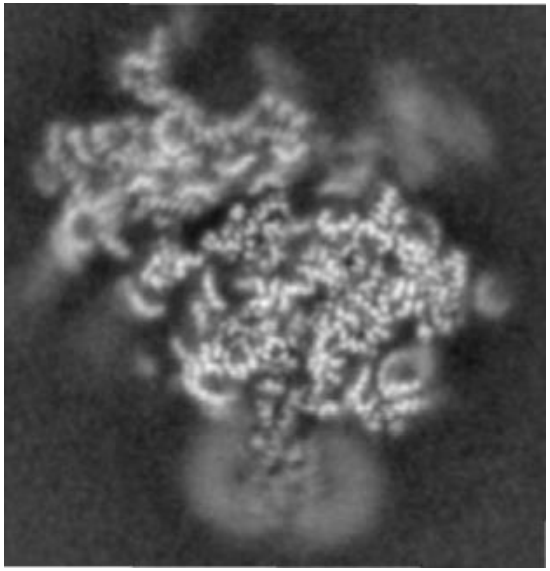


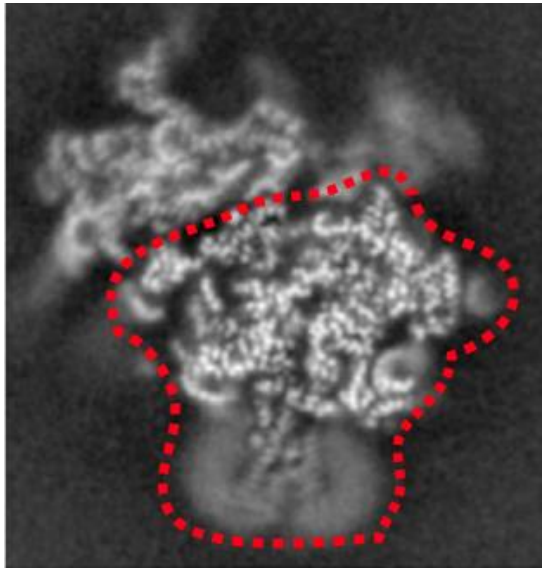

| Model | FSC at 22 Å (σ = 0.016) |
|---|---|
| $\Delta a = -\Delta c$ const. surface | 0.157 |
| $\Delta a = -\sqrt{\Delta c}$ const. volume | 0.145 |
| No deformation | 0.107 |
| $\Delta a = 5\Delta c$ | 0.108 |

**Fotin et al. 2004, Xing et al. 2010**

# Alignment With Masks

80S ribosome + Sec61



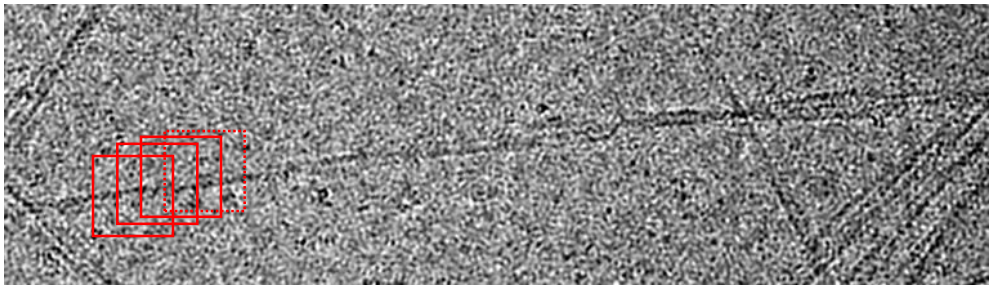


60S ribosome + Sec61

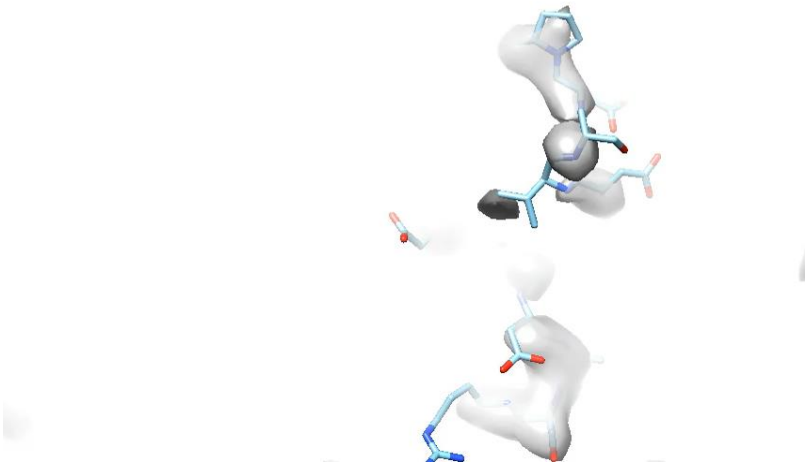
**Voorhees et al. 2014**

# Flexible Helical Filaments

Phage tubulin (PhuZ)



20 nm

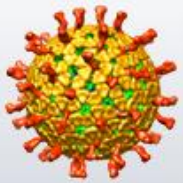


## Frealix

Software for
helical processing

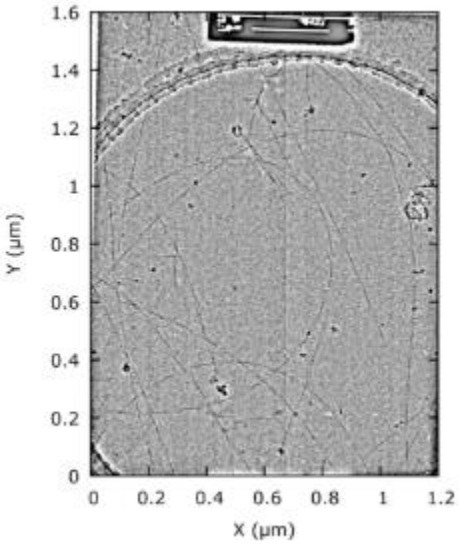
CTF estimation (ctffind4)
Motion correction (unblur)
Frame selection
Subframe motion (Frealix)
3.6 Å resolution

**Elena Zehr, Alexis Rohou, David Agard**

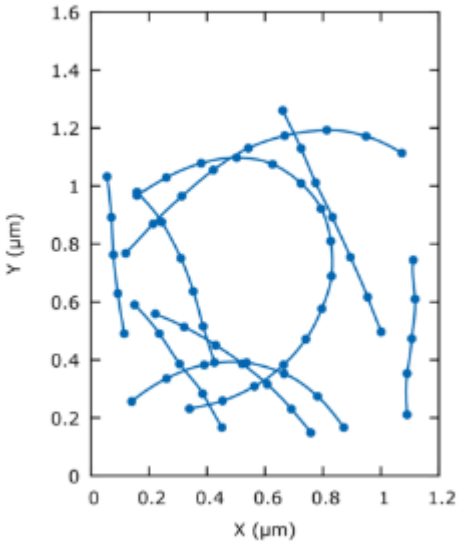


# Full Filament Approach

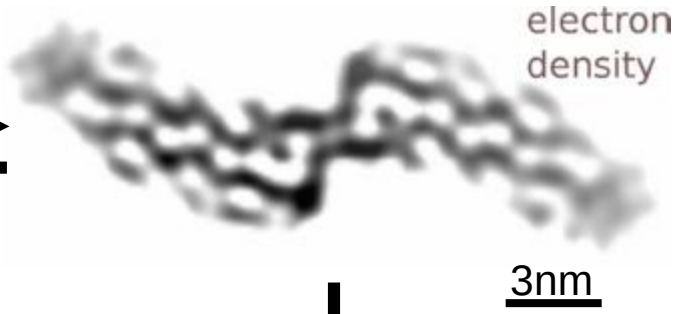
Cryo-EM

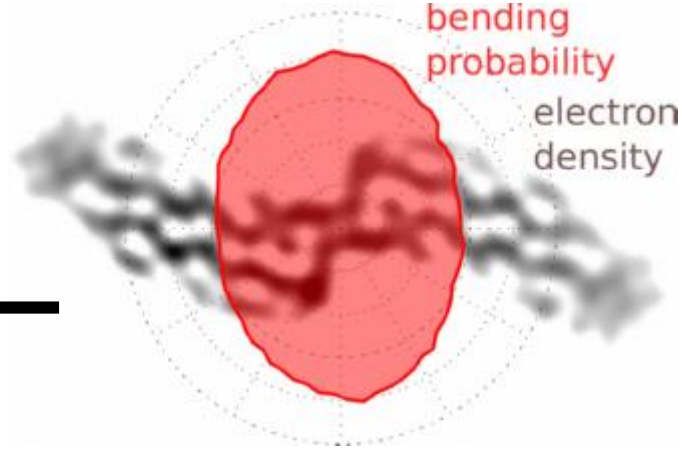



3D filament tracking

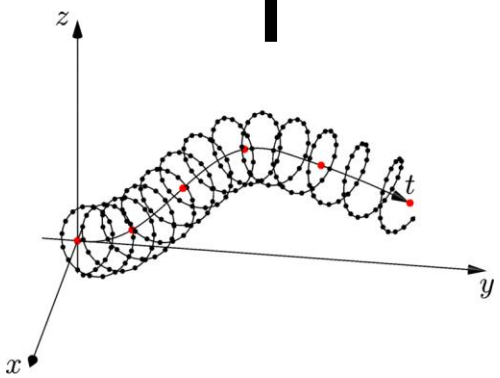



Structure refinement




Deformation statistics




Improved mechanical model

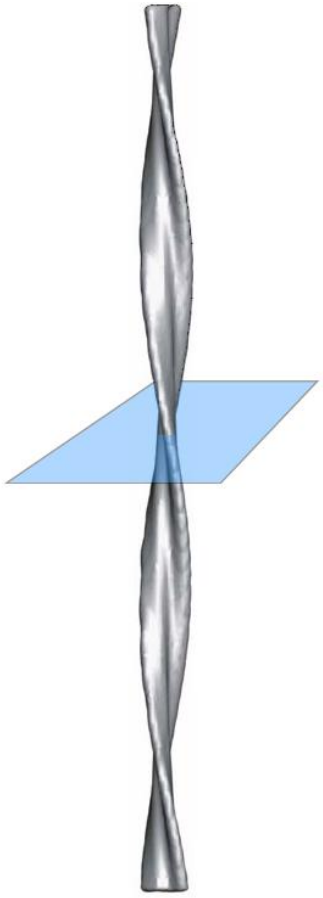
**Frealix**

Software for helical processing

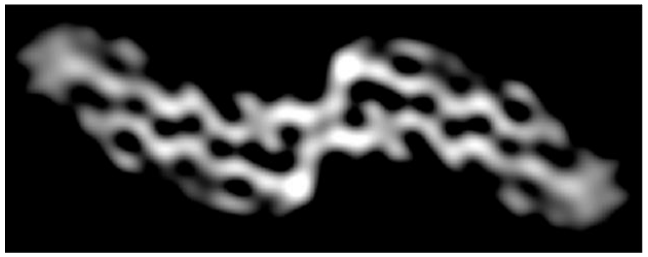
**Rohou & Grigorieff 2014**

# Amyloid Fibrils

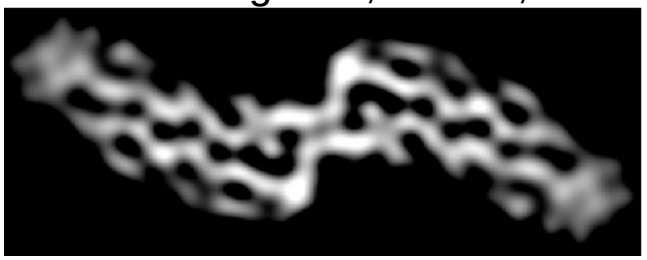
Aβ(1-40)

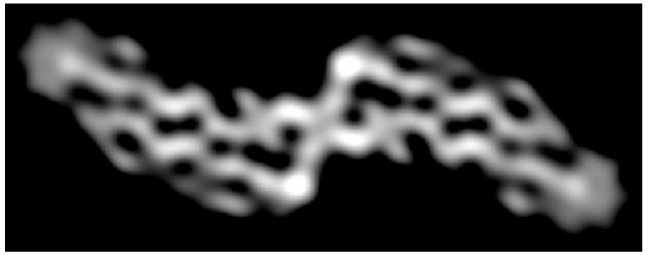


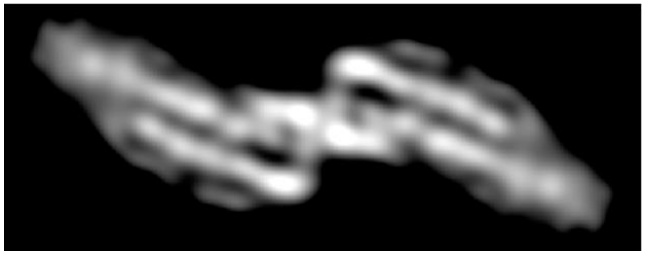
450 total, Frealix, 7.5 Å



188 straightest, Frealix, 7.1 Å

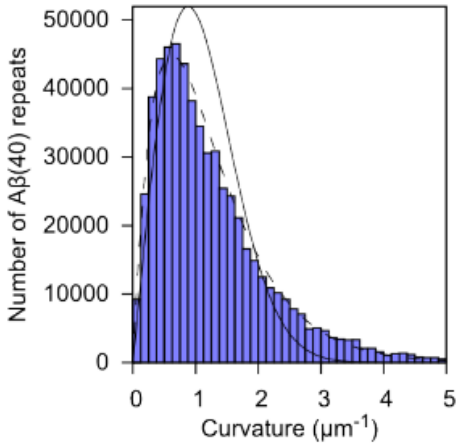


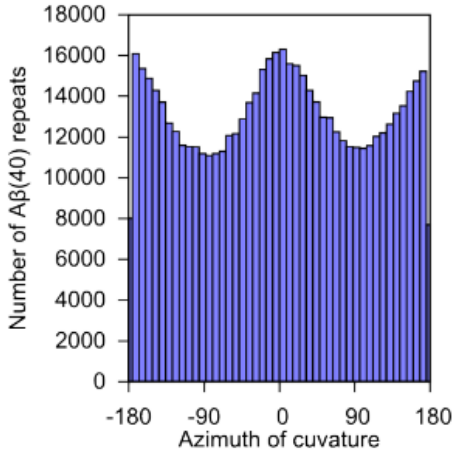




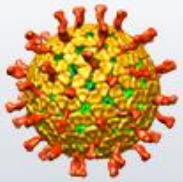
188 most curved, Frealix, 8.3 Å

188 most curved, Frealign, 8.9 Å




**Rohou & Grigorieff 2014**

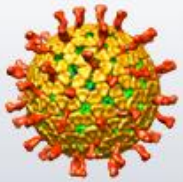


# Summary and Questions

- 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 many 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 (Klaholz/Penczek).
- What are the prospects for getting to atomic resolution for a small and heterogeneous particle?
  - Guess: 200 kDa particle with 20-50 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.

Bottom line

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



# Acknowledgements


- EF-G ribosome — Axel Brilot, Andrei A. Korostelev, Dmitri N. Ermolenko
- Spliceosome — Anna Loveland, Melissa Moore
- VSV polymerase — Bo Liang, Zongli Li, Simon Jenni, Tim Grant, Steve Harrison, Sean Whelan, Tom Walz
- Phage tubulin — Elena Zehr, Alexis Rohou, David Agard, Joe Pogliano
- Frealix — Alexis Rohou

- Cryo-EM facility — Chen Xu (Brandeis), Zhiheng Yu (Janelia)
- **Financial Support:** **HHMI, NIH**

# Thank You!


"Mr. Osborne, may I be excused? My brain is full."