(S9716)

Visualizing ATP-Dependent Substrate-Processing Dynamics of the Human 26S Proteasome at Near-Atomic Resolution

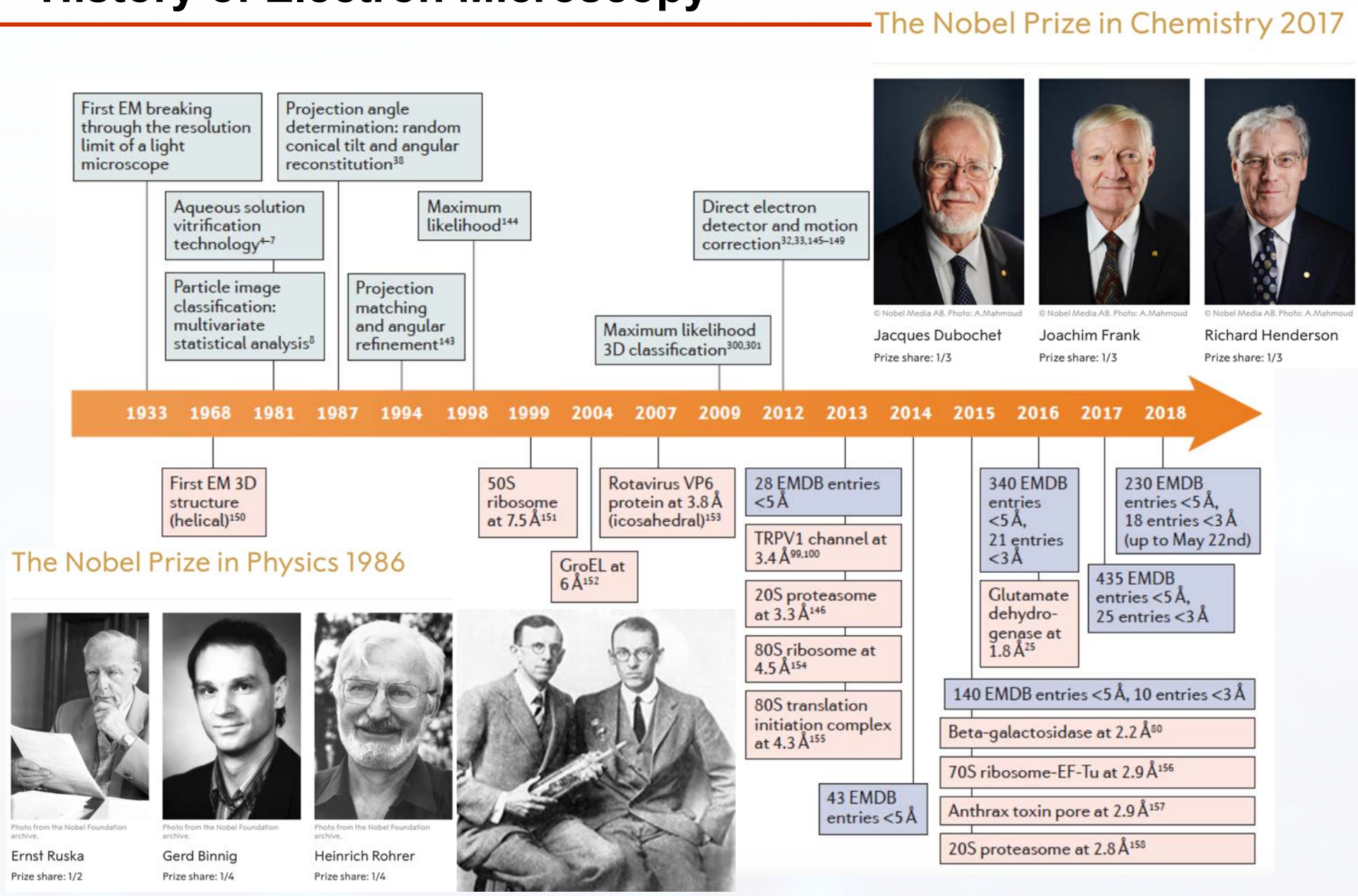
Youdong Mao

Peking University School of Physics

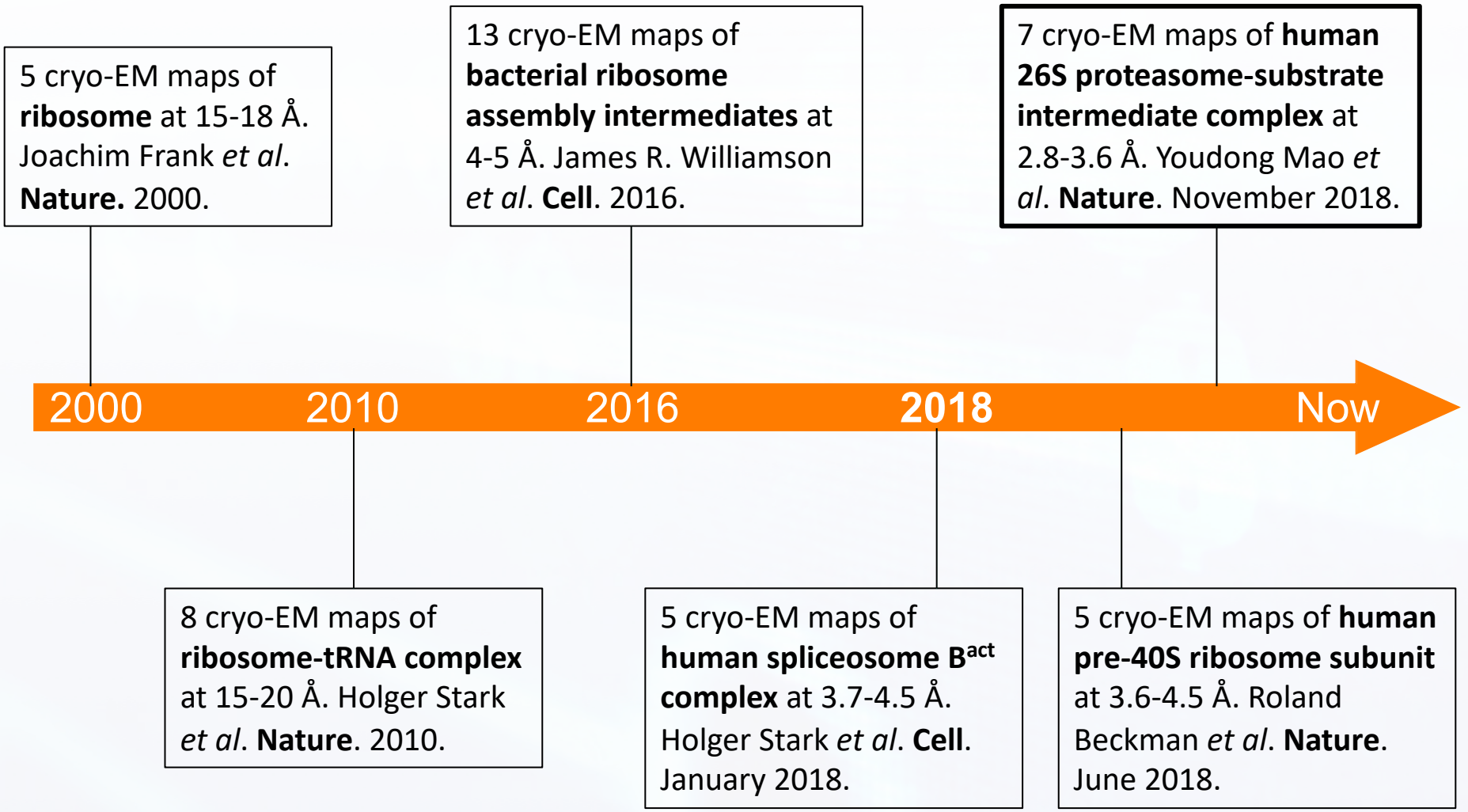
Dana-Farber Cancer Institute, Harvard Medical School

GTC 2019, San Jose

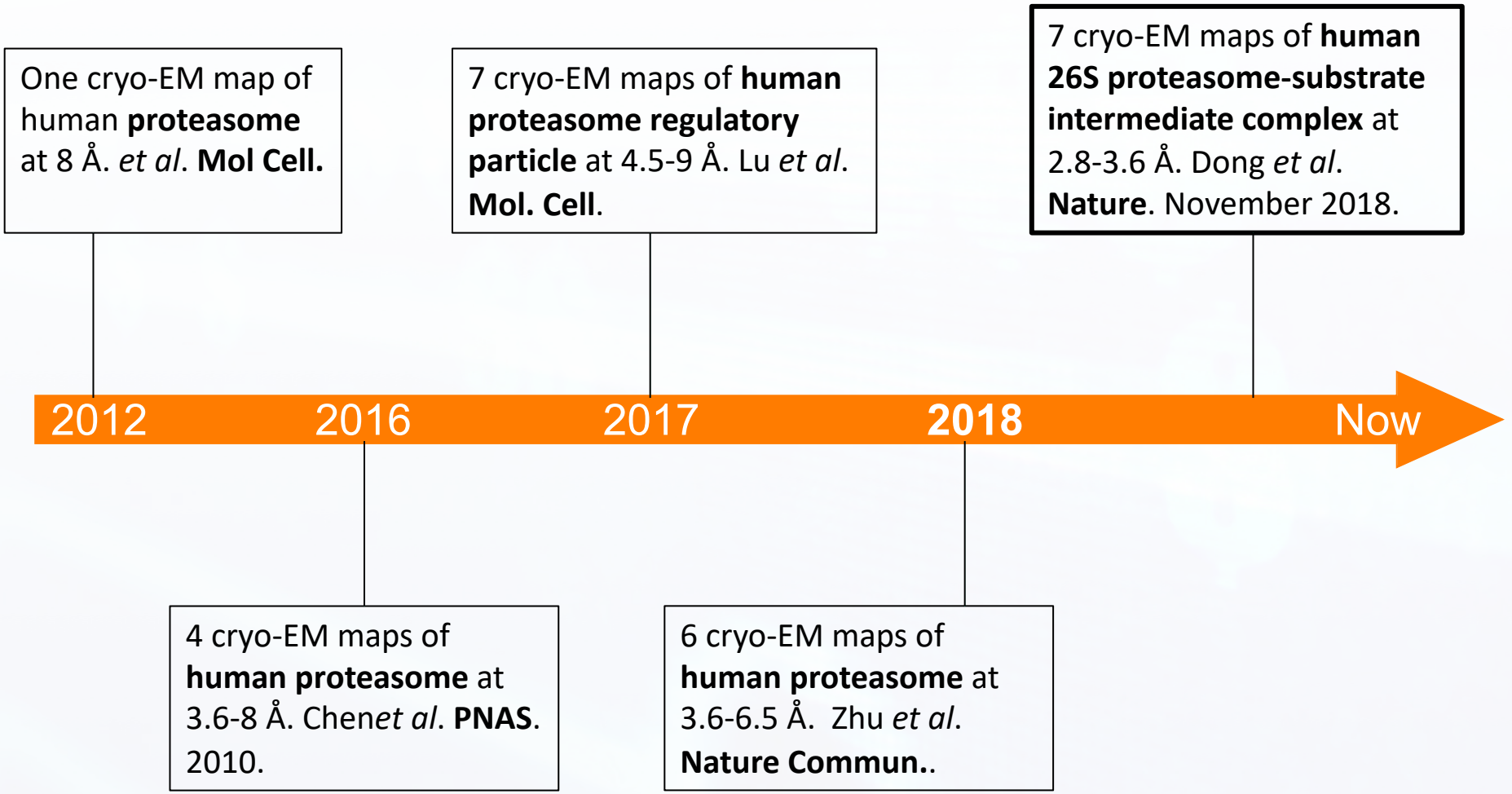
History of Electron Microscopy



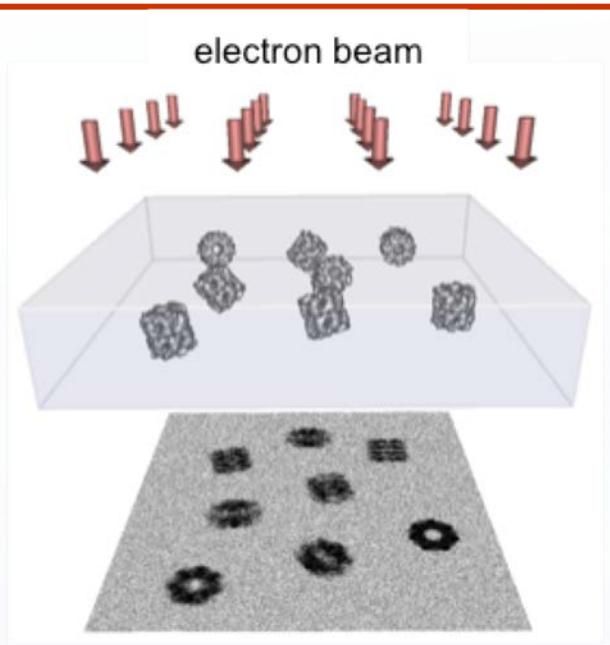
Timeline of Cryo-EM for Complex Dynamics

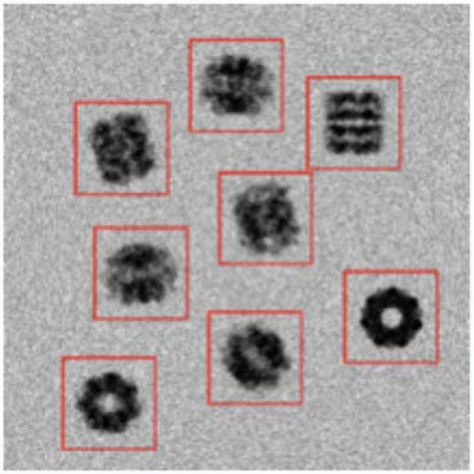


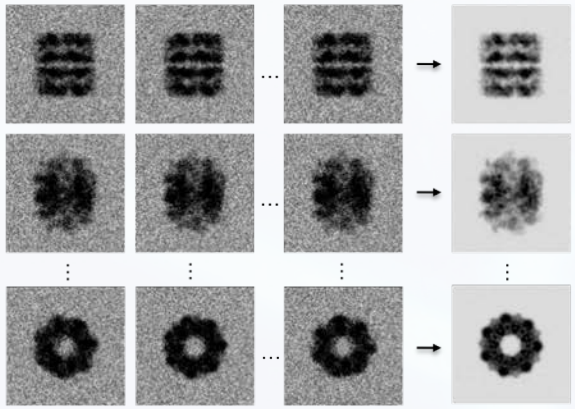
Cryo-EM for Complex Dynamics of Human Proteasome



Single-Particle Reconstruction of Cryo-EM







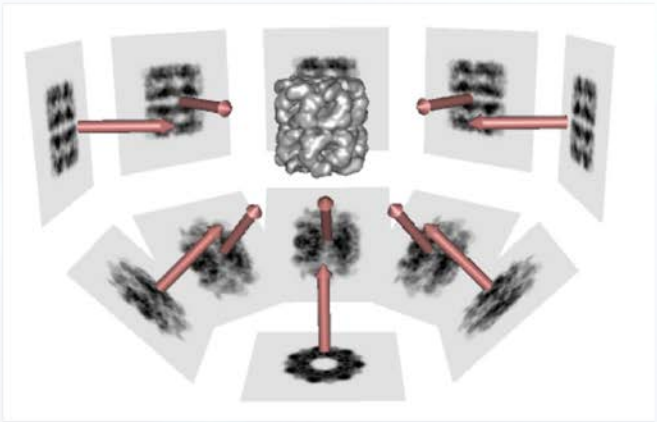
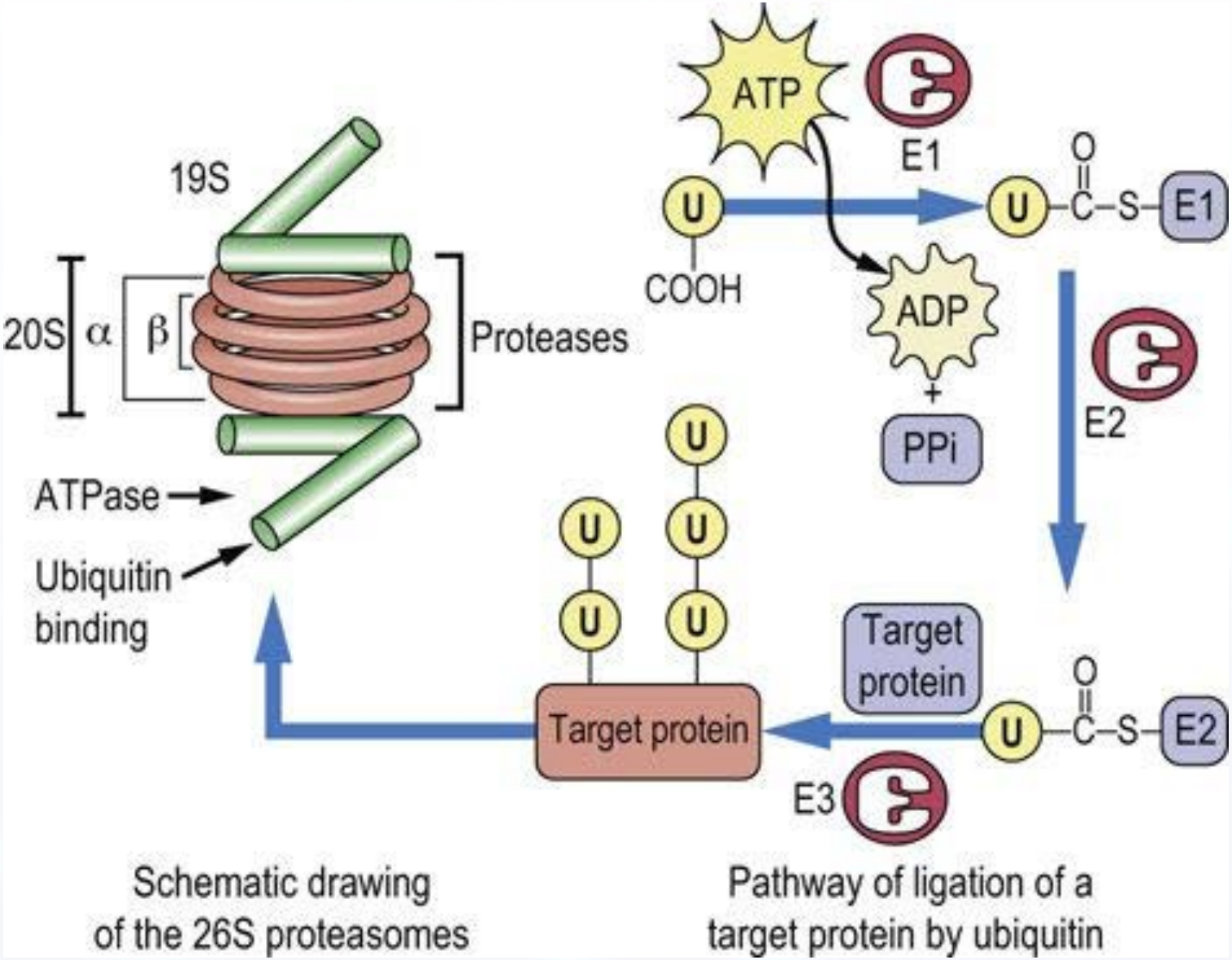


Image Alignment and Classification → Averaging → Orientation Search and Reconstruction

Ubiquitin-Proteasome Pathway



For the discovery of ubiquitin-mediated protein degradation

The Nobel Prize in Chemistry 2004

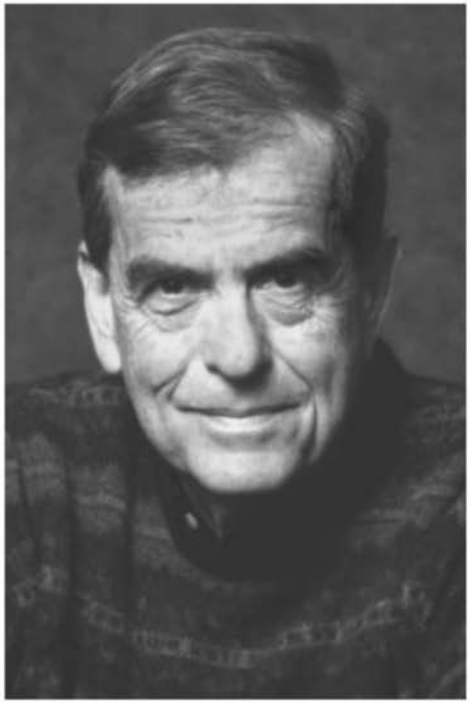


Photo: D. Porges

Aaron Ciechanover

Prize share: 1/3

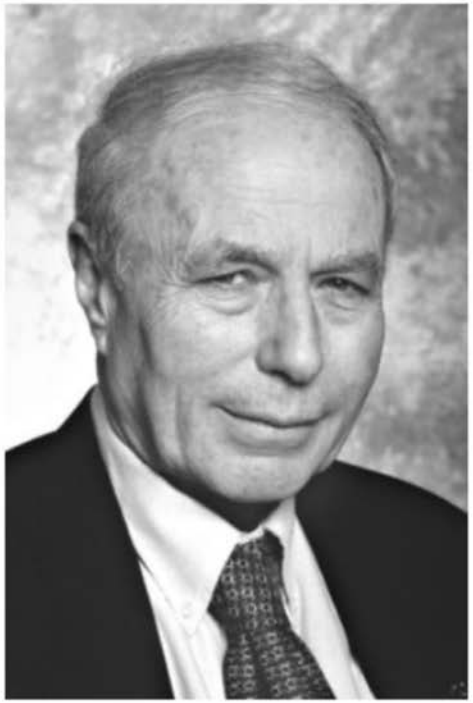


Photo: D. Porges

Avram Hershko

Prize share: 1/3

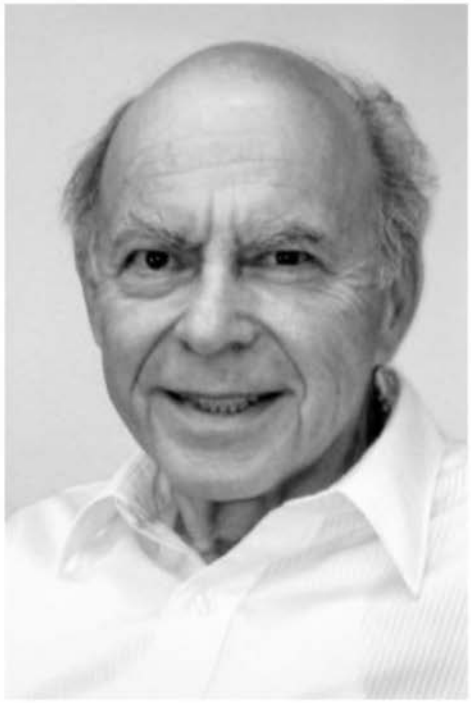


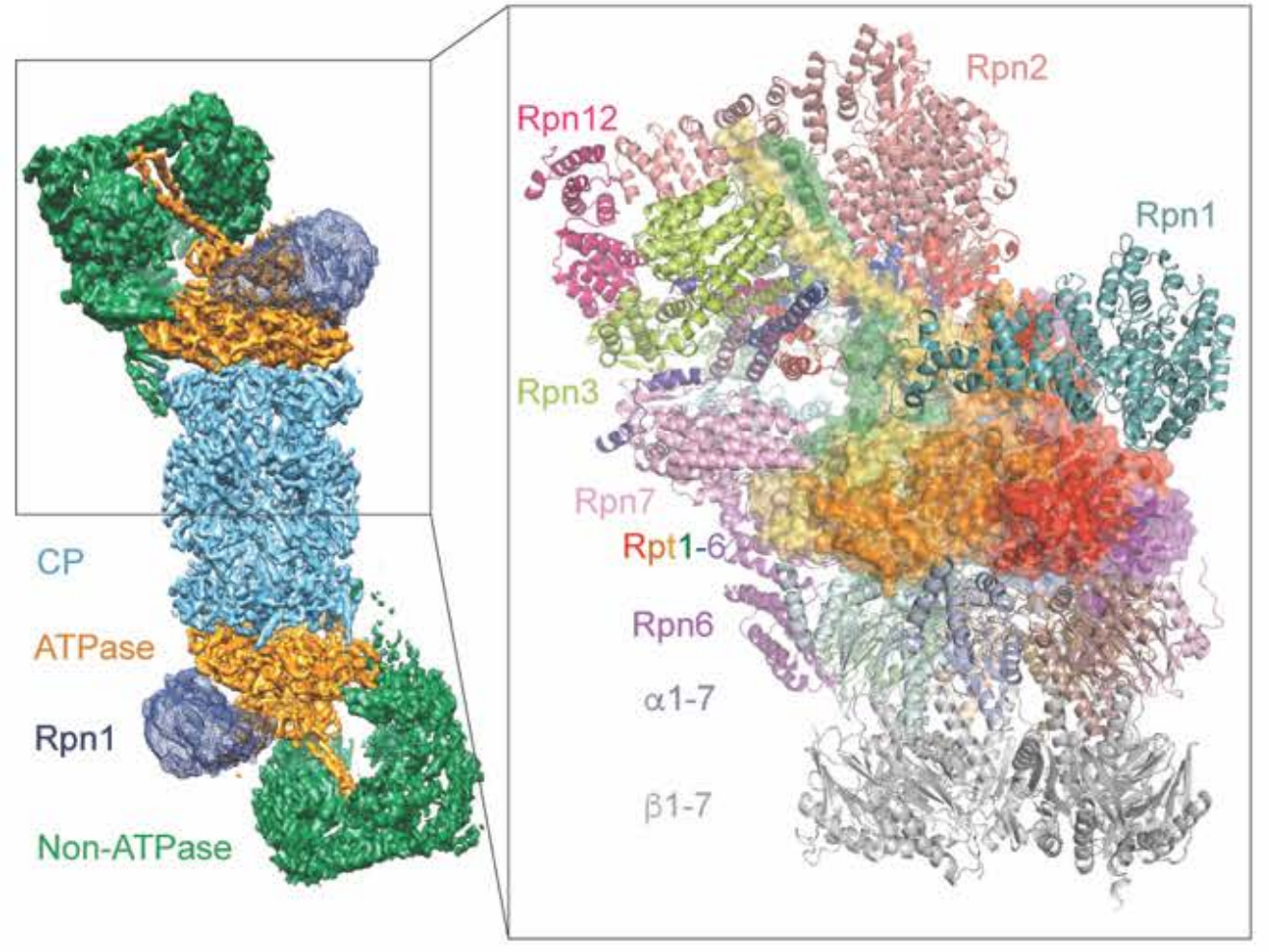
Photo from the Nobel Foundation archive.

Irwin Rose

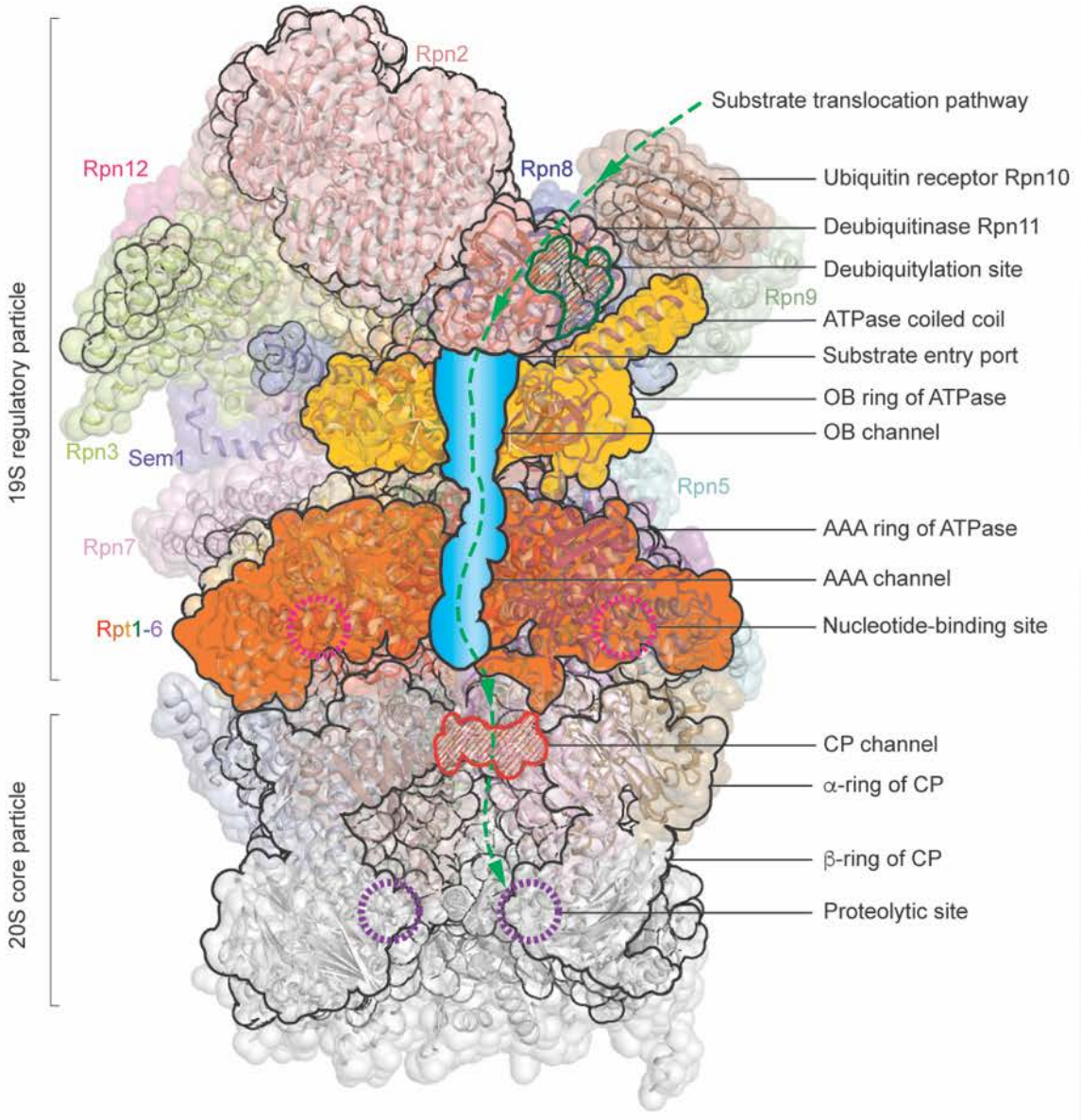
Prize share: 1/3

The largest molecular machine for degradation

- ❑ 28 subunits in 20S core particle (CP)
- ❑ 19 subunits in 19S regulatory particle (RP)

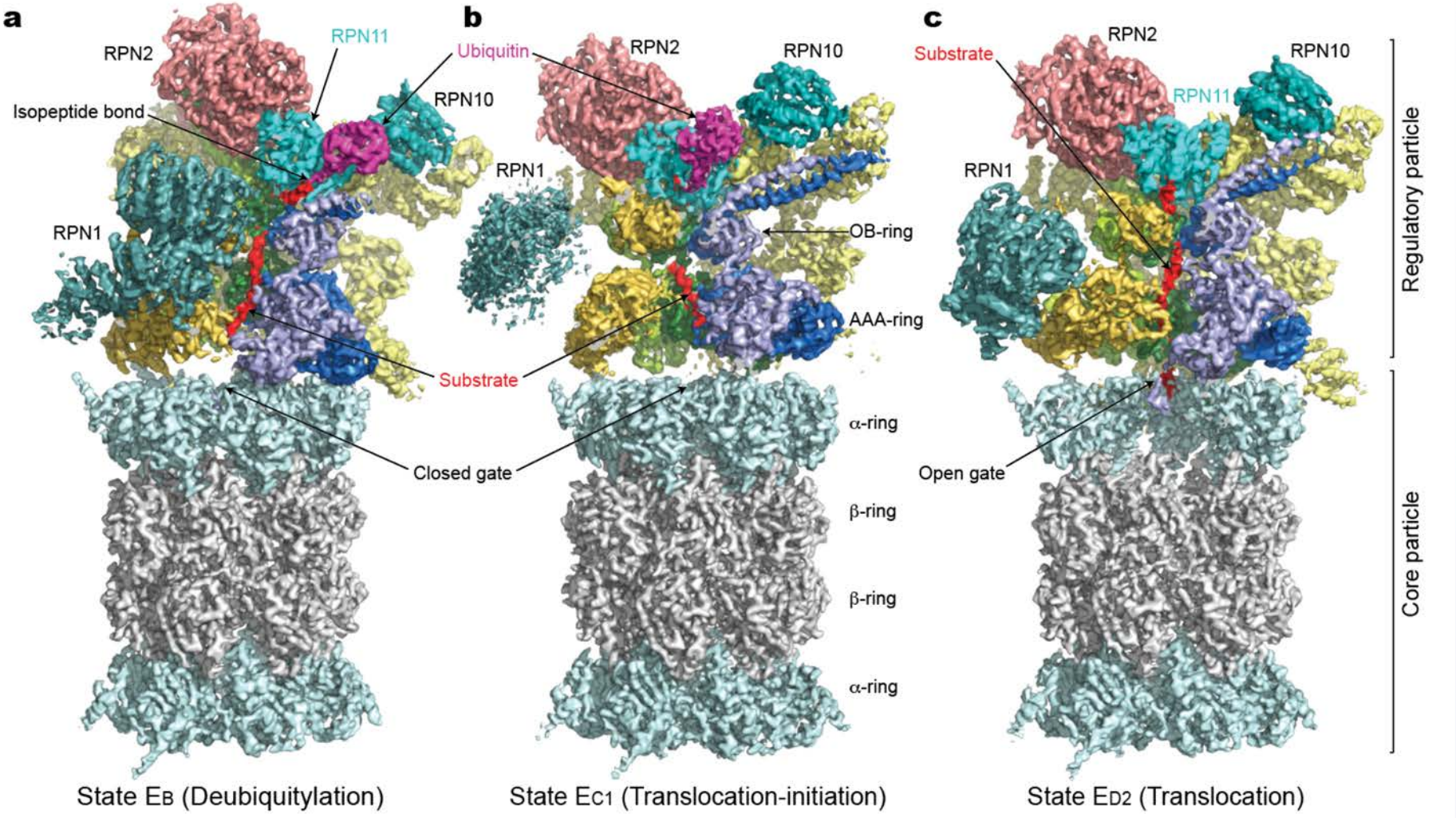


An assembly of multiple molecular sub-machines



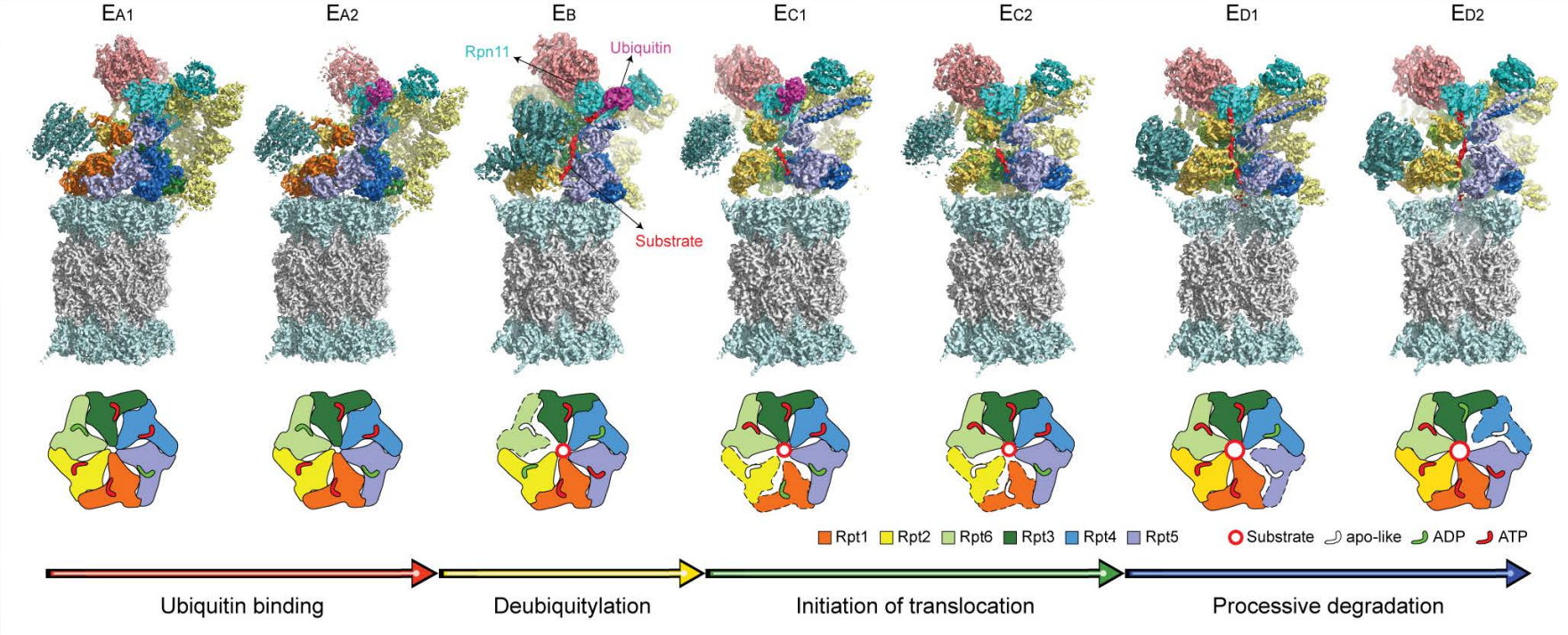
- ❑ **Ubiquitin receptor: capture ubiquitin-tagged substrates**
- ❑ **Deubiquitinase: remove ubiquitin tags from the substrate**
- ❑ **AAA-ATPase Unfoldase: unfold the substrate**
- ❑ **Degradation chamber: degrade the substrate**

Substrate-engaged proteasome



Y. Dong, S. Zhang, …, Y. Mao. *Nature* 565: 49-55 (2018).

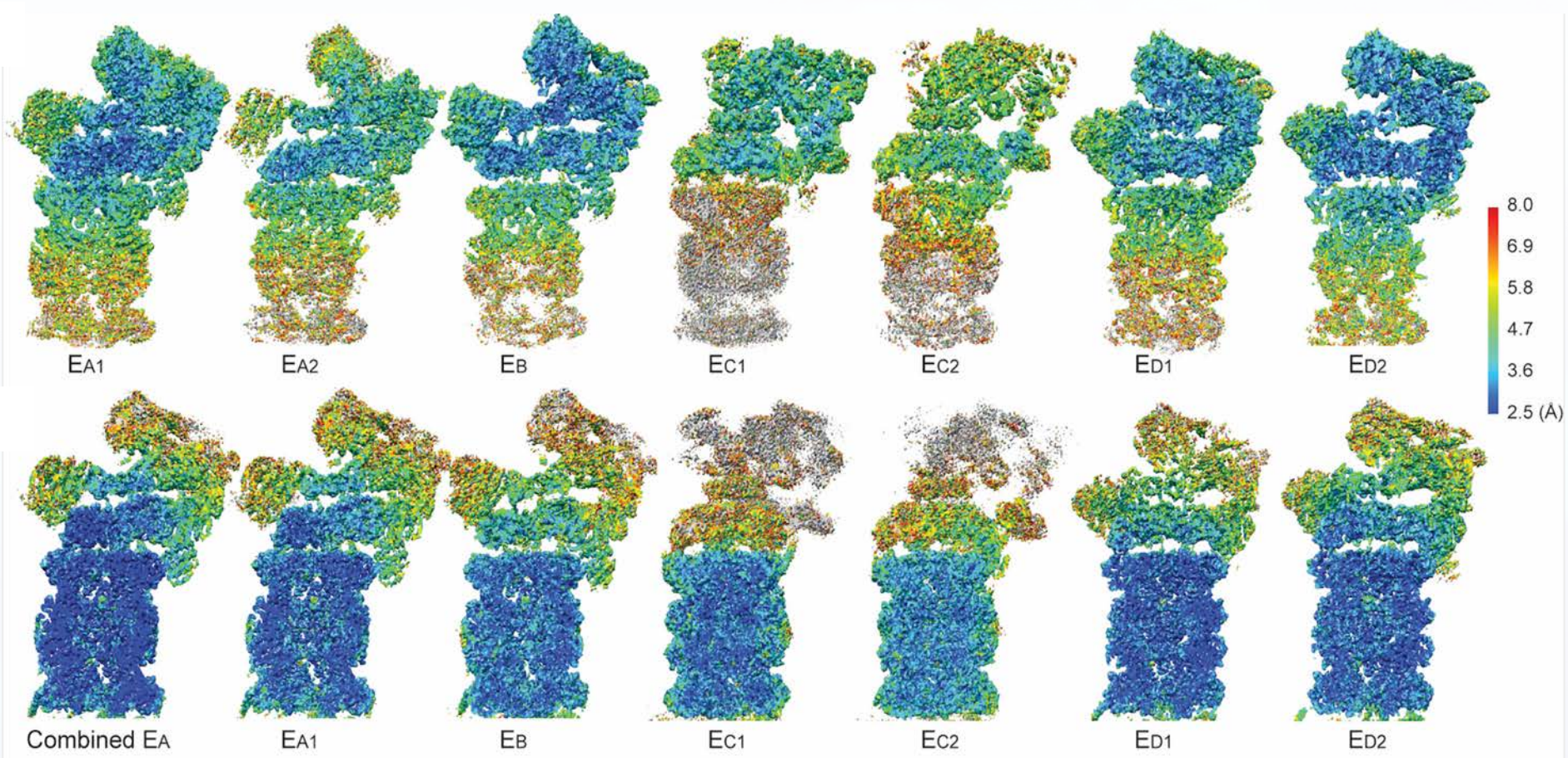
Substrate-engaged human proteasome



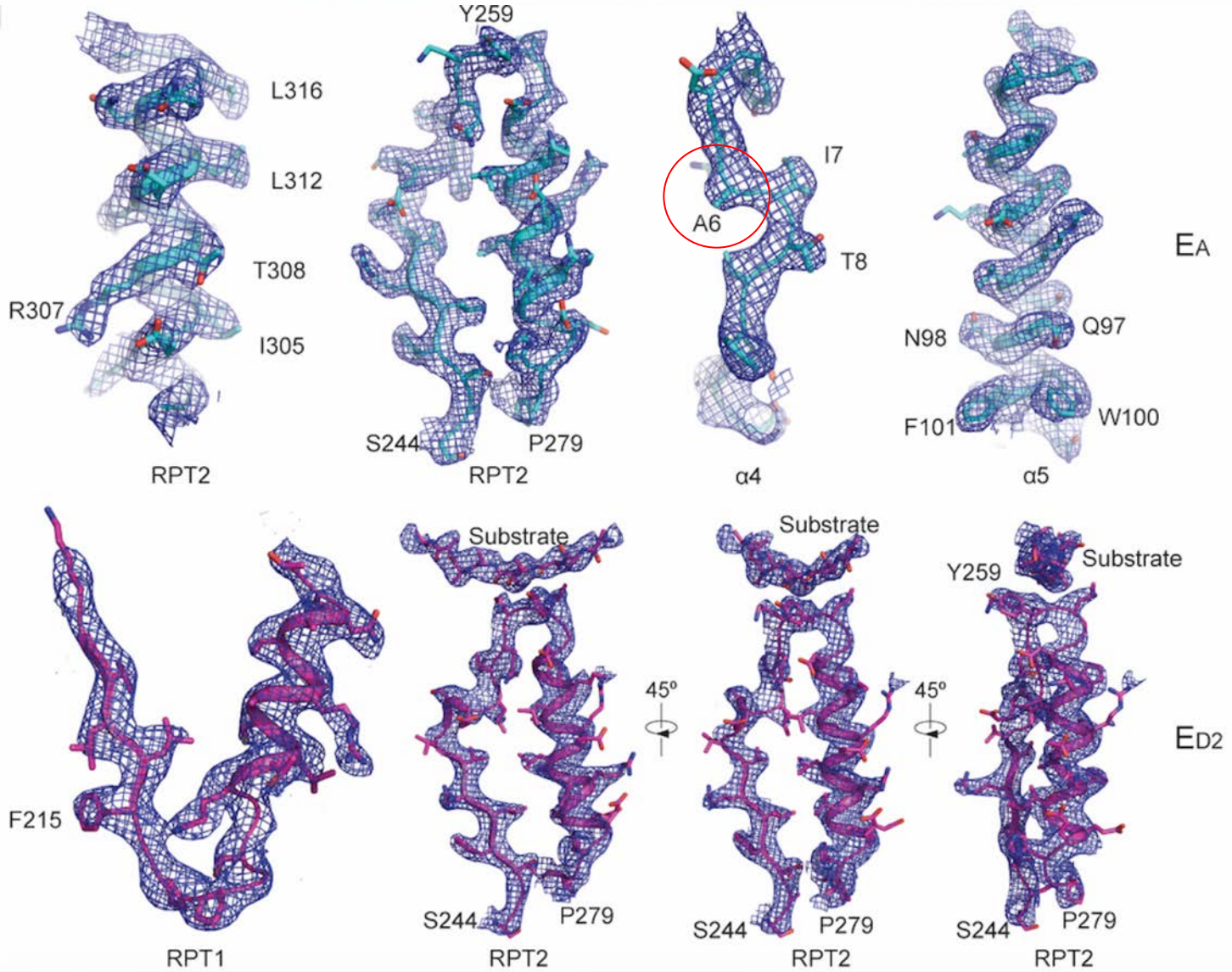
Final particle images (no.)	184,848	105,157	79,691	242,965	112,776	71,651	288,915	348,646
Map resolution (Å)	2.8	3.0	3.2	3.3	3.5	3.6	3.3	3.2
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143
Map resolution range (Å)	2.5-5.8	2.5-5.8	2.5-6.9	2.5-4.7	2.8-5.8	3.0-6.9	2.5-5.8	2.5-4.7

Y. Dong, S. Zhang, …, Y. Mao. *Nature* 565: 49-55 (2018).

Local resolution

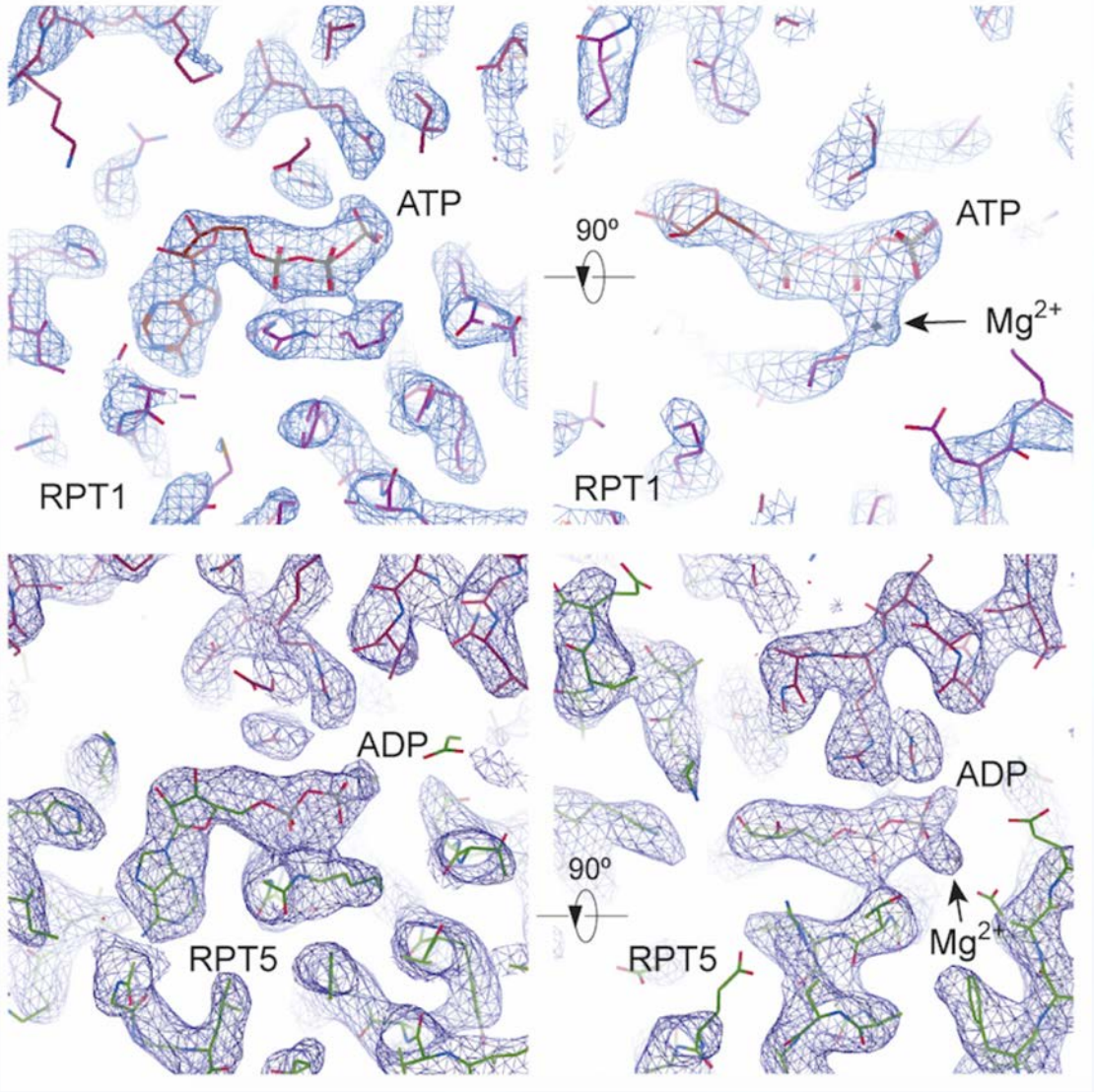


High-resolution density maps



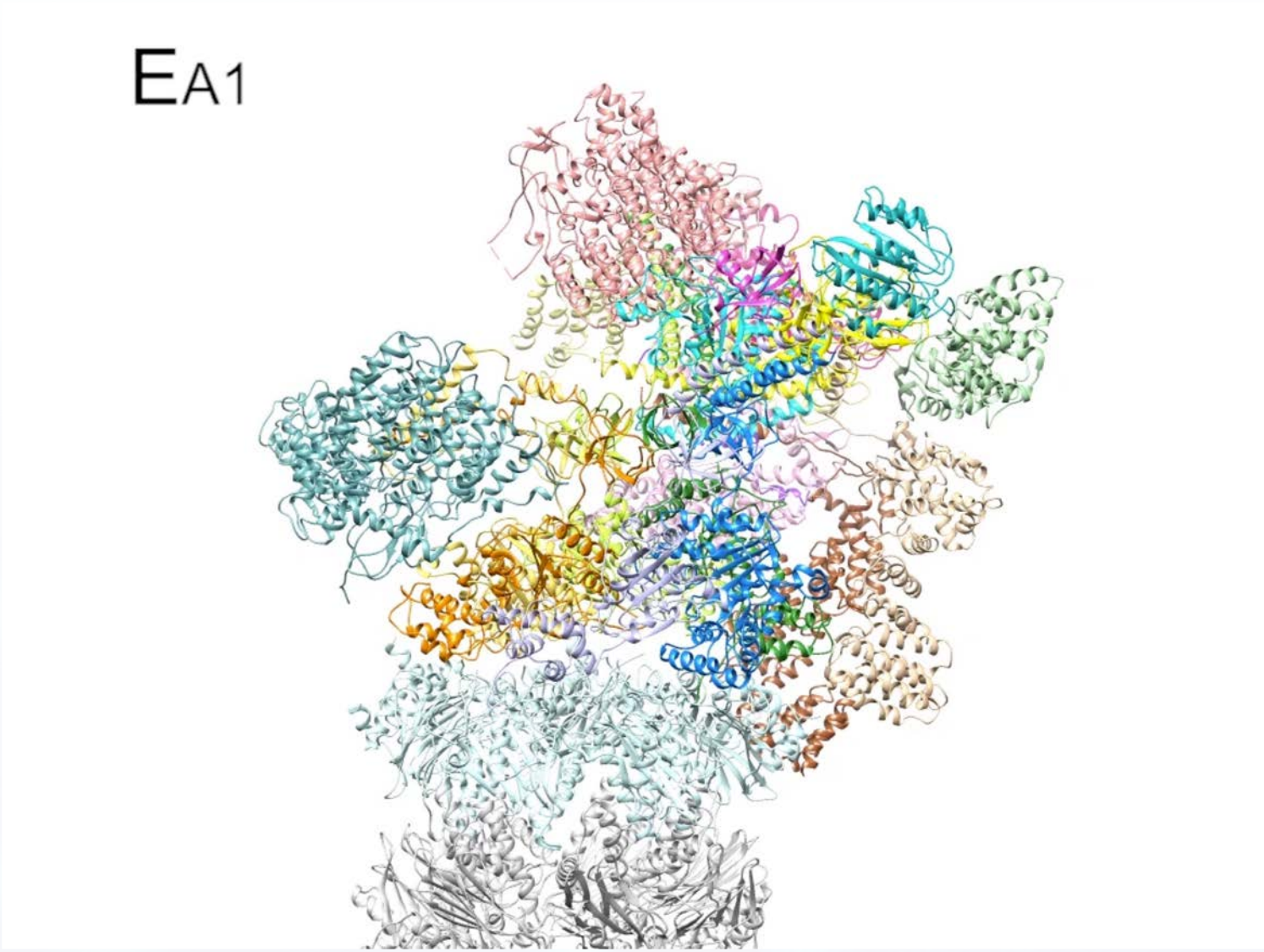
Y. Dong, S. Zhang, …, Y. Mao. *Nature* doi: 10.1038/s41586-018-0736-4 (2018).

Observation of single magnesium ions

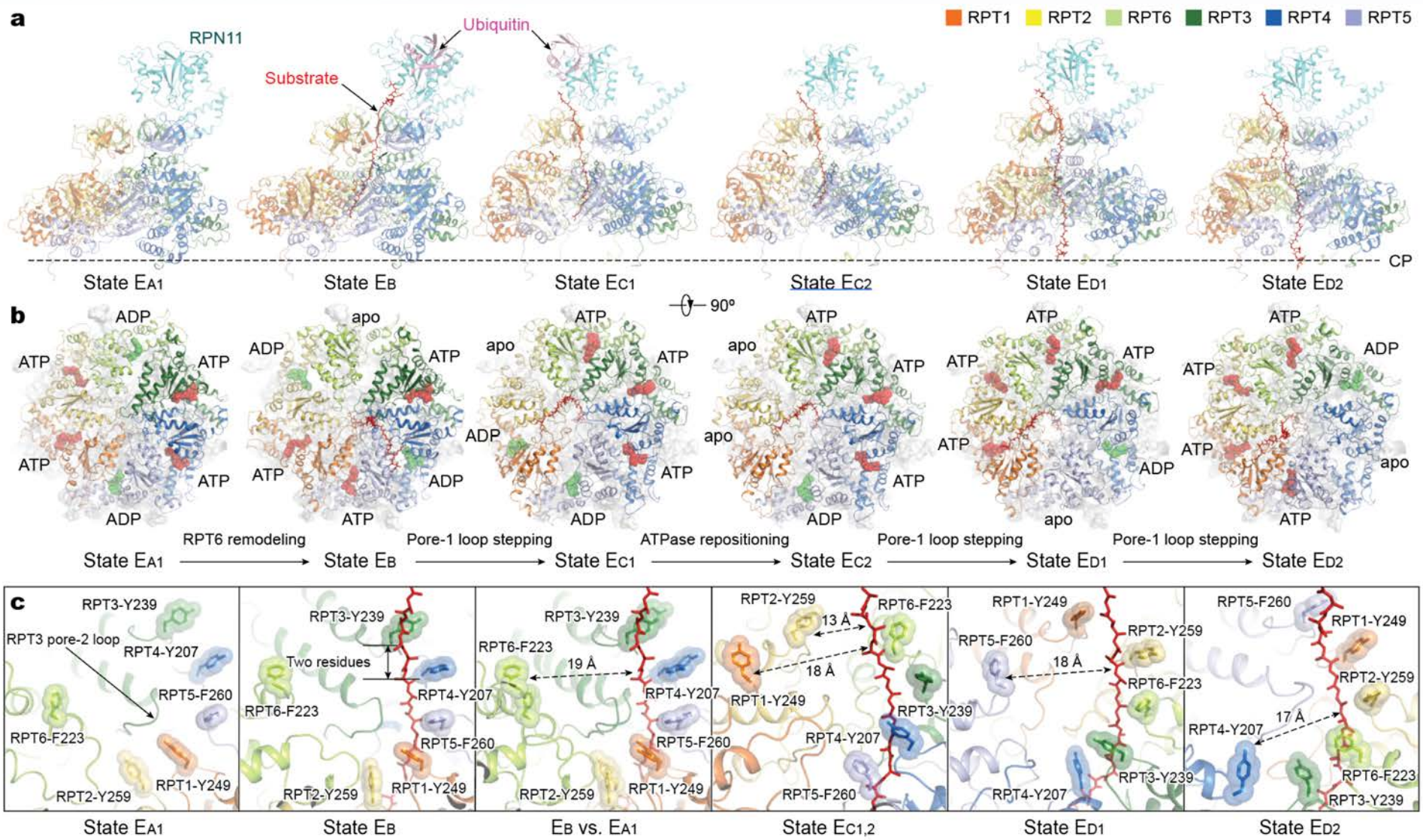


Y. Dong, S. Zhang, …, Y. Mao. *Nature* doi: 10.1038/s41586-018-0736-4 (2018).

Substrate-processing dynamics

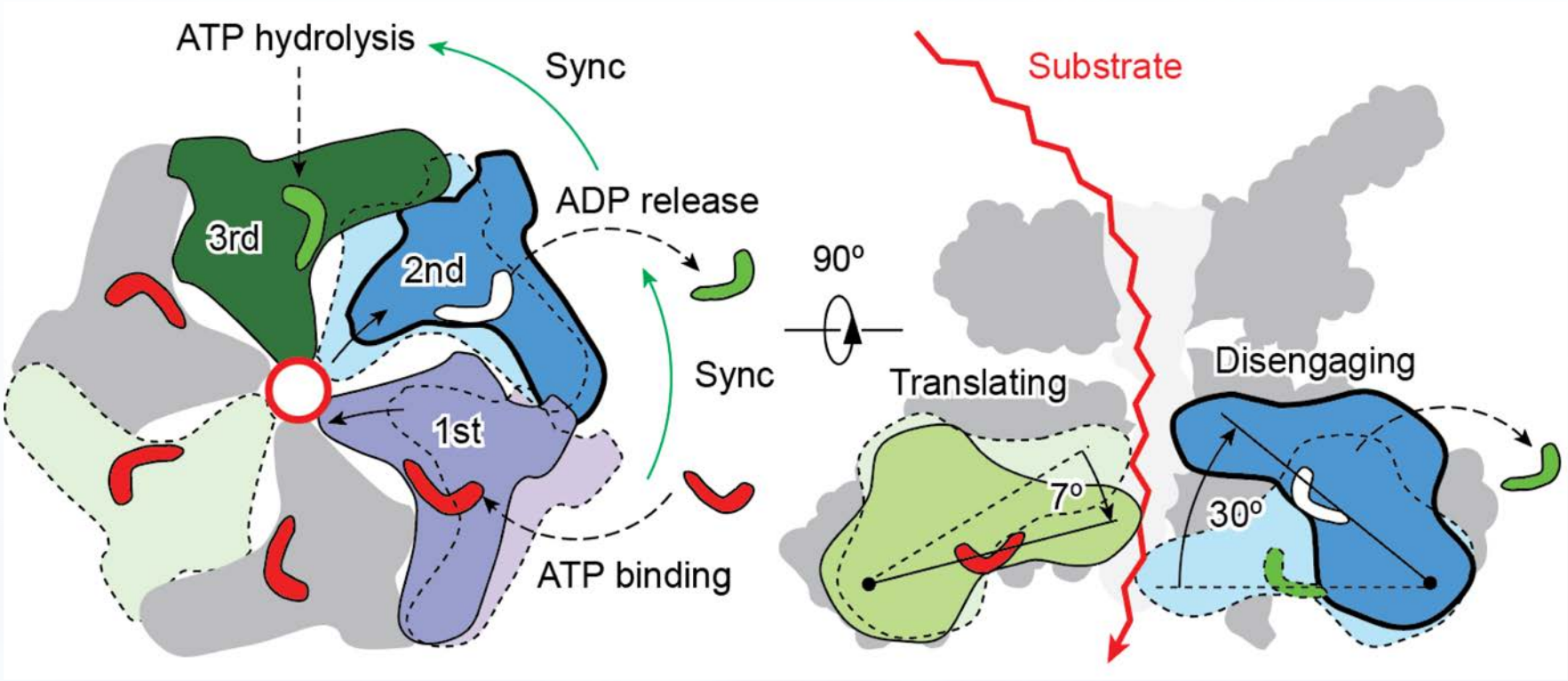


ATP hydrolysis powers the molecular motor

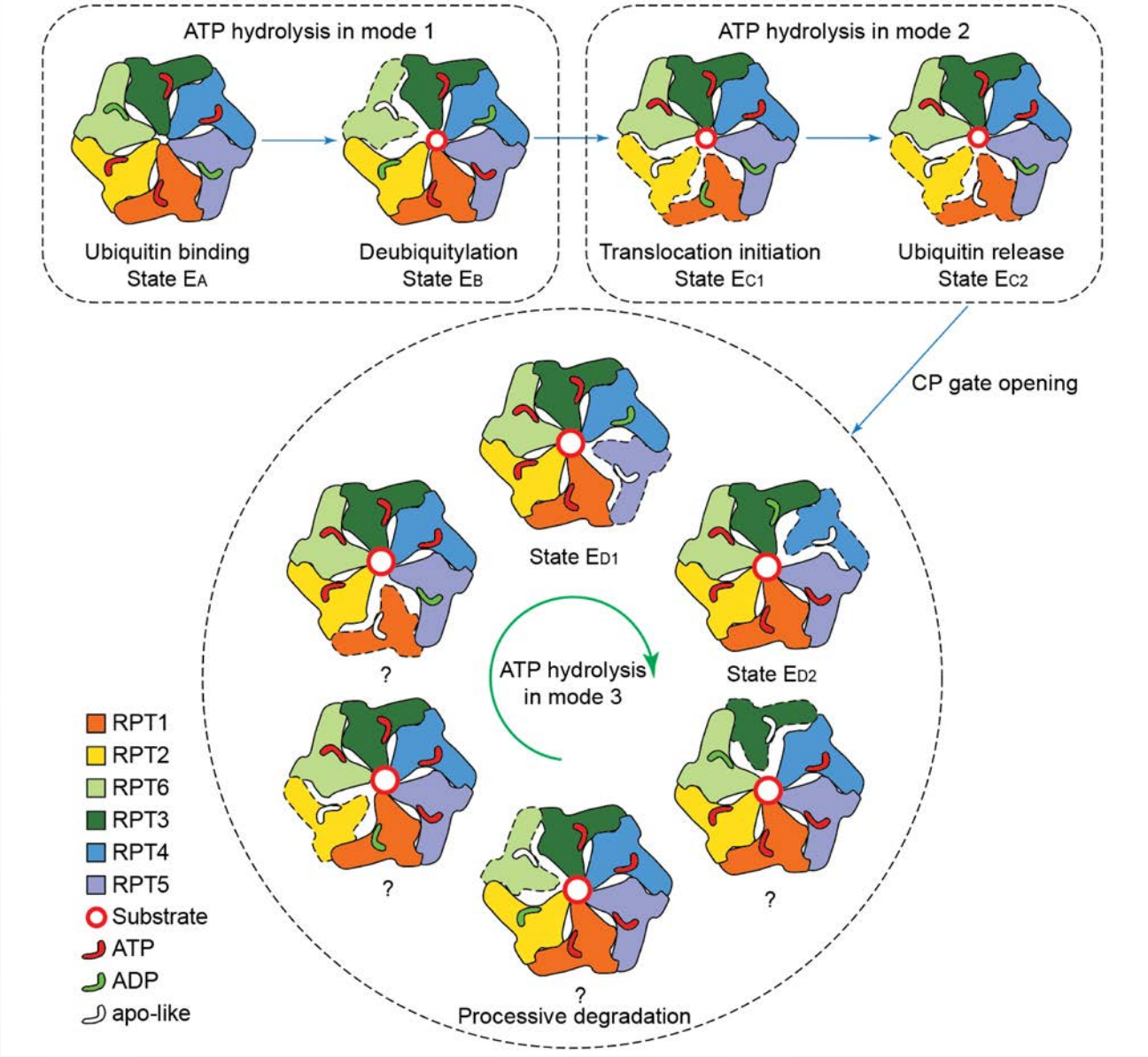


Dynamics of the motor unfolding substrate

- Coordination of three adjacent ATPases



Three modes of ATP hydrolysis

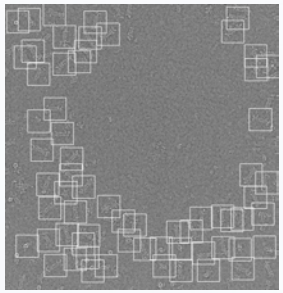


Workflow of Single-Particle Image Analysis

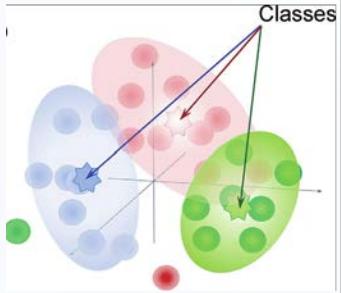
Micrograph Pre-processing:
Gain/drift correction
Quality control

Density Map Post-processing:
B-factor correction
Local resolution measure

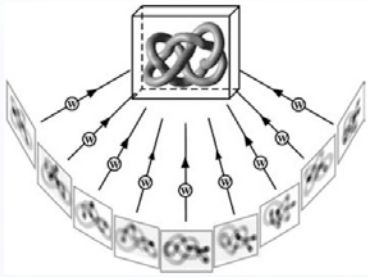
Particle Selection



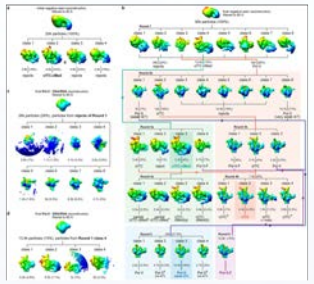
2D Clustering



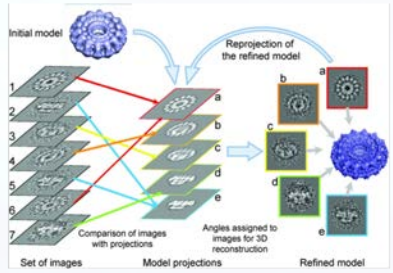
Generating Initial Model



3D Classification



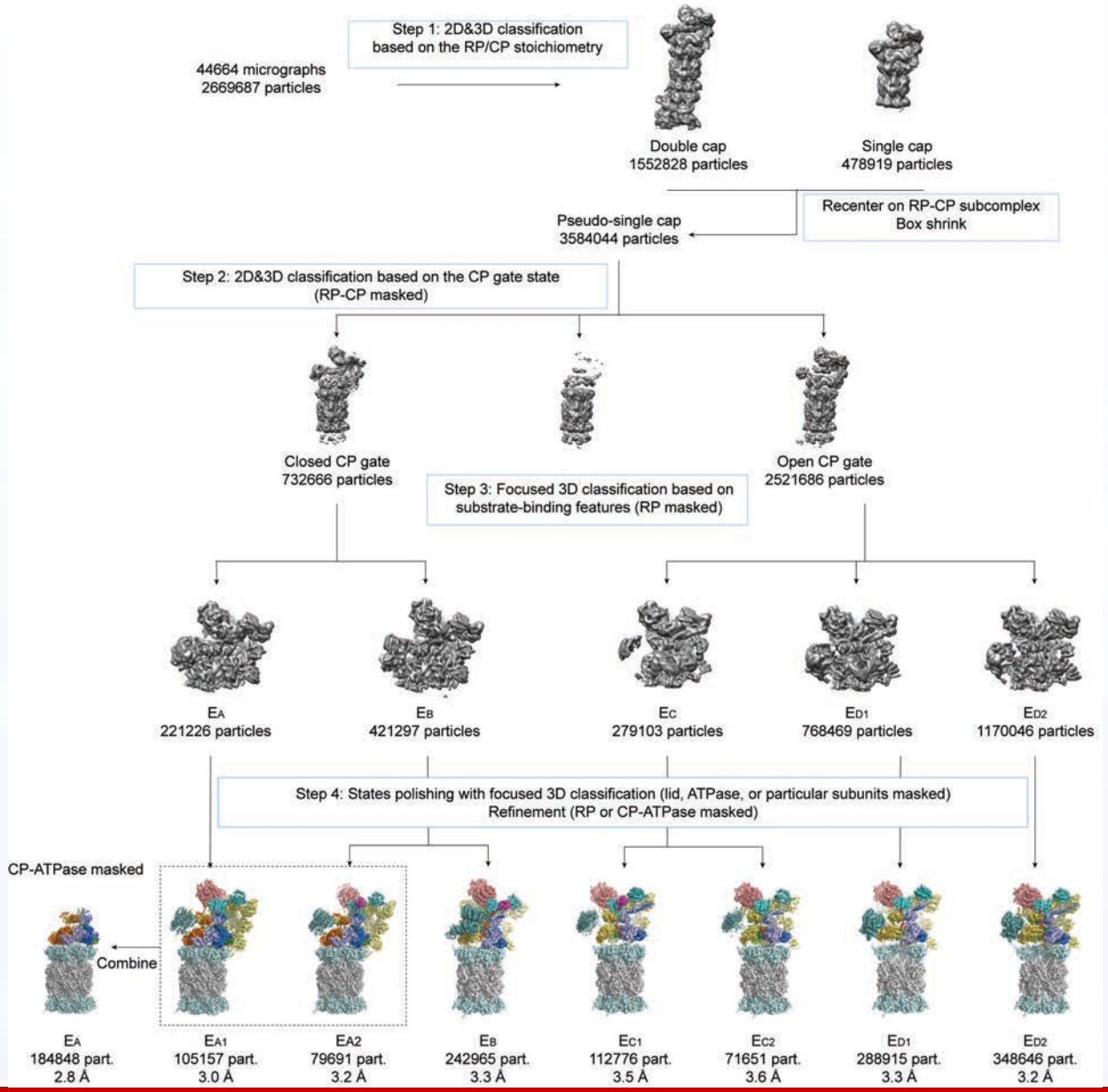
High-Resolution Refinement



Particle Purification

Class Verification

Hierarchical focused 3D classification



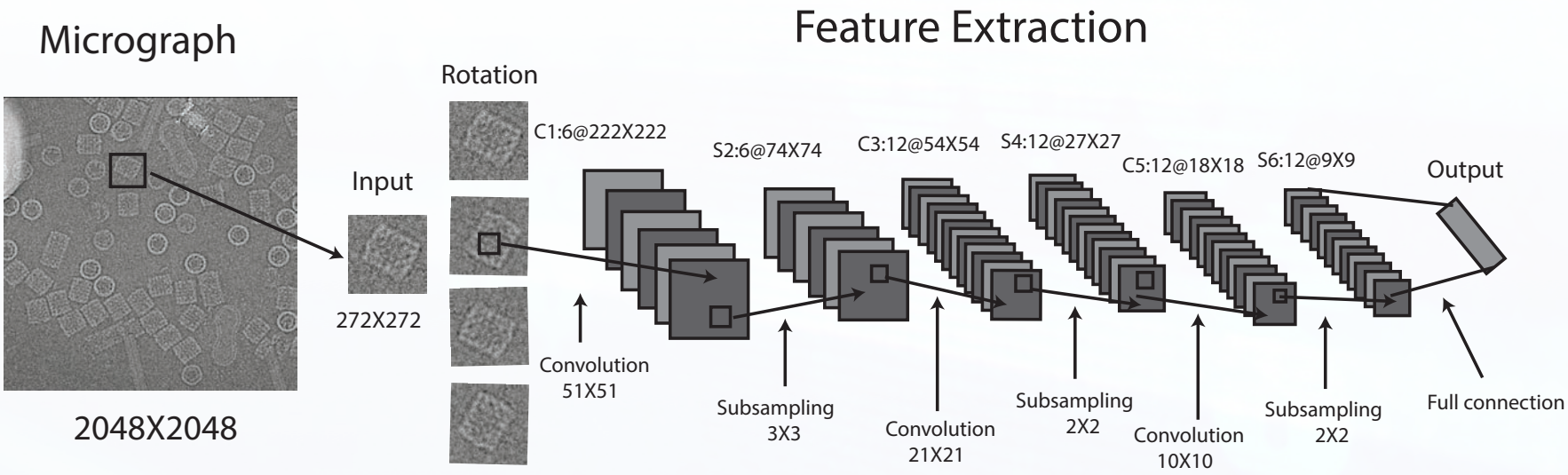
Comparison of approaches in analyzing protein dynamics

	NMR	Single-molecule FRET	Small-angle X-ray scattering	Cryo-EM
Time resolution/scale	fs	ms	ms	ms to seconds
Real-time	Yes	Yes	Yes	Pseudo
Spatial resolution	angstrom	>1 angstrom	nanometer	> 2.5 angstrom
Full atomic modelling	Yes	No	No	Yes
Protein size	<300kDa	Any	Any	>100 kDa
Labeling	Yes	Yes	No	No
Energy landscaping	Complete	Partial	Low-resolution and partial	Complete

Summary of GPU acceleration for cryo-EM

	Software	GPU Acceleration	Based on Deep Learning
Motion Correction	MotionCor2	Yes	ms
CTF parameters	GCtf	Yes	Yes
Particle picking	DeepEM	Yes	nanometer
Deep 2D classification	ROME	**No**	No
3D classification	RELION	Yes	Any
Deep 3D classification	ROME	**No**	No
Cryo-EM refinement	RELION CryoSPARC	Yes	Low-resolution and partial

Convolutional Neural Network

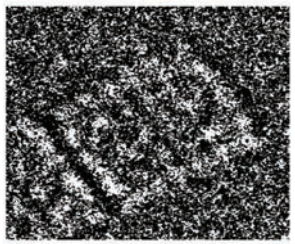


The 6C-3S-12C-2S-12C-2S CNN we designed for KLH dataset.

Y. Zhu, Q. Ouyang, Y. Mao. *BMC Bioinformatics* 18, 348 (2017).

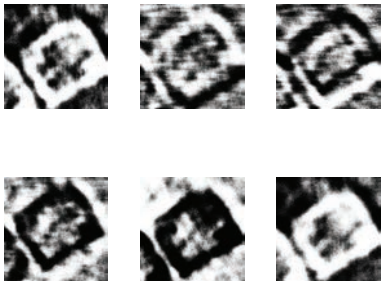
.

The learned feature maps of an example input image

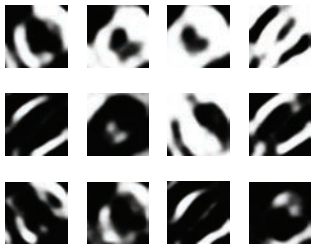


Input Image

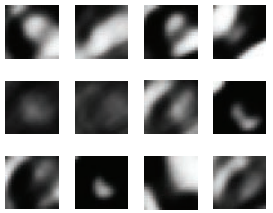
Layer C1 Feature Maps



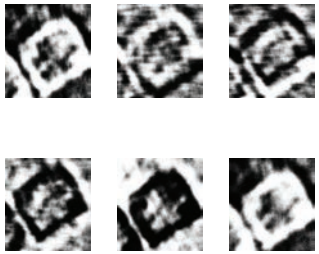
Layer C3 Feature Maps



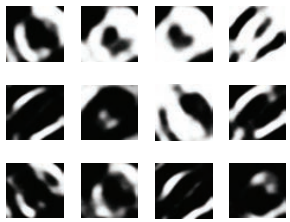
Layer C5 Feature Maps



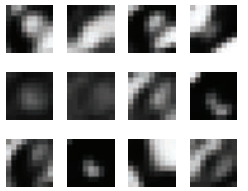
Layer S2 Feature Maps



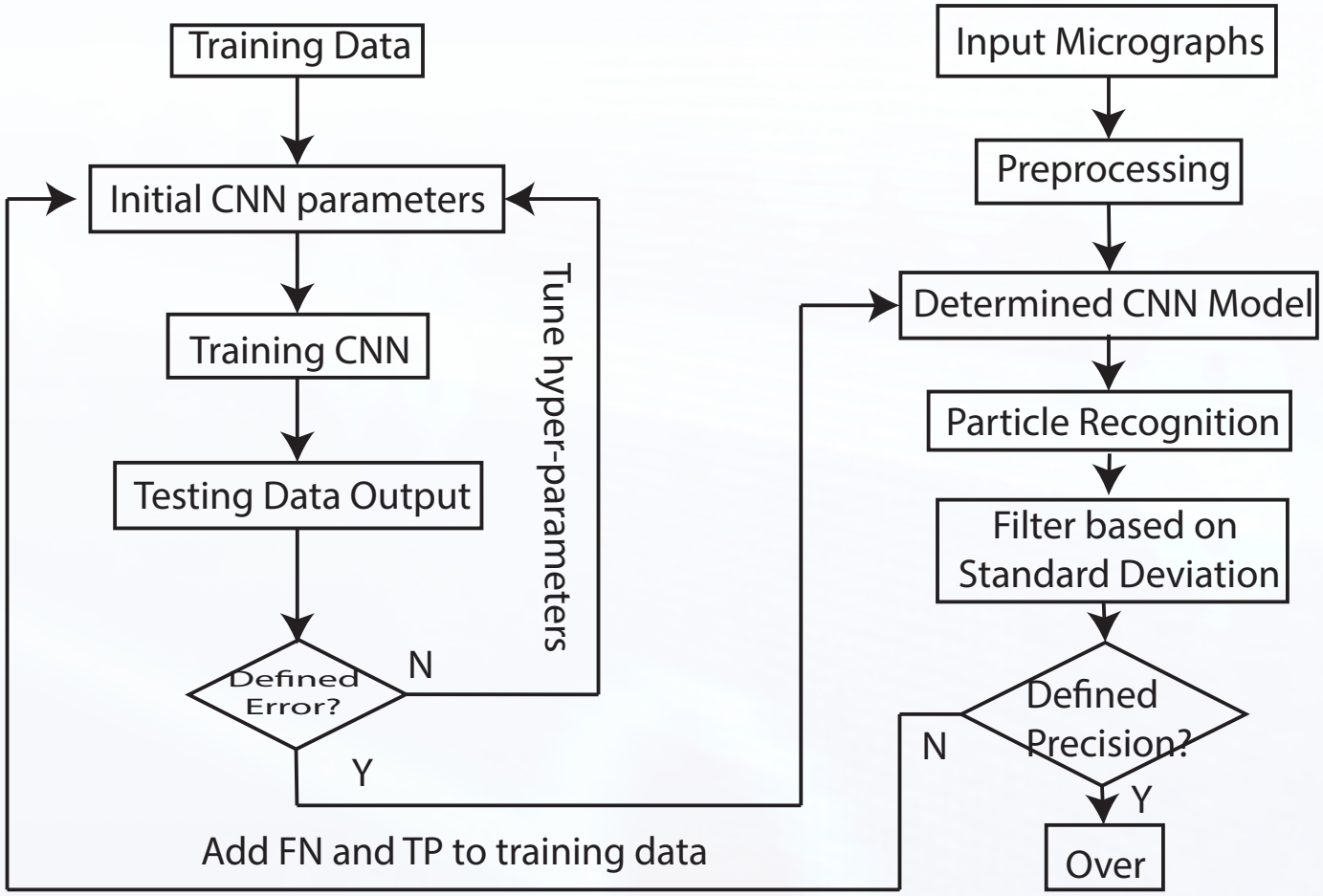
Layer S4 Feature Maps



Layer S6 Feature Maps

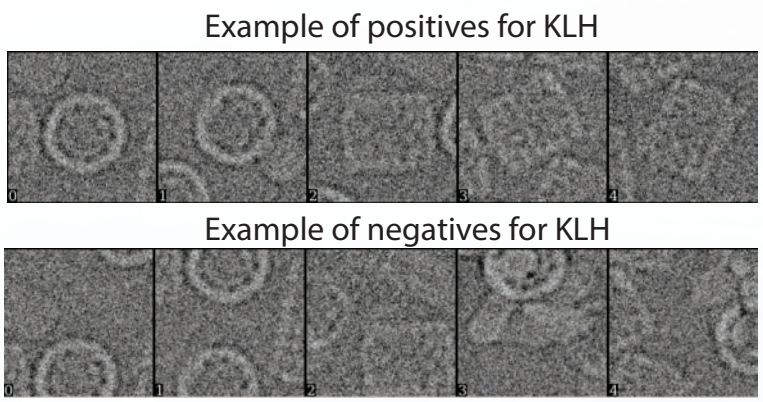


DeepEM Algorithm Workflow

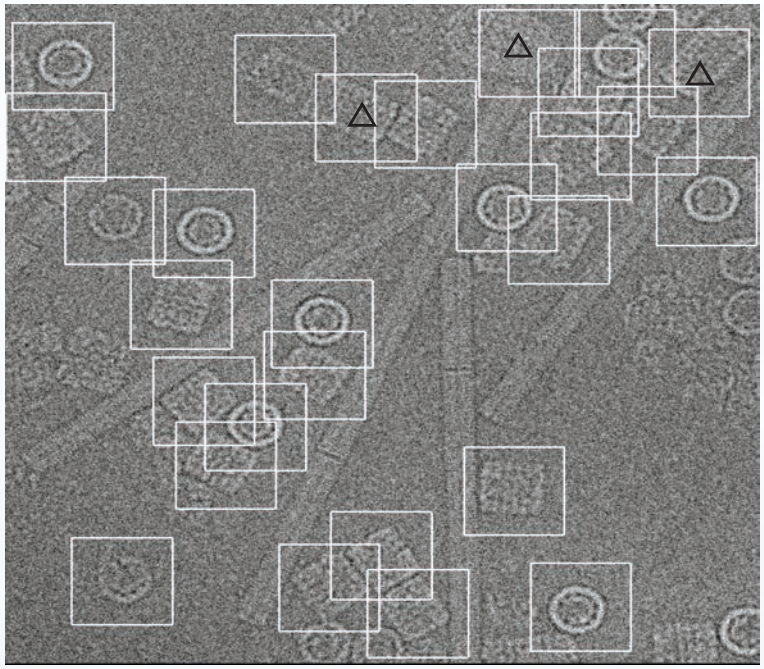


Keyhole Limpet Hemocyanim (KLH)

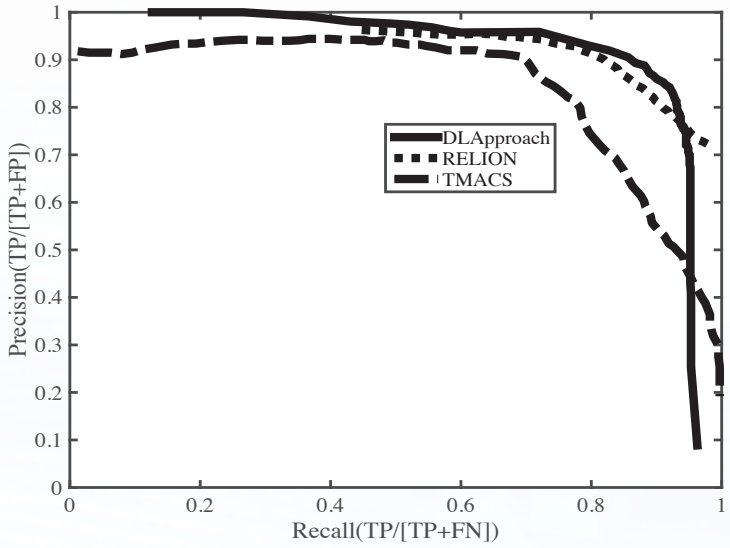
(a)



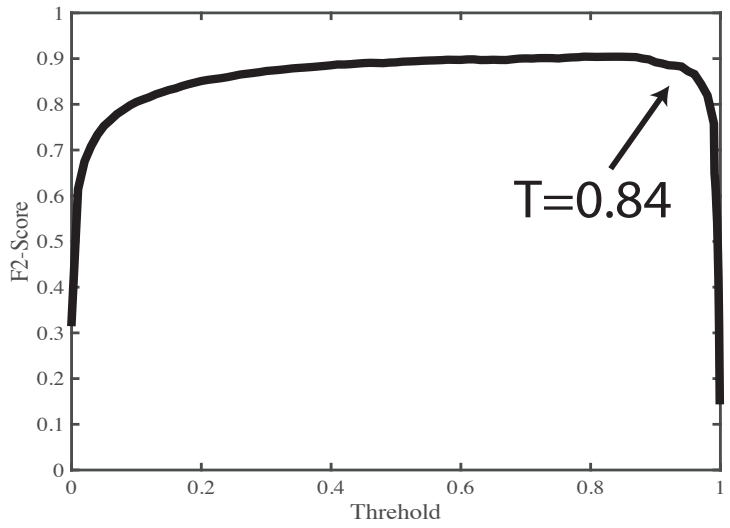
(b)



(c)



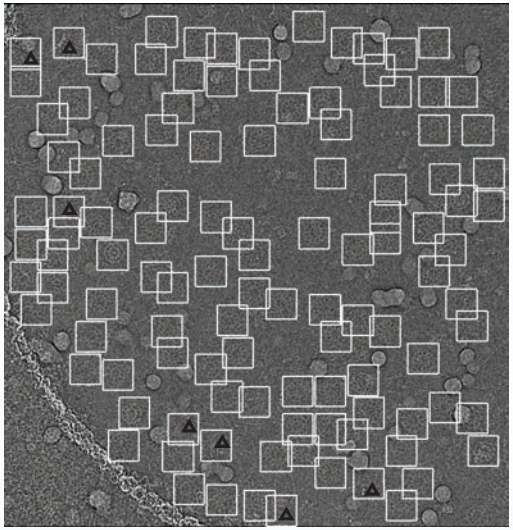
(d)



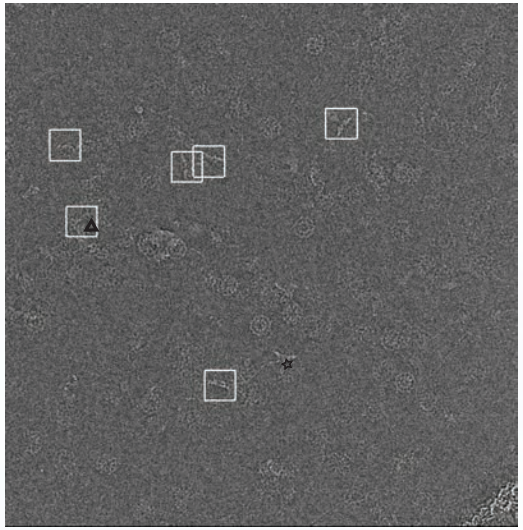
Zhu, Y., *et al*. *IEEE Trans. Med. Imaging* 22, 1053-1062 (2003)

Testing on some challenging datasets

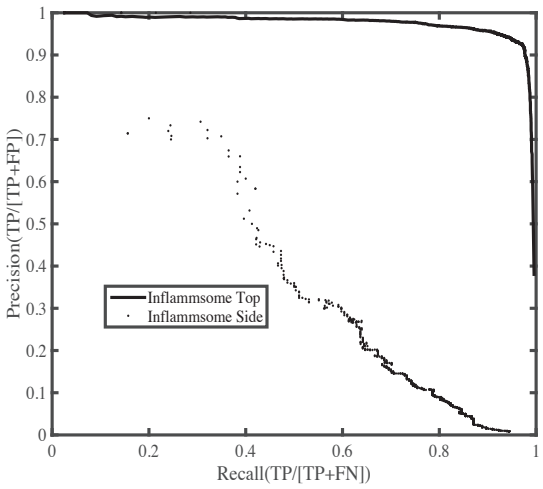
(a)



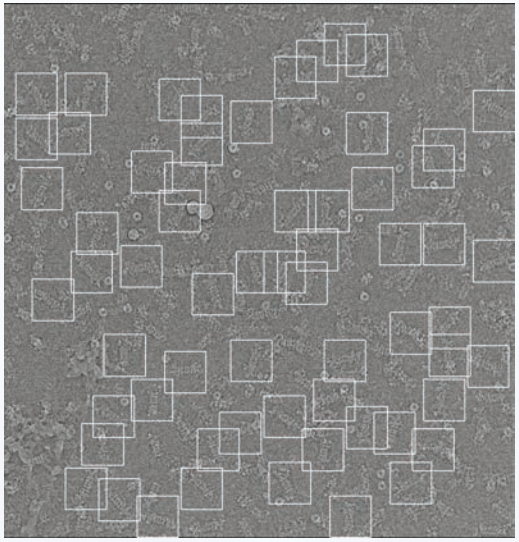
(b)



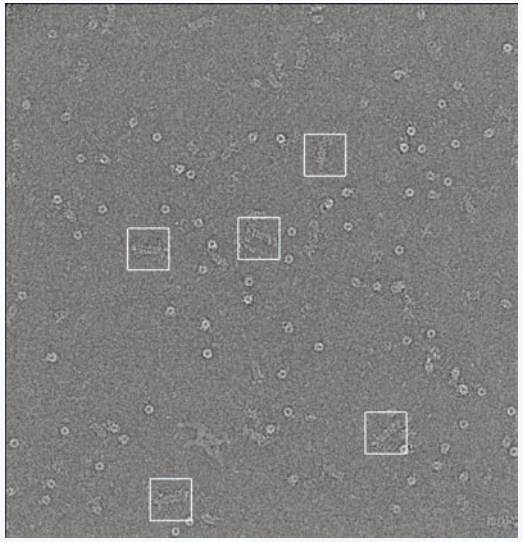
(c)



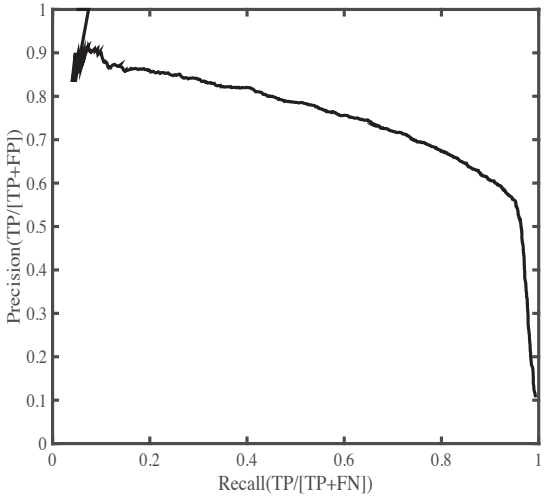
(d)



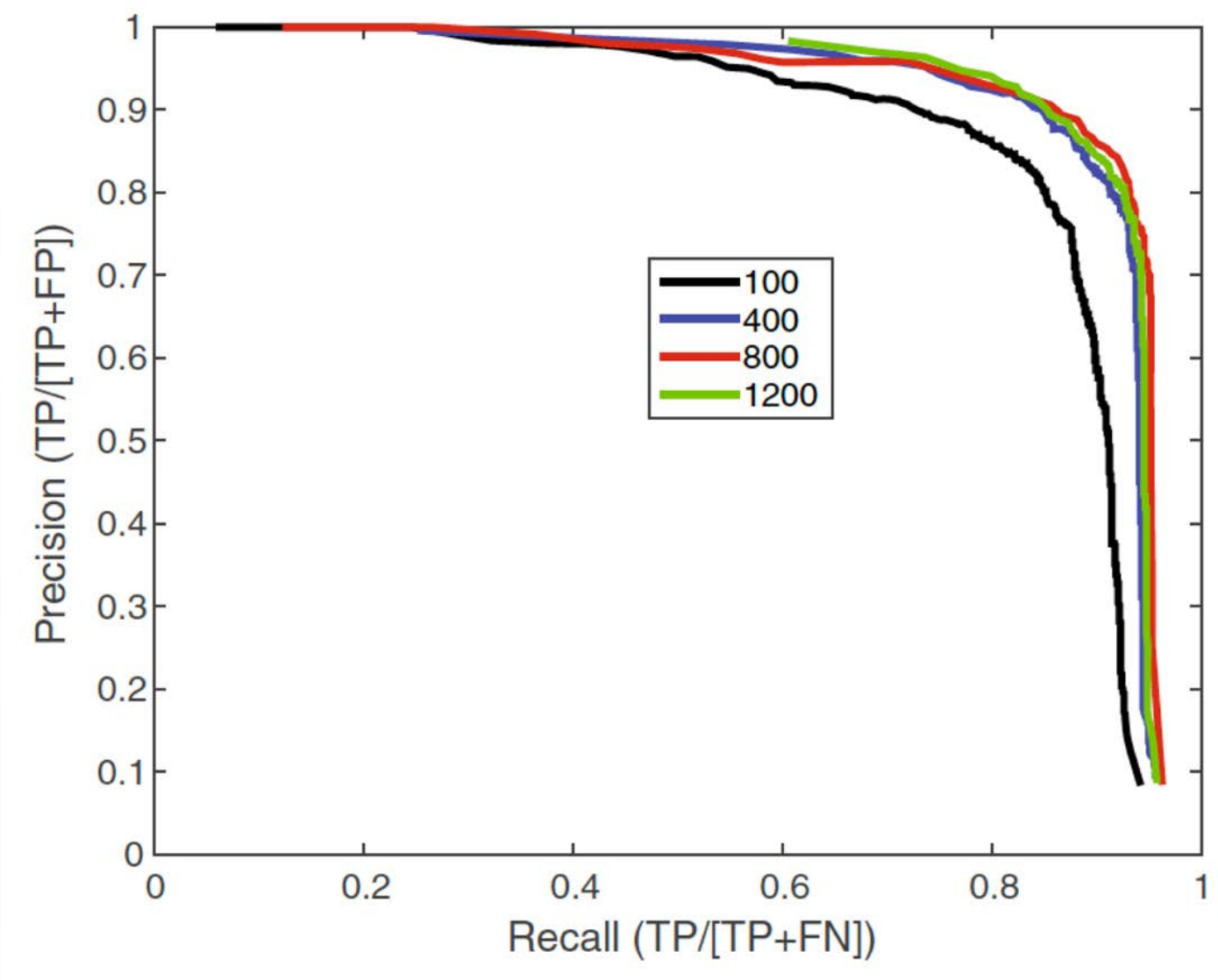
(e)



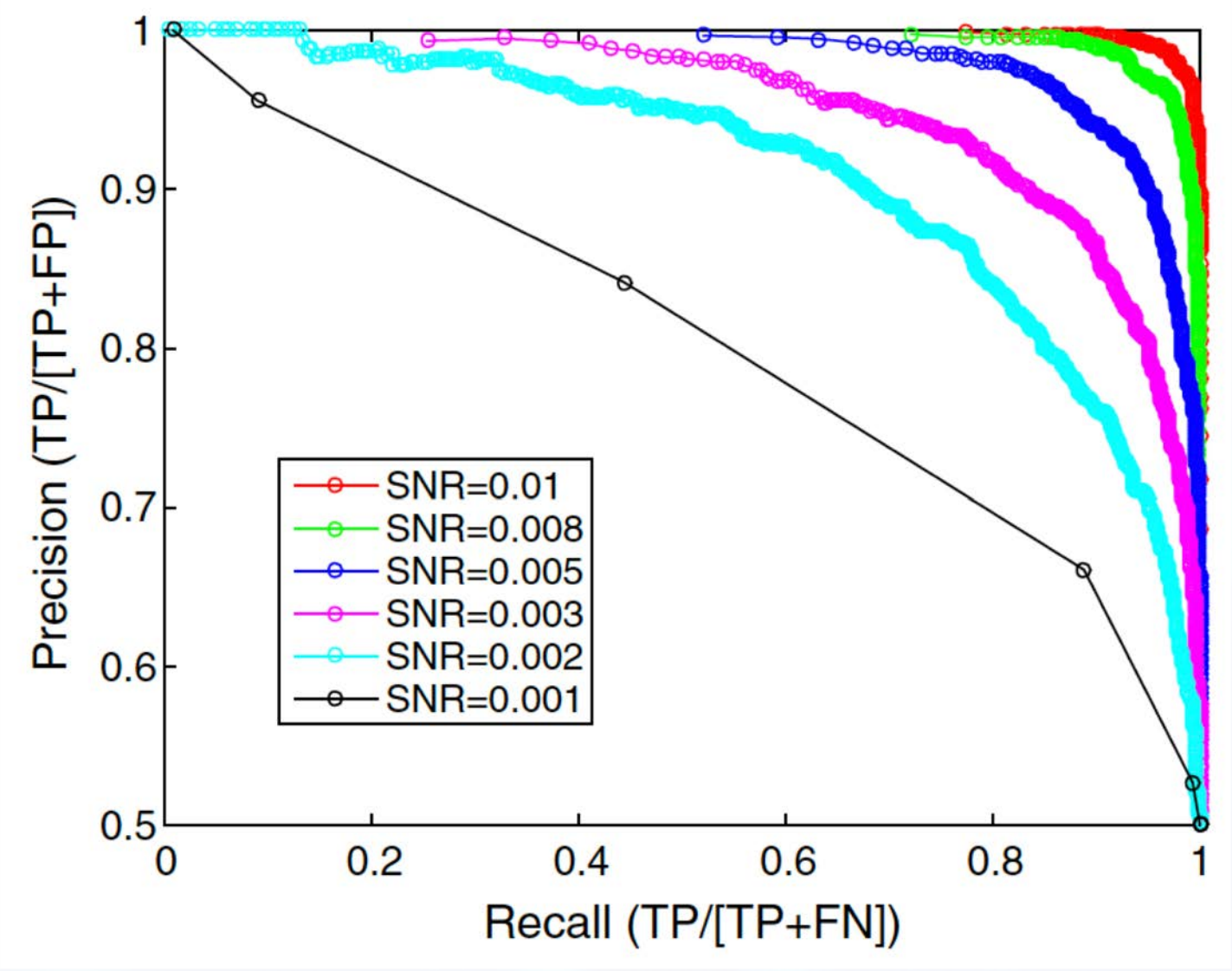
(f)



Effect of training dataset sizes



DeepEM exhibits good performance on low SNR



DeepEM on GPU

- Geforce GTX 970, running Matlab 2016a and CUDA 8.0: **40-190 seconds per micrograph**
- Tesla K20c GPU with CUDA 8.0:

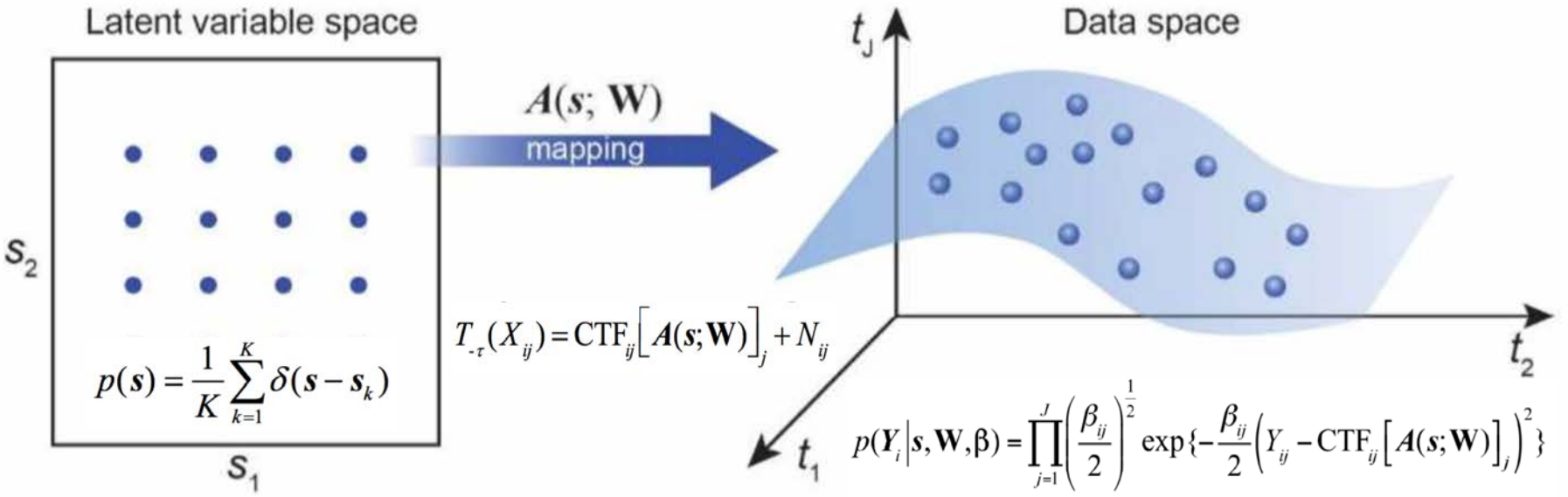
	Pdb1f07	KLH	10,005	10,184	10,075
Micrograph Size	1024*1024	2048*2048	3710*3710	3838*3710	4096*4096
Particle Size	100*100	272*272	180*180	256*256	300*300
Preprocessing	0.17	0.57	3.08	3.72	3.75
Segmentation	0.55	3.64	9.92	9.58	11.23
Classification	2.26	3.12	8.63	6.84	2.75
Postprocessing	0.34	4.59	20.03	6.99	30.38
Total Time of PIXER	3.32	11.94	41.67	47.17	48.07
DeepPicker	10.47	23.75	80.76	81.34	95.43
DeepEM	40.56	80.54	65.47	39.75	54.38

Y. Zhu, Q. Ouyang, Y. Mao. *BMC Bioinformatics* 18, 348 (2017).
J. Zhang, …, F. Sun, F. Zhang. *BMC Bioinformatics* 20, 41 (2019).

Summary

- DeepEM outperforms existing programs in particle recognition, allowing particle picking, selection and verification in an integrated fashion.
- With better training dataset or iterative training, one can improve the accuracy of particle recognition. In other words, the program can be trained to be “smarter”.
- DeepEM can be easily applied to batch processing.
- We expect that DeepEM marks the inception of applications of modern AI technology in expediting cryo-EM structure determination.

Statistical Manifold Learning: Generative Tomographic Mapping



$$\left[A(s;\mathbf{W})\right]_j = \sum_{m=1}^{M}\varphi_m(s)W_{mj}$$

$$\varphi_m(s) = \begin{cases} \exp\{-\frac{\| s - \mu_m \|^2}{2\sigma^2}\}, m \le M_{NL} \\ 1, m = M_{NL}+1 \end{cases}$$

- **The shortest line between two points s_i and s_j in the latent space is mapped to geodesic line on a manifold between $A(s_i)$ and $A(s_j)$ in the data space.**
- **Latent space: the structural difference resulting from changes in the image orientation and/or the molecular conformations**
- **Data space: the Fourier transform of particle images**

J. Wu, Y. Ma, C. Congdon, B. Brett, S. Chen, Q. Ouyang, Y. Mao. *PLoS ONE* 12, e0182130 (2017).

SML optimizes a MAP objective function

- The maximum-likelihood estimator $\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(\boldsymbol{Y}_i \mid \Theta)$
- The maximum-a-posterior (MAP) estimate

$$\hat{\Theta} = \operatorname{argmax}\left[\sum^{N} \log p(\boldsymbol{Y}_i \mid \Theta) + \log p(\Theta)\right]$$

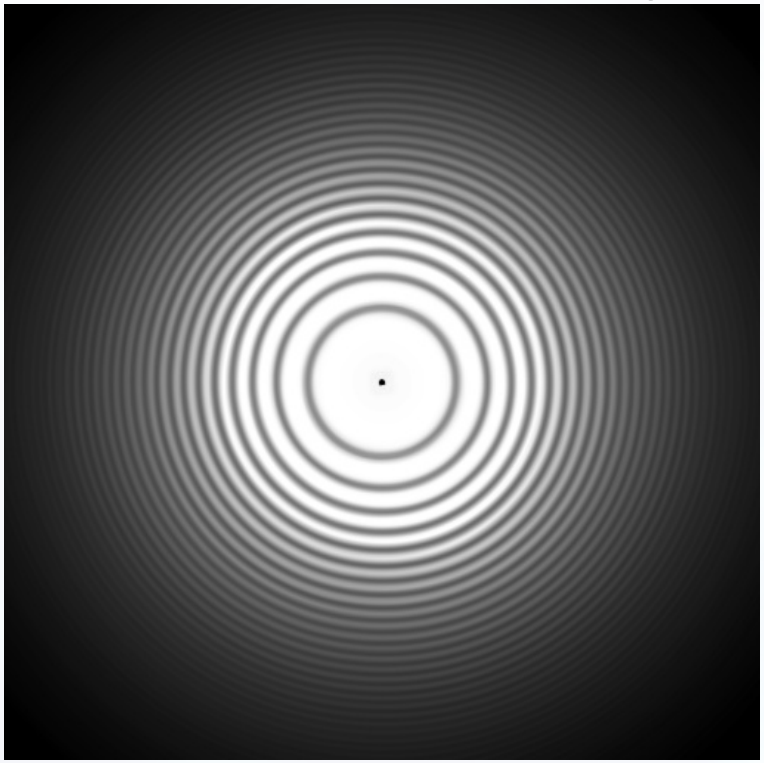
$$p(\mathbf{W}) = \left(\frac{\alpha}{2\pi}\right)^{\frac{MJ}{2}} \exp\left(-\frac{\alpha}{2}\sum_{m=1}^{M}\sum_{j=1}^{J} W_{mj}^2\right)$$

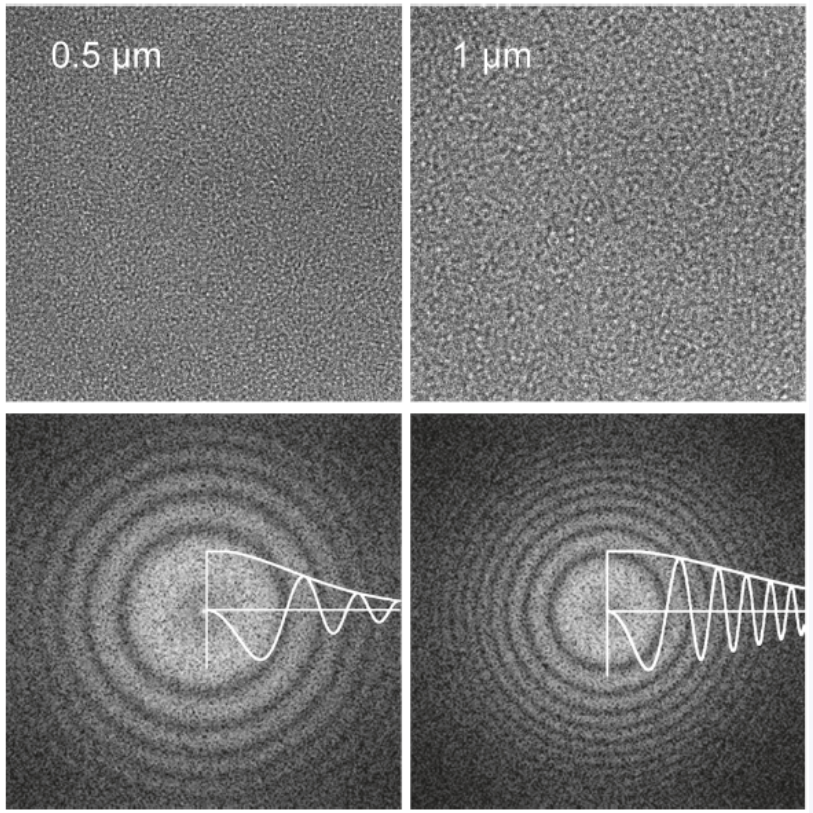
- Maximization of the objective allows a numeric solutions to the model parameters that define the mapping of latent variables to the manifold in the data space.
- We developed Expectation-Maximization algorithm with CTF embedded for numerical solution.

J Wu et al. *arXiv* : 1604.04539 (2016).

Contrast Transfer Function (CTF) and Aberration

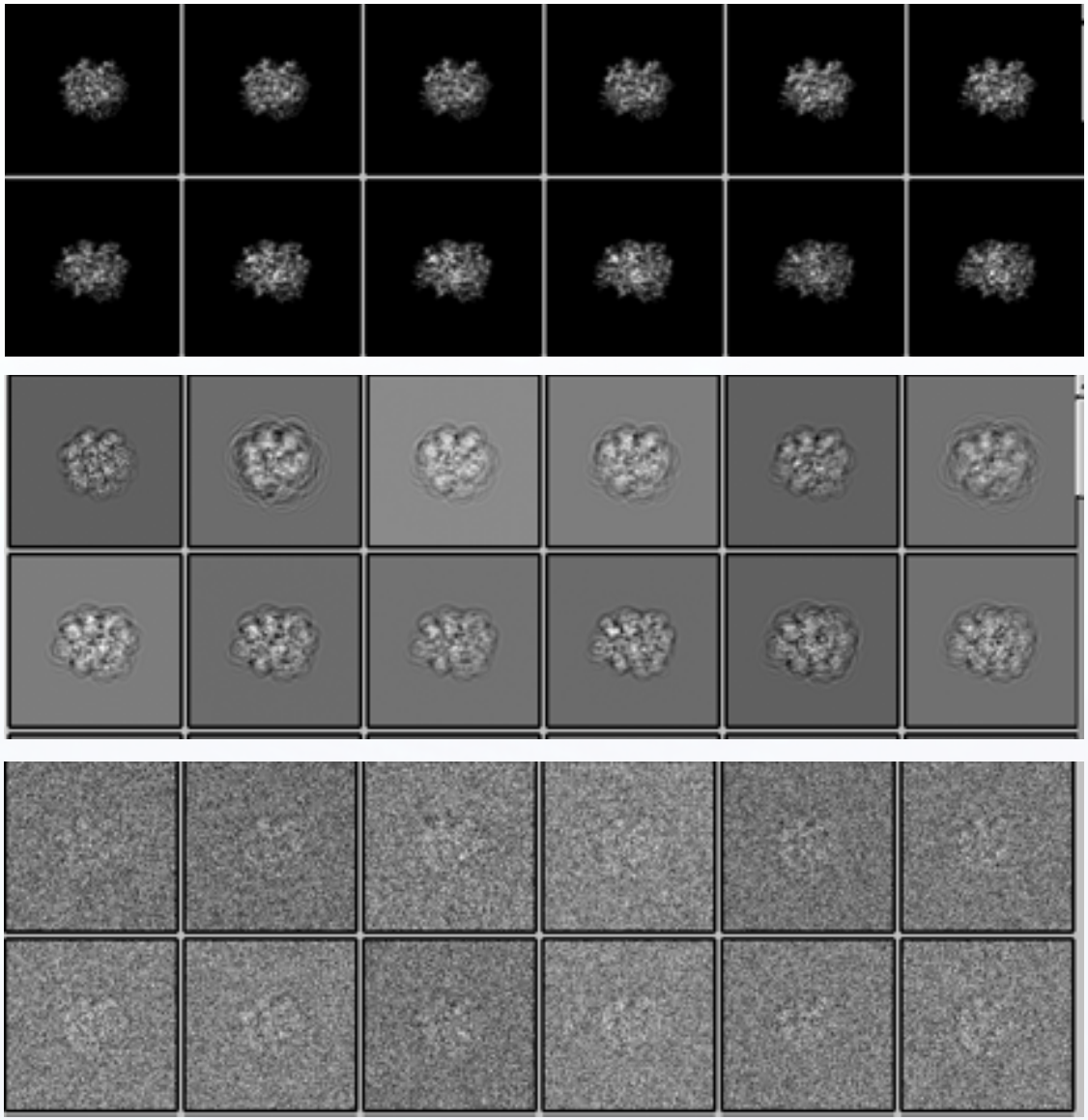
- CTF is the Fourier transform of the point spread function of transmission electron microscope
- CTF characterize how the images are distorted by the aberration of objective lens





Combined Effect of CTF and Noise

- Noiseless and no CTF
- Noiseless with CTF
- Heavy noise with CTF



Expectation-maximization algorithm

- In the E-step, we evaluate the posterior probability

$$R_{ki}(\Theta^{[n]})=p(\boldsymbol{s}_k\left|\boldsymbol{Y}_i,\mathbf{W}^{[n]},\beta^{[n]}\right.)=\frac{p(\boldsymbol{Y}_i\left|\boldsymbol{s}_k,\mathbf{W}^{[n]},\beta^{[n]}\right.)p(\boldsymbol{s}_k)}{\sum_{k'=1}^{K}p(\boldsymbol{Y}_i\left|\boldsymbol{s}_{k'},\mathbf{W}^{[n]},\beta^{[n]}\right.)p(\boldsymbol{s}_{k'})}$$

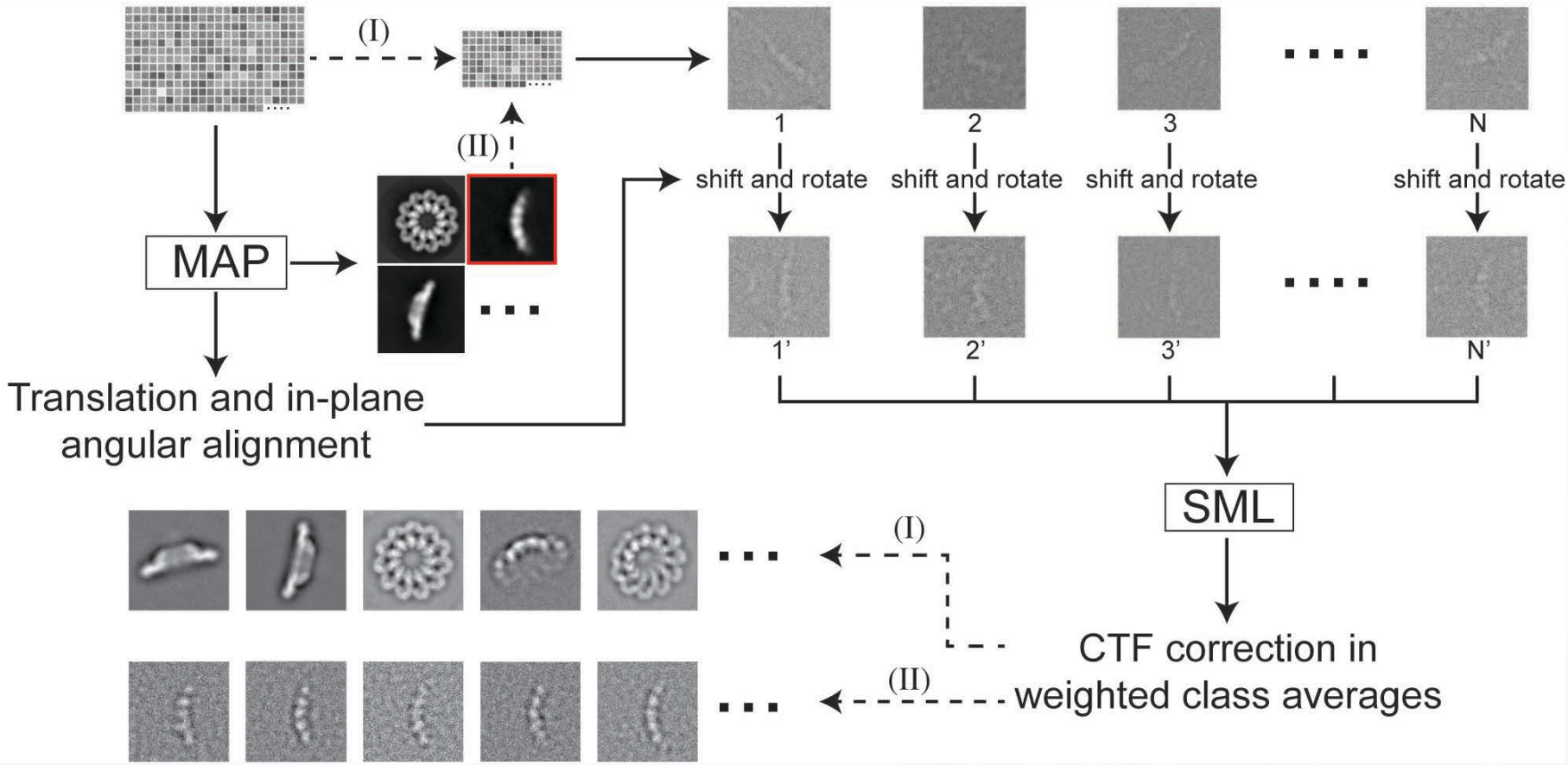
- In the M-step, we maximize the MAP estimator

$$Q(\Theta,\Theta^{[n]})=\sum_{i=1}^{N}\sum_{k=1}^{K}R_{ki}(\Theta^{[n]})\ln p(\boldsymbol{Y}_i\left|\Theta\right.)+\ln p(\Theta)$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K}R_{ki}(\Theta^{[n]})\beta_{ij}^{[n]}\varphi_m(\boldsymbol{s}_k)\mathrm{CTF}_{ij}(Y_{ij}-\mathrm{CTF}_{ij}\sum_{m'=1}^{M}\varphi_{m'}(\boldsymbol{s}_k)W_{m'j}^{[n+1]})-\alpha W_{mj}^{[n+1]}=0$$

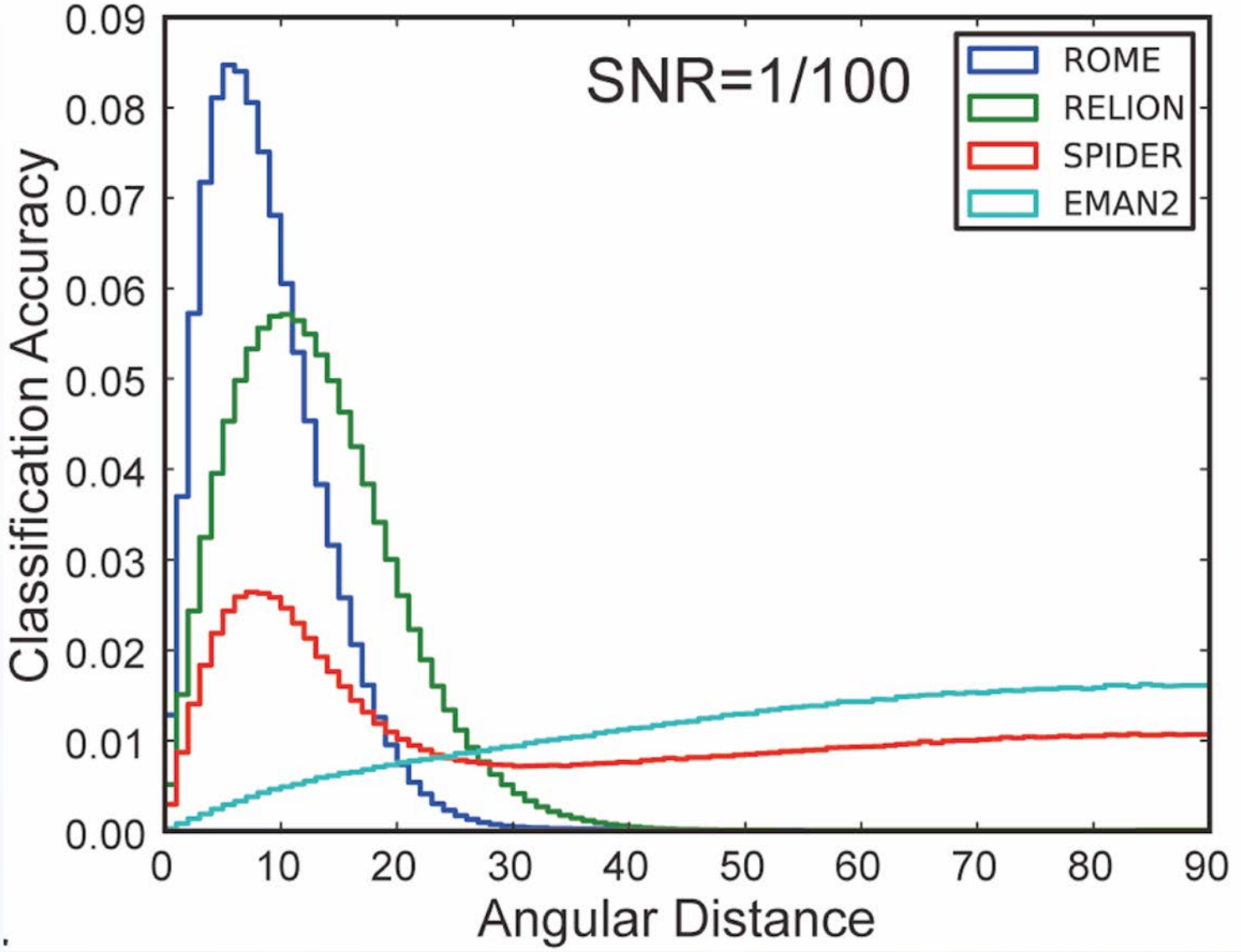
$$\frac{1}{\beta_{ij}^{[n+1]}}=\sum_{k=1}^{K}R_{ki}(\Theta^{[n]})\left(Y_{ij}-\mathrm{CTF}_{ij}\sum_{m=1}^{M}\varphi_m(\boldsymbol{s}_k)W_{mj}^{[n+1]}\right)^2$$

Deep classification by SML



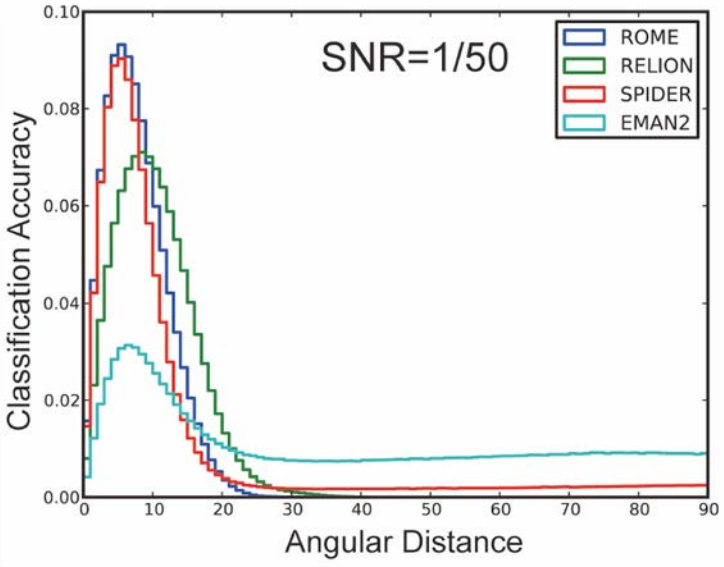
- There is a probability calculated for each image assigned to a given class
- The class average is CTF corrected and is a probability-weighted average.

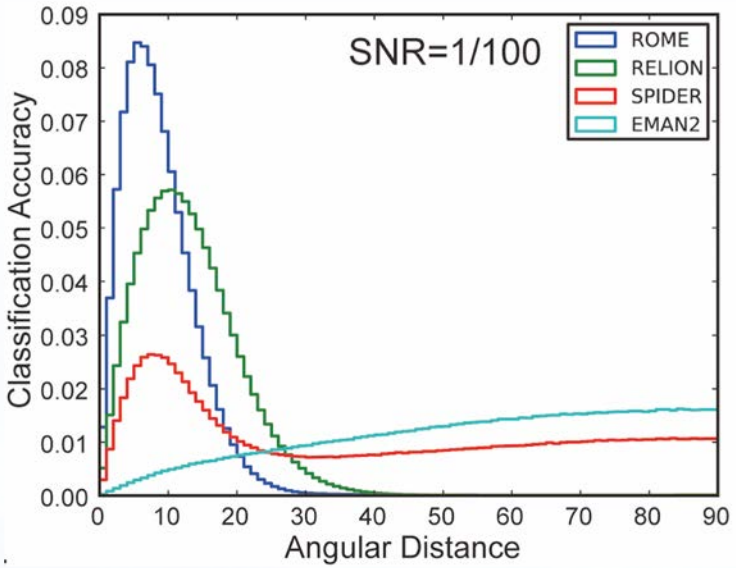
SML vs. MAP: ~40% higher angular accuracy

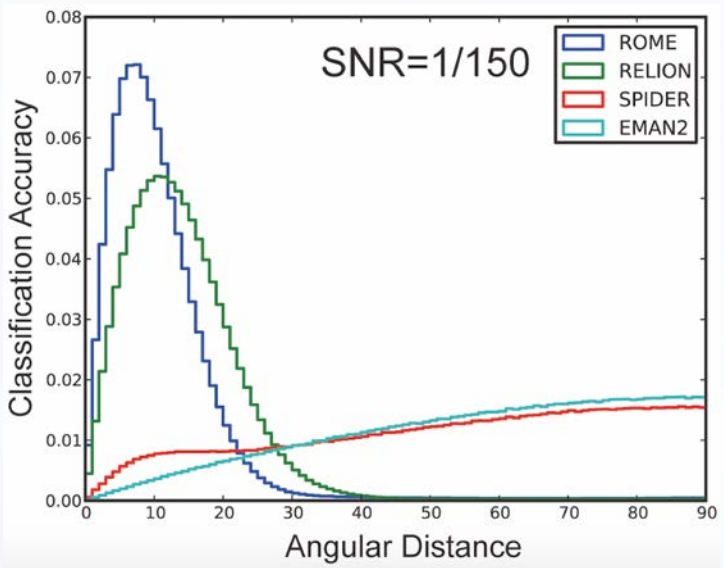


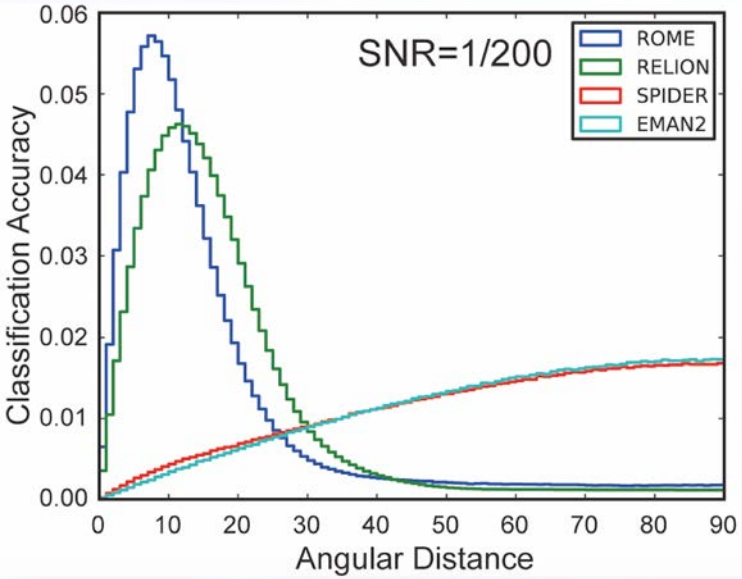
Testing platform configuration:
Intel Xeon processor E5-2670 with dual sockets, 16 cores/socket HT disabled @2.6 GHz 115W 64 GB RAM per compute node, 14 compute nodes, Red Hat Enterprise Linux Server release 6.5

Effect of SNR on angular error

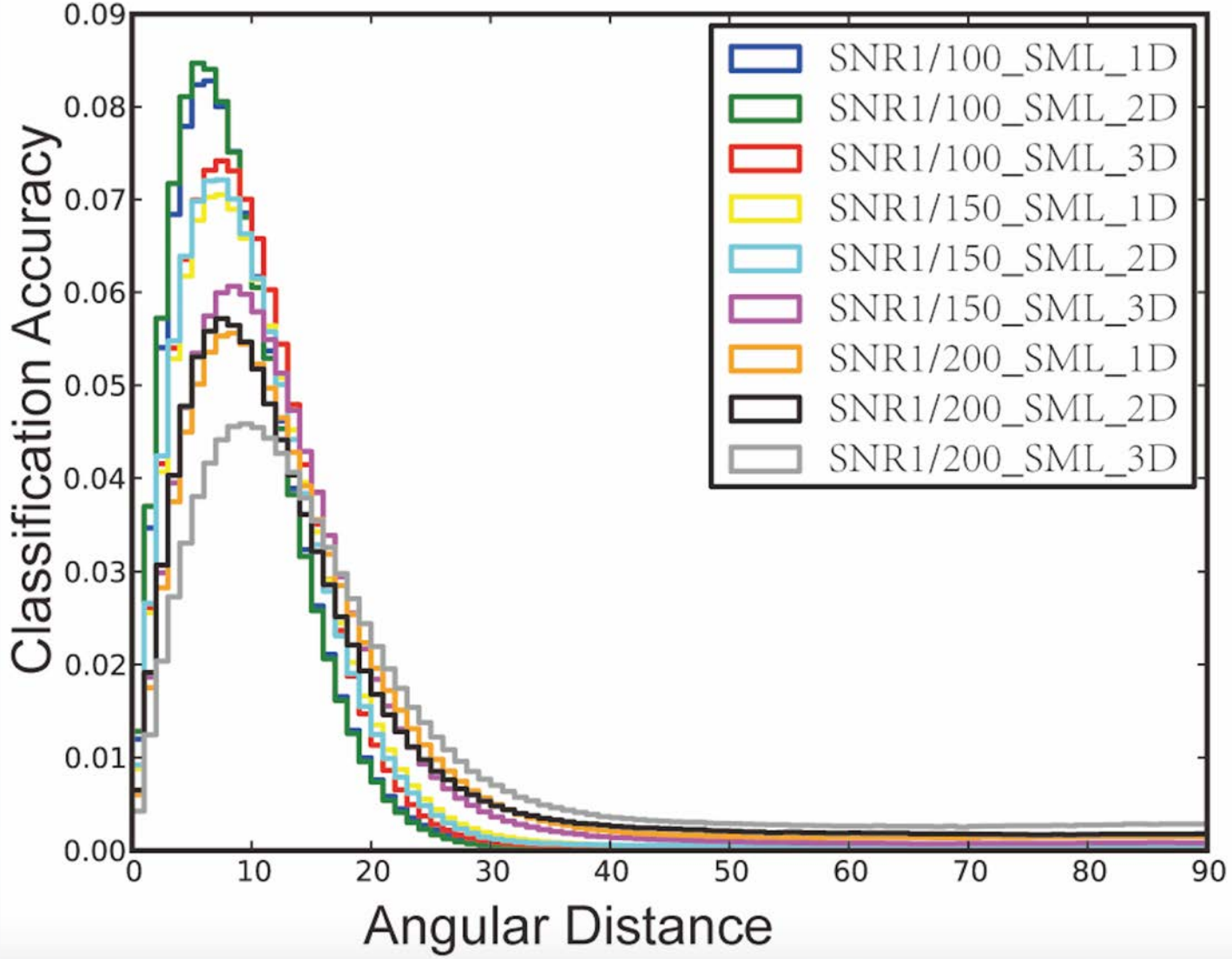




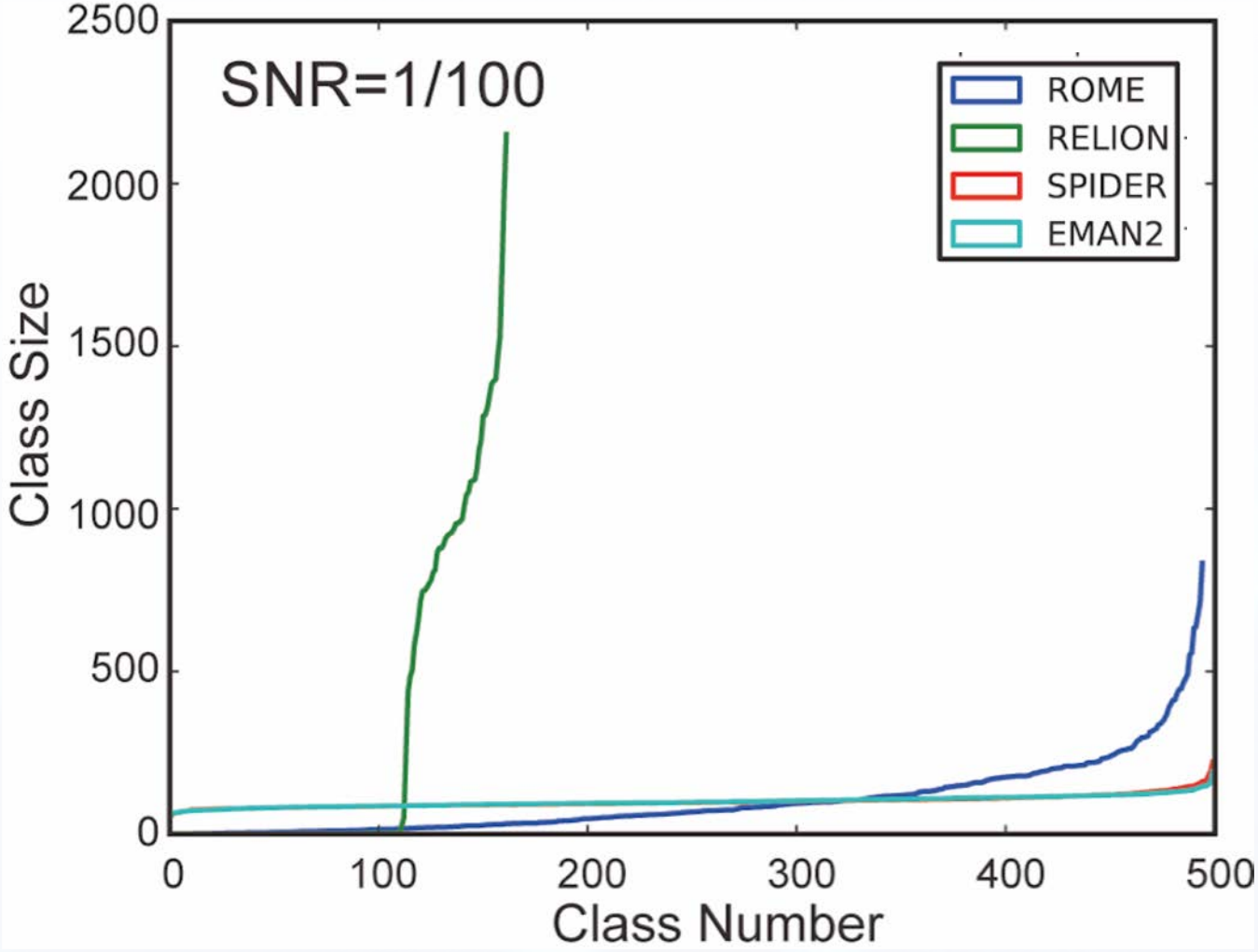




Effect of dimensionality of latent space

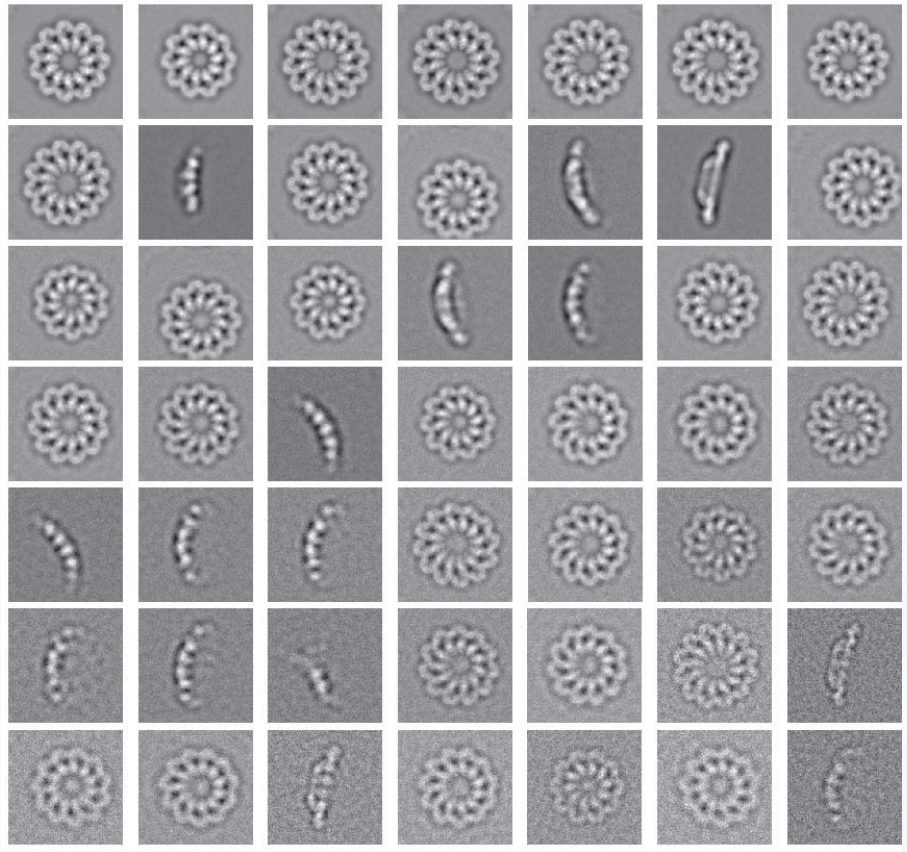


Improved balance of class sizes

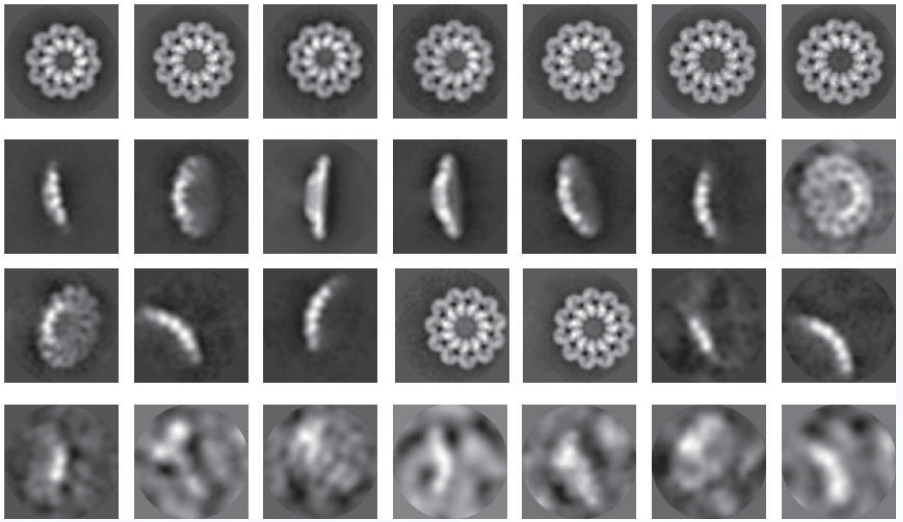


SML vs. MAP classification

SML

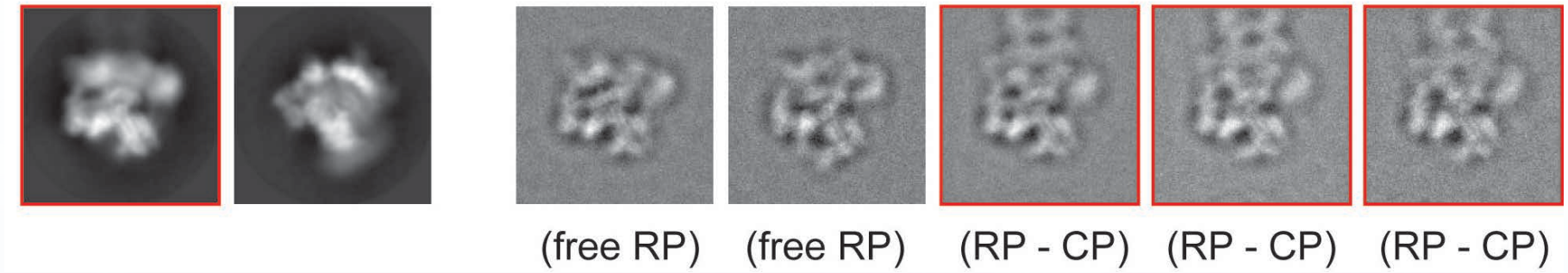


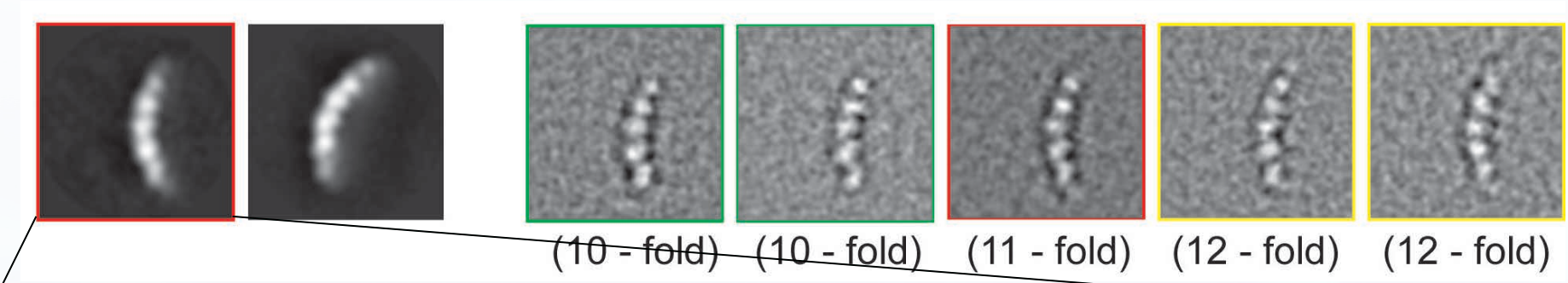
MAP (maximum a posteriori)

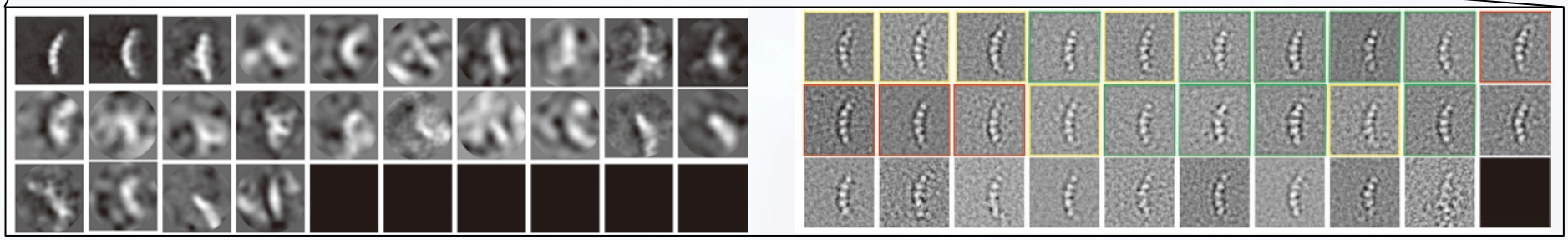


17,103 inflammasome particles are classified into 300 reference-free classes. Only classes whose particle numbers were larger than 9 are exhibited.

Deeper classification identifies hidden heterogeneity



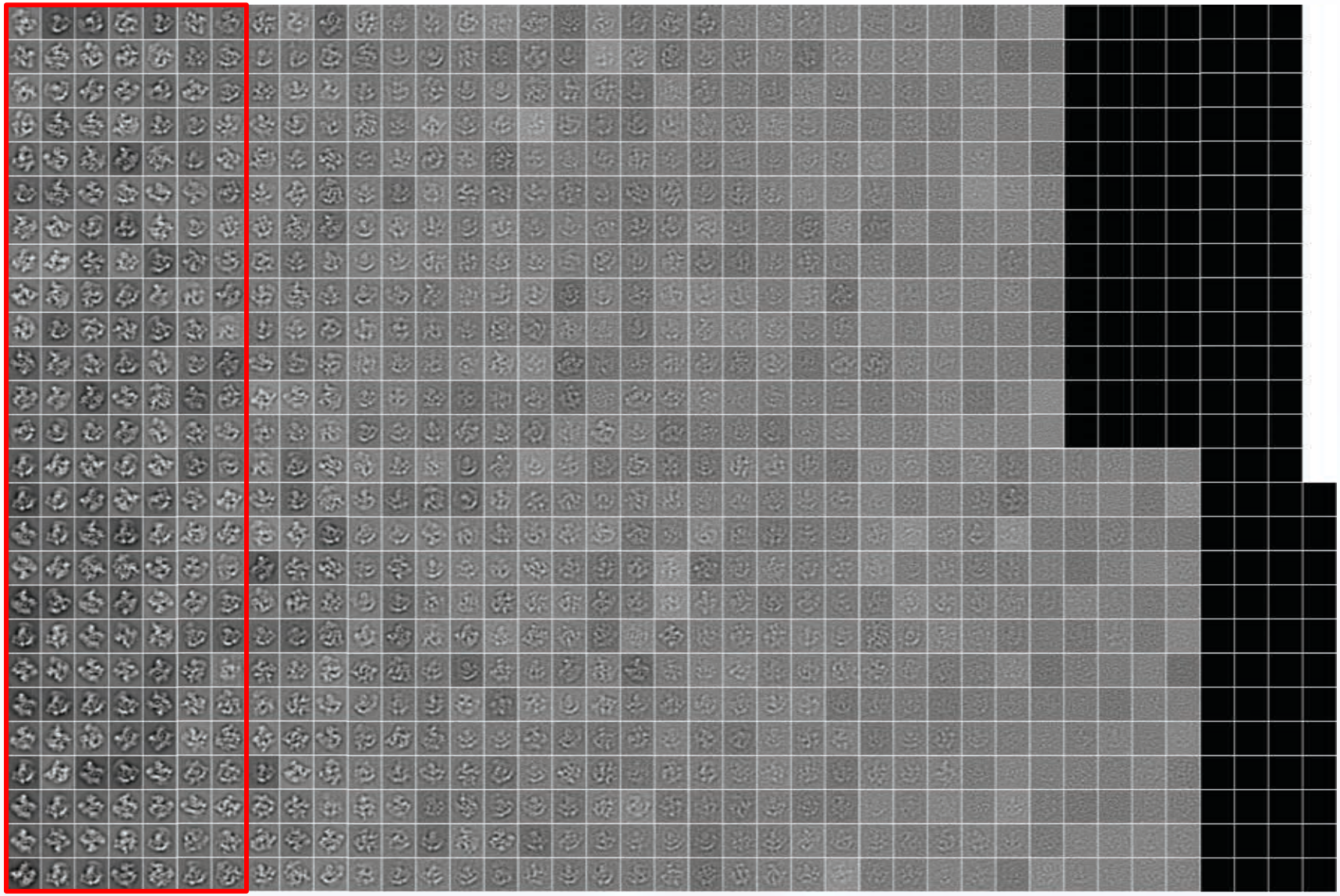




Secondary MAP classification

Secondary SML classification

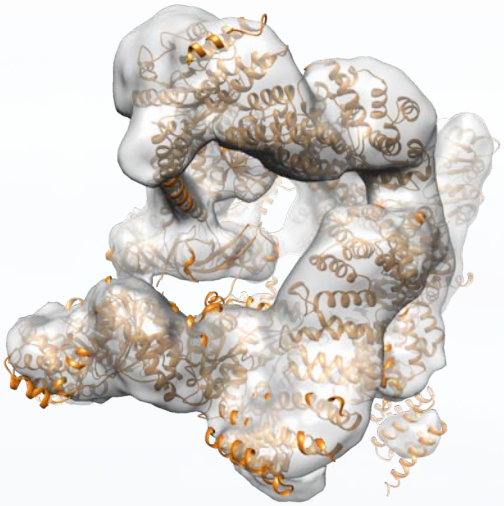
1000 reference-free classes computed in a few hours

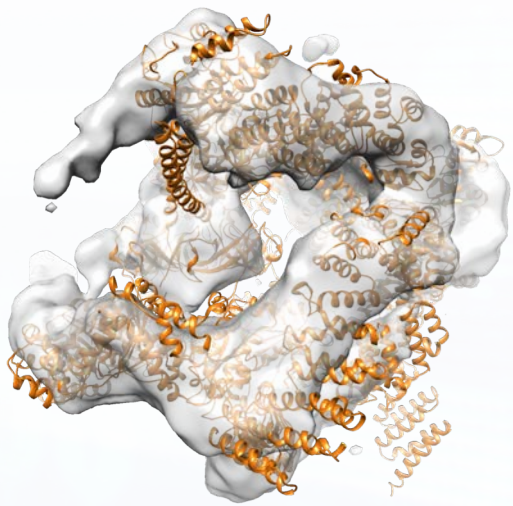


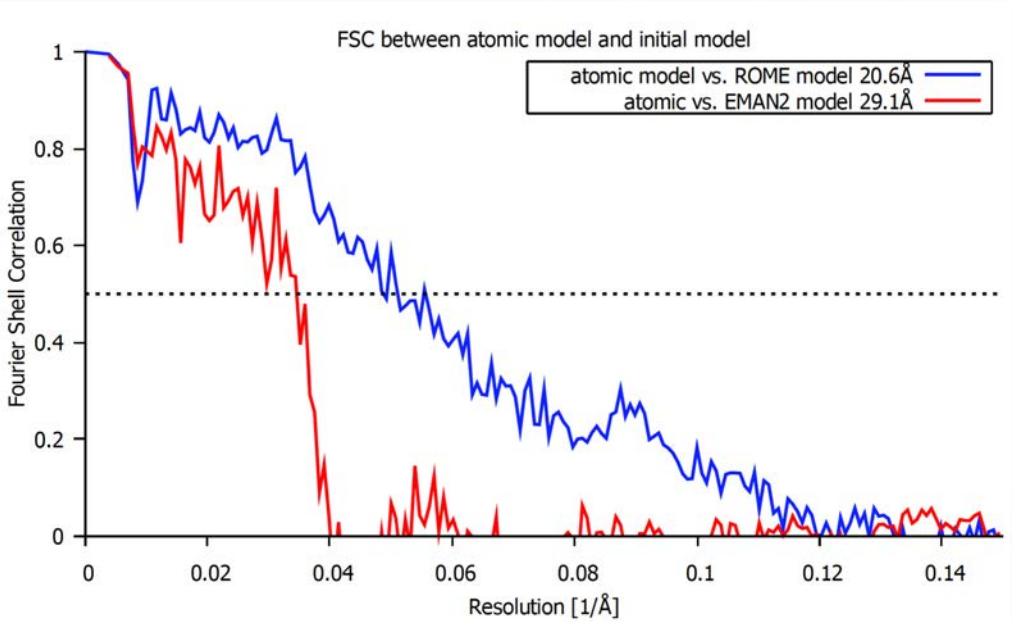
Deep classification improves initial reconstruction

ROME

EMAN2

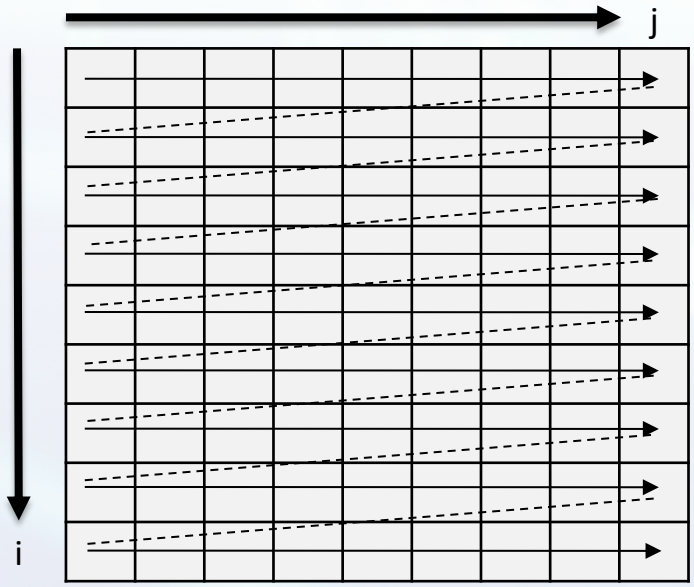




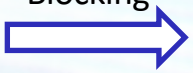


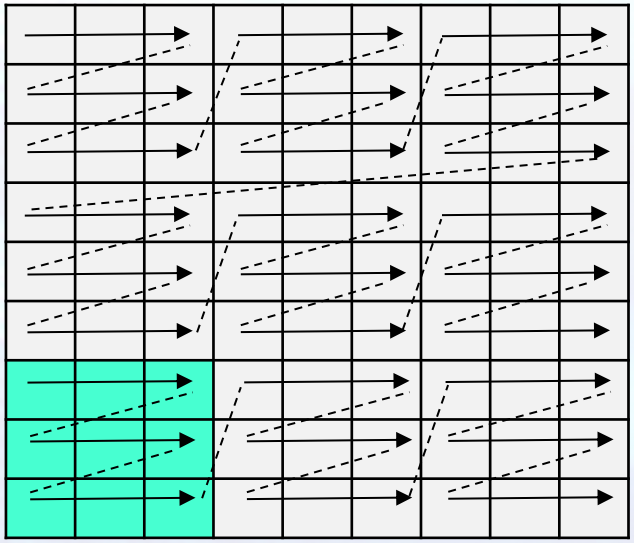
- ROME-based initial model: 234 best class averages selected from 1000 reference-free classes of a cryo-EM dataset (117,471 images) within 3 hours on a cluster of 512 CPU cores (32 nodes)
- EMAN2-based initial model: all 128 ref-free class averages calculated by EMAN2 for >8 hours using the same dataset.
- The resolution of the ROME-based initial model is 10-Å (50%) higher, show in the FSC plots on the left.

Memory/cache optimization-code blocking



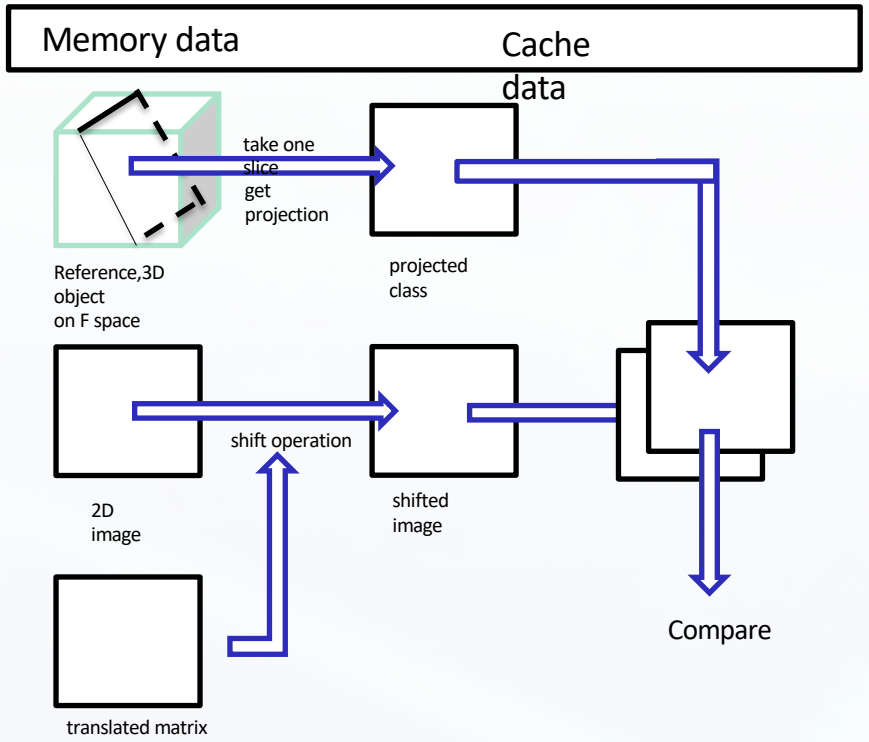
Access pattern

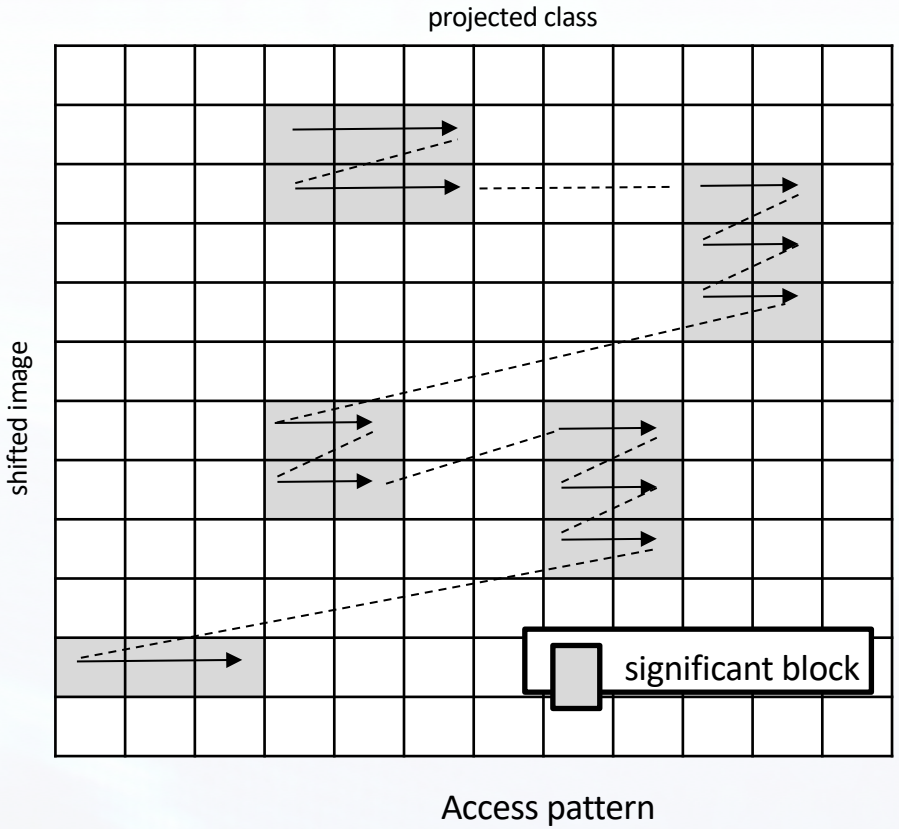




Access pattern after blocking

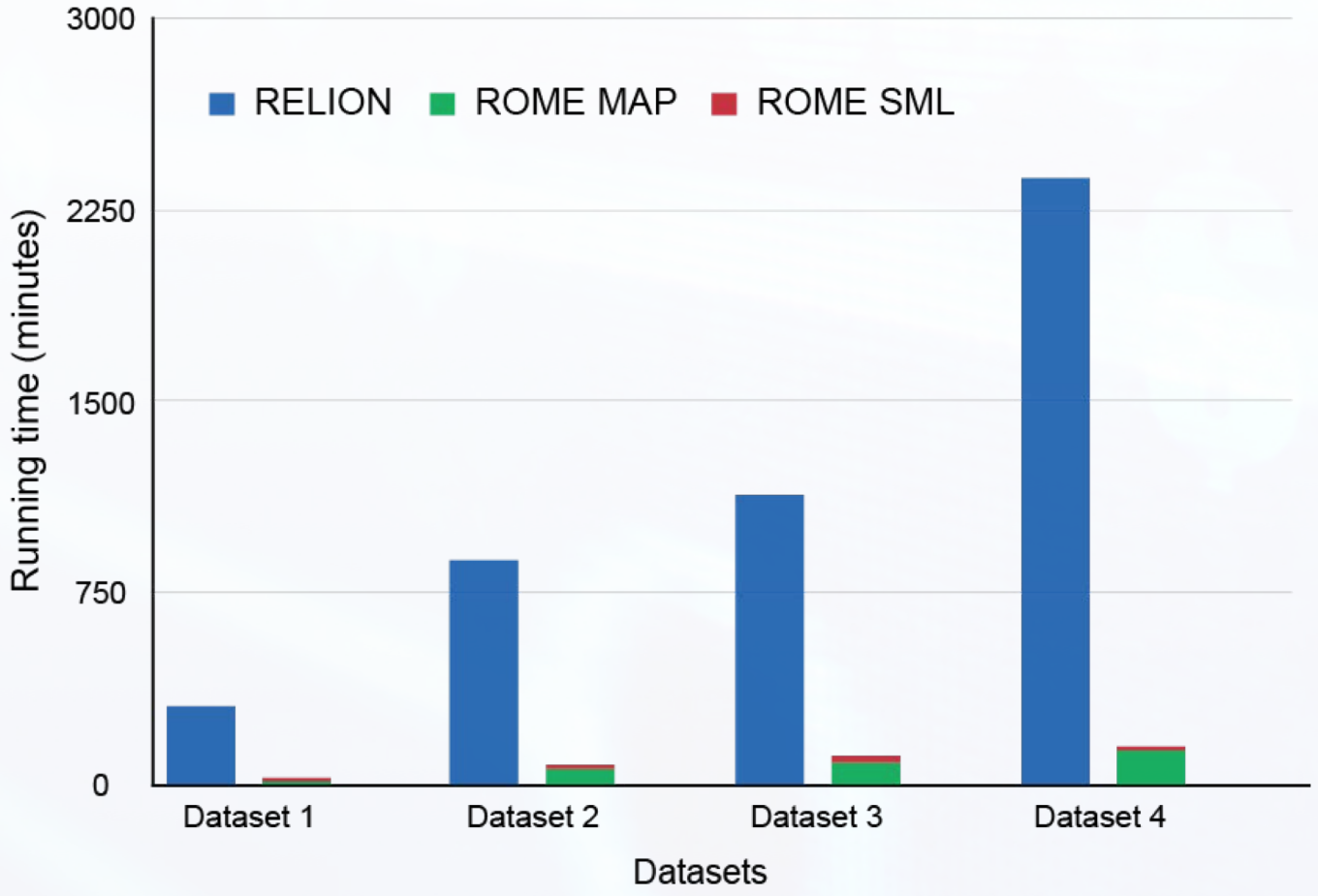
Memory/cache optimization-code blocking for adaptive search





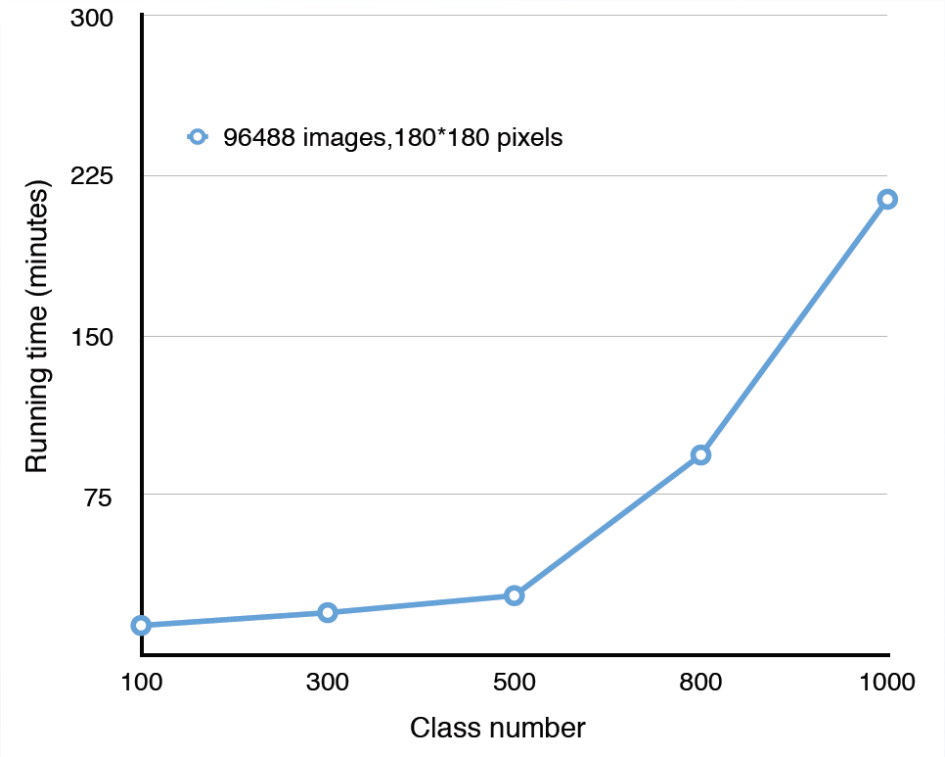
ROME vs. RELION: ~20-fold faster on the same hardware

- When GTM and ML algorithms in ROME are used for unsupervised deep classification, it outperforms RELION by 10-20 times

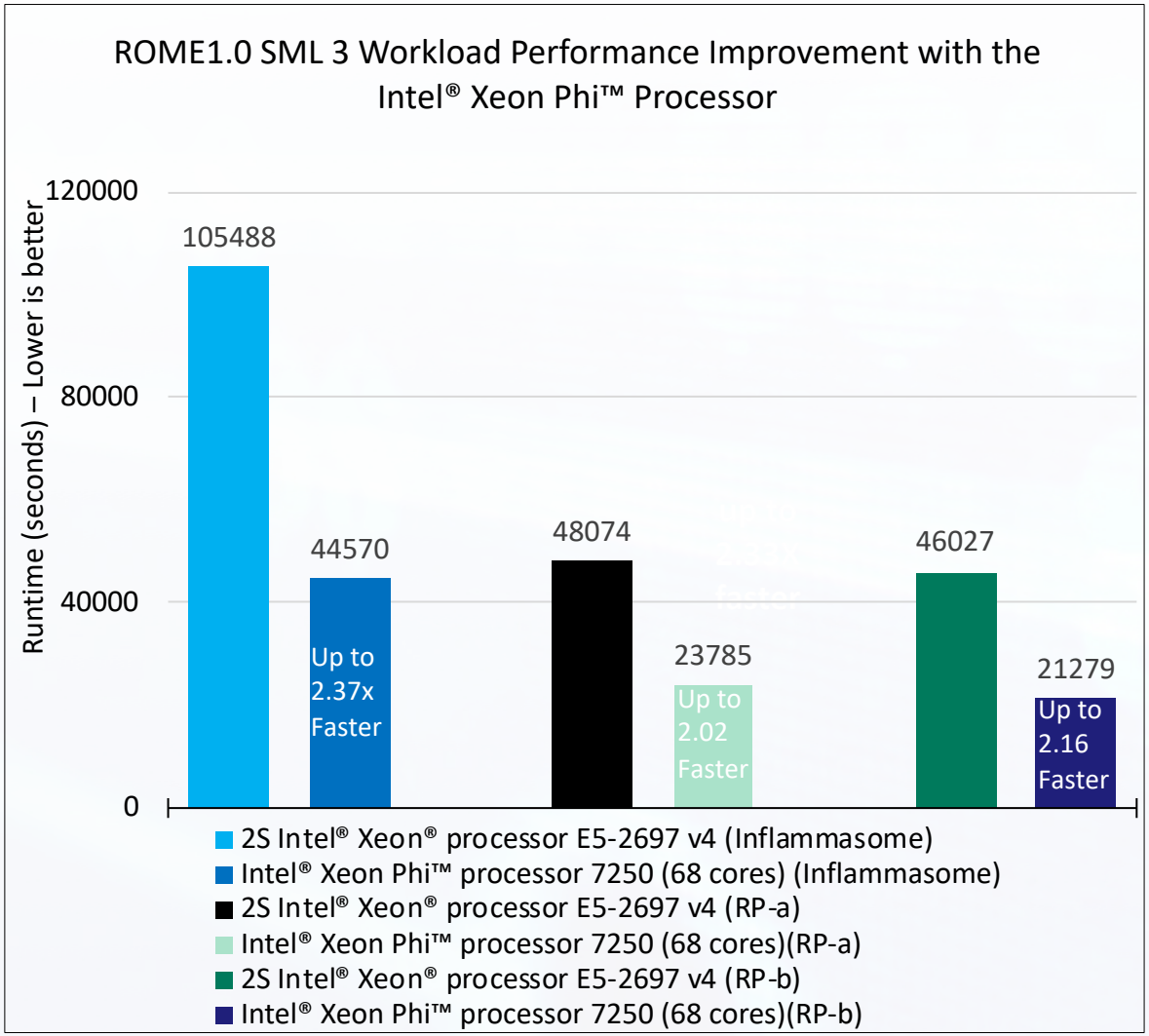


Performance of deep classification by ROME

- ROME can generate 1000 reference-free classes of real cryo-EM data within a few hours on a cluster of 500 CPU cores (32 nodes)
- All other programs tested have crashed on the same dataset in our test clusters when doing this task



Speedup on Intel Xeon Phi Knights Landing processors



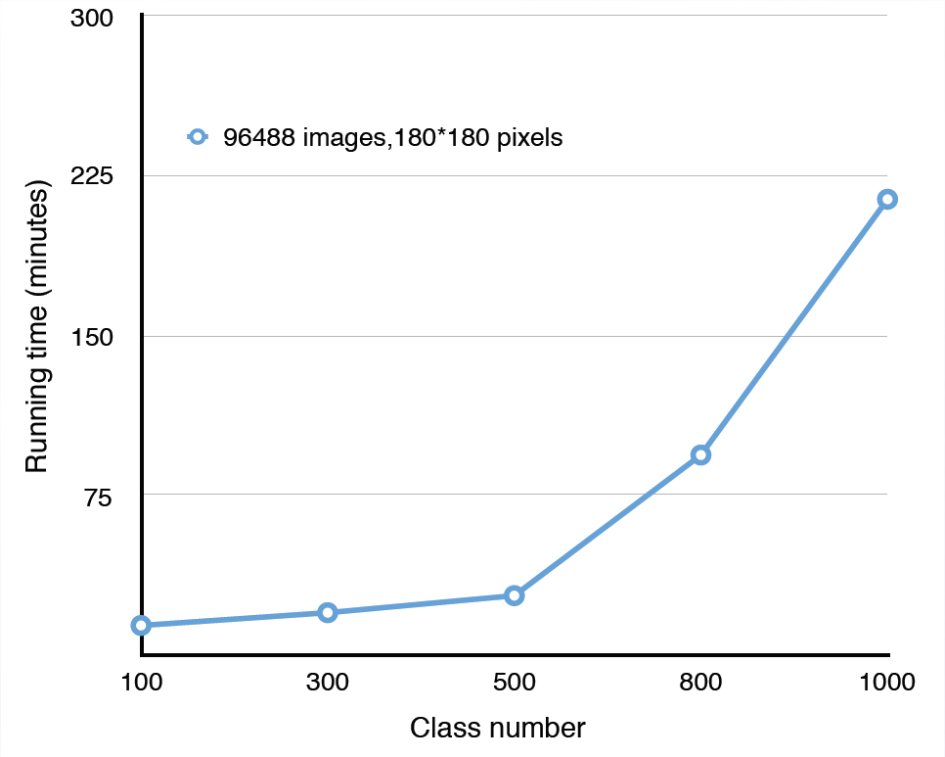
SOURCE: https://software.intel.com/en-us/articles/recipe-rome-1-0-sml-for-the-intel-xeon-phi-processor-7250

Features in our ROME v1.0 system

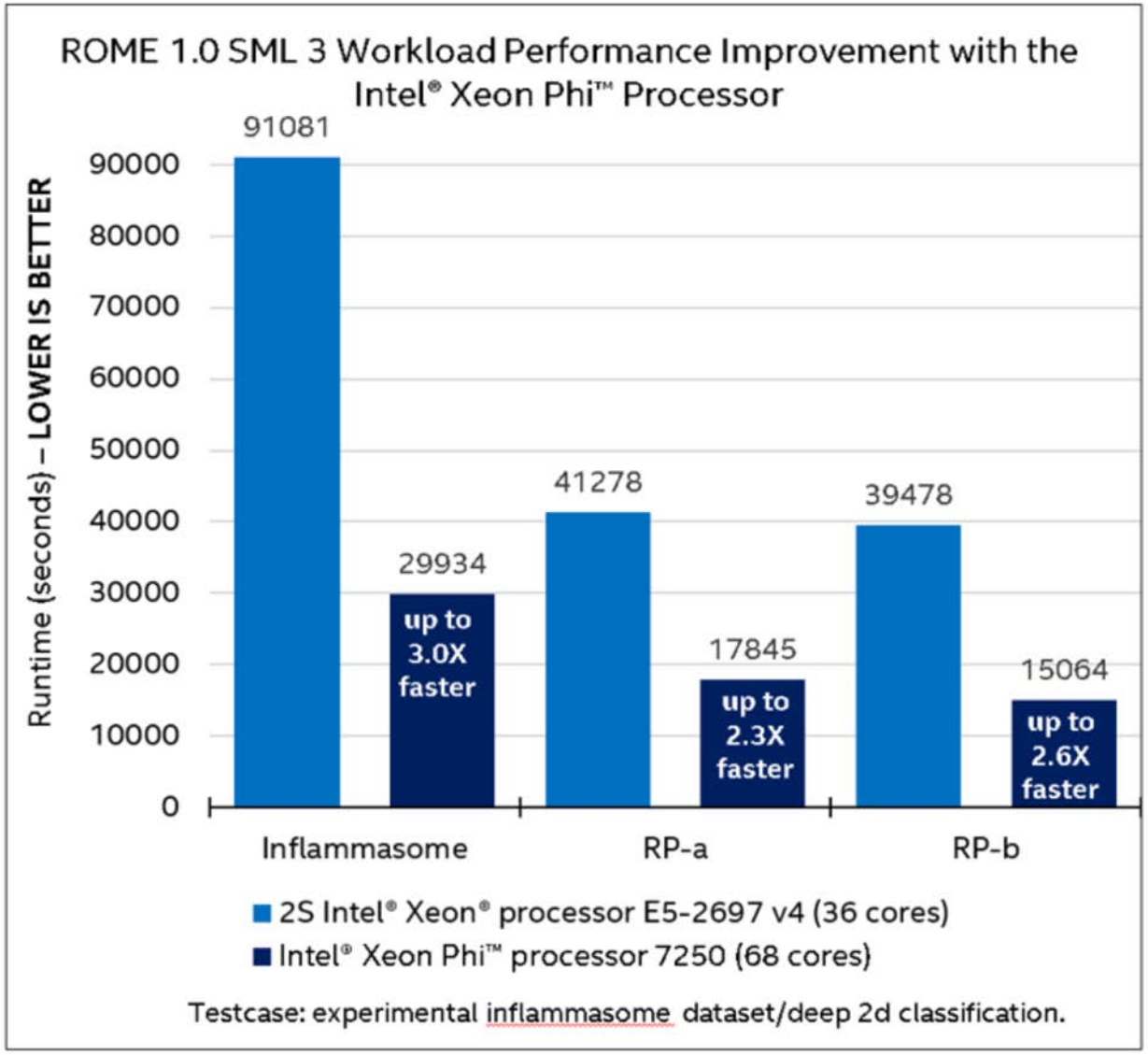
- ROME (Refinement and Optimization with Machine-lEarning) implements SML in HPC.
- Fully compatible with the I/O file format used in other software in structural biology and cryo-EM imaging, including RELION and SPIDER.
- Fully modernized code, designed and parallelized for Intel Xeon CPUs
- Designed to use both OpenMP and MPI.
- It can perform image classification at a scale of thousands of classes in a single run with improved accuracy, which is about 1-2 orders of magnitude greater than existing software in this area.

Performance of deep classification by ROME

- ROME can generate 1000 reference-free classes of real cryo-EM data within a few hours on a cluster of 500 CPU cores (32 nodes)
- All other programs tested have crashed on the same dataset in our test clusters when doing this task



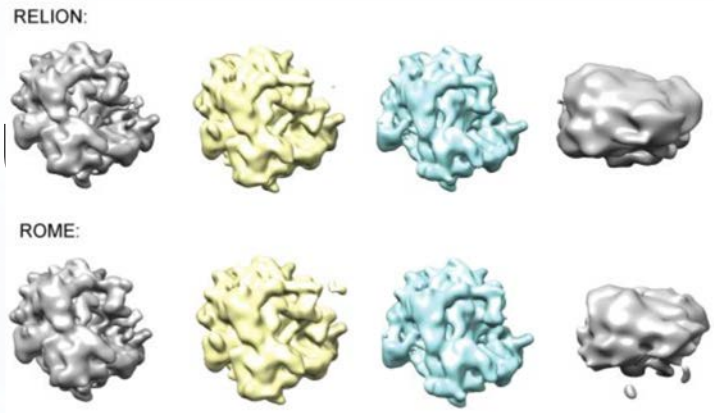
Speedup on Intel Xeon Phi Knights Landing (KNL) processors



SOURCE: https://software.intel.com/en-us/articles/recipe-rome-1-0-sml-for-the-intel-xeon-phi-processor-7250

MAP3D benchmark dataset 1 : ribosomes wi/wo EFG

- 5 nodes Intel(R) Xeon(R) CP
  E5-2697 v4 @ 2.30GHz
- RELION 1.4

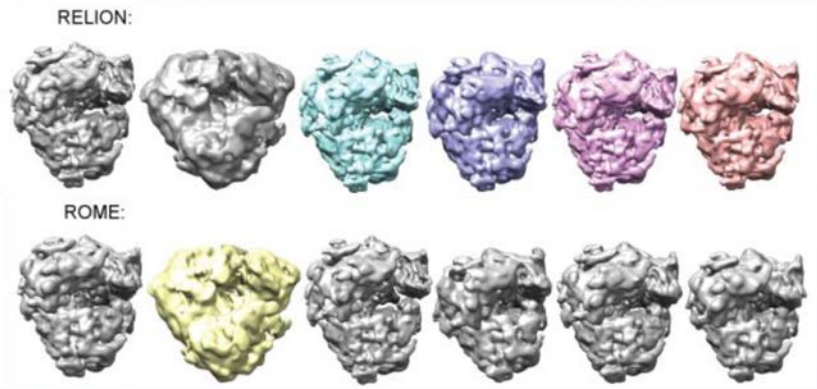


MPI_Rank	Threads	Running_time(minutes)
10	18	222.95
10	36	166.03
20	9	187.03
20	18	139.37
45	4	160.9
45	8	125.68
90	2	151.08
90	4	120.0

- ROME 1.1.0

MPI_Rank	Threads	Running_time(minutes)
10	36	26.1
5	72	21.98

MAP3D benchmark dataset 2 : Plasmodium ribosome



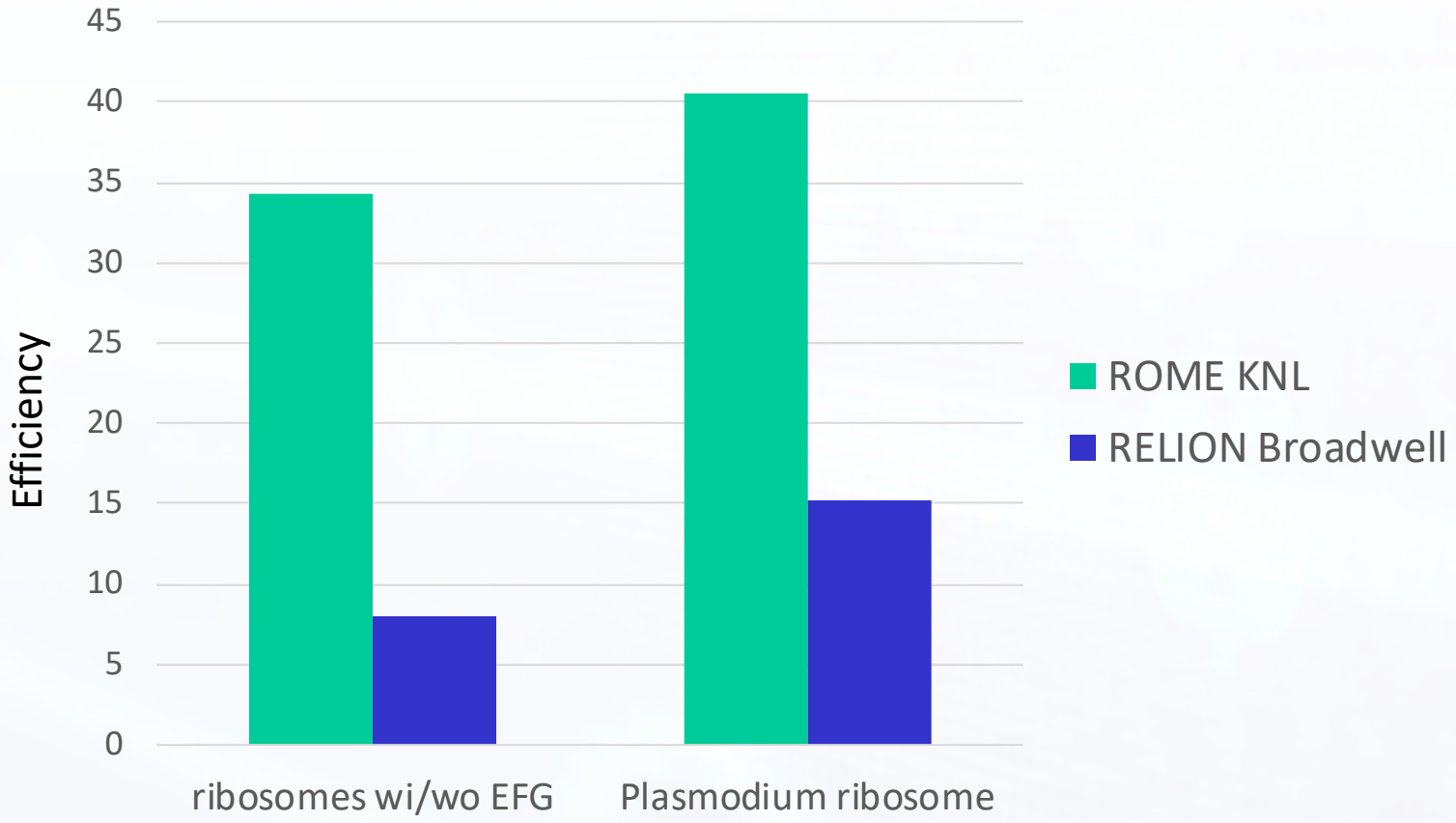
❑ RELION 1.4

MPI_Rank	Threads	Running_time(minutes)
10	18	857.6
10	36	644.02
20	9	769.15
20	18	624.23
45	4	835.37
45	8	658.73
90	2	crash
90	4	crash

❑ ROME 1.1.0

MPI_Rank	Threads	Running_time(minutes)
10	36	180.55
5	72	166.2

MAP3D for KNL benchmark (3D classification/refinement)



platform	CPU	Recommended Customer Price
KNL	Intel® Xeon Phi™ Processor 7230 16GB, 1.30 GHz, 64 core	$1,992.00
Broadwell	Intel® Xeon® Processor E5-2697 v4 45M Cache, 2.30 GHz	$2,702.00

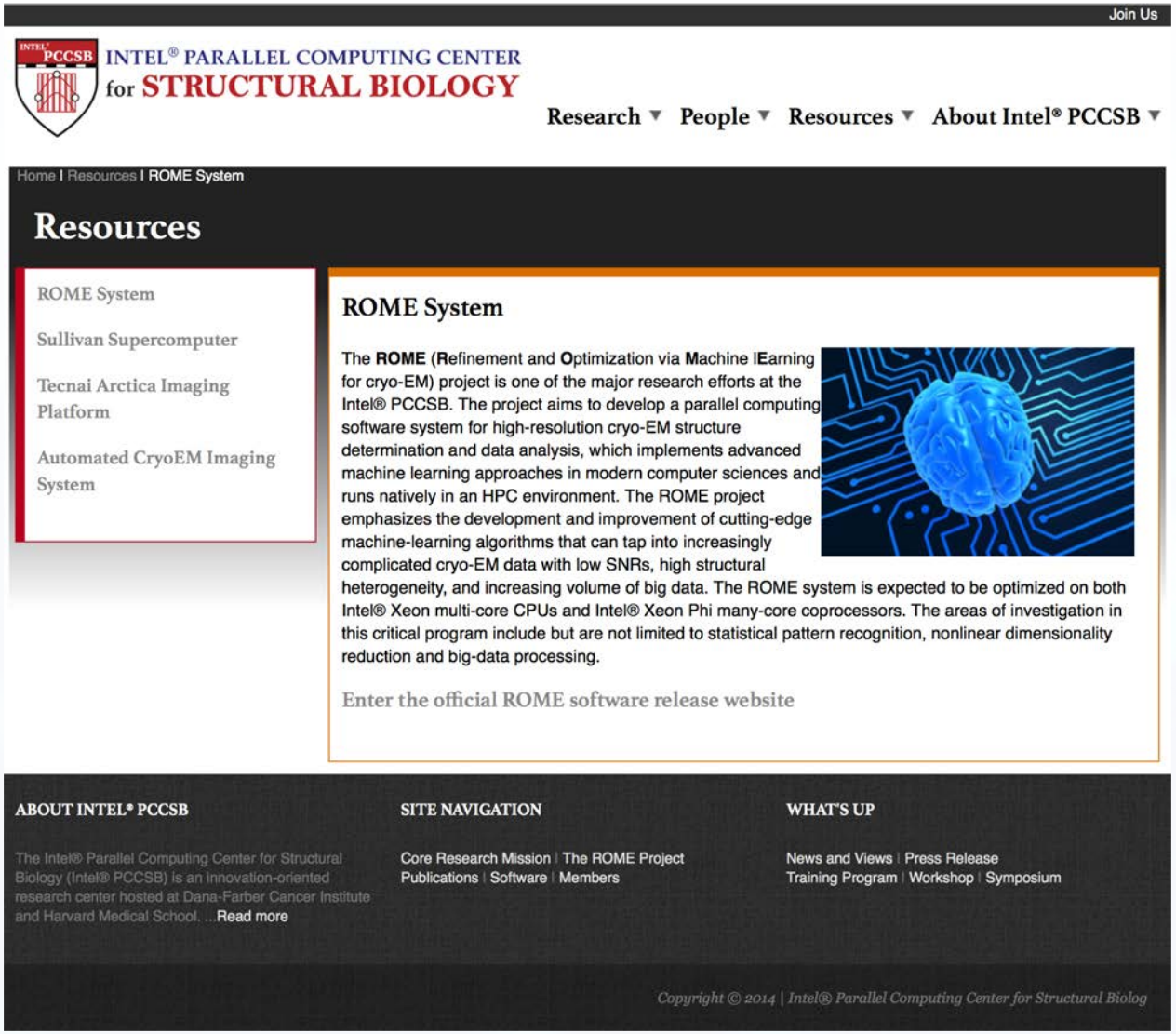
Features in ROME v1.1.0 system

- ROME (Refinement and Optimization with Machine-lEarning) implements SML and MAP3D in HPC.
- Fully compatible with the I/O file format used in other software in structural biology and cryo-EM imaging, including RELION and SPIDER.
- Fully modernized code, designed and parallelized for Intel Xeon CPUs
- Designed to use both OpenMP and MPI.
- It can perform image classification at a scale of thousands of classes in a single run with improved accuracy, which is about 1-2 orders of magnitude greater than existing software in this area.

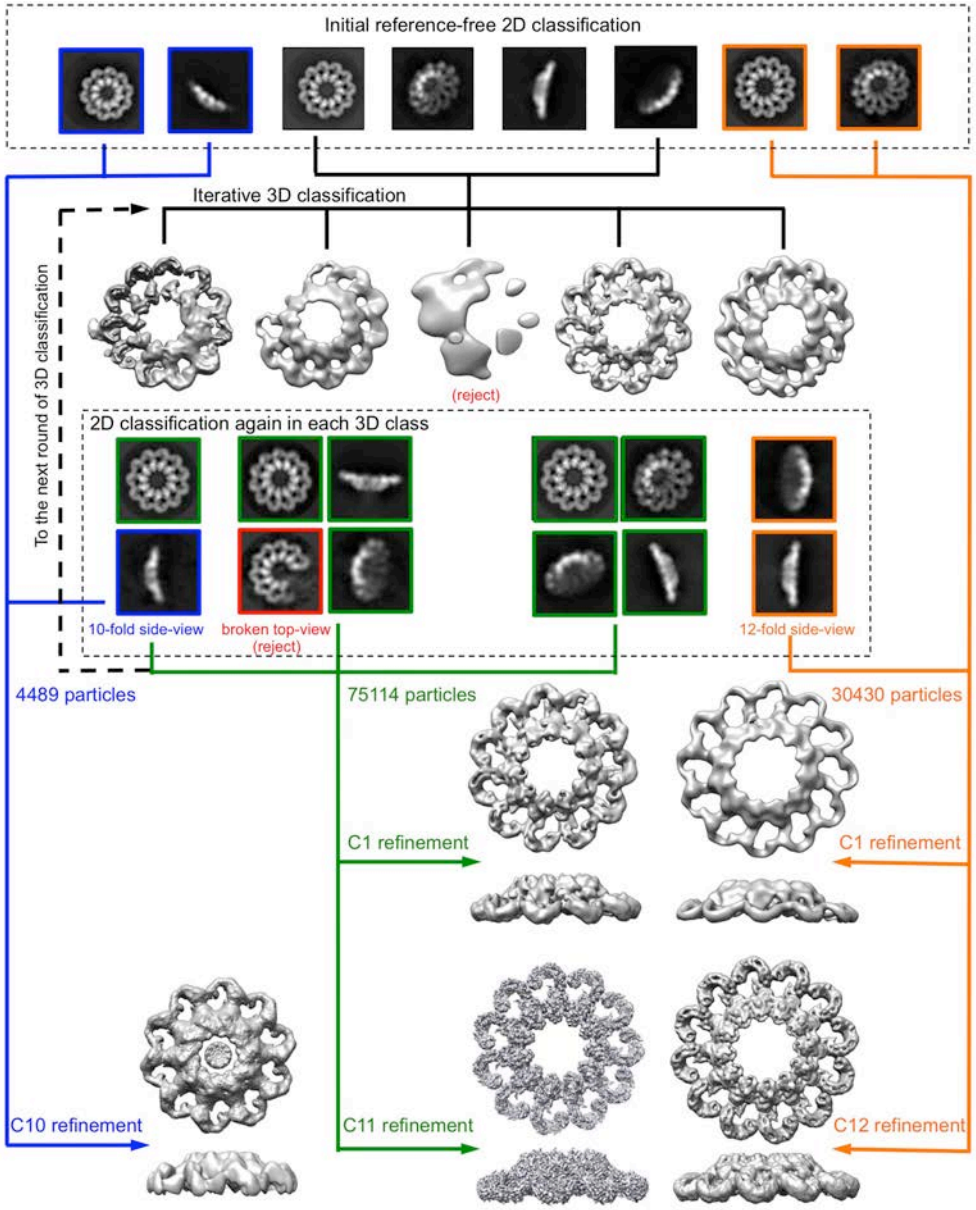
Summary

- SML enables unsupervised deep classification at a greater scale.
- ROME, the first HPC software implementing SML algorithm, is optimized based on the modern standard of parallelization.
- Despite SML classification is mathematically more complicated than MAP classification, it is more suited for parallelization in the latest CPU architecture with a higher throughput of vector processing.

ROME Official Website
http://ipccsb.dfci.harvard.edu/rome



Solution to sample heterogeneity



L. Zhang et al. Science 350: 404-409 (2015).

Acknowledgement

Yuanchen Dong

Shuwen Zhang

Zhaolong Wu

- Early work on substrate-free proteasome was done with collaboration with Prof. Marc Kirschner's group at Harvard.
- Recent data were collected at Electron Microscopy Laboratory and Cryo-EM platform at PKU.
- Most data processing was done at PKU HPC platform.
- Our work was partly funded by NSFC, Young 1000-Talent Plan, Intel Corporation, CLS@PKU and NIH.

Thank you for your attention!