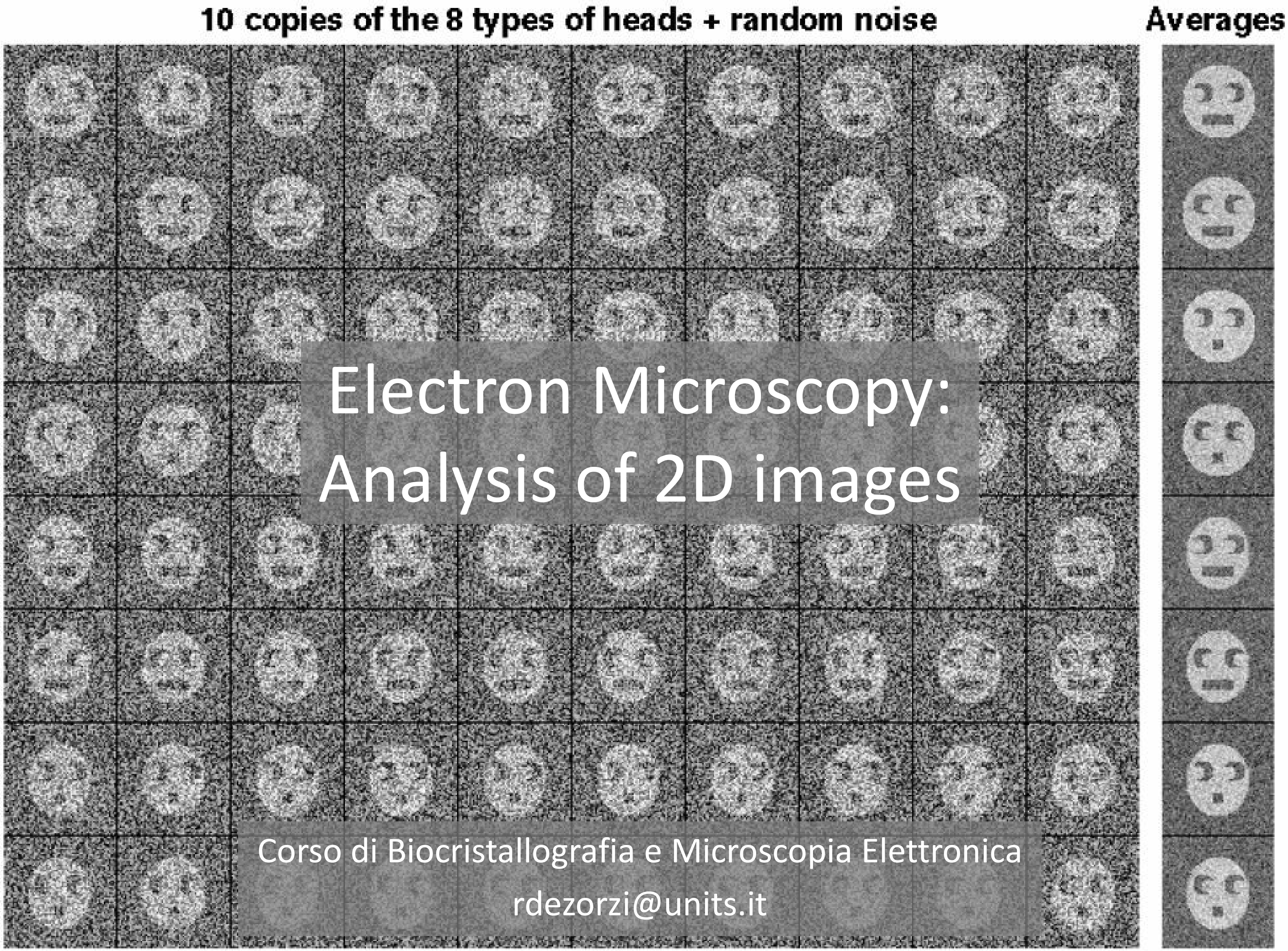
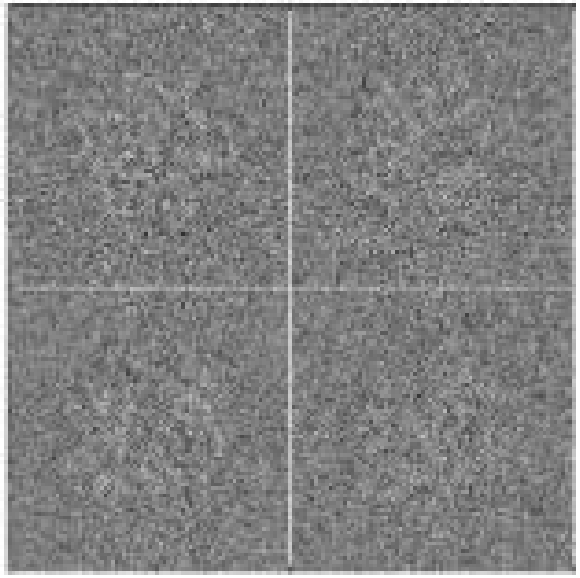


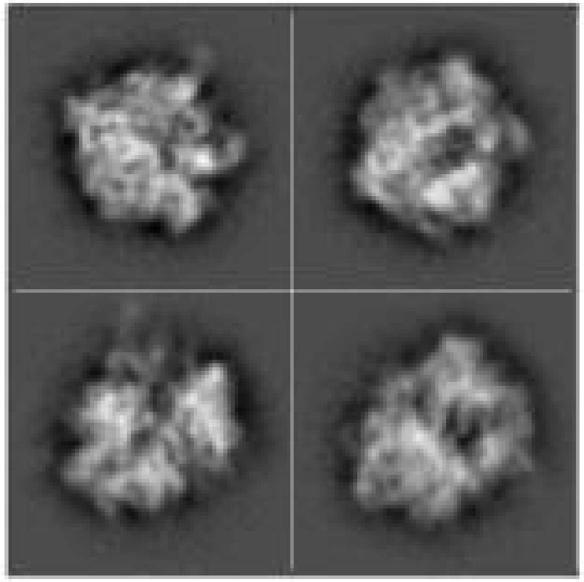
Electron Microscopy: Analysis of 2D images

Corso di Biocristallografia e Microscopia Elettronica

rdezorzi@units.it

Image = signal + noise





Due to low dose, Signal-to-Noise ratio of electron microscopy images is low.

Sources of noise: supporting carbon film, stain, fluctuations of the source, inelastic scattered electrons, lack of homogeneity in camera response, charging of the sample, …

Noise:

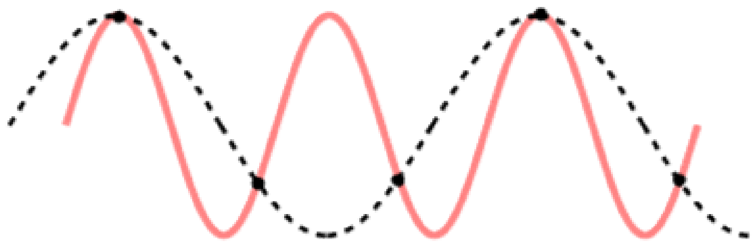
- Fixed pattern (e.g. noise of the camera): can be corrected by subtraction
- Stochastic (e.g. fluctuations of the source): can be corrected if the noise distribution is known

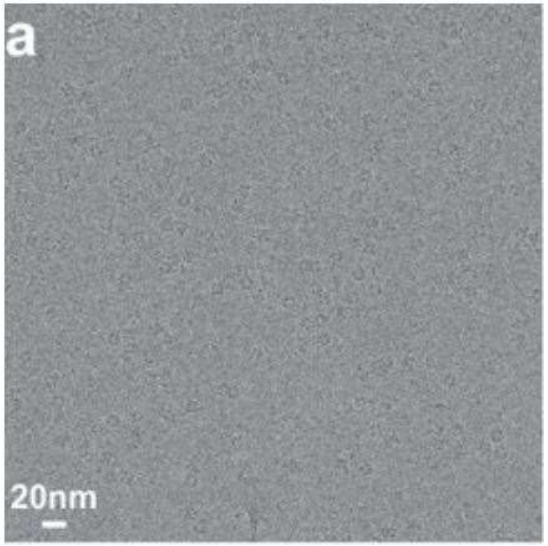
Noise whose effect is additive can be corrected by averaging. More complex corrections for non-additive noise.

Sampling theorem

(or Nyquist-Shannon or Whittaker-Shannon theorem)

A continuous function can be represented as a set of discrete measurements taken at regular intervals.
To describe a function $f(x)$ with a maximum frequency B, the minimum frequency of the sampling has to be 2B.





Magnification: ratio between the dimension of the image of the object and the dimension of the object itself (e.g. 47000x)

Micrograph dimensions: number of pixels in each direction of the micrograph (e.g. 4k x 4k camera)

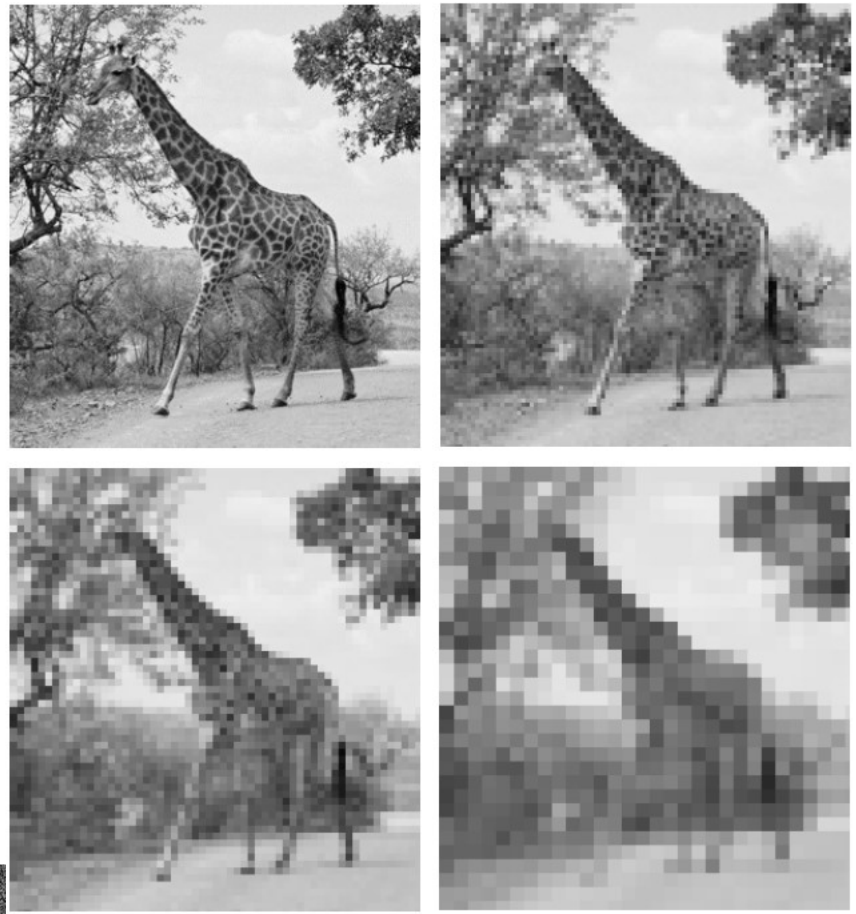
Pixel size: at a defined magnification, dimension of each detector pixel at the object level (e.g. 1.71 Å)

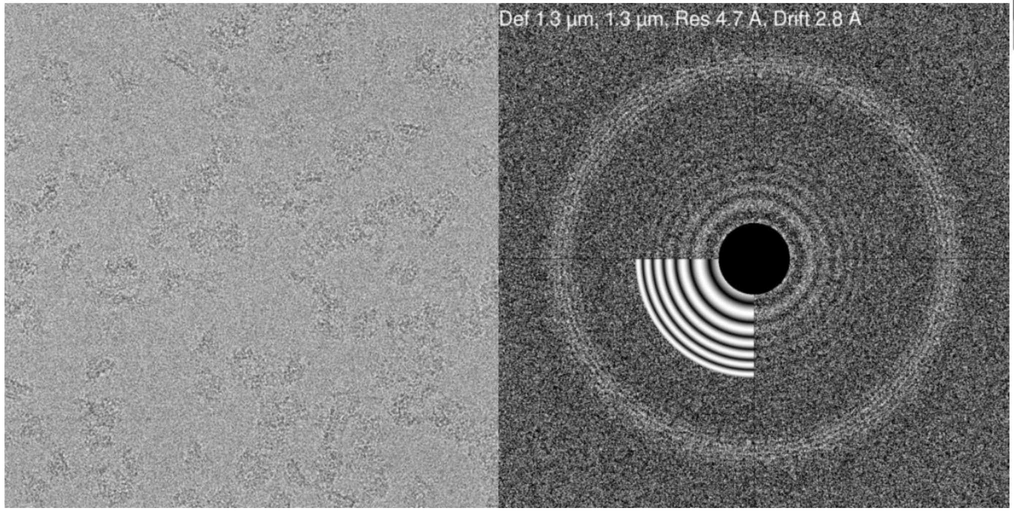
Object: convolution of different features

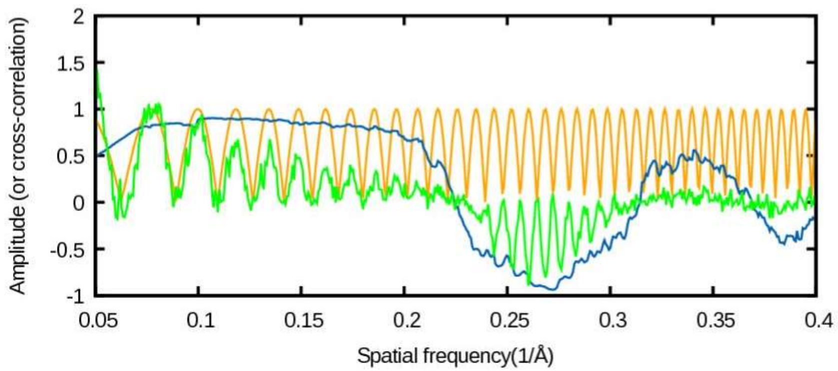
Nyquist limit: maximum frequency that can be observed considering the sampling frequency of the image

$$\nu_N = \frac{1}{2}\nu_{sampling} = \frac{1}{2}\cdot\frac{1}{px\ size}$$

Fourier transform of the object (power spectrum): decomposition of the different spatial frequencies of features forming the object

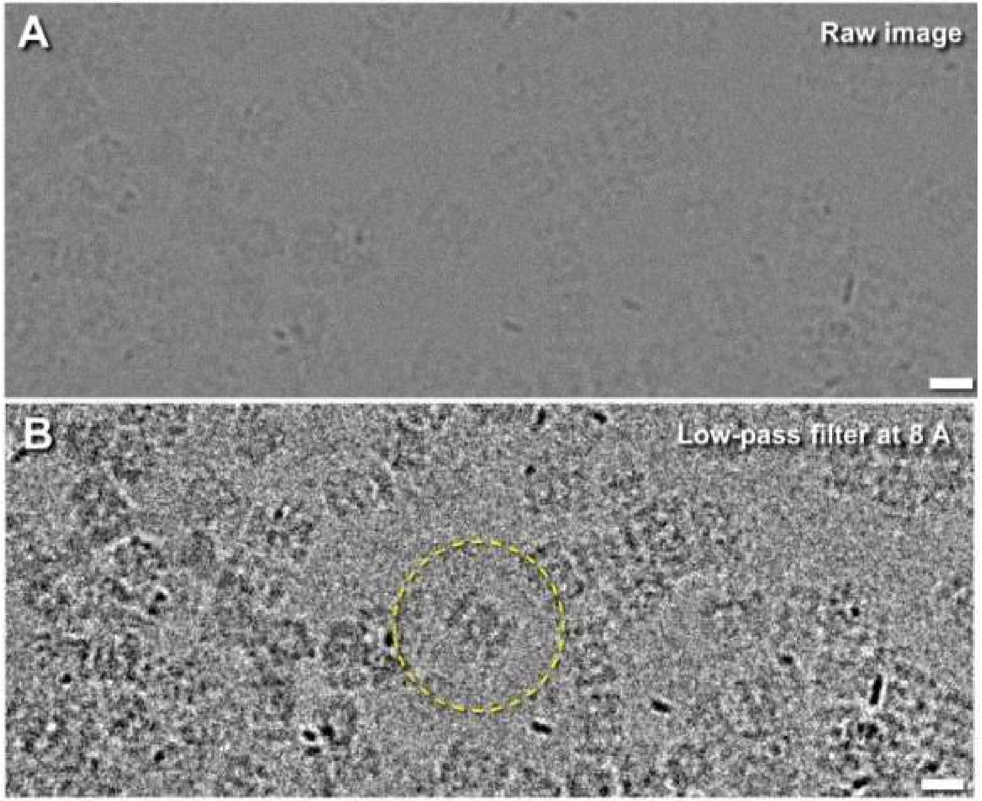




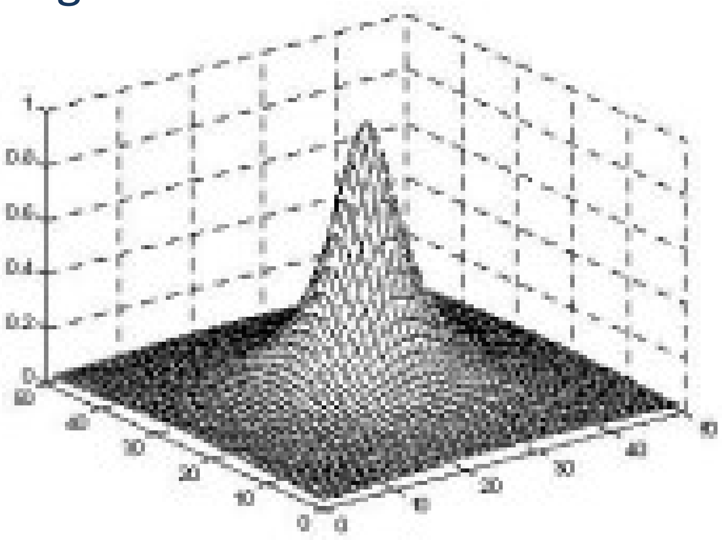


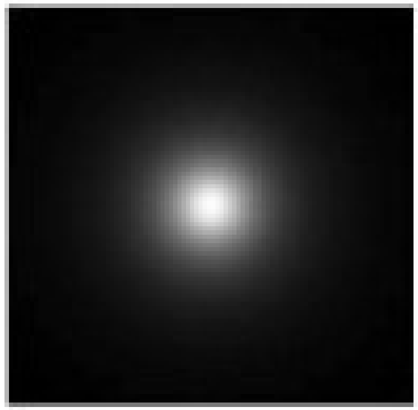
Signal-to-Noise Ratio (SNR):
variance of the signal divided by the variance of the noise

$$SNR = \frac{\int_B |O(\boldsymbol{s})|^2 \cdot |H(\boldsymbol{s})|^2 \, d\boldsymbol{s}}{\int_B |N(\boldsymbol{s})|^2 d\boldsymbol{s}}$$



To improve contrast, **low-pass filtering** of the micrograph:
in the frequency space, convolution of a Gaussian function with the signal





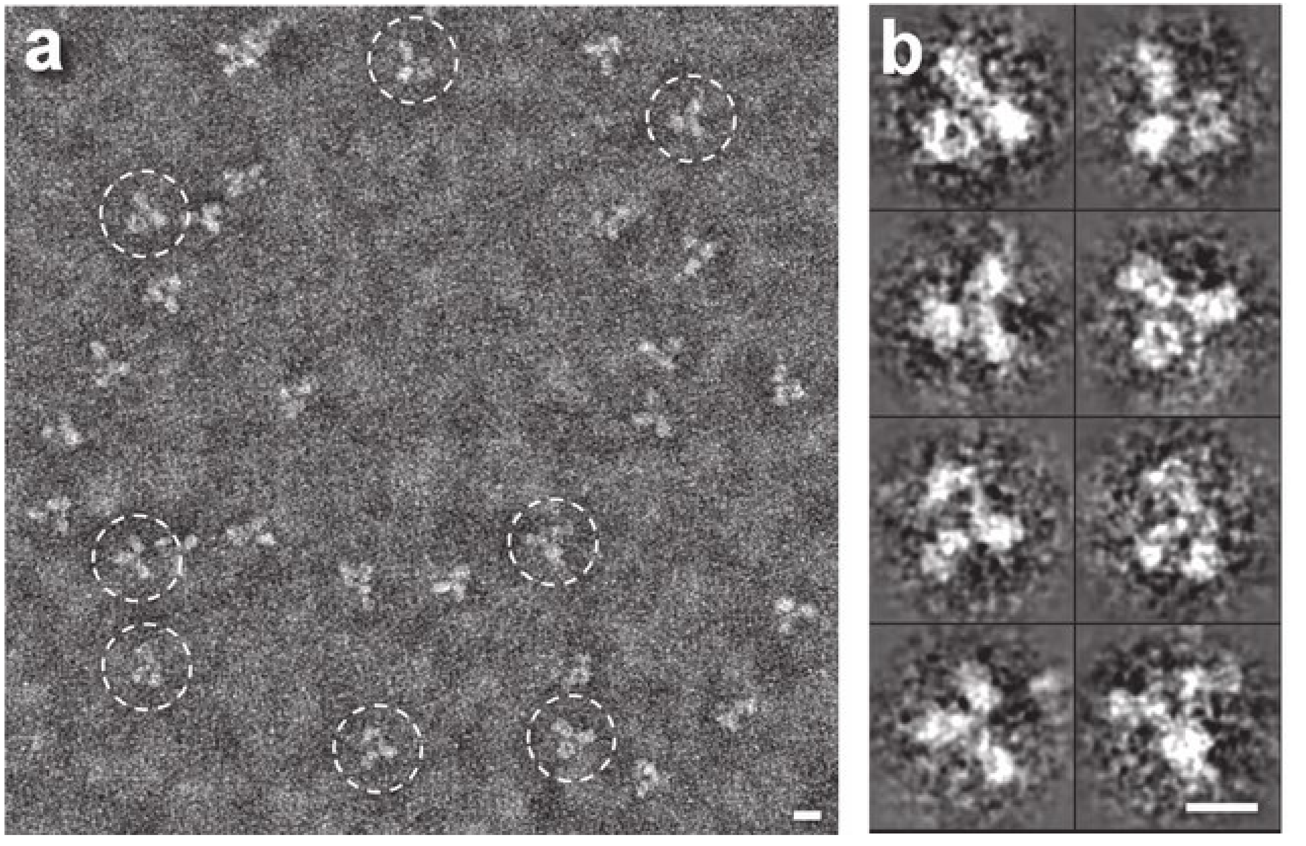
Loss of signal at high frequency, but improved SNR:

$$for\ B' < B$$

$$SNR_{B'} > SNR_B$$

1^{st} task: boxing and masking of particles

Identify positions of the particles and cut them out of the image.



Application of a mask to the particle to remove noise outside particle boundary.

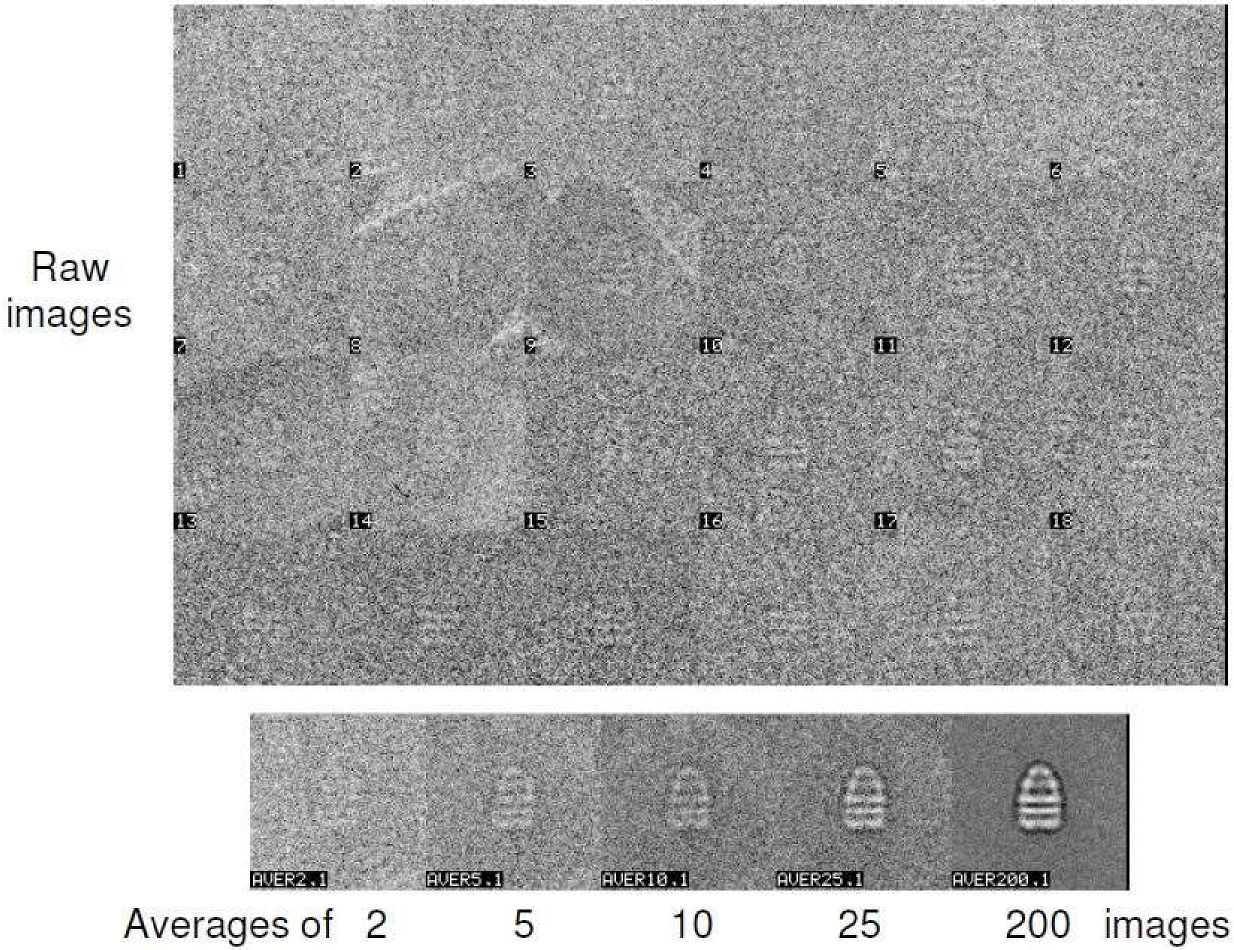
Mask with sharp borders introduces features at high resolution, due to sharpness of the mask border.

Optimal: **Gaussian mask**

Reduction of noise by averaging

Additive noise can be reduced by averaging multiple particles

$$I(\boldsymbol{r}) = O(\boldsymbol{r}) \otimes H(\boldsymbol{r}) + N(\boldsymbol{r})$$



Averaging: for each pixel (*i*) of the *N* images:

$$p_i = \frac{1}{N}\sum_{j=1}^{N} p_{ij}$$

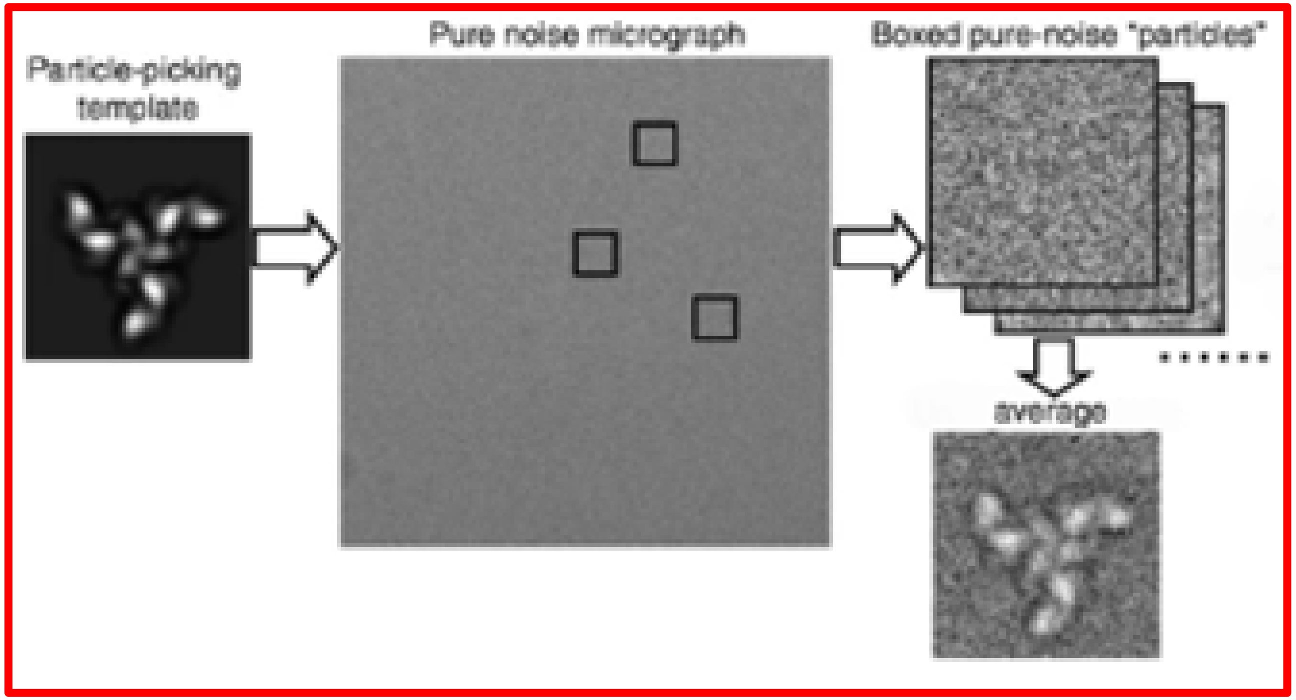
The improvement of the SNR due to averaging is proportional to $\sqrt{N}$.

But images have to be aligned otherwise averaging produces blurring of the image features.

Automated particle picking

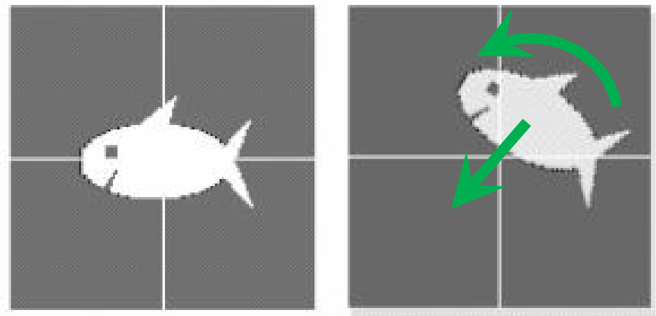
When very large datasets of images are available (*N > 1* million), automated particle picking is essential.

For automated particle picking, use of a template to recognize features of the particle in the micrograph.



To avoid pitfalls of automated particle picking, use as a template averages from alignment & classification on a smaller number of manually identified particles.

2nd task: alignment



Homogeneous images (same representation of the particle) have to be aligned before averaging.

E.g. negative staining images, with preferential orientation of the particle.

Alignment: search for transformations that bring each particle in register with reference

$$I'(\boldsymbol{r}_j) = \boldsymbol{R}I(\boldsymbol{r}_j) + \boldsymbol{t}$$

For a 2D image, the rotation matrix is just one rotation angle (α), the translation vector has 2 components (t_x, t_y).

To obtain optimal α and $\boldsymbol{t}$, minimize Euclidean distance between two images:

$$E_{12}(\alpha, \boldsymbol{t}) = \sum_j^J \left[I_1(\boldsymbol{r}_j) - I_2(\alpha \boldsymbol{r}_j + \boldsymbol{t})\right]^2$$

$$= \underbrace{\sum_j^J I_1(\boldsymbol{r}_j)^2 + \sum_j^J I_2(\alpha \boldsymbol{r}_j + \boldsymbol{t})^2}_{\text{Invariant with respect to transformations}} - 2\boxed{\sum_j^J I_1(\boldsymbol{r}_j)\, I_2(\alpha \boldsymbol{r}_j + \boldsymbol{t})}$$

Maximize cross-correlation coefficient, after normalization

Invariant with respect to transformations

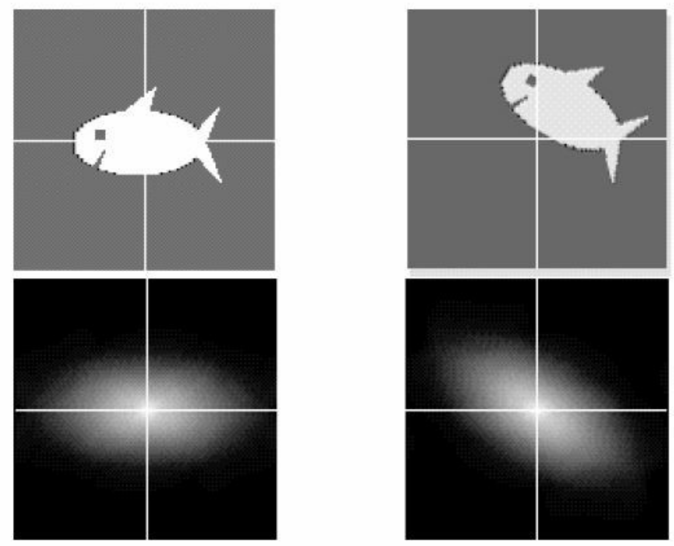
Image is discretized and interpolation is required.
Interpolation causes reduction of resolution.

Rotation-translation procedure

1. **Rotation:** Calculate autocorrelation function of each image (insensitive to translation) and compare them using cross-correlation

$$ACF_i(\boldsymbol{t}) = \sum_j^J I_i(\boldsymbol{r}_j) \cdot I_i^*(\boldsymbol{r}_j - \boldsymbol{t})$$

$$CC_{12}(\alpha) = \sum_k ACF_1(\boldsymbol{t}_k)\, ACF_2(\boldsymbol{t}_k, \alpha)$$



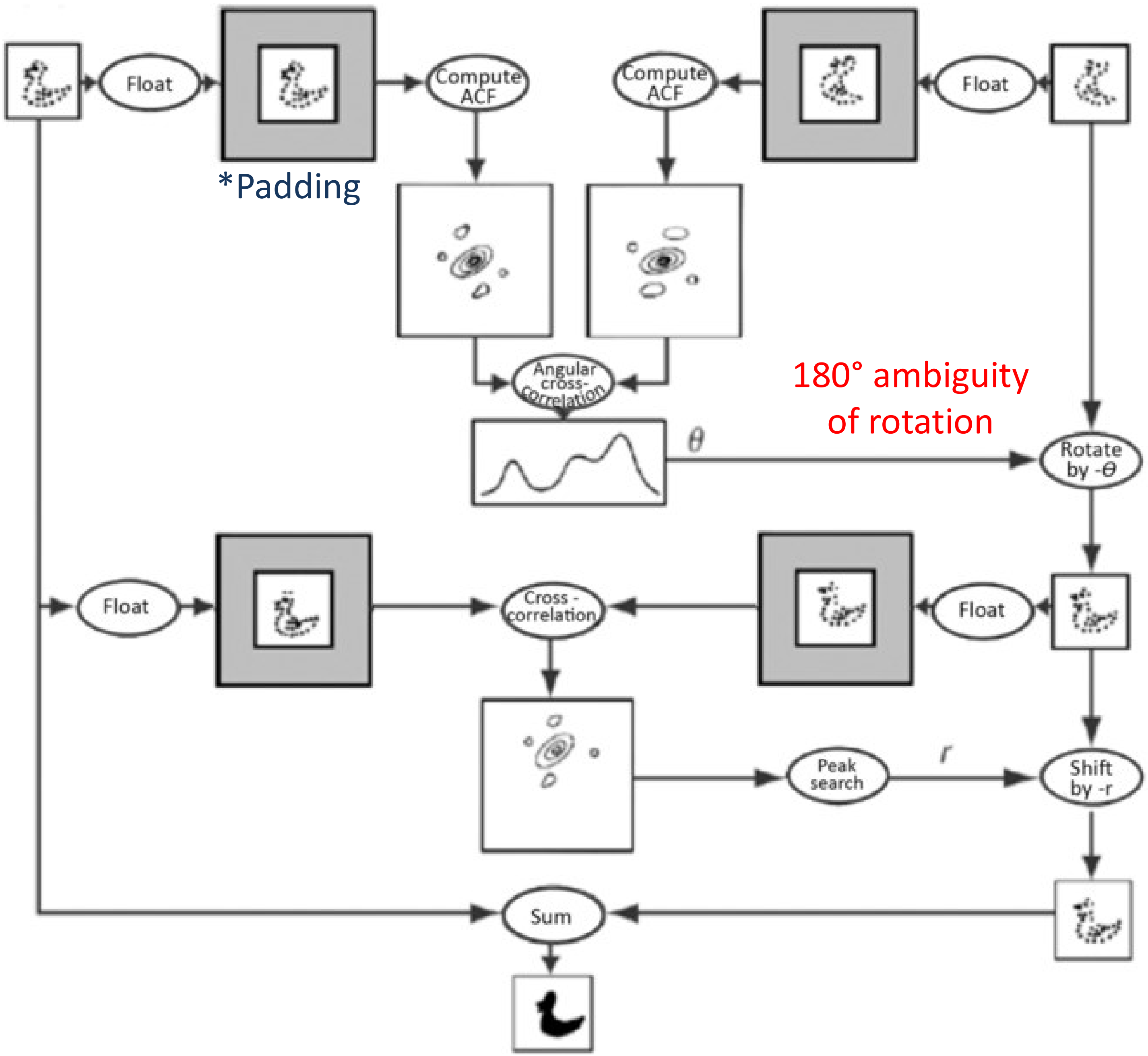
2. **Translation:** Maximize cross-correlation function to obtain $\boldsymbol{t}$

$$CC_{12}(\boldsymbol{t}) = \frac{\sum_j^J [I_1(\boldsymbol{r}_j) - \langle I_1 \rangle][I_2(\alpha \boldsymbol{r}_j + \boldsymbol{t}) - \langle I_2 \rangle]}{\sum_j^J [I_1(\boldsymbol{r}_j) - \langle I_1 \rangle]^2 \sum_j^J [I_2(\alpha \boldsymbol{r}_j + \boldsymbol{t}) - \langle I_2 \rangle]^2}$$

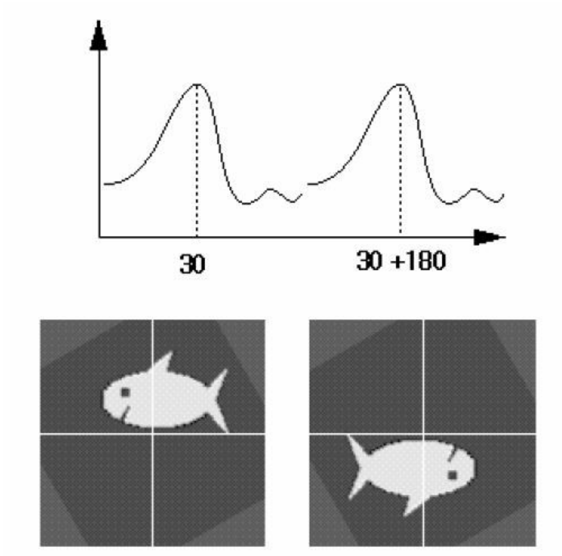
Usually performed in the Fourier space: $\phi_{12}(\boldsymbol{t}) = FT^{-1}\{FT[I_1(\boldsymbol{r})]FT[I_2(\boldsymbol{r} + \boldsymbol{t})]^*\}$

Rotation-translation procedure

Often performed through iterative algorithms

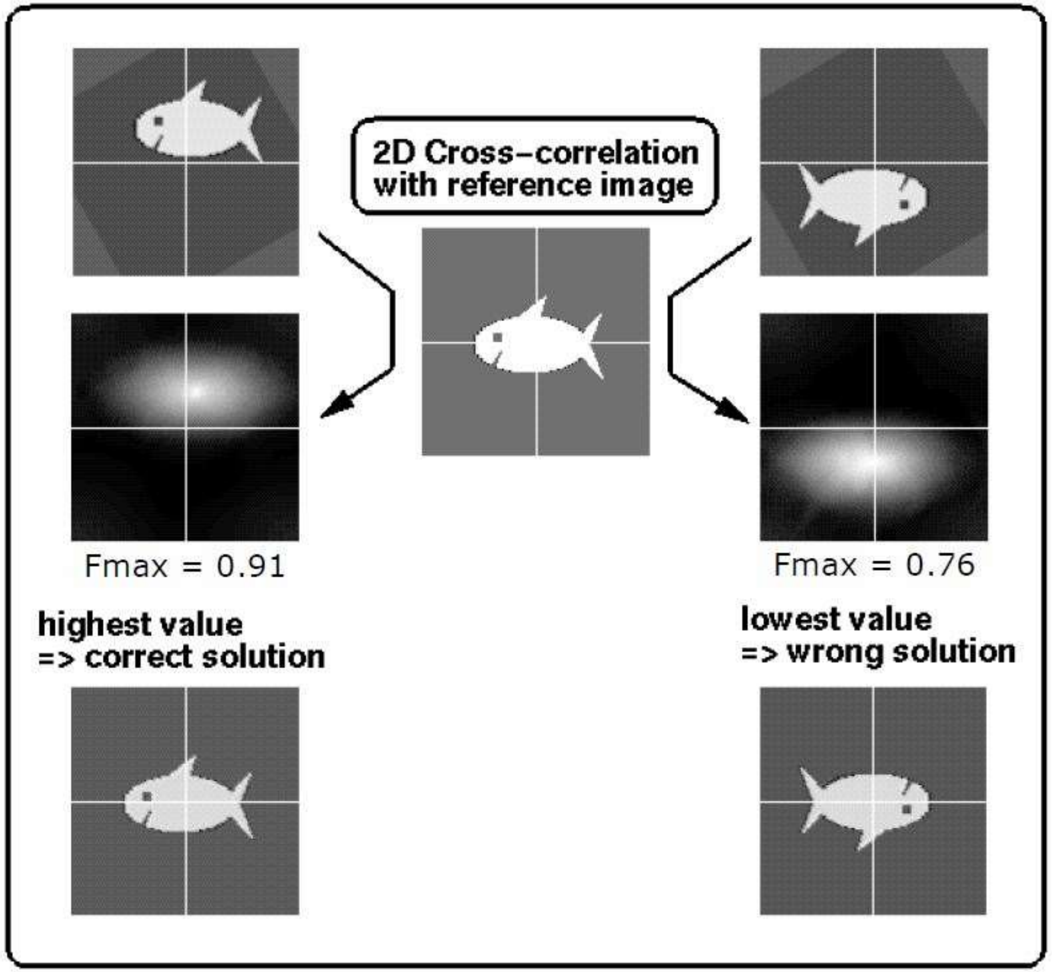


*Padding: to overcome reduction of the space due to interpolation and translation



Autocorrelation function is centrosymmetric: 180° ambiguity of rotation

To solve ambiguity, cross-correlation of the two possible solutions with the original image: the highest cross-correlation value yields the correct solution.

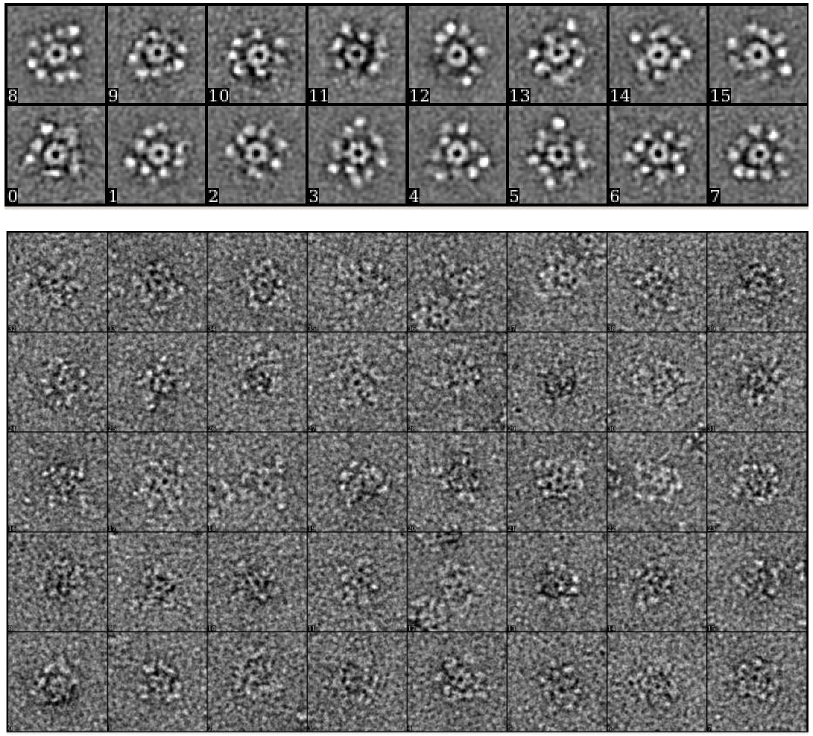


Heterogeneity

Heterogeneous images:

- cryo images showing different orientations of the particle
- conformationally heterogeneous negative staining particles
- presence of ligands or protein components in a fraction of the particles
- presence of contaminants

Classification is required in parallel with alignment.



Ca^{2+}/calmodulin-dependent protein kinase II
~8000 particles

Multi-reference alignment: L references for N images.
In the first run, each image is assigned to the correct reference using cross-correlation function. Assignment may change while alignment is improved.

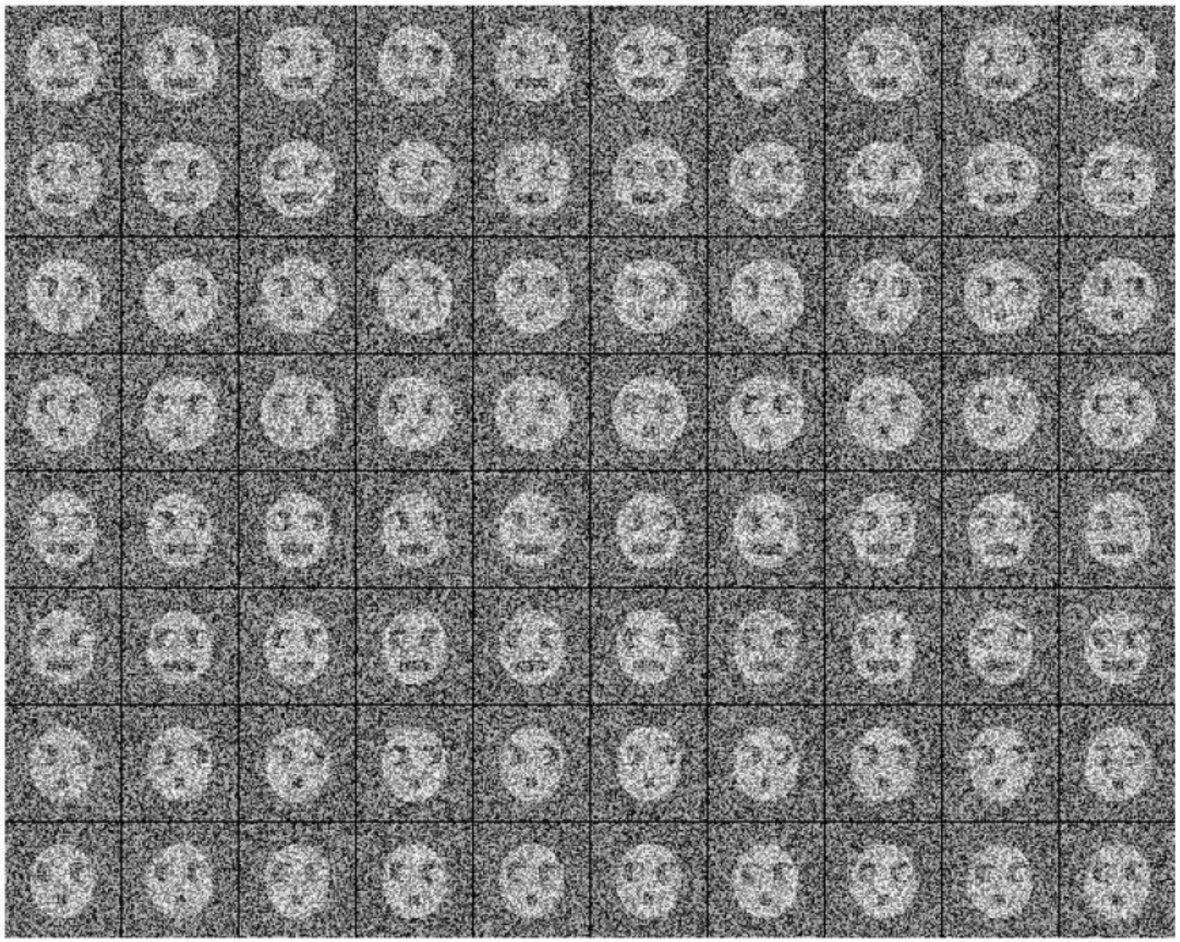
Reference-free alignment with the use of invariants.
Use of the Double Auto-Correlation Function (DACF) to avoid model bias. DACF is insensitive to both translation and rotation. Requires iterations.

3rd task: classification

Classification: Create homogeneous sets of images that can be averaged to improve SNR

Test case:
faces, with
(a) different mouth (large/small),
(b) eyes in opposite directions (left/right),
(c) different shape (round/oval)

Tasks:
(1) Identify how many different face types are present
(2) assign each image to the right class



Principal Component Analysis (PCA)

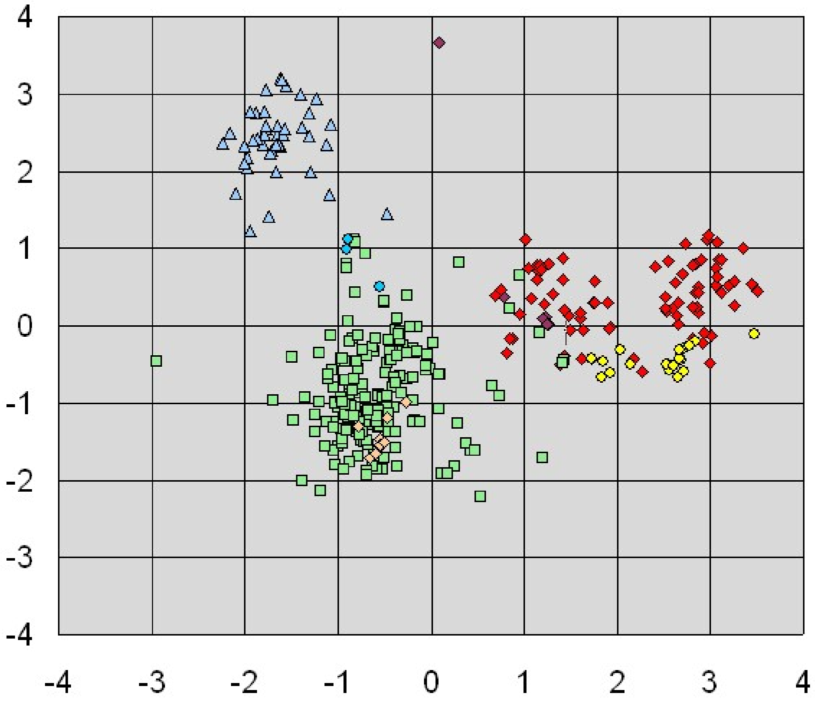
For images of $N \; x \; N$ pixel, each image is defined as a point (vector) in the hyperspace of N^2 dimensions:

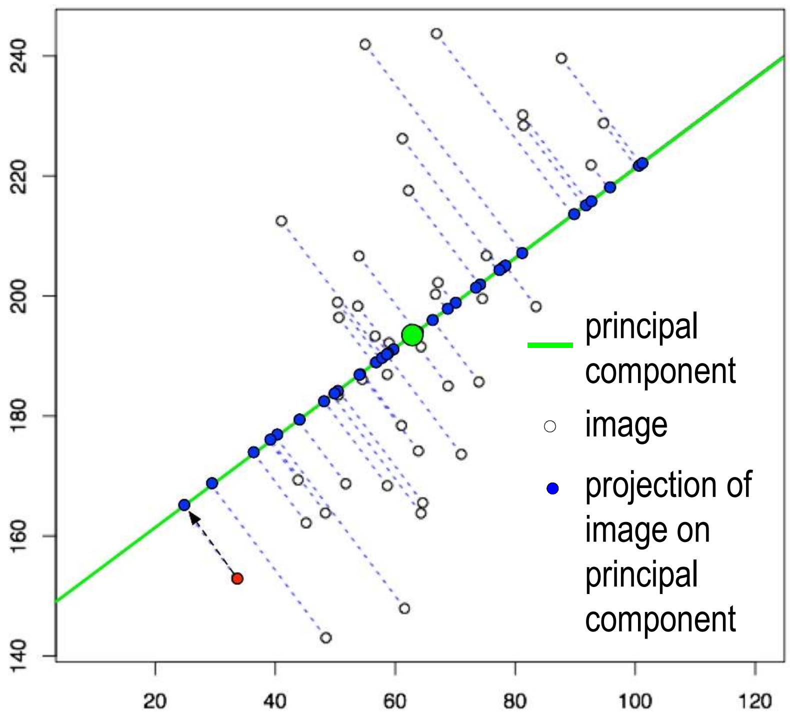
$$I(px_1, px_2, px_3, \dots px_{NxN})$$

To compare images, calculate distance between points representing the images:

$$\overline{I_1 I_2} = \sqrt{\sum_{n}^{NxN} (px_{n,1} - px_{n,2})^2}$$

Similar images form clouds in the hyperspace (their vectors are close, have small distances)





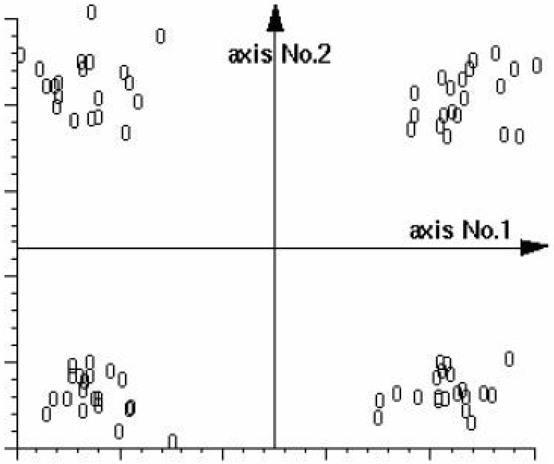
Objective of PCA is to identify independent directions of maximum extension of the clouds by a least-squares minimization

For the test case, 3 independent directions of maximum extension are obtained from analysis (highest variability of the cloud), corresponding to the different conformations present in the population.
[Identifing the number of different conformations is often non trivial...]

3 different features

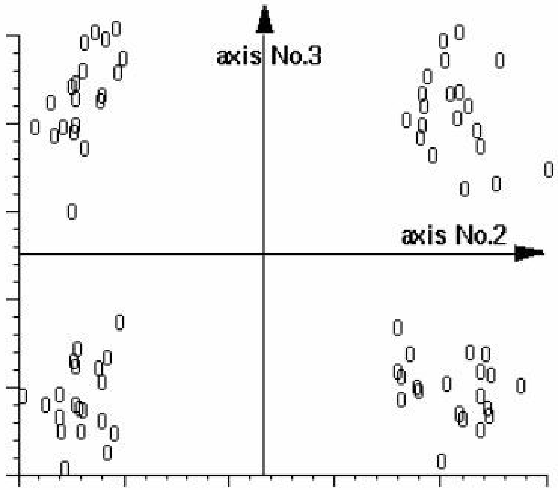
Independent directions are compared:

(2) eyes direction



(1) face shape

(3) mouth



In the 3-dimensional space of the 3 principal components, a total of 8 groups can be distinguished

8 different conformations, 8 classes

2D classification methods

Hard classification

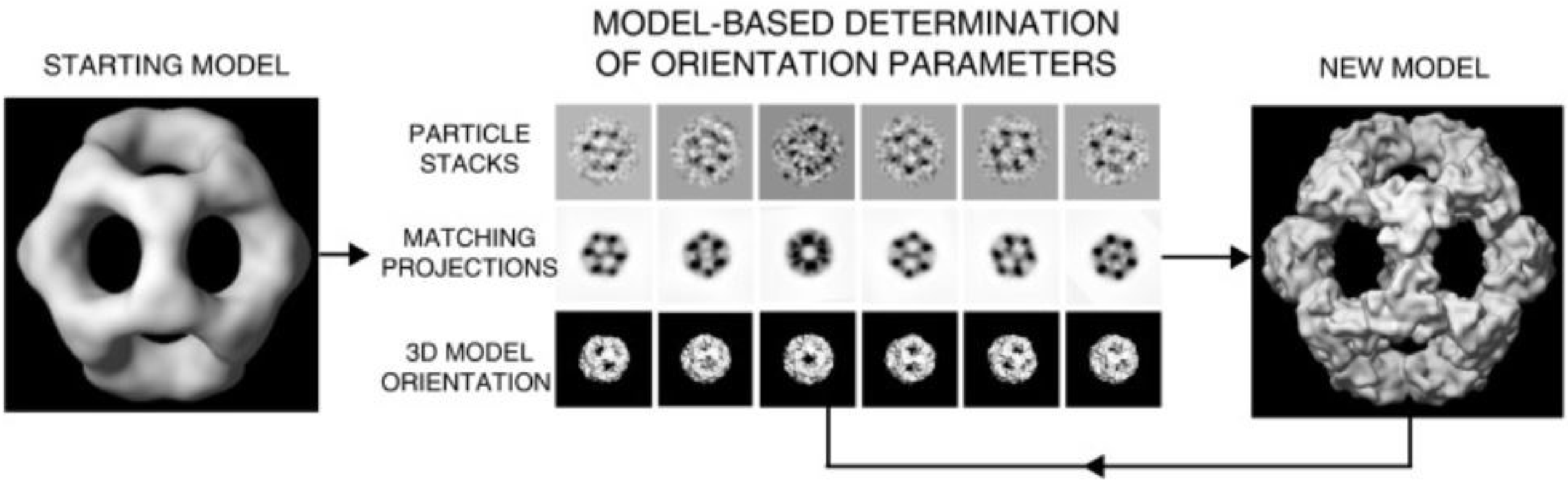
Each image is assigned to a single class. Class might be changed during iterative refinement.

Fuzzy classification

For each image, a coefficient is determined for each class, representing the contribution of the classes to the image. Particularly useful when probability distributions are used in classification (Bayesian approach).

Supervised classification: Uses templates to classify images, i.e. assign each particle to a class (or to determine coefficients for fuzzy classification).

Used also for reference-based orientation determination in 3D reconstruction methods. Affected by model bias.

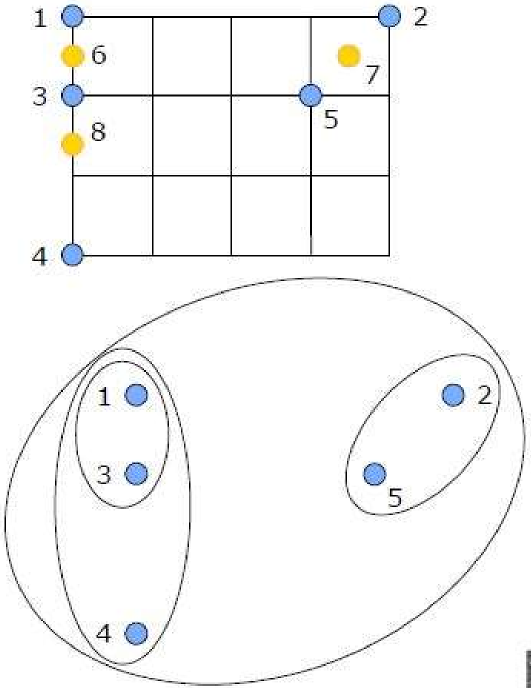


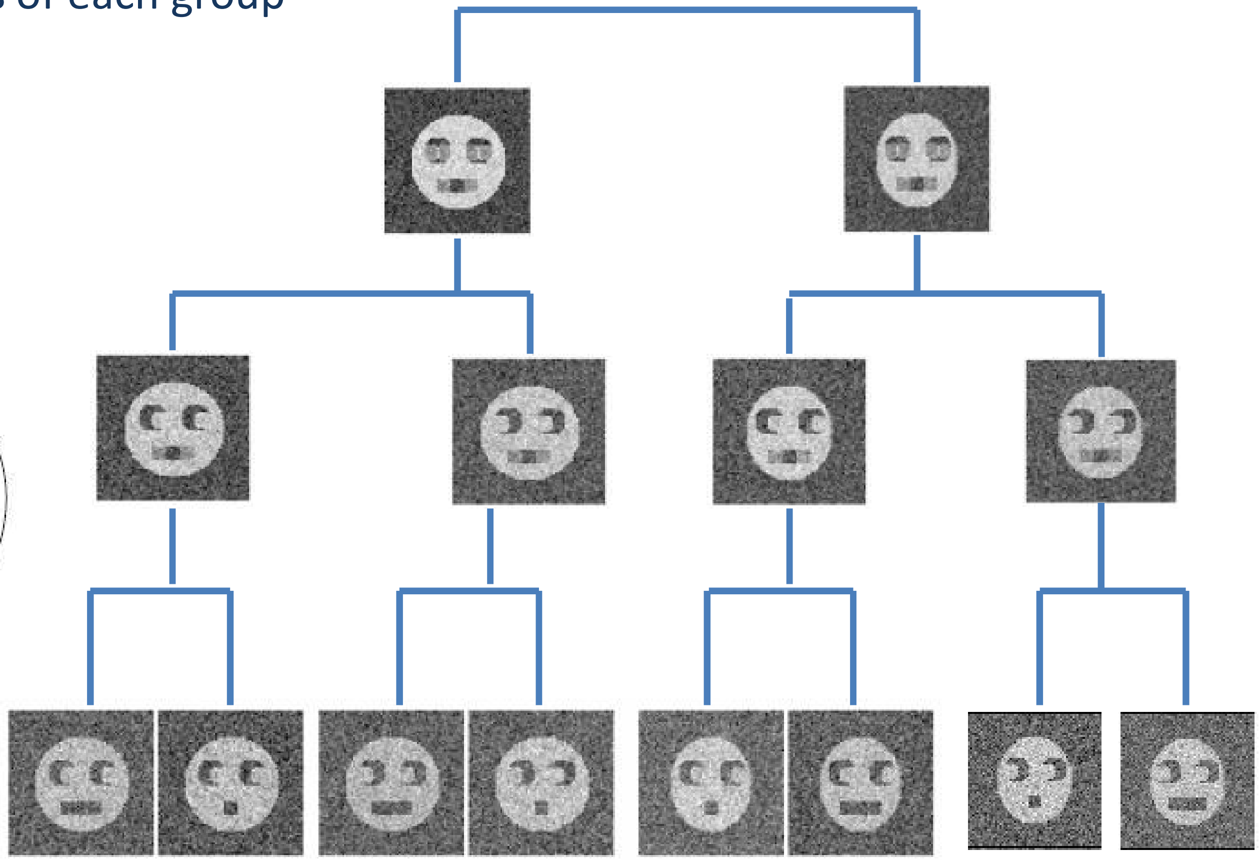
Unsupervised classification: No template is used; images are divided in groups according to statistical evaluation of their distance. Uses Principal Component Analysis.

No model bias!!

Hierarchical Ascendant Classification

Based on PCA, a dendrogram is obtained analyzing distances between averages of each group

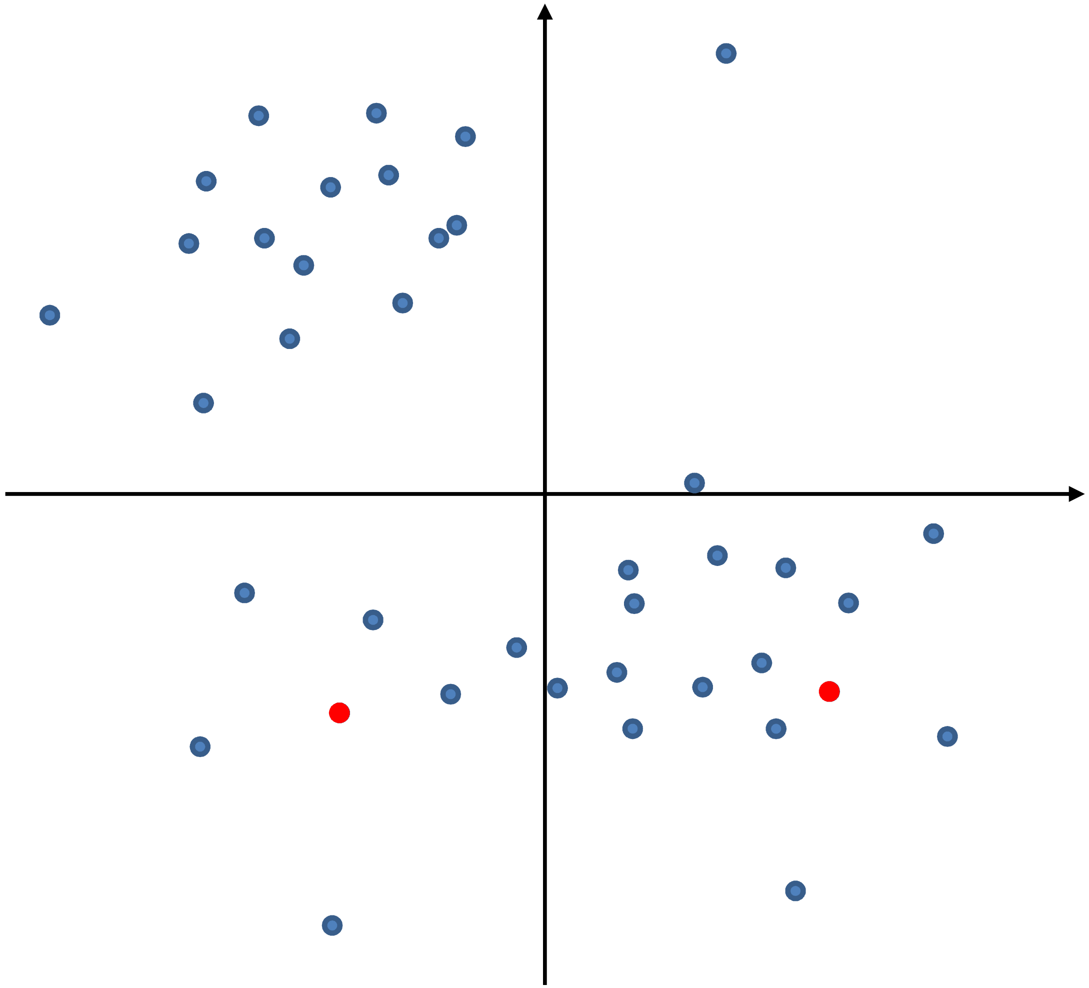




K-means Classification

Considering results of PCA, a number of classes K is set at the beginning of the classification

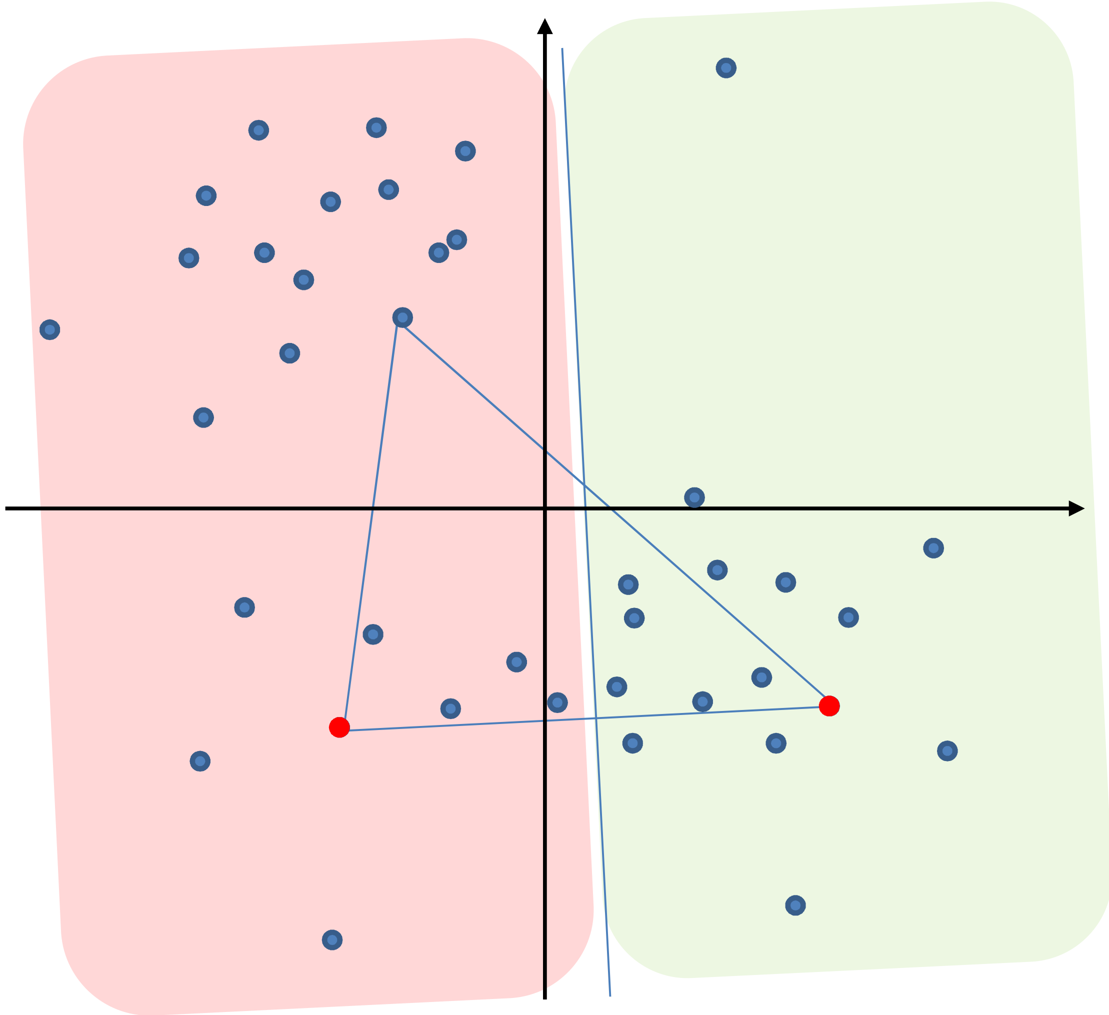
1) Chose K random “seeds” (one for each class)



K-means Classification

Considering results of PCA, a number of classes K is set at the beginning of the classification

1) Chose K random “seeds” (one for each class)
2) Each image is assigned to the class of the closest seed (in hyperspace)



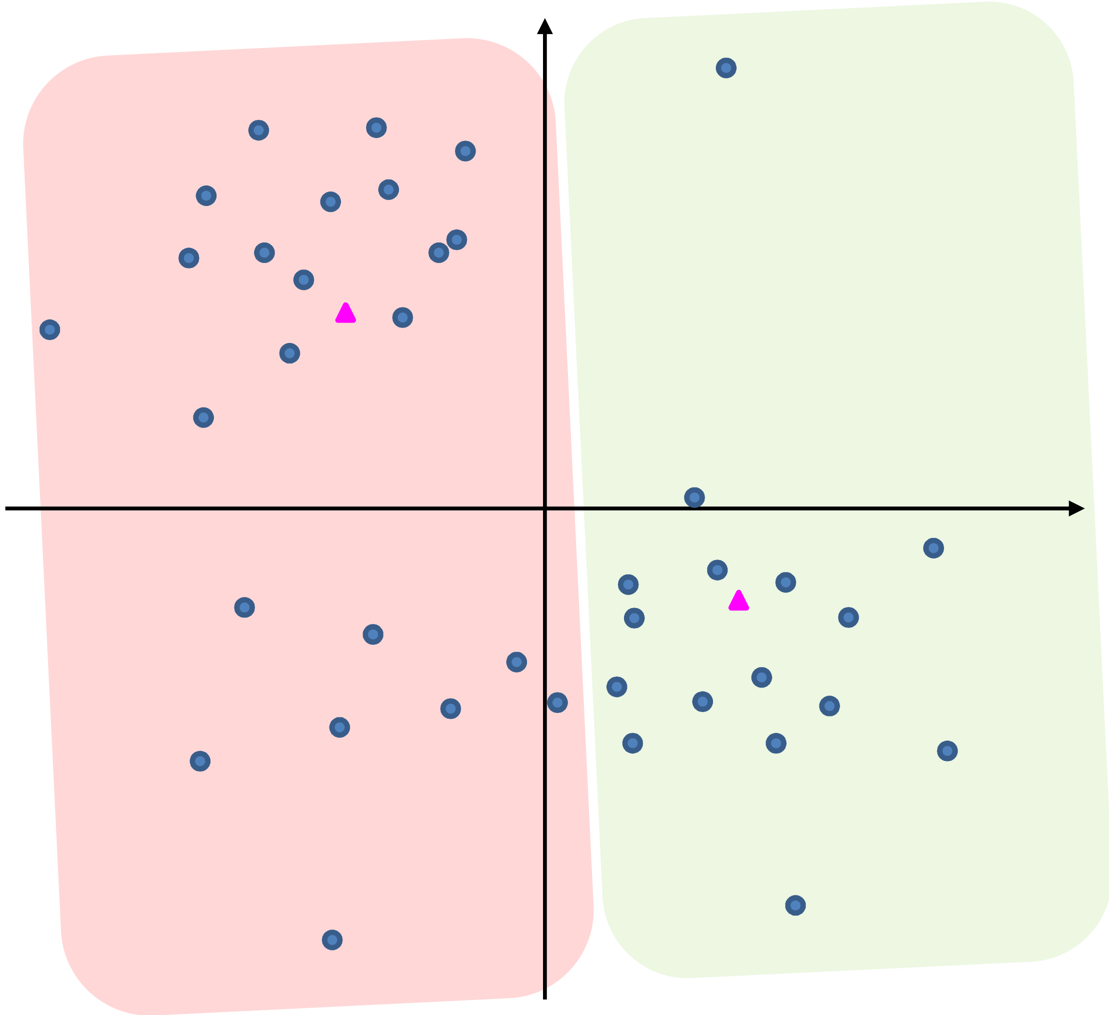
K-means Classification

Considering results of PCA, a number of classes K is set at the beginning of the classification

1) Chose K random “seeds” (one for each class)

2) Each image is assigned to the class of the closest seed (in hyperspace)

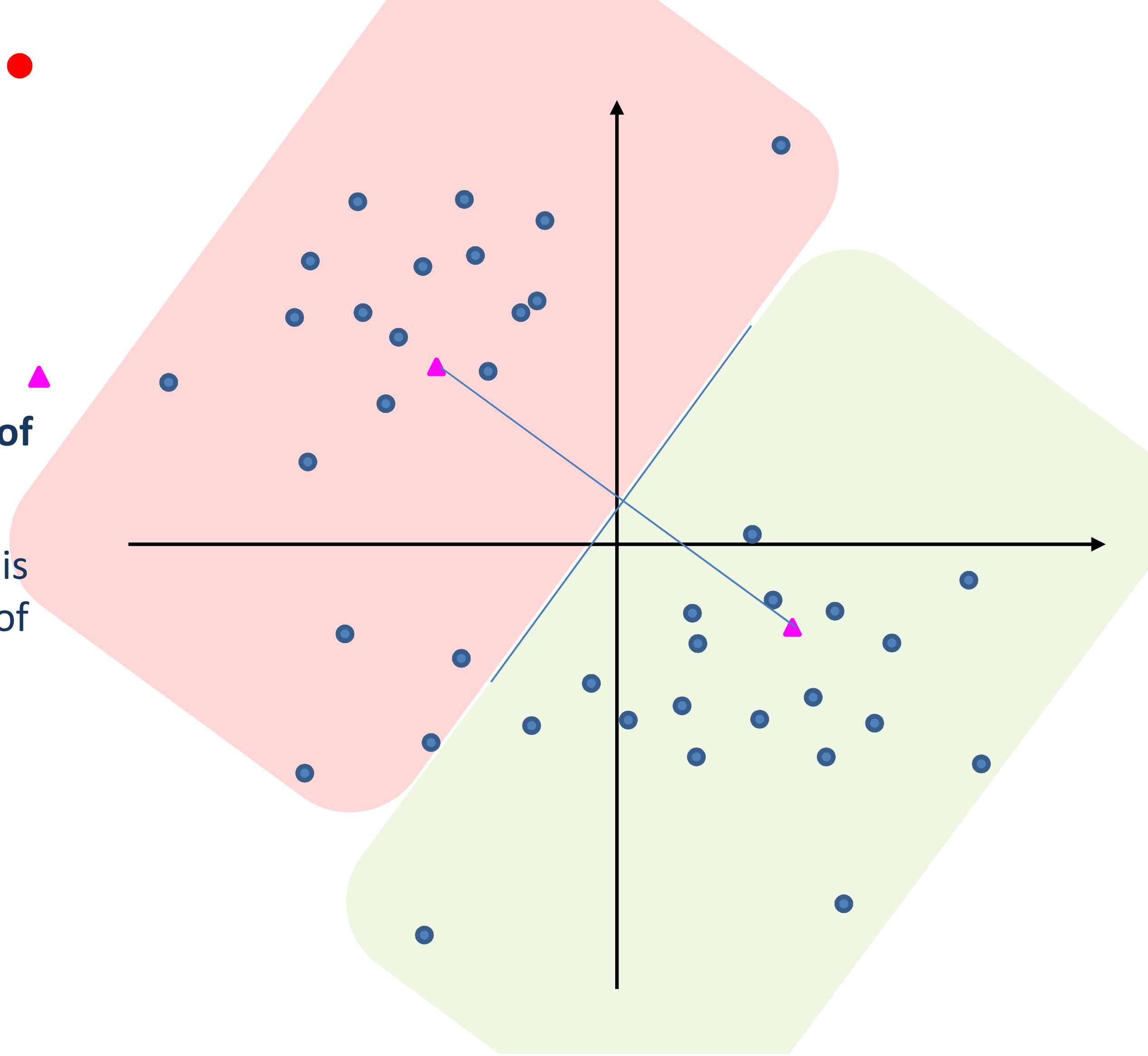
3) For each class, new seeds are calculated as **centers of gravity** of the whole class



K-means Classification

Considering results of PCA, a number of classes K is set at the beginning of the classification

1) Chose K random “seeds” (one for each class)
2) Each image is assigned to the class of the closest seed (in hyperspace)
3) For each class, new seeds are calculated as **centers of gravity** of the whole class
4) Iterate until classification is stable (no more changes of images between classes)



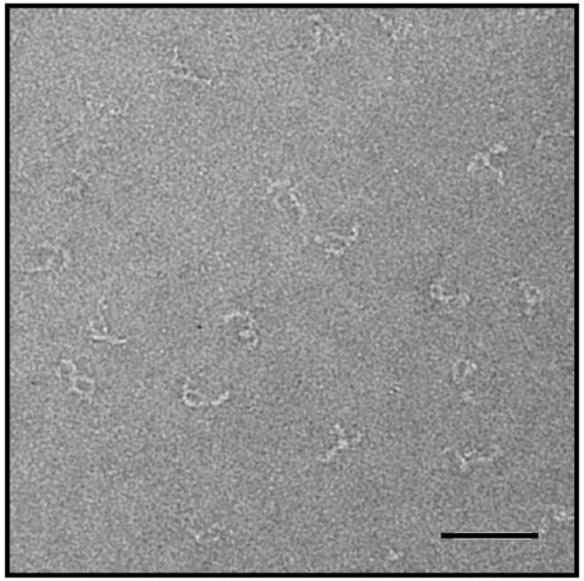
* This method has the tendency to yield spherical classes, while elongated classes are usually not identified

Negative staining

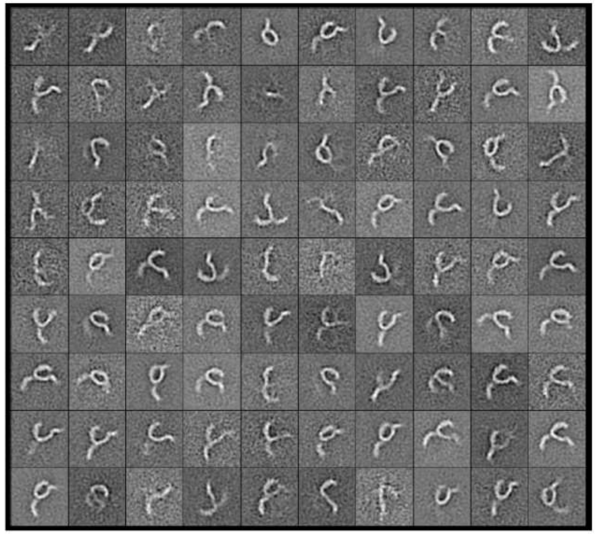
Conserved Oligomeric Golgi complex (COG): complex of 4 subunits (Cog1, Cog2, Cog3 and Cog4)

Walz group
(HMS - Boston)
2010

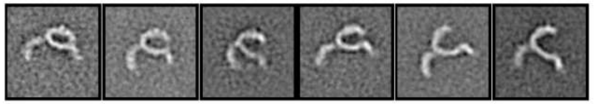
Cog1-4 sub-complex of COG

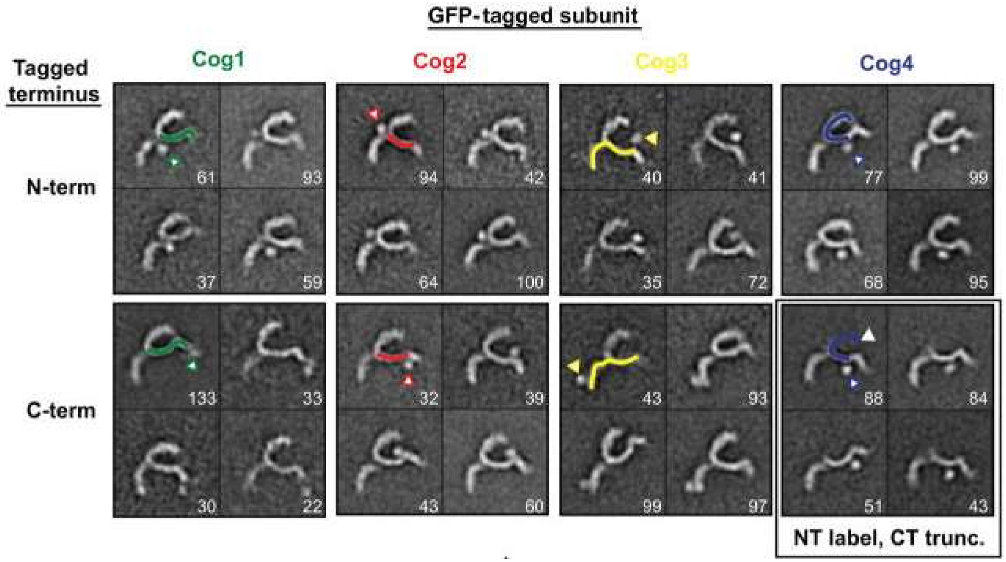


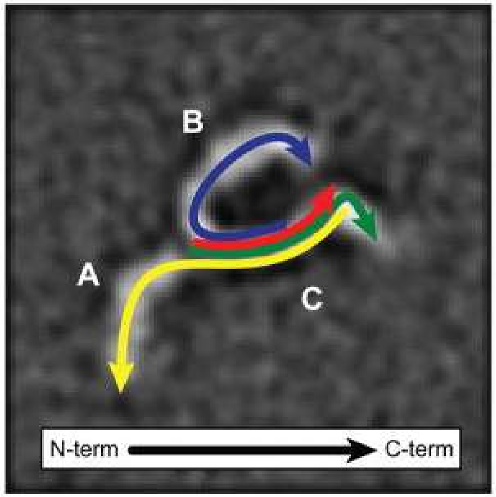
raw image (negative stain)



class averages

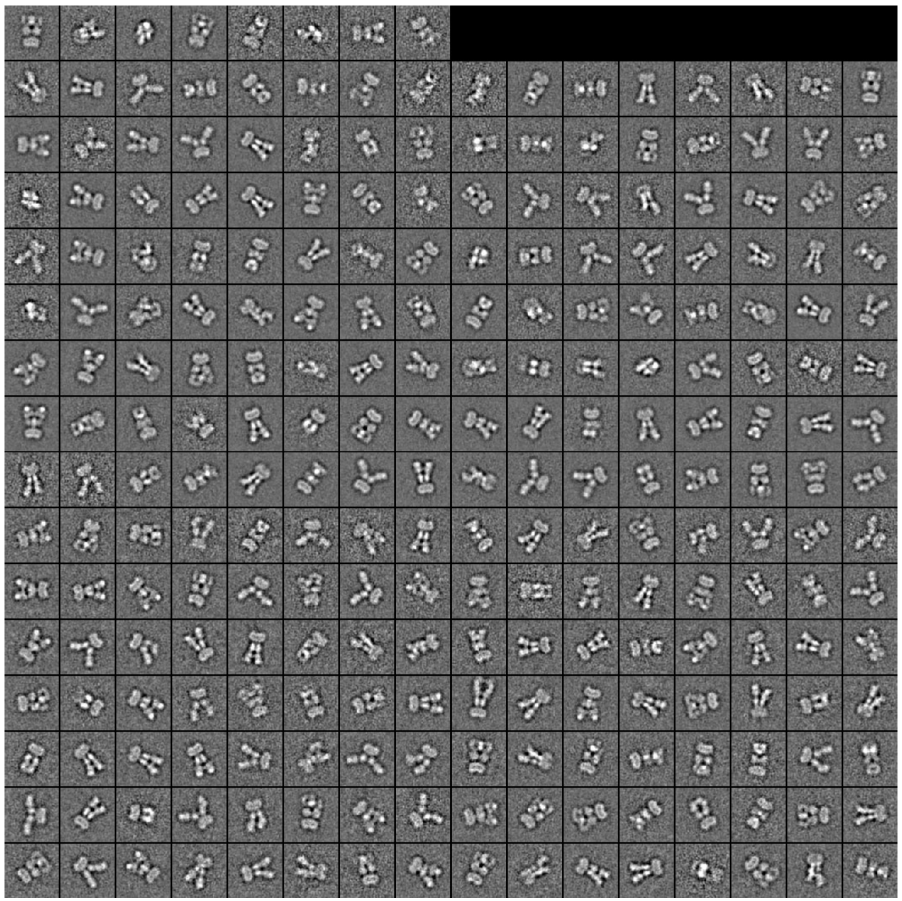


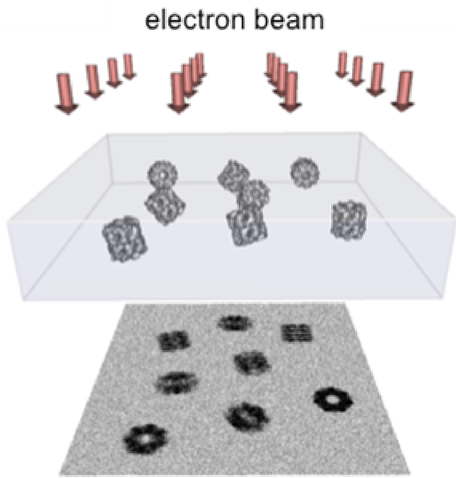


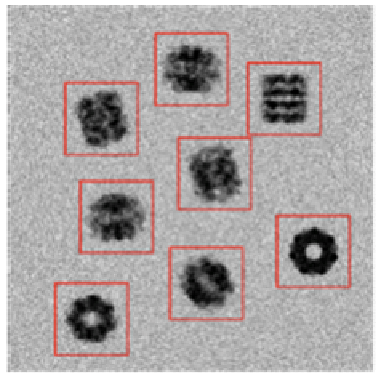


Cryo EM

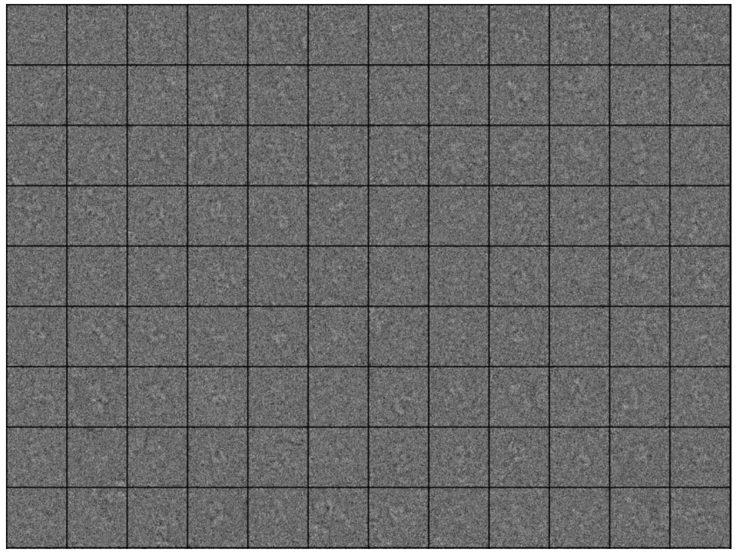
AMPA Receptor images (F20, 200kV, DDD camera)







1. Particle picking
2. Normalization and CTF correction



3. Class averaging showing large conformational heterogeneity