EXPLORING THE CONFORMATIONAL FLEXIBILITY OF MACROMOLECULAR NANOMACHINES

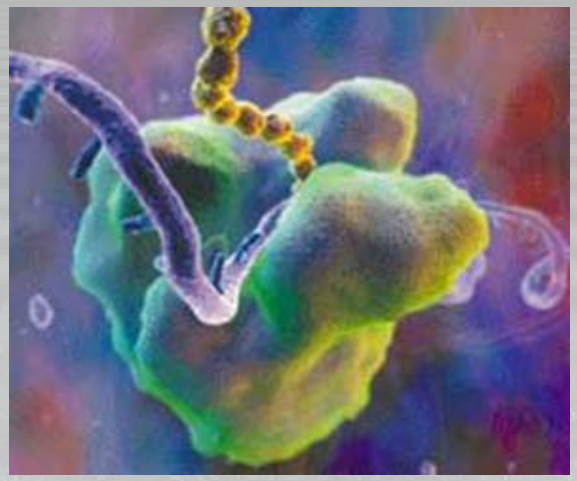
Jose-Maria Carazo

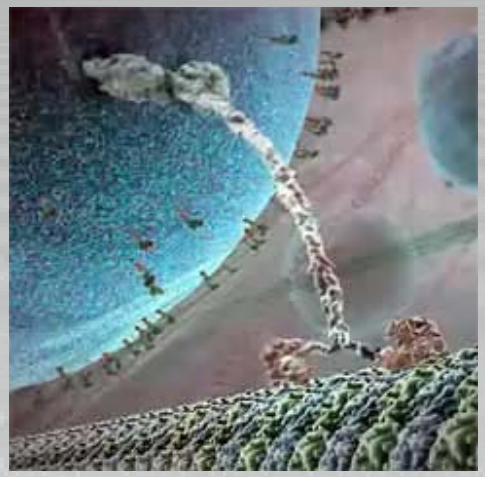
Sjors Scheres, Paul Eggermont & Gabor Herman

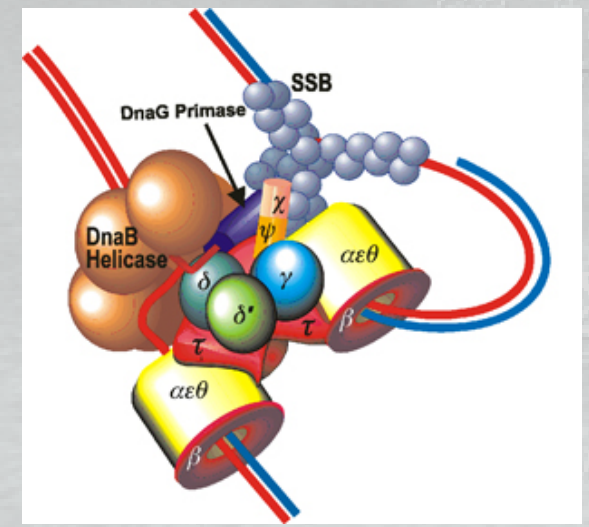
National Center for Biotechnology – CSIC

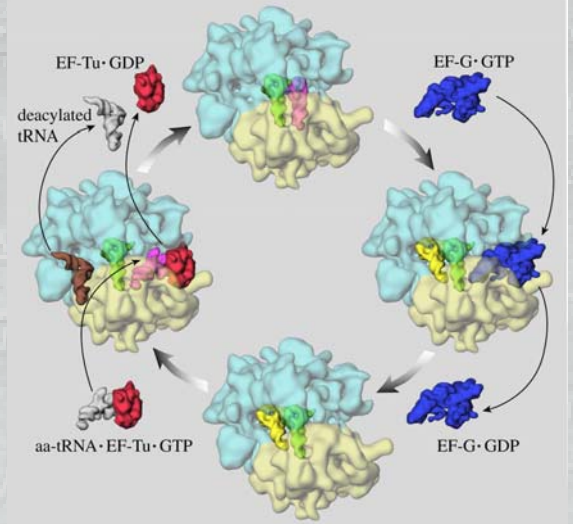
Madrid, Spain

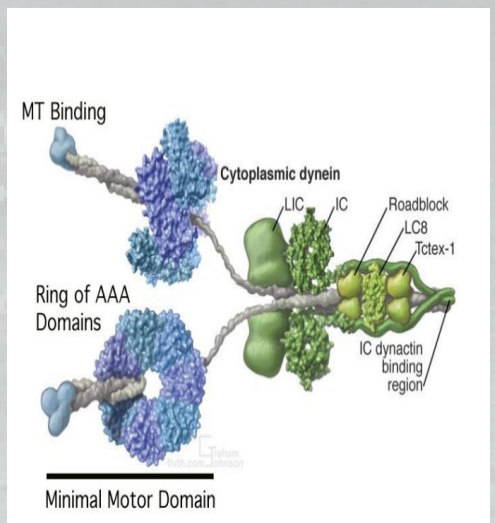
Life based on molecular machines



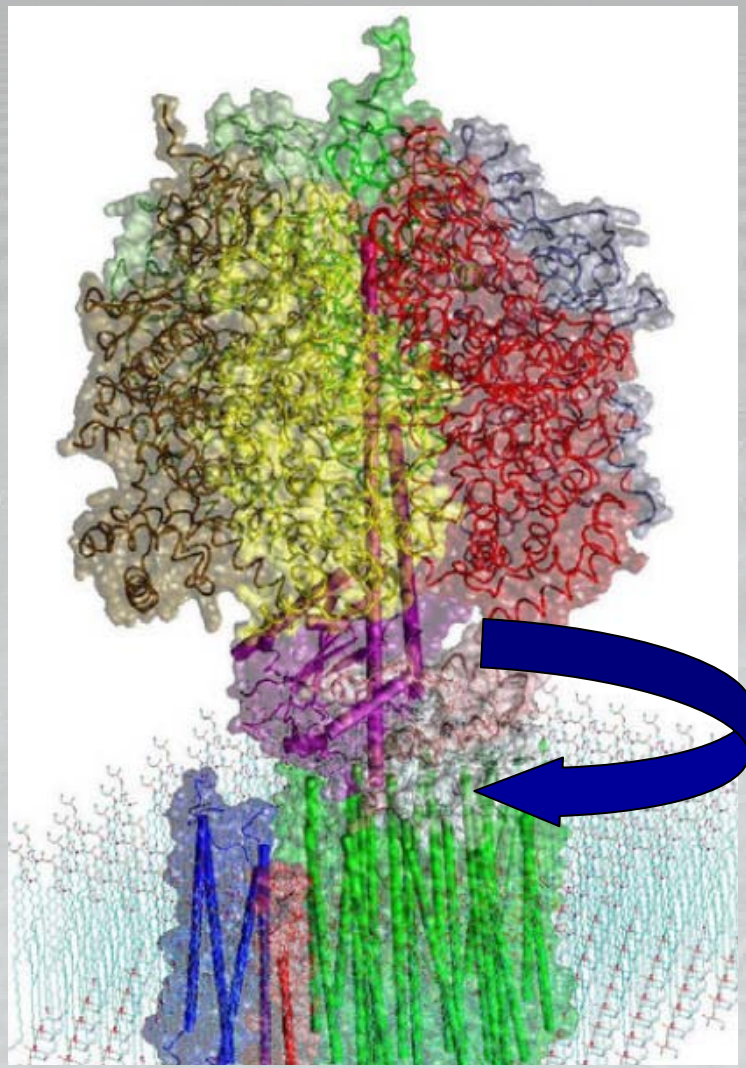






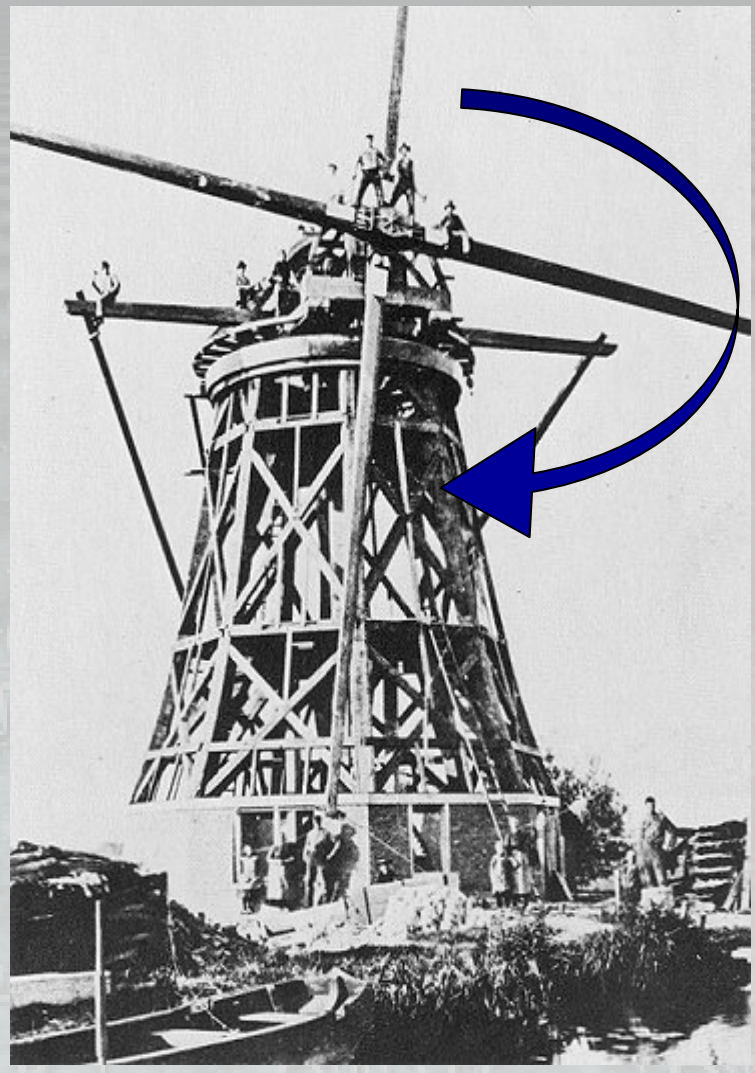


Molecular machines



15 10^{-9} m

F1-ATPase: Abrahams et al., 1994



15 m

Dutch windmill

Studying these machines

- The different states tell much about the way these machines work!

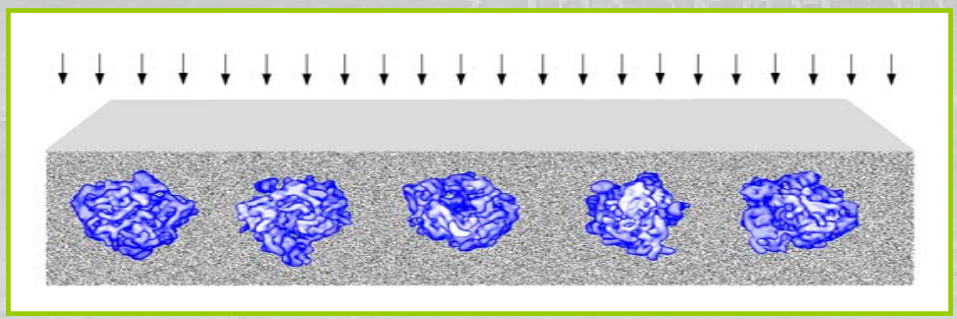
- Different conformations of (chemically identical) molecules are very hard to purify

- Biophysical techniques that study the bulk, “average-out” information about these conformations

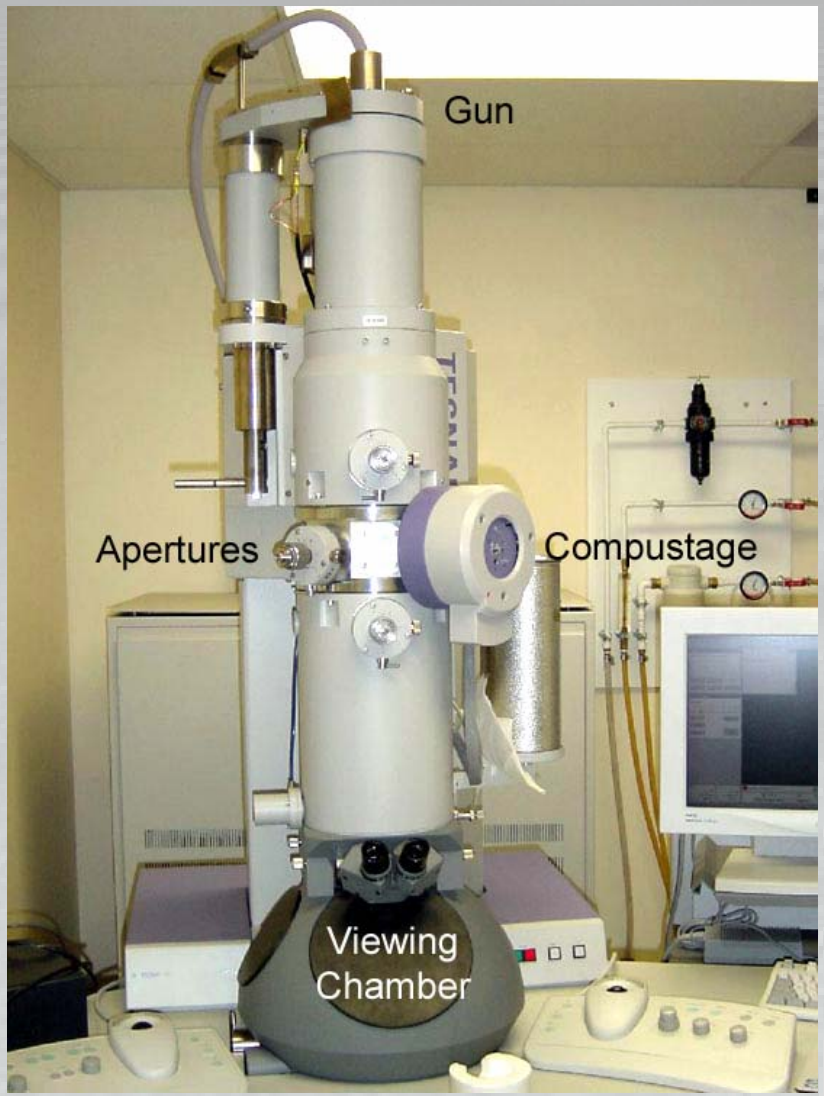
The promise of 3D-EM

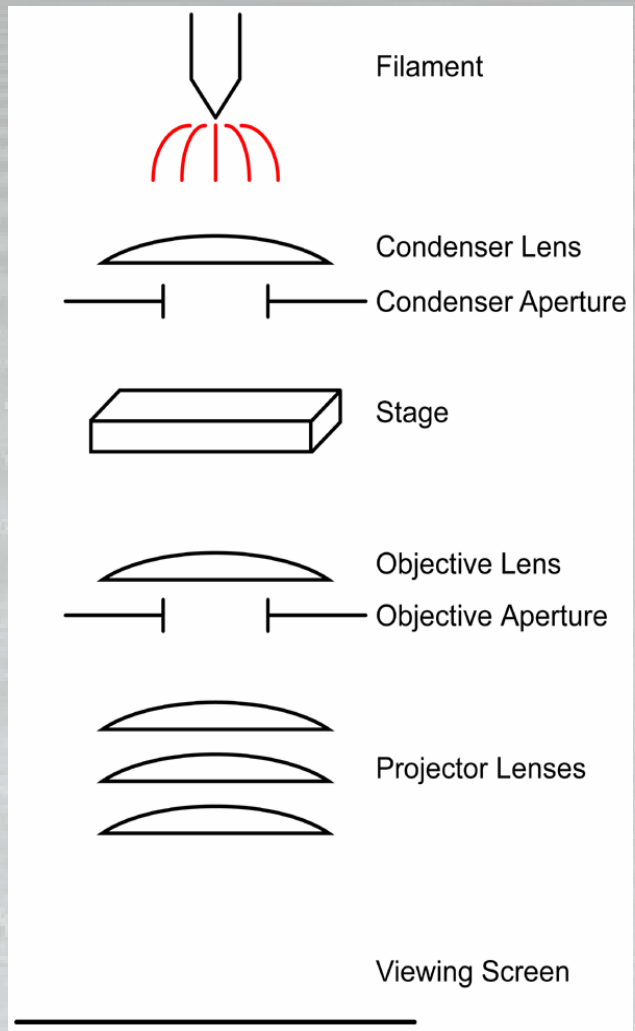
- In 3D Electron Microscopy *individual molecules* are visualized

- Trapped in ice, these molecules are free to adapt many conformations



An electron microscope



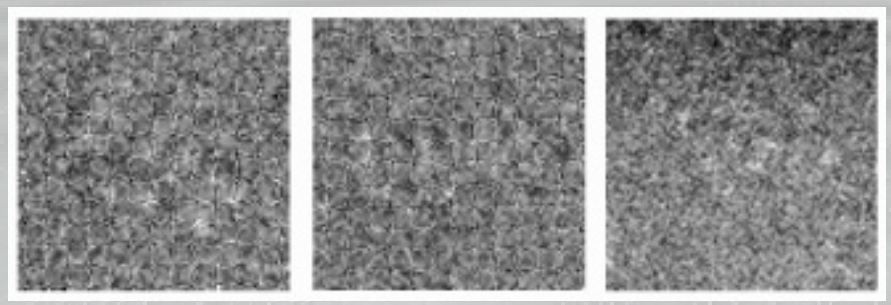


Inconveniences in 3D-EM

- The experimental signal-to-noise ratio is ~1/10
- We collect 2D-images, while often we want to know about our molecules in 3D
- The molecules adopt unknown orientations on the experimental support
- The molecules may adopt distinct conformations

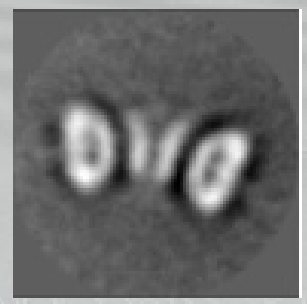
Quite a problem

- Fight the noise by averaging



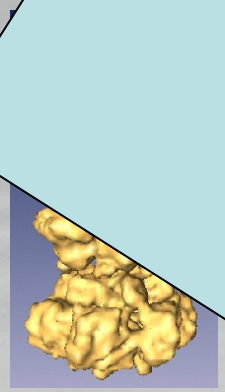
average over

2,000 copies



- BUT, this requires:
  - alignment: determine the unknown orientations
  - classification: separate distinct conformations

Classification



noise forms a serious problem!

strongly intertwined!

alignment & class

Structural heterogenity

- Our approach:
  - Combine classification & alignment in a single optimization process
    - multi-reference refinements
  - Use maximum-likelihood principles

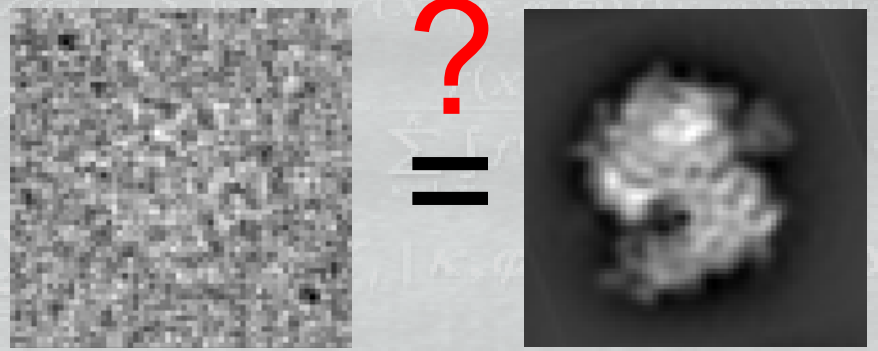
Why maximum likelihood?

Conventional data models

No noise term considered

Maximum cross-correlation (~least squares)

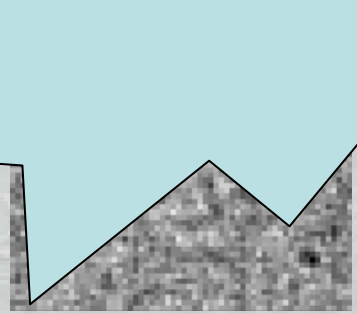
$$X_i = P_\varphi V_k$$

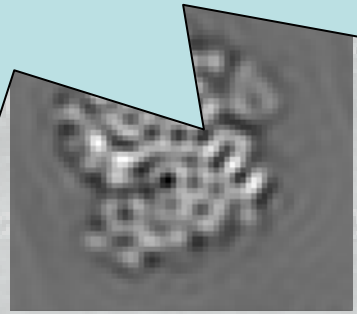


Conventional data models

Maximum cross-co ~least squares)

But what about experimental noise ?!



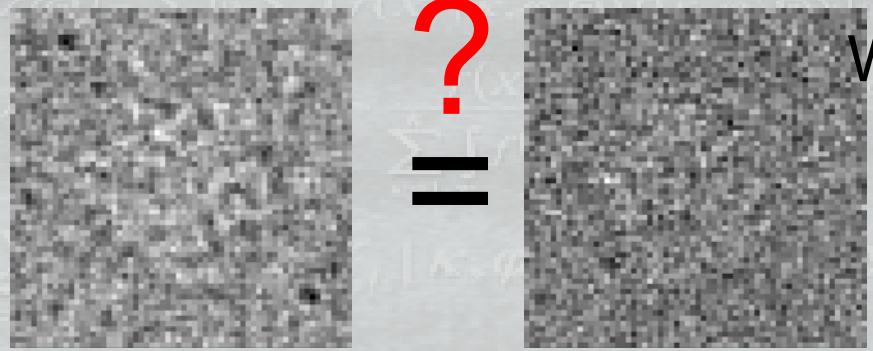


Statistical data models

Introducing a “simple” additive noise term

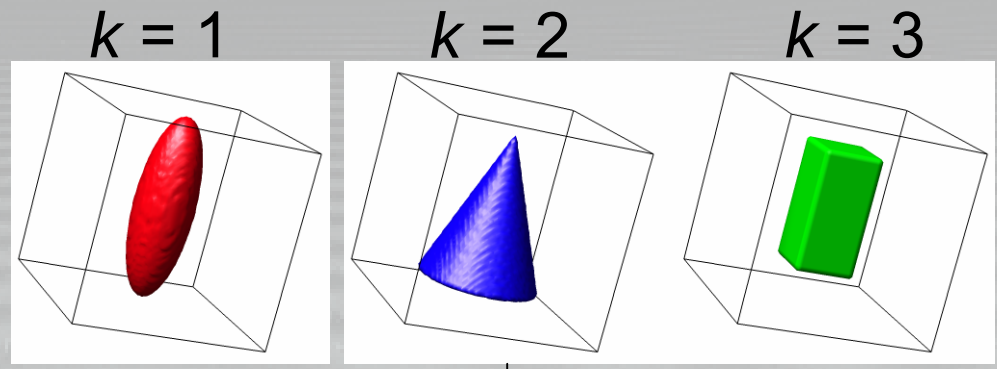
Maximum likelihood

$$X_i = P_\varphi V_k + N_i$$



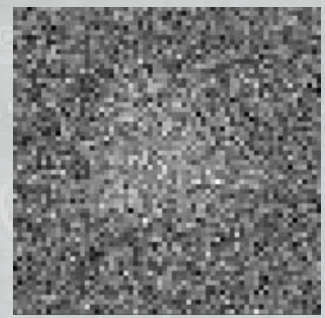
White, stationary, Gaussian noise

Statistical model



Each image is a projection of one of *K* underlying 3D objects *k*

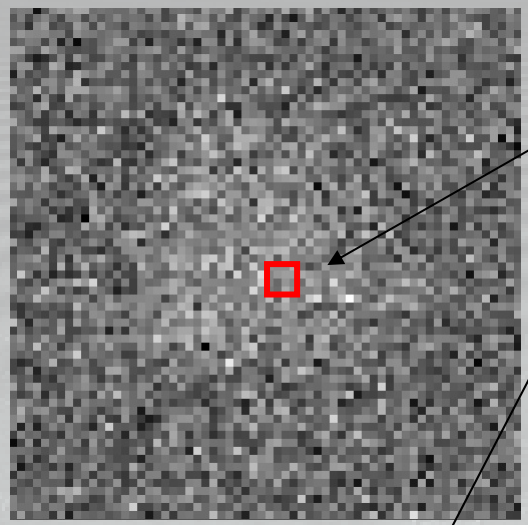
with addition of white Gaussian noise



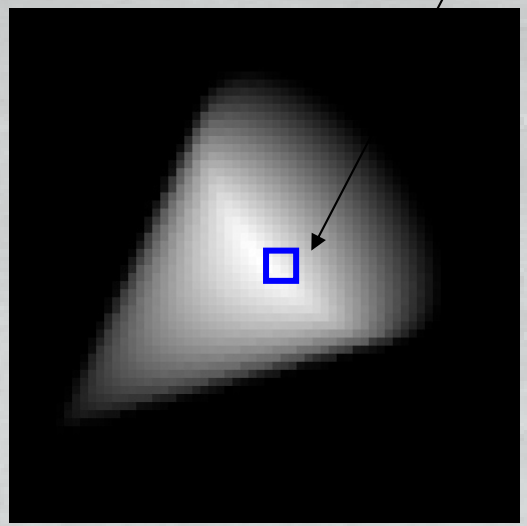
Unknowns: the 3D objects *k*, orientations

Statistical model

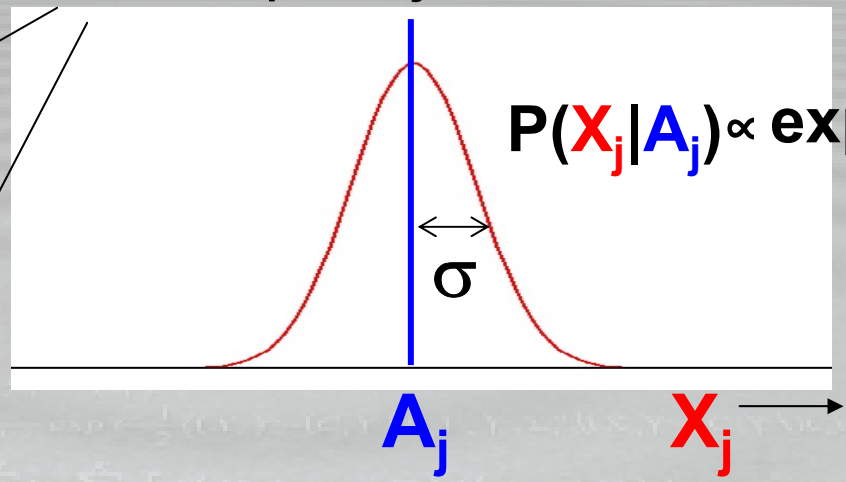
data: X



model: A



for each pixel j:



$$P(X_j|A_j) \propto \exp\left(\frac{(X_j - A_j)^2}{-2\sigma^2}\right)$$

White noise =
independence between pixels!

P(data image|model image) ~

$$\prod_j P(X_j|A_j)$$

Log-likelihood function

- Adjust model to maximize the log-likelihood of observing the entire dataset:

$$L(\text{model}) = \sum_{i=1}^{N} \ln P(\text{image}_i \mid \text{model})$$

$$= \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \sum_{orient} P(\text{image}_i \mid k, \text{orient.}, \text{model}) P(k, \text{orient.} \mid \text{model})$$

The **model** comprises:
- estimates for the underlying objects
- estimate for the amount of noise (σ)
- statistical distributions of k & orient.

Optimization algorithm: Expectation Maximization

Two cases

- Alignment & classification in 2D:
  - align images and calculate 2D averages for the distinct classes

- Alignment & classification in 3D
  - align images and calculate 3D reconstructions for the distinct classes

The 2D algorithm

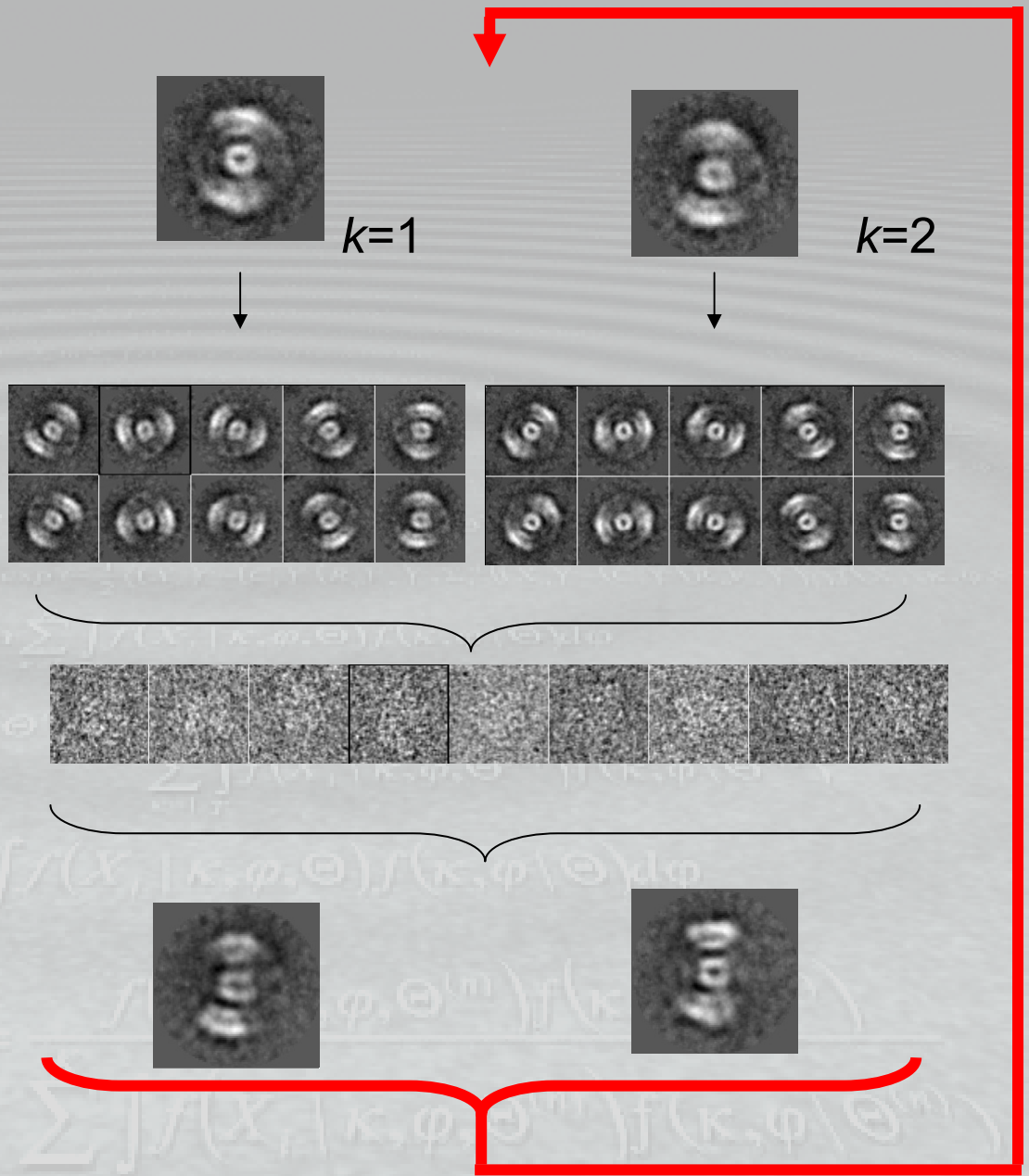
estimates for *K* 2D objects

sampled rotations 360°

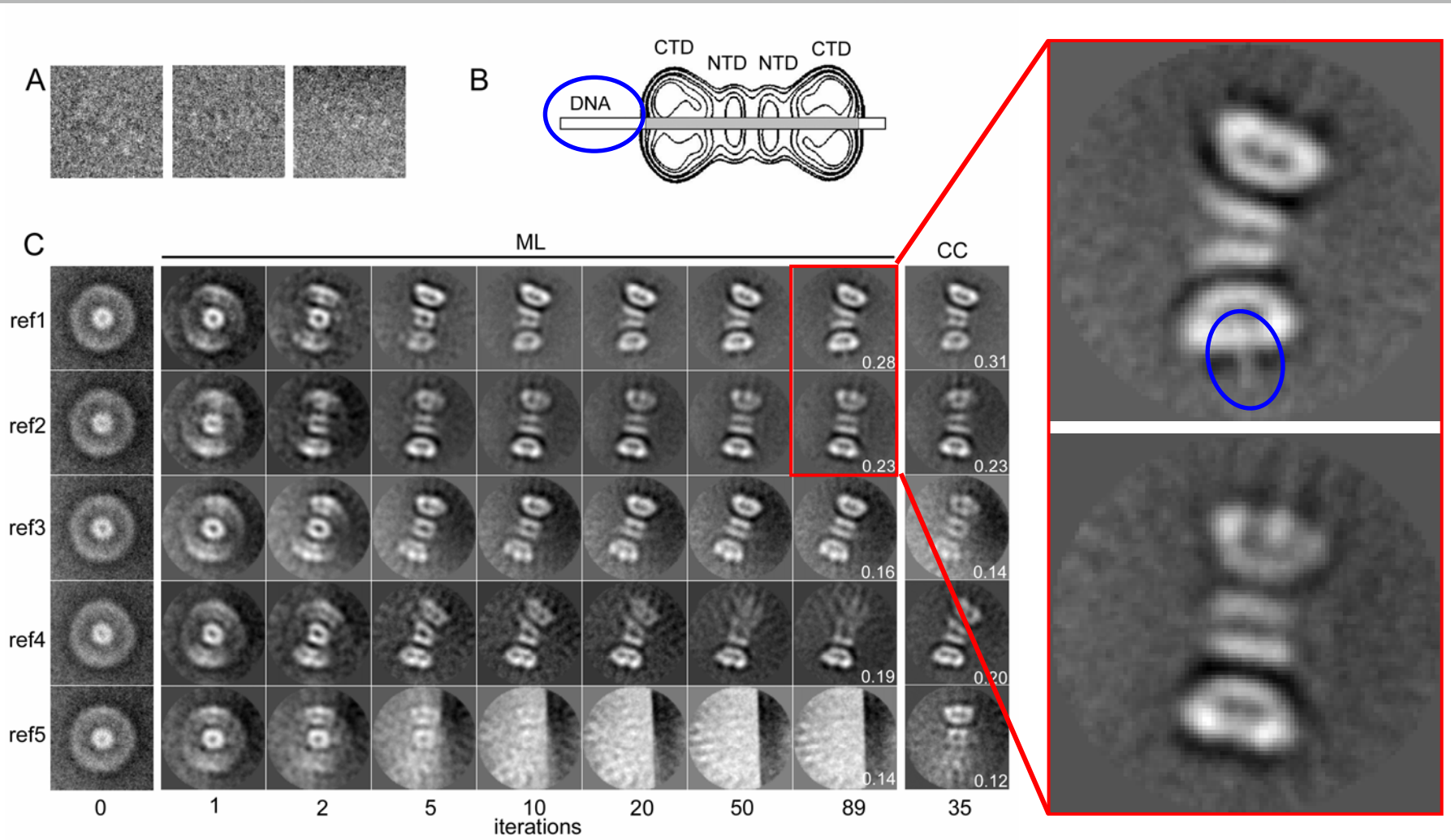
for each image, calculate all

$$P(\text{image}_i \mid k, \text{rot})$$

calculate new 2D average as *probability weighted averages*



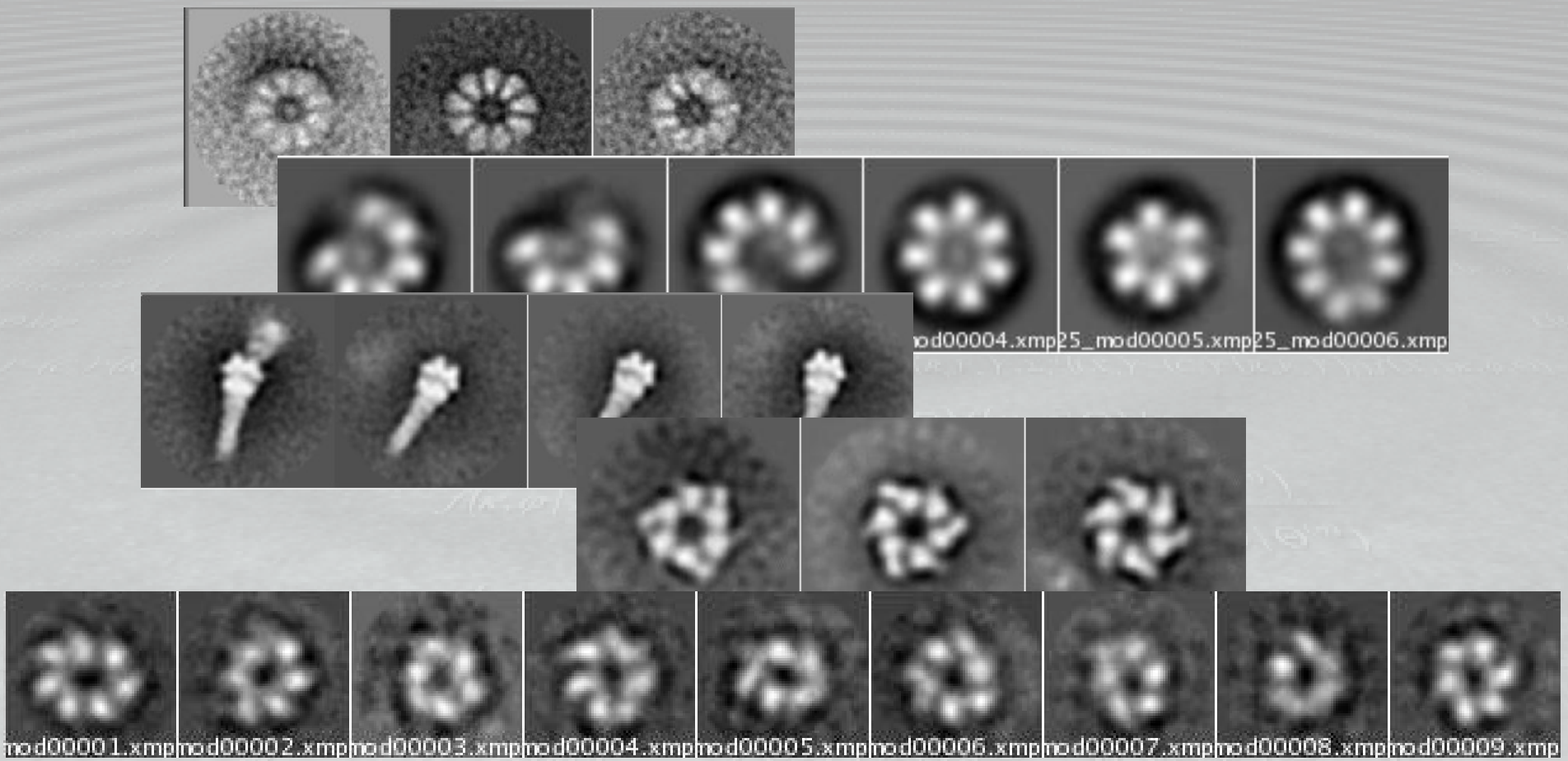
ML2D classification



Scheres *et al.* (2005) *J. Mol. Biol.*, **348**, 139-149

Scheres *et al.* (2005) *Bioinformatics* **21** (Suppl. 2), ii243-ii244

ML2D classification



The 3D algorithm

estimates for *K* 3D objects

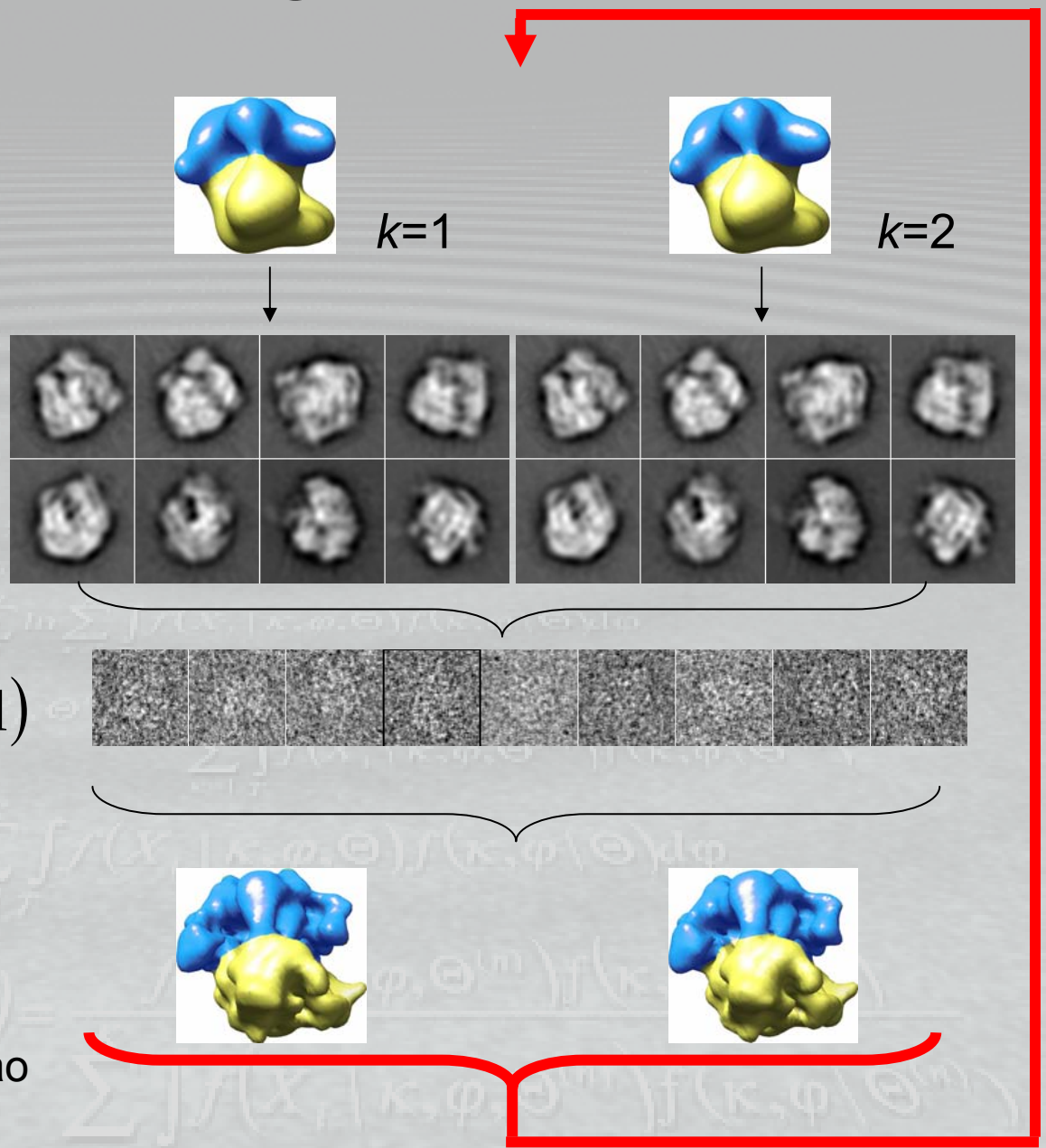
project into all (discretely sampled) orientations

for each image, calculate all

$$P(\text{image}_i \mid k, \text{orient.}, \text{model})$$

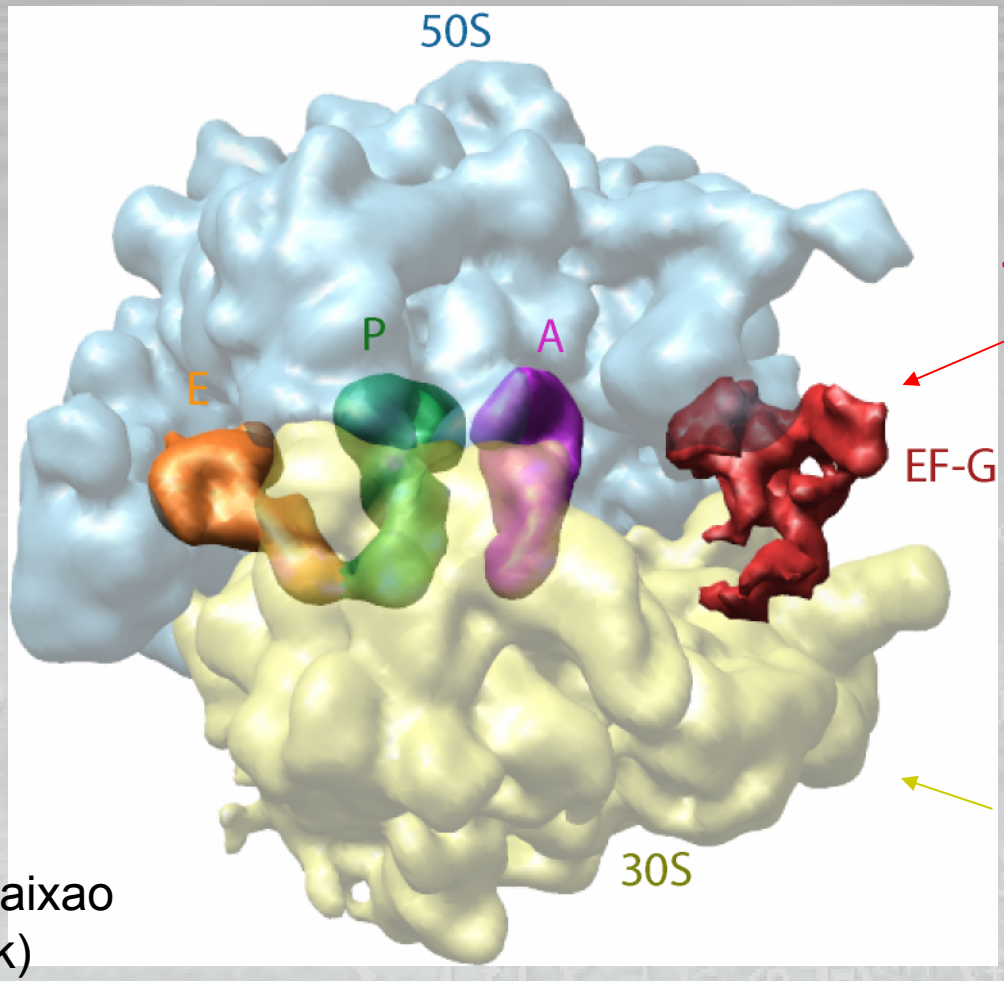
calculate new 3D estimates as *probability weighted 3D reconstructions*

(kindly provided by Haixao Gao & Joachim Frank)



Prelim. ribosome reconstruction

91,114 particles; 9.9 Å resolution

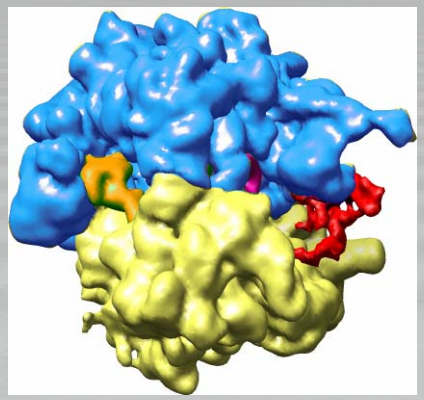


fragmented
(depicted at a lower threshold)

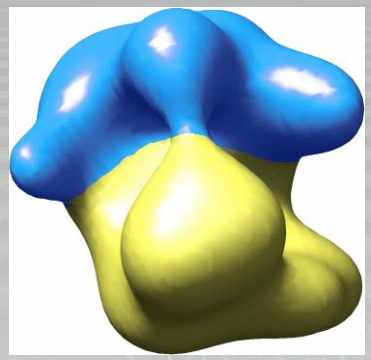
blurred

(kindly provided by Haixao Gao & Joachim Frank)

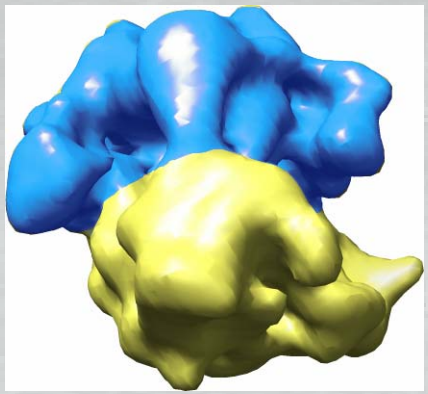
Seed generation

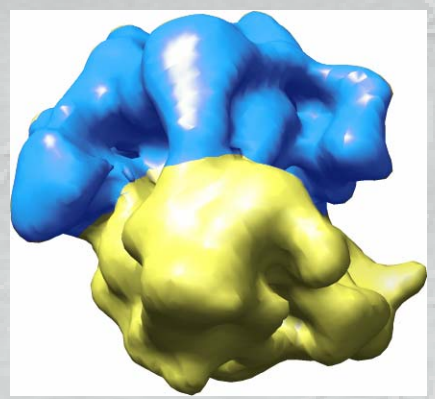


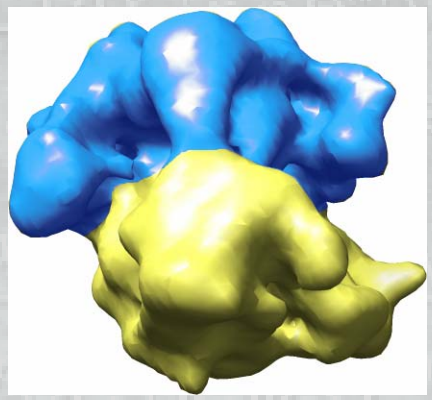
80 Å
filter

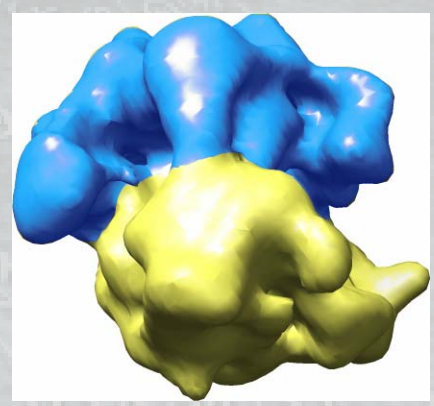


4 random subsets; 1 iter ML



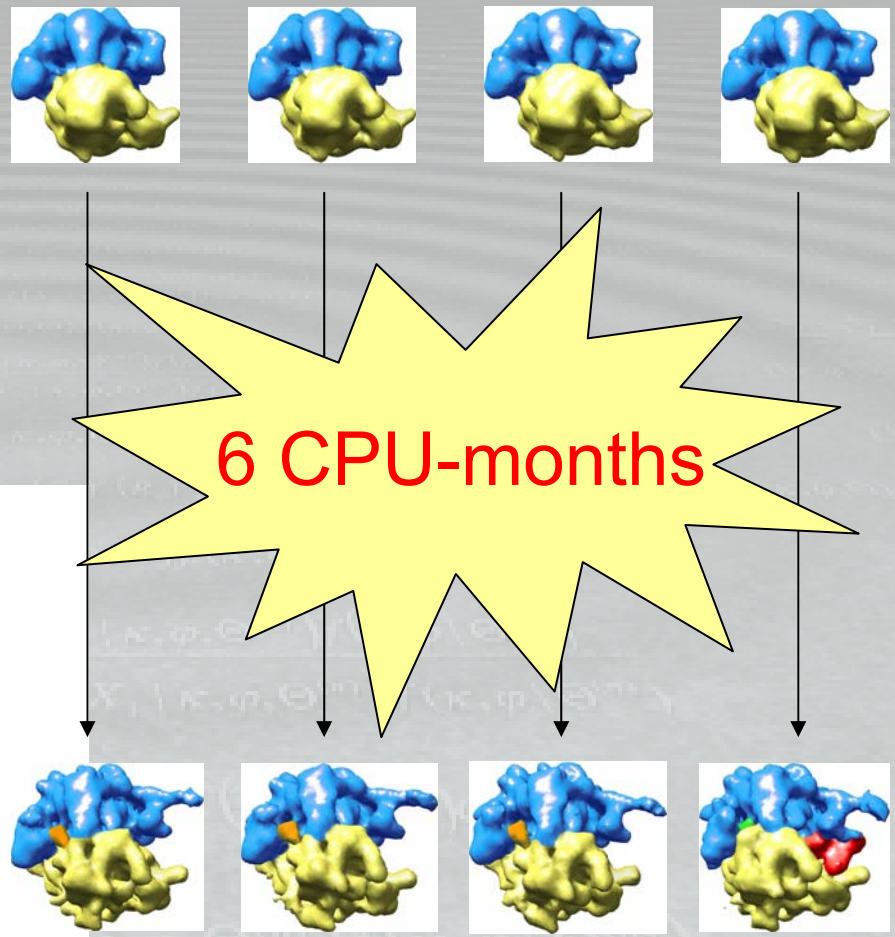


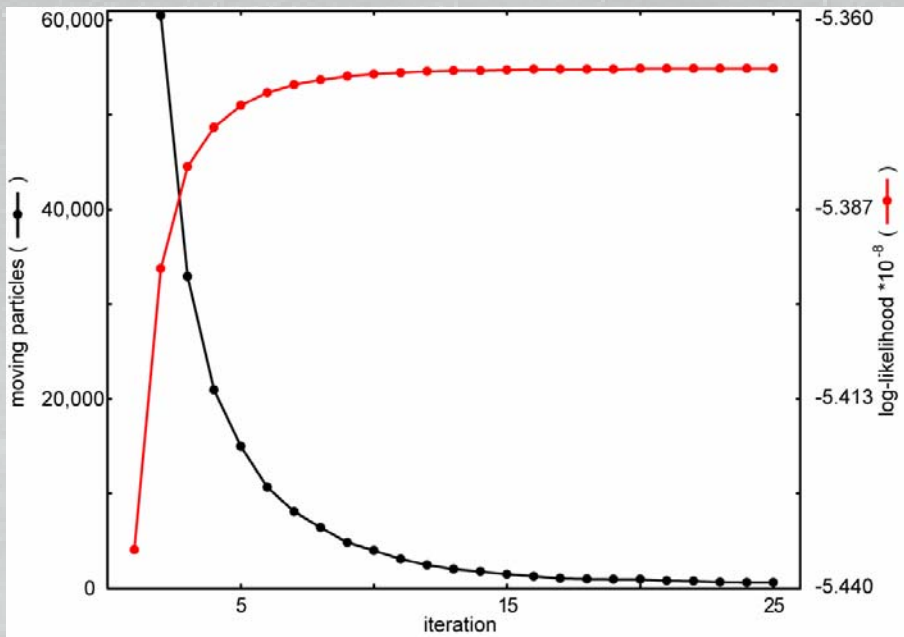




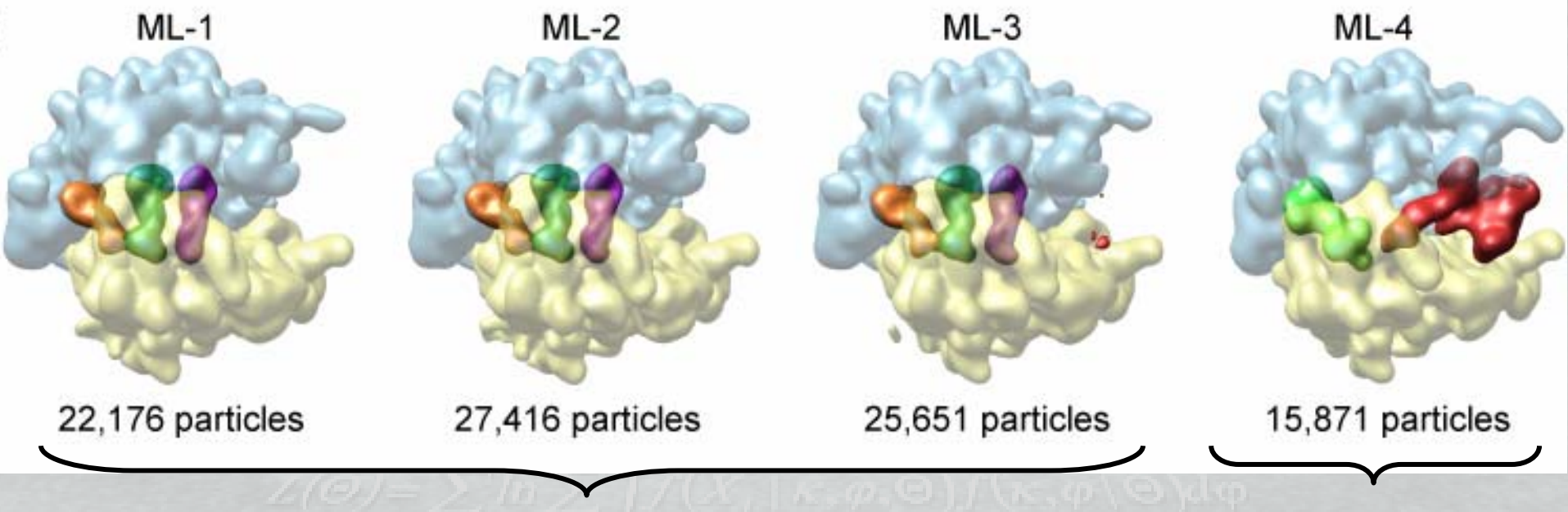
ML3D-classification

- 4 references
- 91,114 particles
- 64x64 pix (6.2Å/pix)
- 25 iterations
- 10° angular sampling





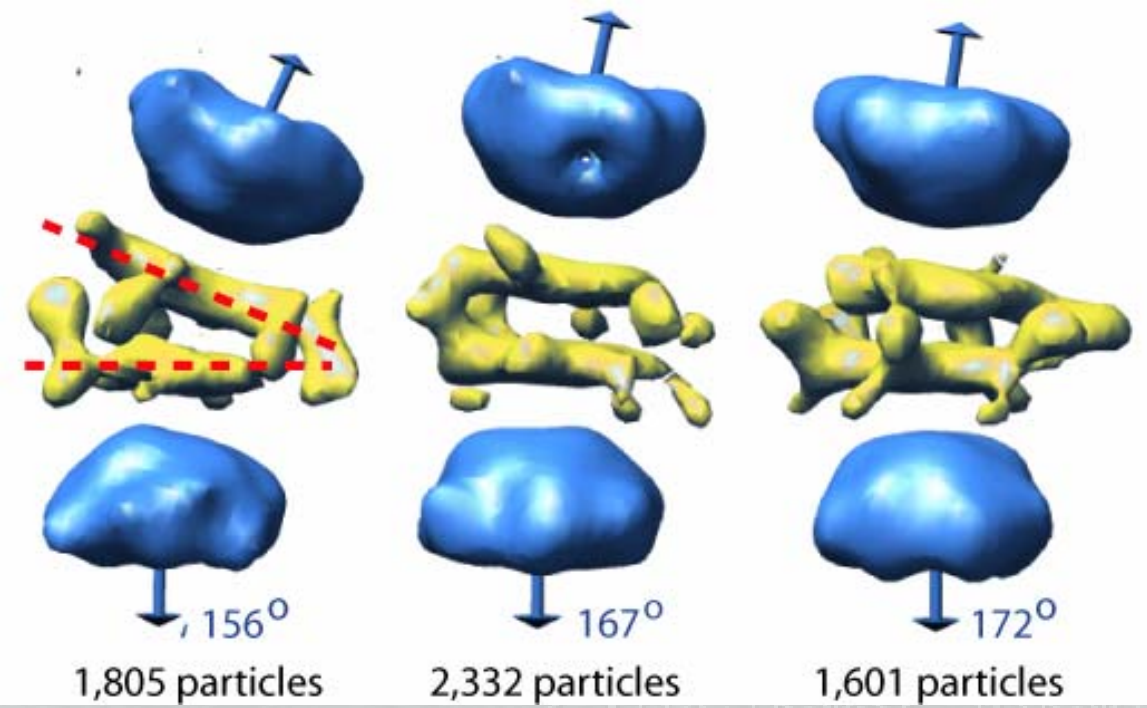
ML-derived classes



no ratcheting; no EF-G; 3 tRNAs
differences: overall rotations

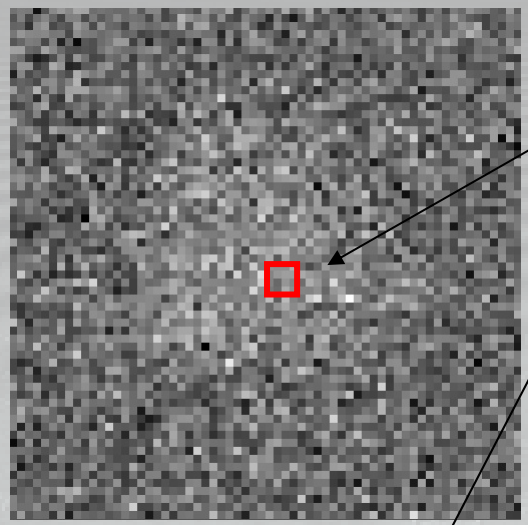
ratcheting,
EF-G, 1 tRNA

ML3D classification

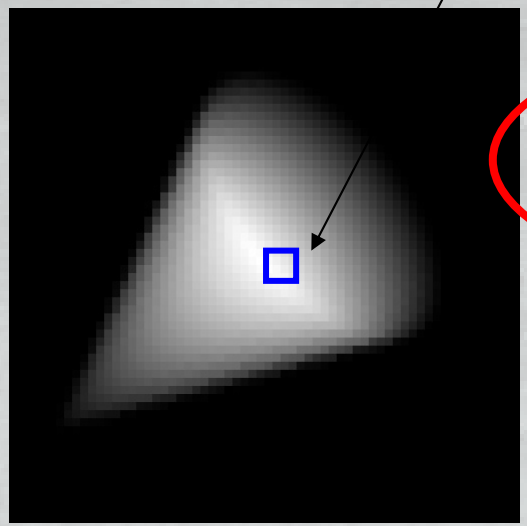


Scheres *et al.* (2007) *Nat Methods*, 4, 27-29

Statistical model

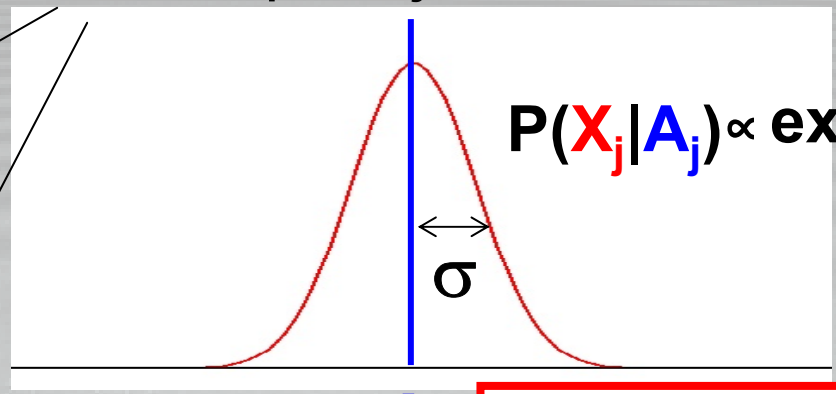


data: X



model: A

for each pixel j:



$$P(X_j|A_j) \propto \exp\left(\frac{(X_j - A_j)^2}{-2\sigma^2}\right)$$

A_j

NOT TRUE!

White noise =
independence between pixels!

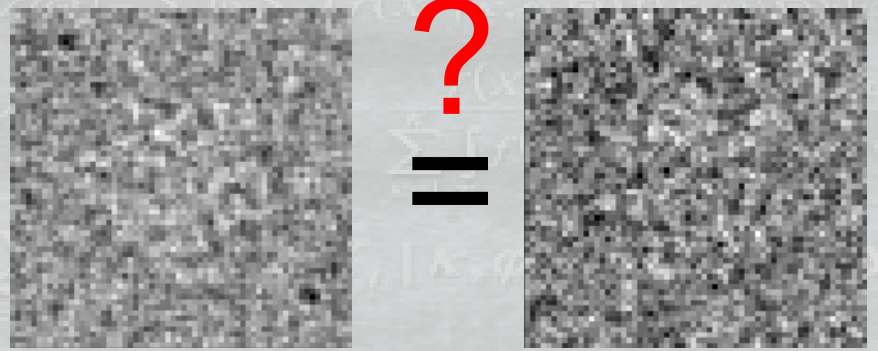
P(data image|model image) ~

$$\prod_j P(X_j|A_j)$$

An improved data model

Maximum likelihood

$$X_i = CTF_i * P_\varphi V_k + N_i$$

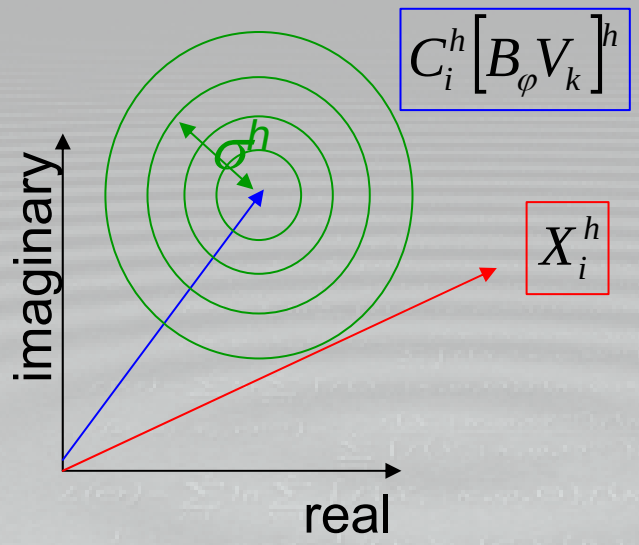


spatially stationary Gaussian noise,

Coloured noise!,…

Coloured noise model

for each Fourier pixel *h* 2D-Gaussian in complex plane

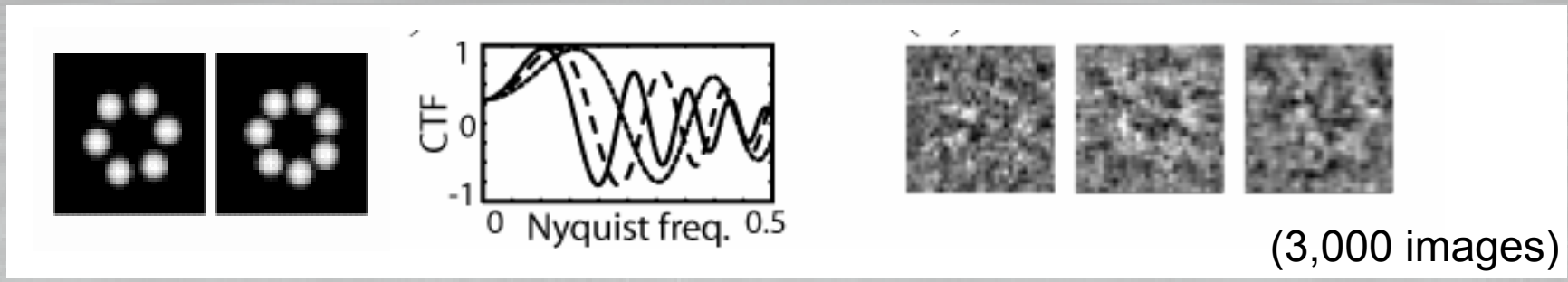


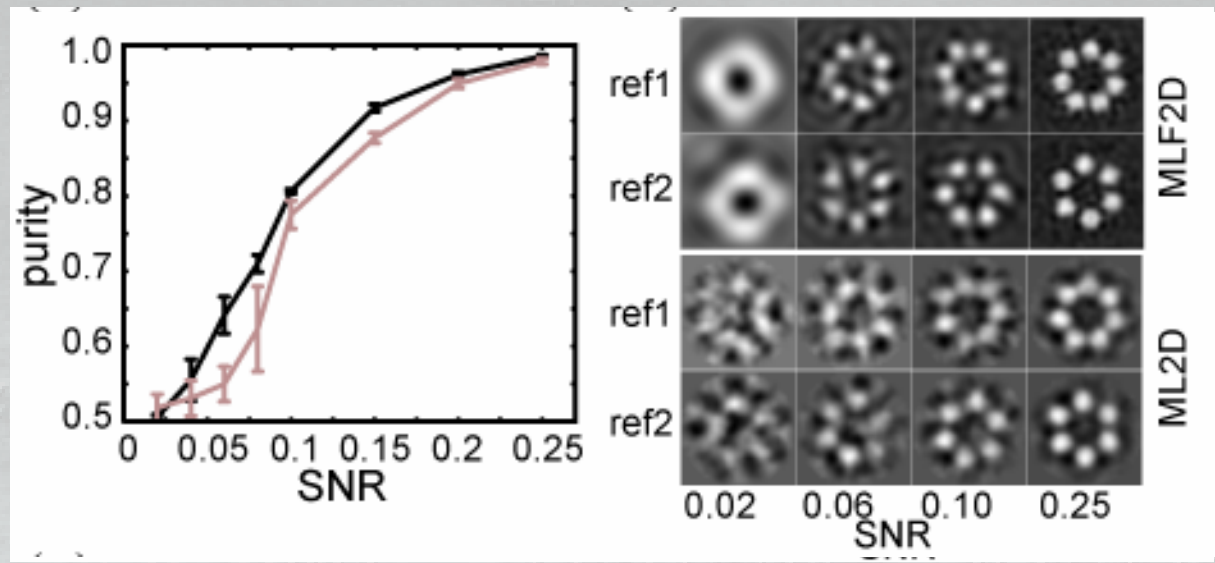
Assuming independence of noise between all Fourier terms:

$$P\left(X_i \mid k,\varphi,\Theta\right)=\prod_{h=1}^{H}\frac{1}{2\pi\left(\sigma^h\right)^2}\exp\left(\frac{\left|CTF_i^h\left[P_\varphi V_k\right]^h-X_i^h\right|^2}{-2\left(\sigma^h\right)^2}\right)$$

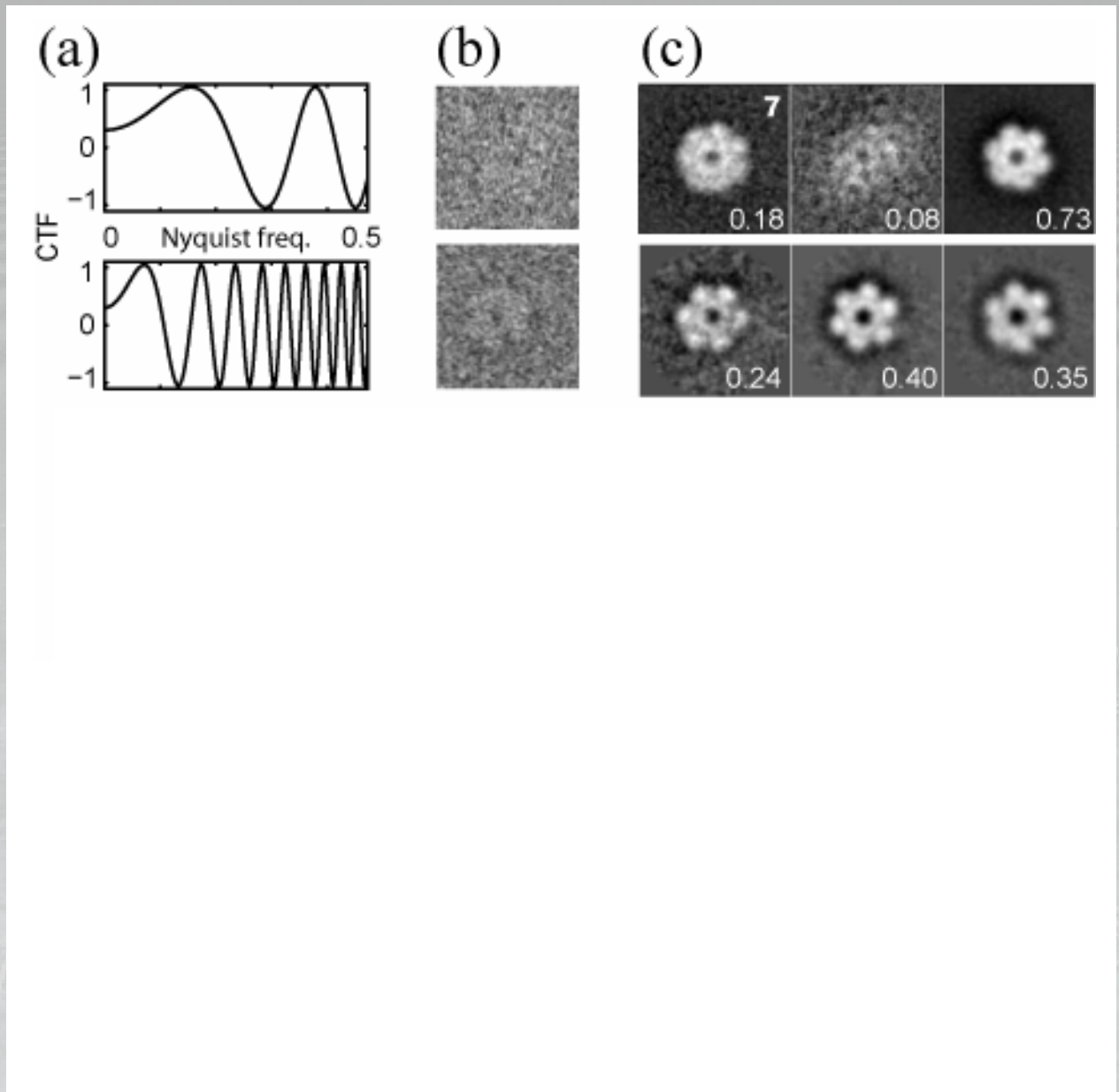
resolution-dependent noise model!

Simulated data



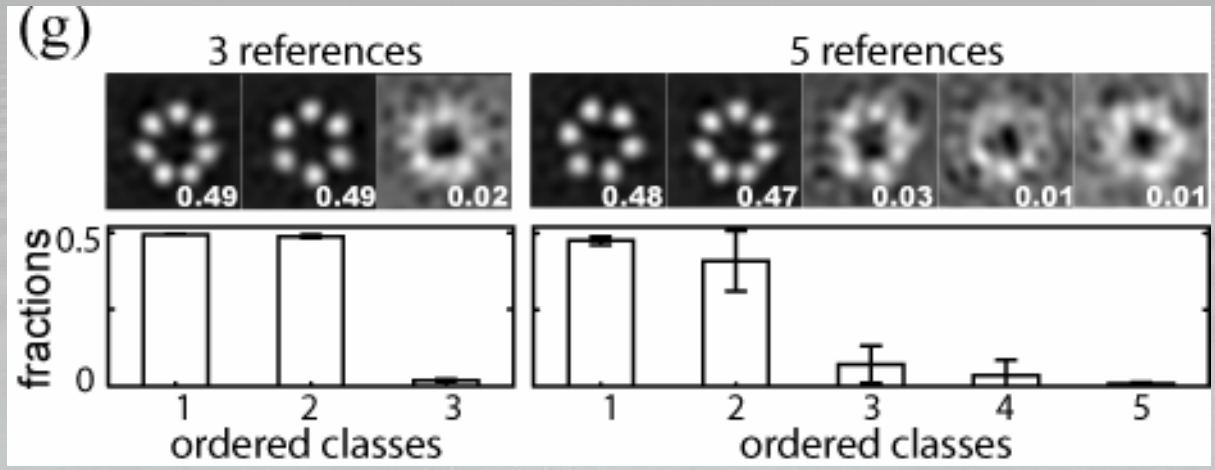


Archaeal helicase MCM

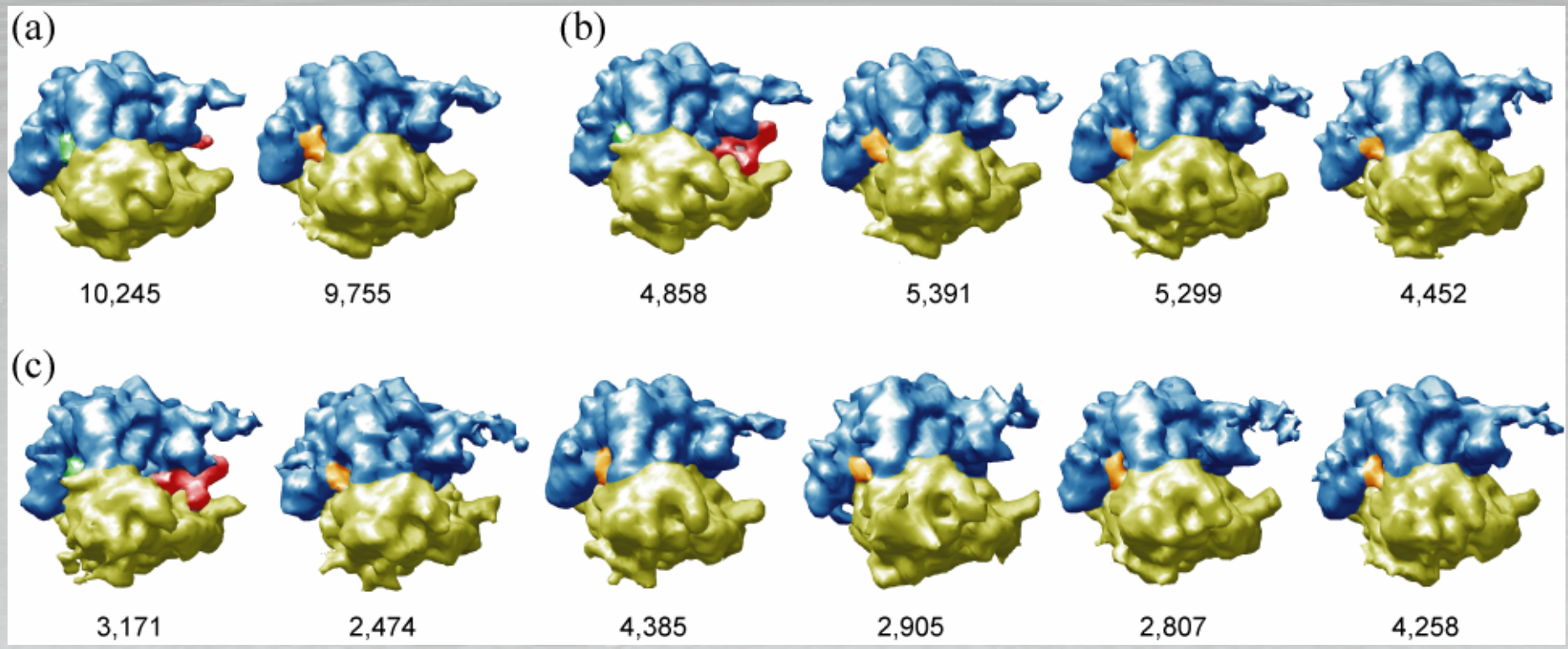


(4,042 images)

Simulated data



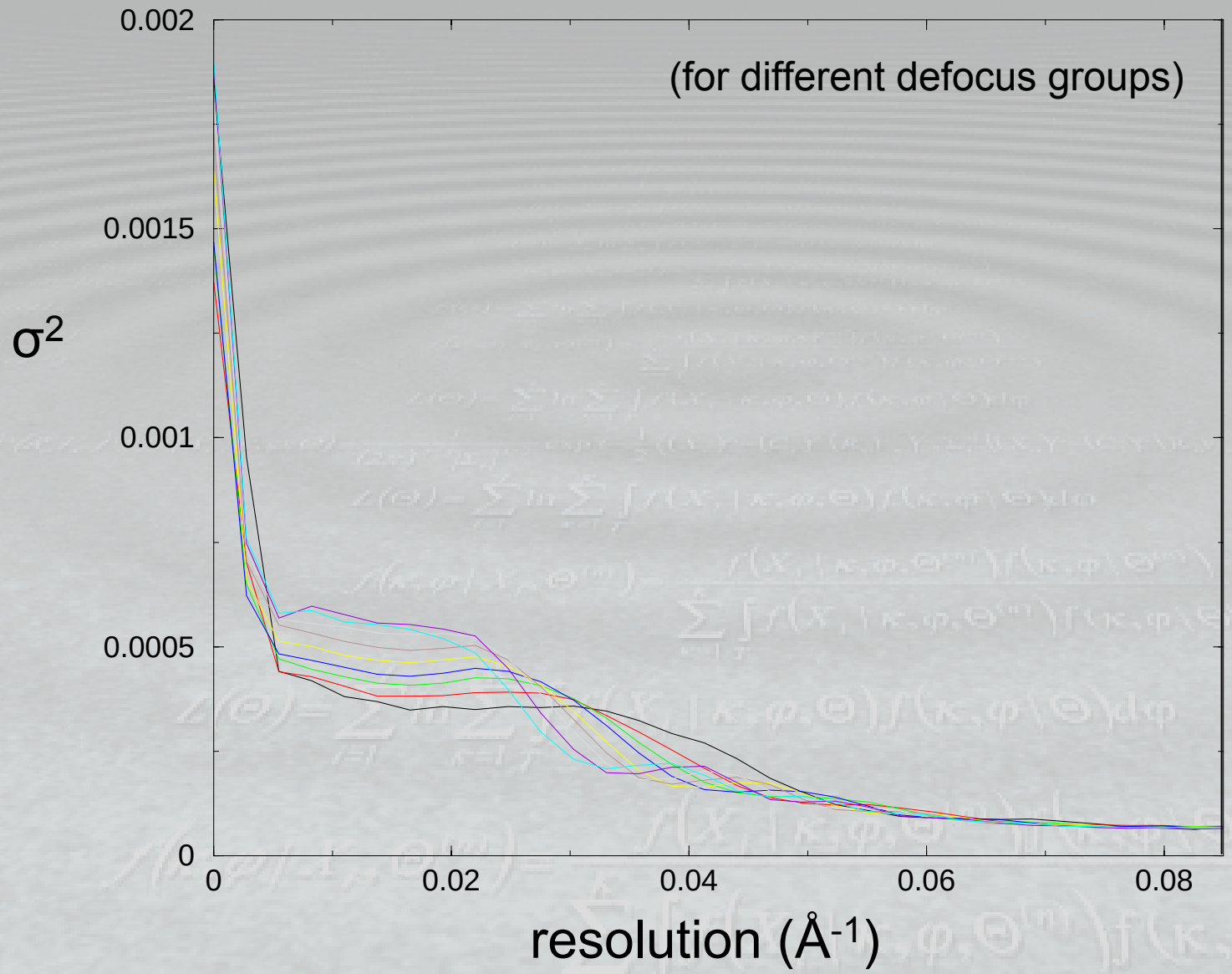
70S *E.coli* ribosome



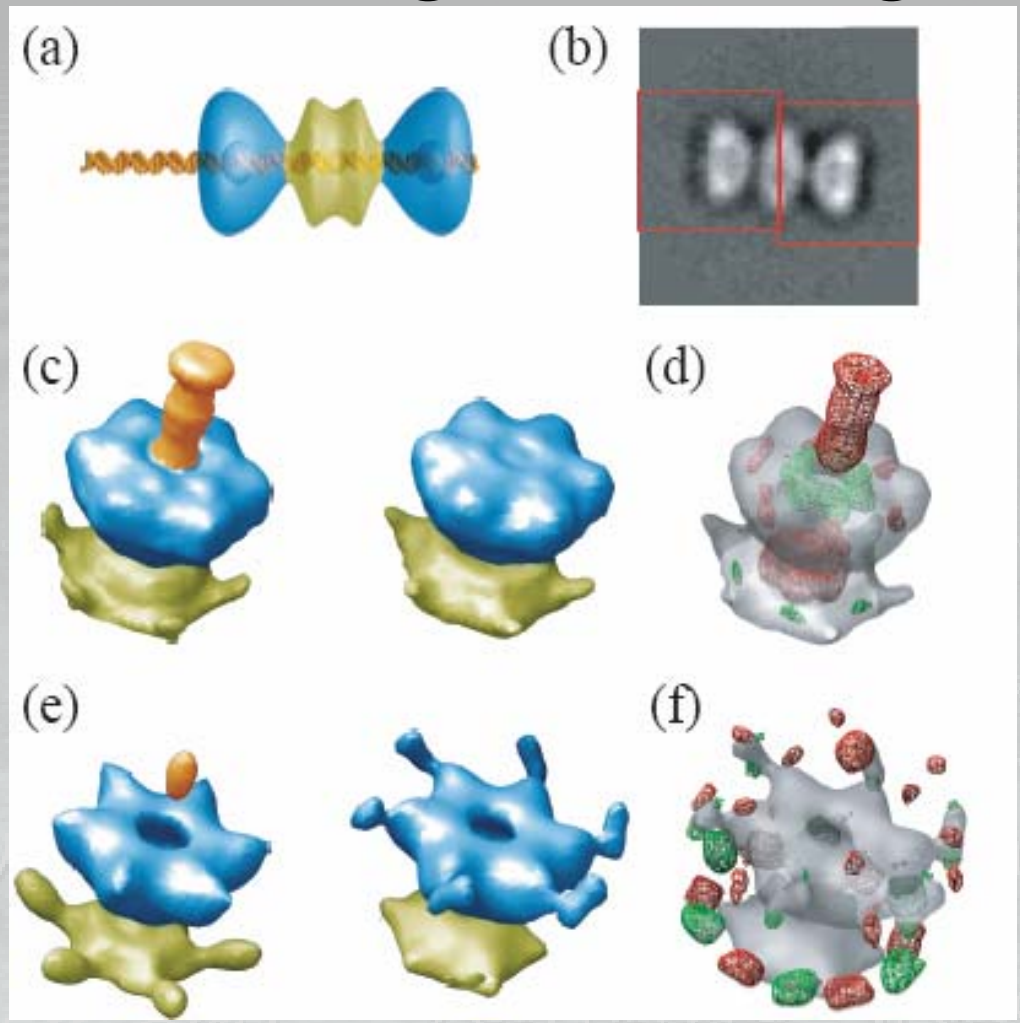
(kindly provided by Haixao Gao & Joachim Frank)

(20,000 images)

Coloured noise!!



SV40 large T-antigen



(7,718 sub-images)

Future plans

- Improve robustness: Outliers!
- Decrease computational burdens
- Overcome model bias!!!!!!!!!
  - *One of the most serious problems in the field*

MLF3D: A new approach that complements previous methods

- Like 2D/3D classification
  - by “Quantitative Self Organizing Maps” (KerDenSOM)
- Like new factorization schemes oriented to provide “factors” more directly understandable than PCA factors:
  - non-smooth Non-Negative Matrix Factorization (ns-NMF)

Exploring data: Smoothly Distributed Kernel Probability Density Estimator

- In the context of "Exploratory Data Analysis", it would be interesting to work with a new SOM optimized to preserve the estimation of the pdf of the input in the mapped (output) space

$$\max\left\{\sum_{i=1}^{c}\ln\left(\frac{1}{c}\sum_{j=1}^{c}K\left(X_i-V_j;\alpha\right)\right)-\frac{\vartheta}{2\alpha}tr\left(V^T DV\right)\right\}$$

- Results:

$$K_\alpha;Kernel(Parzen)$$

Maximum Likelihood

Pascual et al., Pattern Rec. 2001, J.Struct.Biol. 2002, 2003,
PCT and US Patents

Calculation of U_{ij}

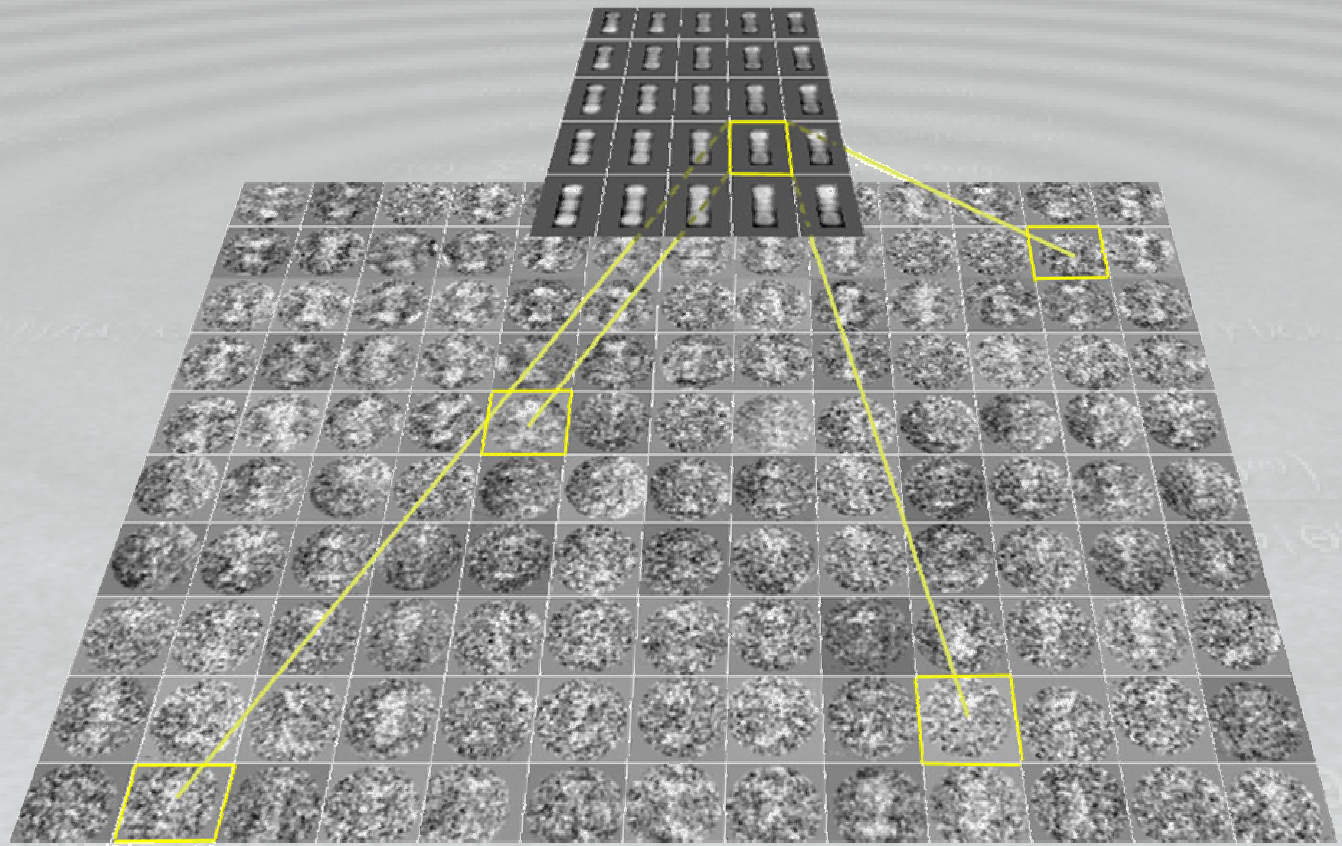
$$U_{ji}=\frac{K(X_i-V_j;\alpha)}{\sum_{k=1}^{c}K(X_i-V_k;\alpha)}$$

Iterative calculation of V_j

$$\mathbf{V}_j=\frac{\sum_{i=1}^{n}U_{ji}^{m}\mathbf{X}_i+\vartheta\overline{V}_j}{\sum_{i=1}^{n}U_{ji}^{m}+\vartheta}$$

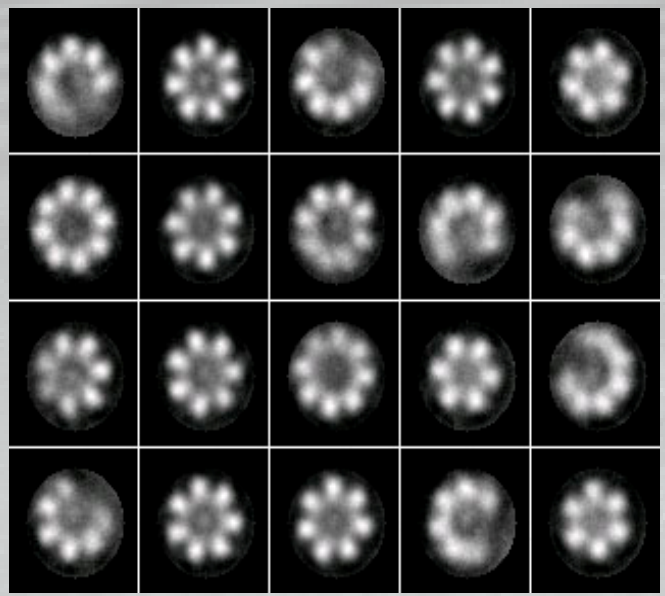
Application in 2D analysis:

Original T-Antigen double hexamers cryo-electron single particle images.

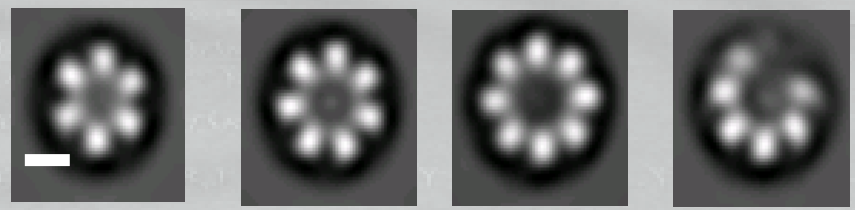


KerDenSOM in 2D

Self-organizing map
***mt* MCM**



Class average images



6-fold 7-fold 8-fold Open Ring

Gómez-Llorente et al, J.Biol.Chem., 2005

KerDenSOM in 3D

Subtomogram averaging: Insect Flying Muscle
(K.Taylor collaboration)

Non-negative matrix factorization

$$\mathbf{V} \approx \mathbf{WH}$$

$$(\mathbf{V})_{i\mu} \approx (\mathbf{WH})_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu}$$

V： Data matrix

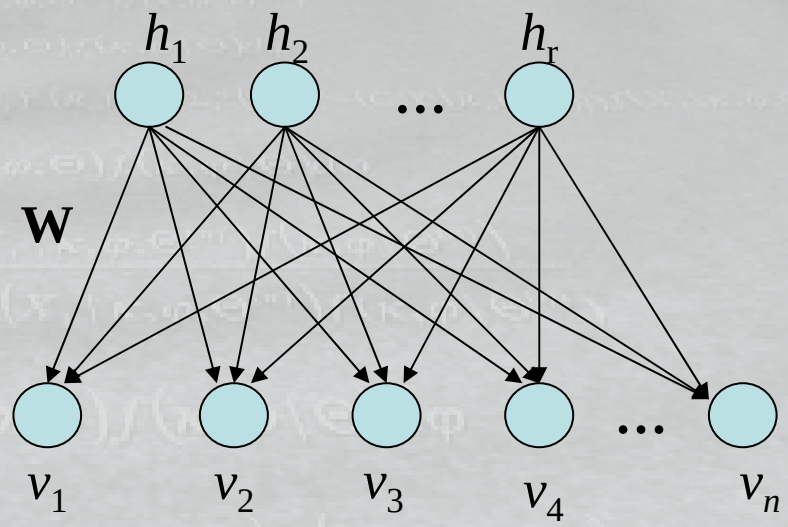
W： basis matrix (prototypes)

H： encoding matrix (in low dimension)

CONSTRAINTS:

$$V_{i\mu}, W_{ia}, H_{a\mu} \geq 0$$

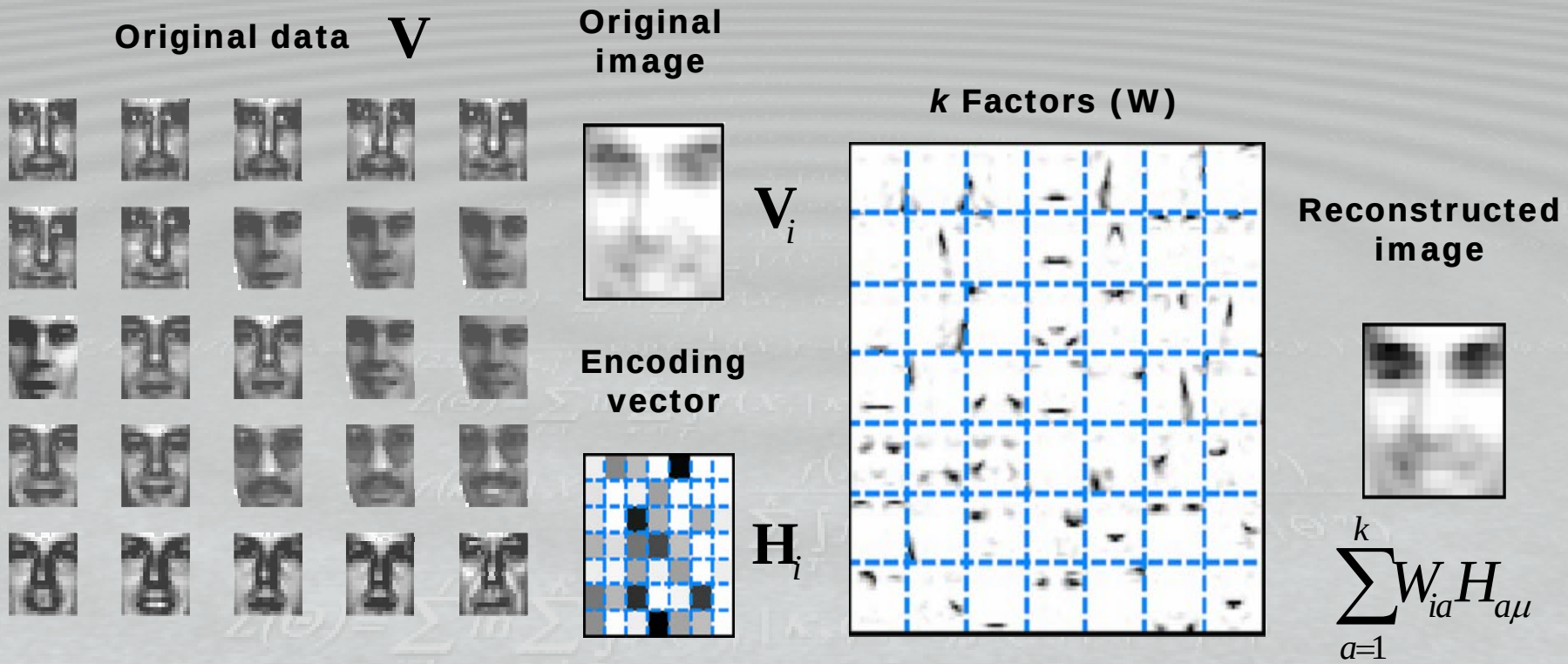
- NMF as a latent variable model



Daniel D. Lee & H. Sebastian Seung. NATURE |VOL 401 | 21 OCTOBER 1999

Pascual-Montano et al., IEEE PAMI, 2006

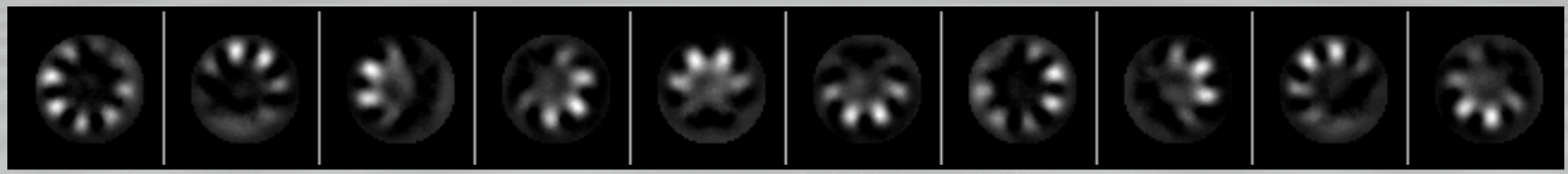
Example with NMF:



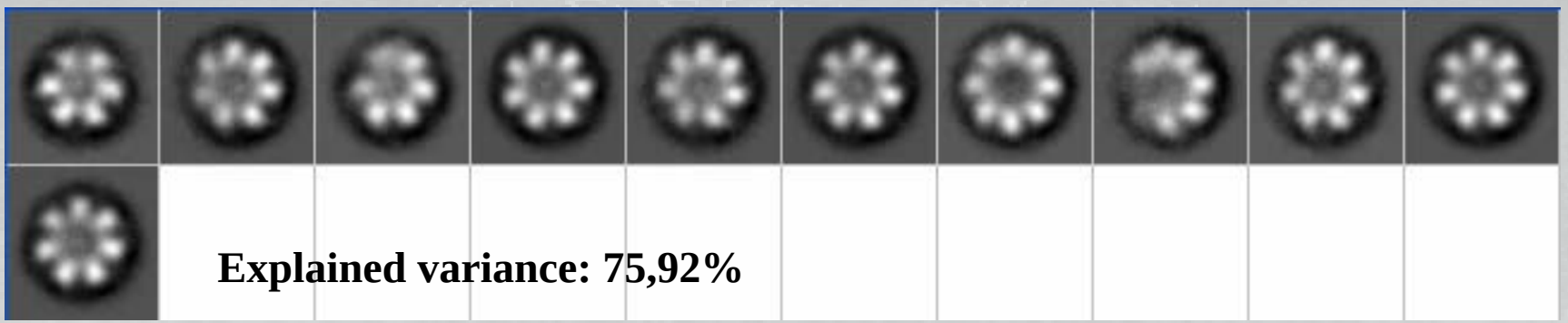
Lee, D.D. and Seung, H.S., Nature, 1999. 401 (6755): p. 788-91

ns-NMF (on *mt* MCM)

ns-NMF factors:



Classes after classification:



Pascual-Montano el al. IEEE PAMI, 2006; Chagoyen et al., BMC Bioinformatics, 2006a,b; Carmona et al., BMC Bioinformatics, 2006

The Biocomputing "cluster"

J.M.Carazo (CNB)

- ***Structural biology of helicases***
  - **Dr.Mikel Valle (CNB) * (Biogune-Bilbao)**
    - **Roberto Melero (CNB)**
  - **Dr.Carmen San Martín (CNB)**
    - **Yacob Gómez (CNB)**
    - **Marta Rajkiew (CNB)**
  - **Dr. Rafael Núñez (CNB)**
  - **Dr. Isabel Cuesta (CNB)**

- ***Structural biology of the centrosome***
  - **Dra.Rocio González (CNB)**
  - **Dr.Johan Busselez (CNB)**

Methods development in 3DEM

- **Dr. Sjors Scheres (CNB)**
- **Dr. Javier Velázquez (CNB) *UCSF**
- **Dr. Roberto Marabini (UAM)**
  - **Ignacio Arganda (UAM)**
  - **Ana Iriarte (UAM)**
- **Dr. Carlos Oscar Sánchez (CEU)**

- ***Computational and data Grid***
  - **Dr. José Ramón Macías (PCM/INB)**
  - **Eng. Alfredo Solano (CNB)**

- ***National Institute for Bioinformatics***
  - **Dr.Natalia Jiménez-Lozano**
  - **Joan Segura**

- ***Gene Expression Data Analysis-UCM (Dr. Alberto Pascual)***
  - **Dr. Federico Abascal**
  - **Dr. Monica Chagoyen (CNB)**

- ***Main external collaborators***
  - **Prof. Gabor Herman (NYU)**
  - **Prof. Ellen Fanning (Vanderbilt)**
  - **Prof. Xiojiang Cheng (USC)**
  - **Prof. Juan Carlos Alonso (CNB)**
  - **Prof. J. Frank (Albany/Columbia)**
  - **Dr.Sergio Marco (Curie)**

- **Integromics S.L., Integromics International Inc., Varbanov Soft Ltd**

  - Madrid, Granada, Rousse and Philadelphia