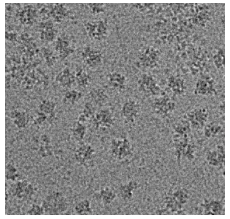
Methods in orientation determination

K.R.Vinothkumar
MRc Laboratory of Molecular Biology

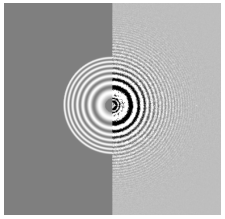
MRC Laboratory of Molecular Biology

LMB cryo EM course

Image processing workflow

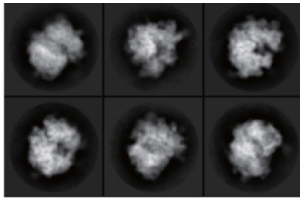


Micrographs



CTF estimation

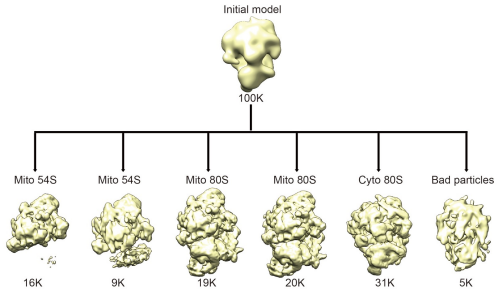
Particle picking



2D class averages

Initial model

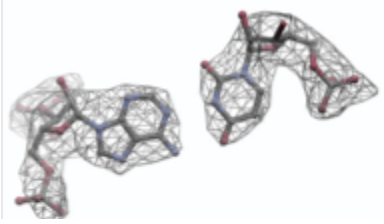
3D refinement



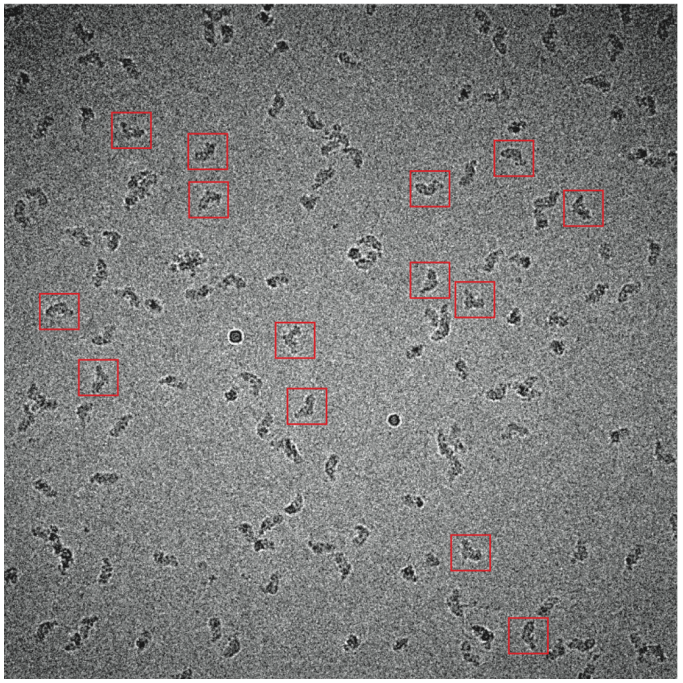
3D classification

B-factor sharpening

Map interpretation

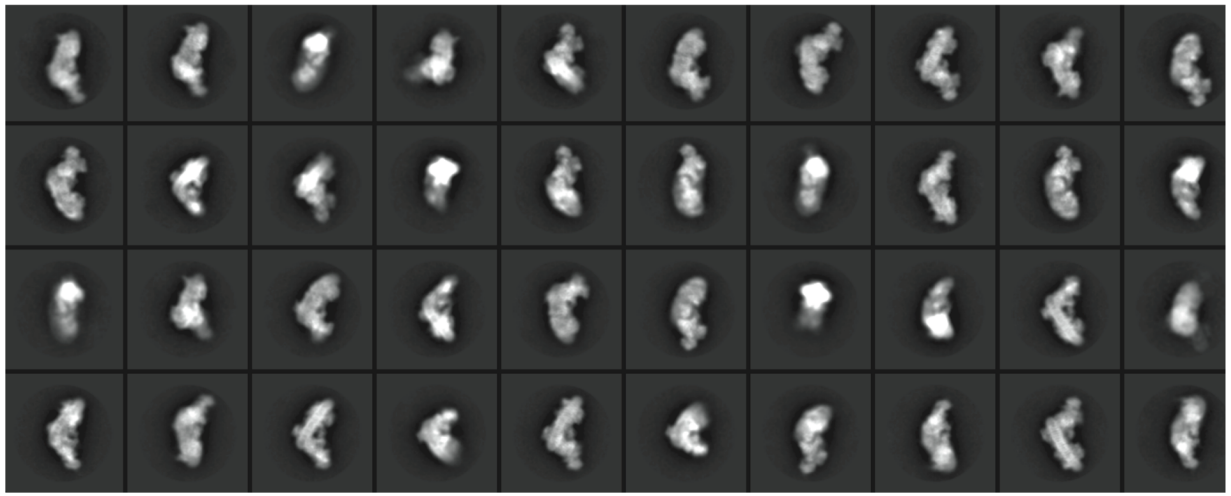


Average to increase SNR!

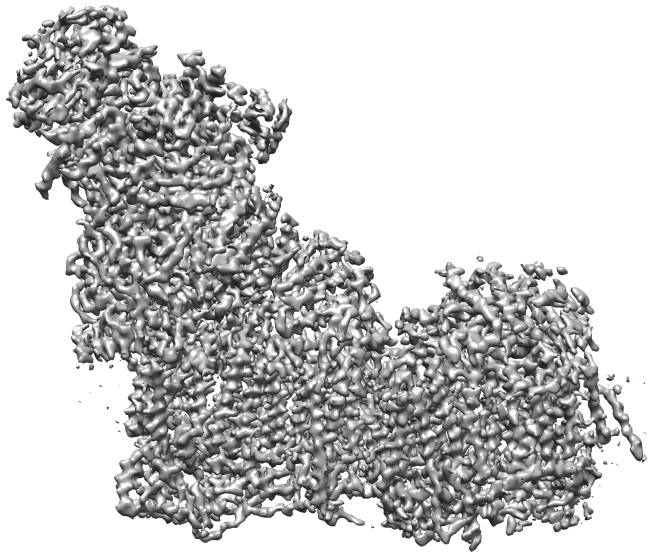


Micrograph

Reference free 2D class average

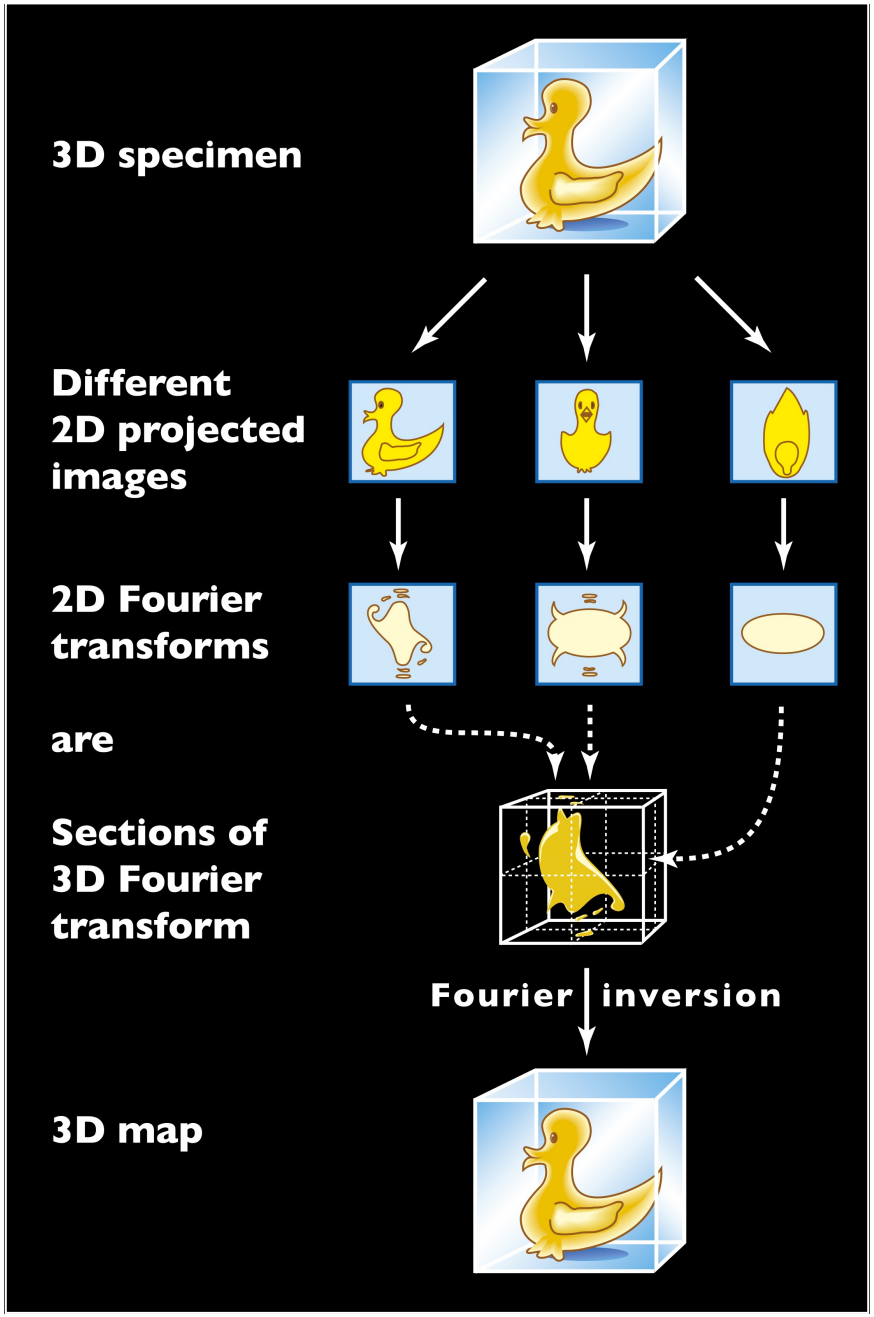


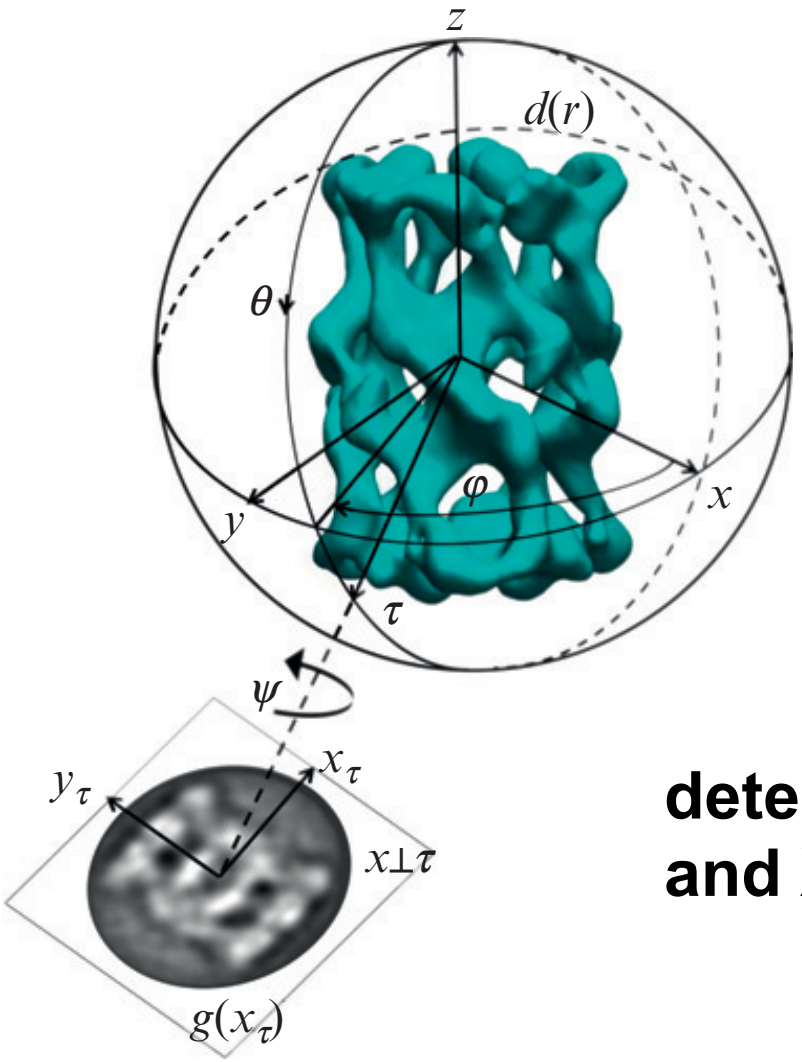
orientation determination, refinement
reconstruction



3D map

Projection slice theorem

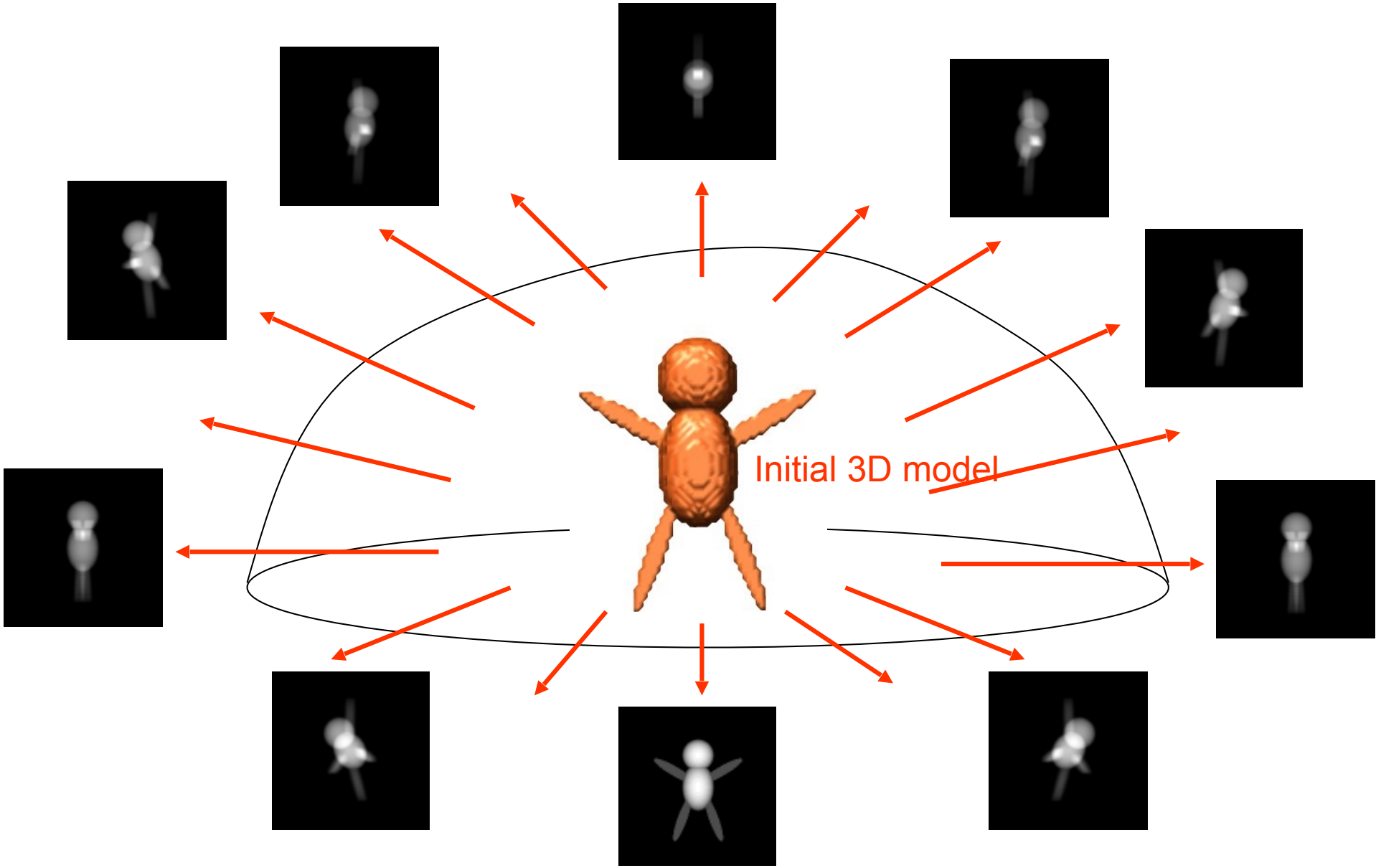




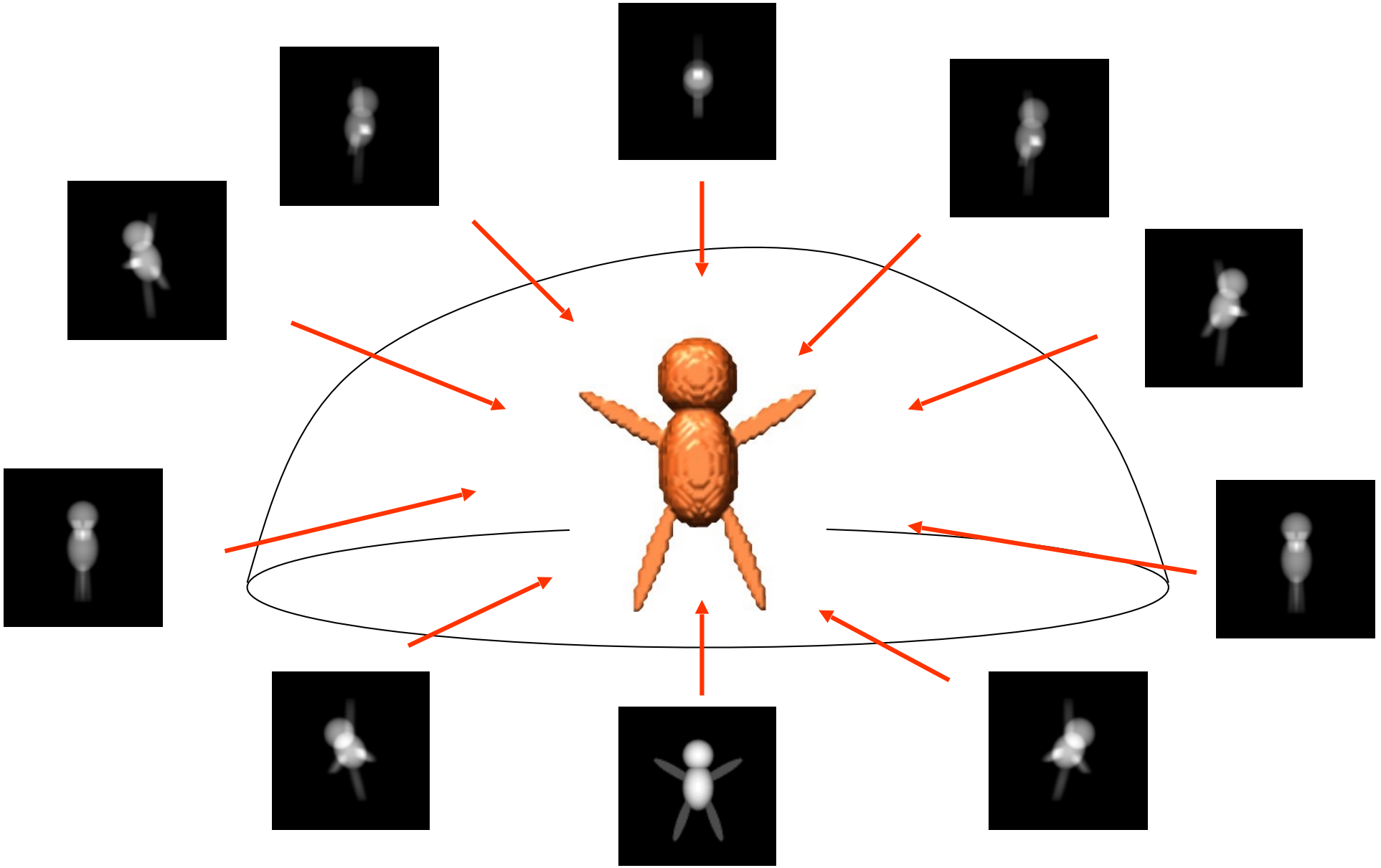
determine 3 euler angles and X+Y shifts

Penczek 2010

Projection matching



3D Reconstruction



Cryo-EM images - reality

- Electrons damage biological material
  - Low dose: large amounts of noise!

- We need to defocus to get contrast
  - Strong artefacts (CTF) [a good phase plate will solve this]

- We can't control how the particles fall on the grid
  - Unknown orientations & classes

Inverse, Incomplete, Ill-posed problems

Inverse problem

The forward model (in real-space)

$$X_i = \mathrm{CTF}_i \otimes \mathbf{P}_{\varphi} V_k + N_i$$

Given V, φ and CTF, we can simulate X very well.

SPEMS - Single particle electron microscopy simulation (Greg McMullan unpublished)

But the opposite is very difficult.

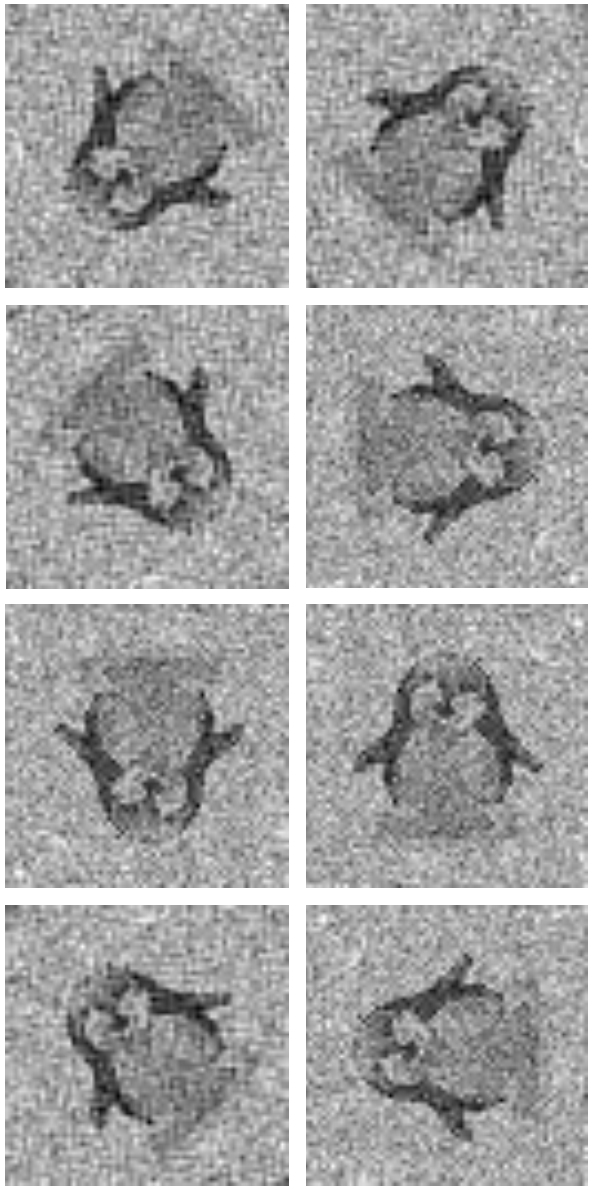
Incomplete data problems

- Part of the data was not observed experimentally
  - Orientations
  - Class assignments

- Difficult to solve!
  - Iterative methods?

- Complete data problem would be very easy to solve

Incomplete data problems

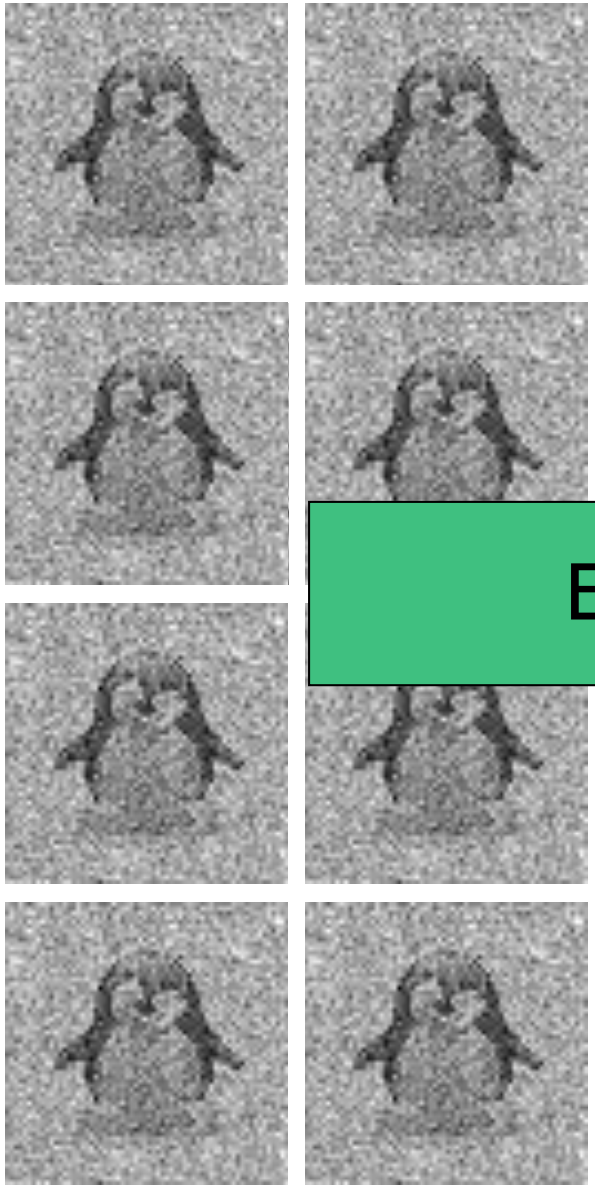


Not easy

Observed data (*X*): images

Missing data (Y): orientations

Complete data problems



white Gaussian noise

$$L(\Theta) = P(X \mid \Theta)$$

Easy

$$A^{MLE} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Observed data (*X*): images

Incomplete data problems

- Option 1: add *Y* to the model → Maximum cross-correlation / least-squares

$$L(Y,\Theta) = P(X \mid Y,\Theta)$$

- Option 2: marginalize over *Y* → Maximum Likelihood

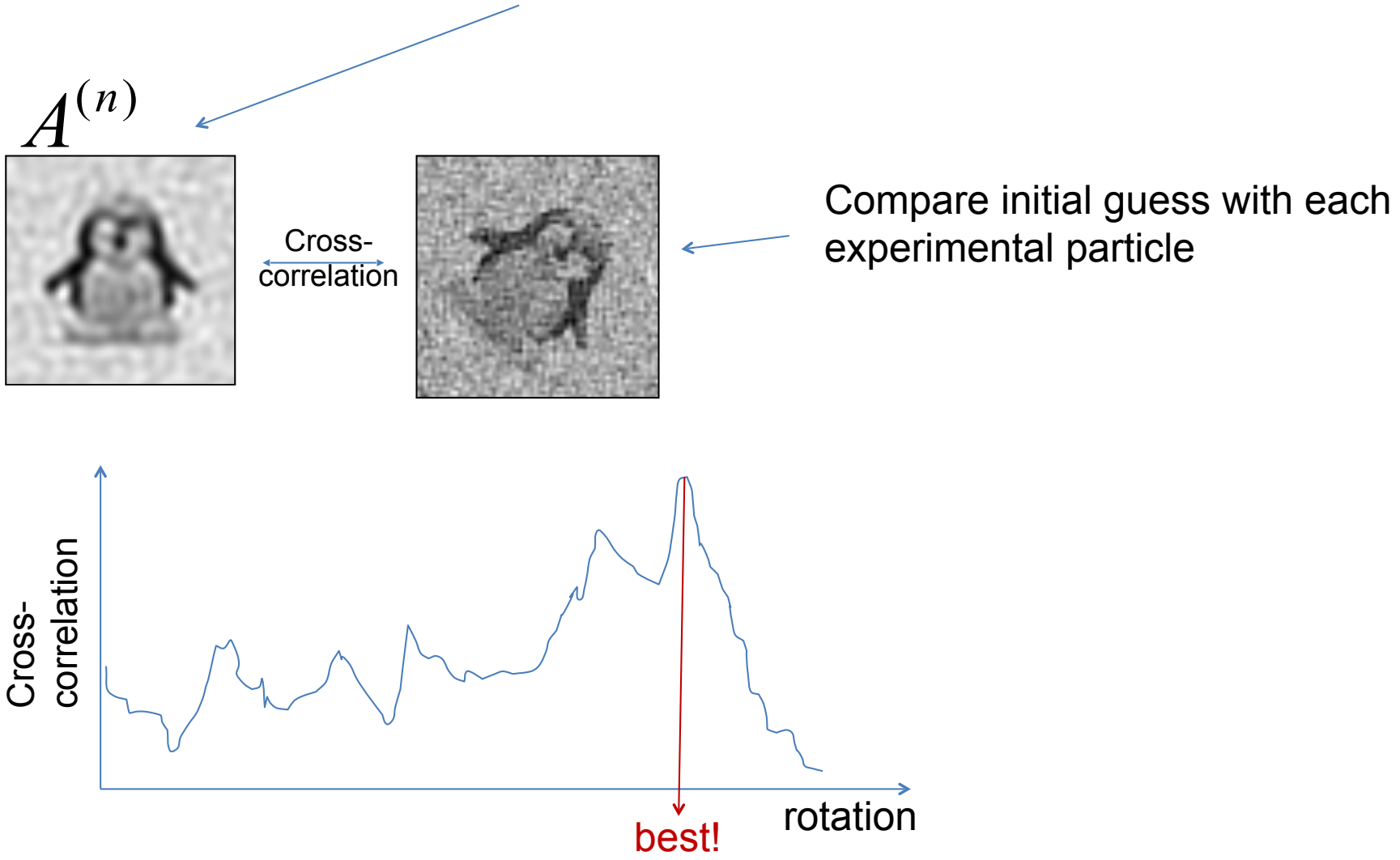
$$L(\Theta) = P(X \mid \Theta) = \int_Y P(X \mid Y,\Theta) P(Y \mid \Theta) d\phi$$

Probability of X, regardless Y

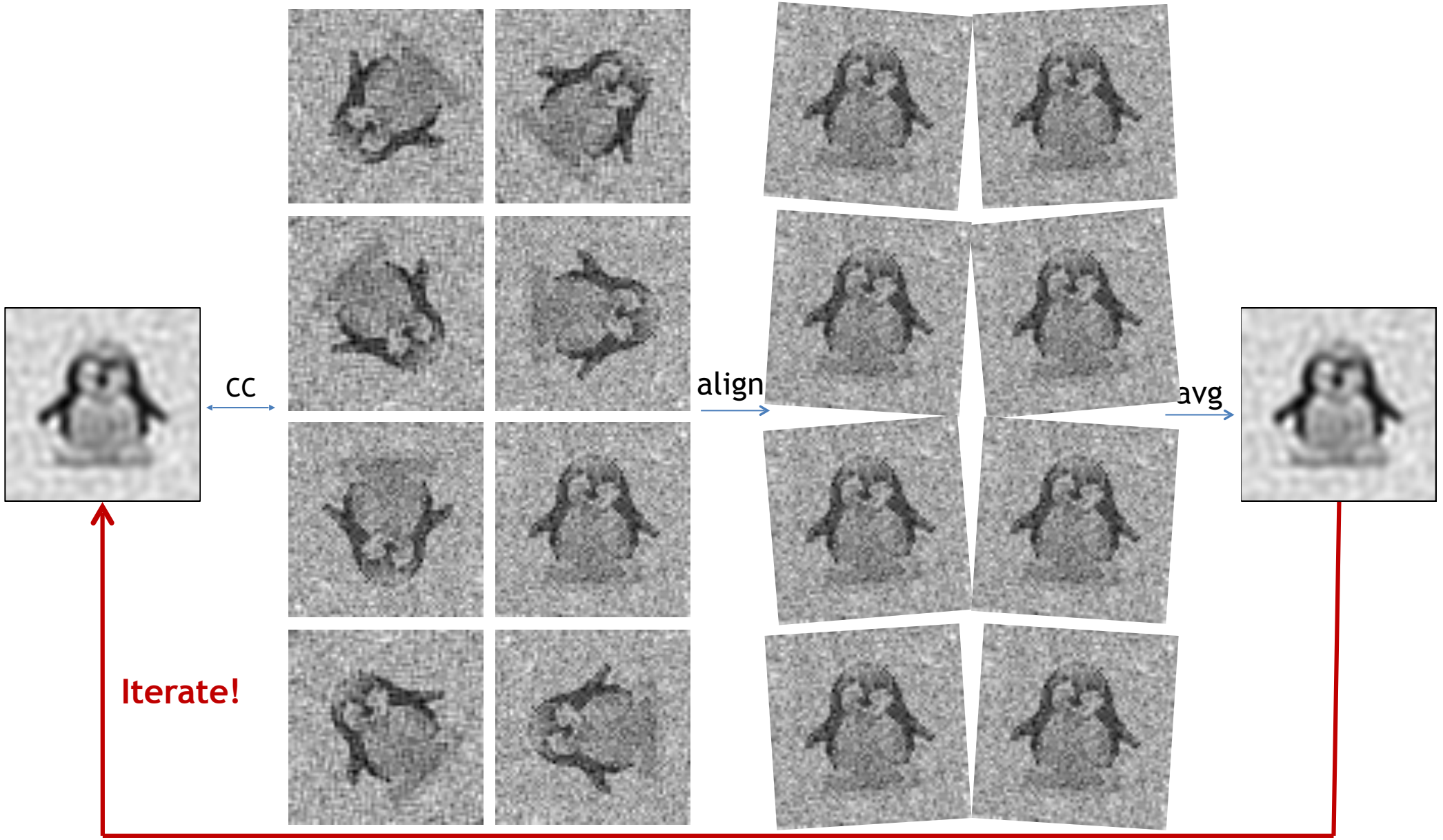
The maxCC approach

Reference-based alignment

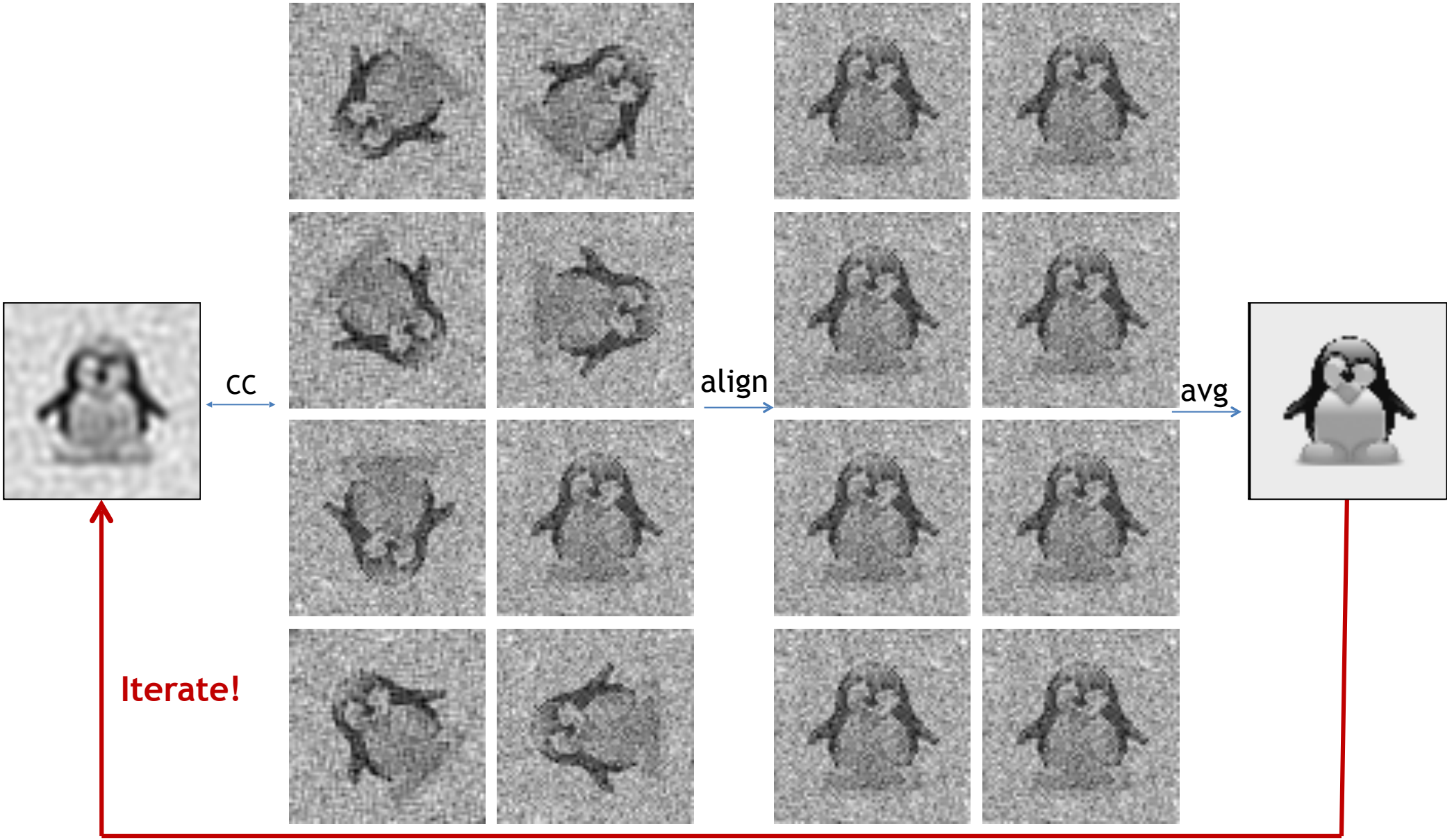
- Starts from some initial guess about the structure



Align and average



Align and average

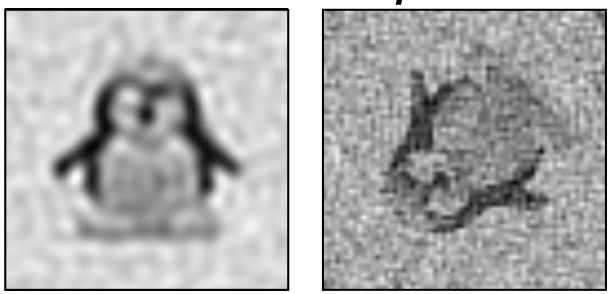


The ML approach

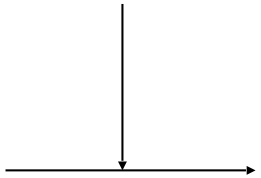
Single-reference alignment in real-space
Sigworth, J. Struct.Biol. 1998

Maximum likelihood

$A^{(n)}$ X_i

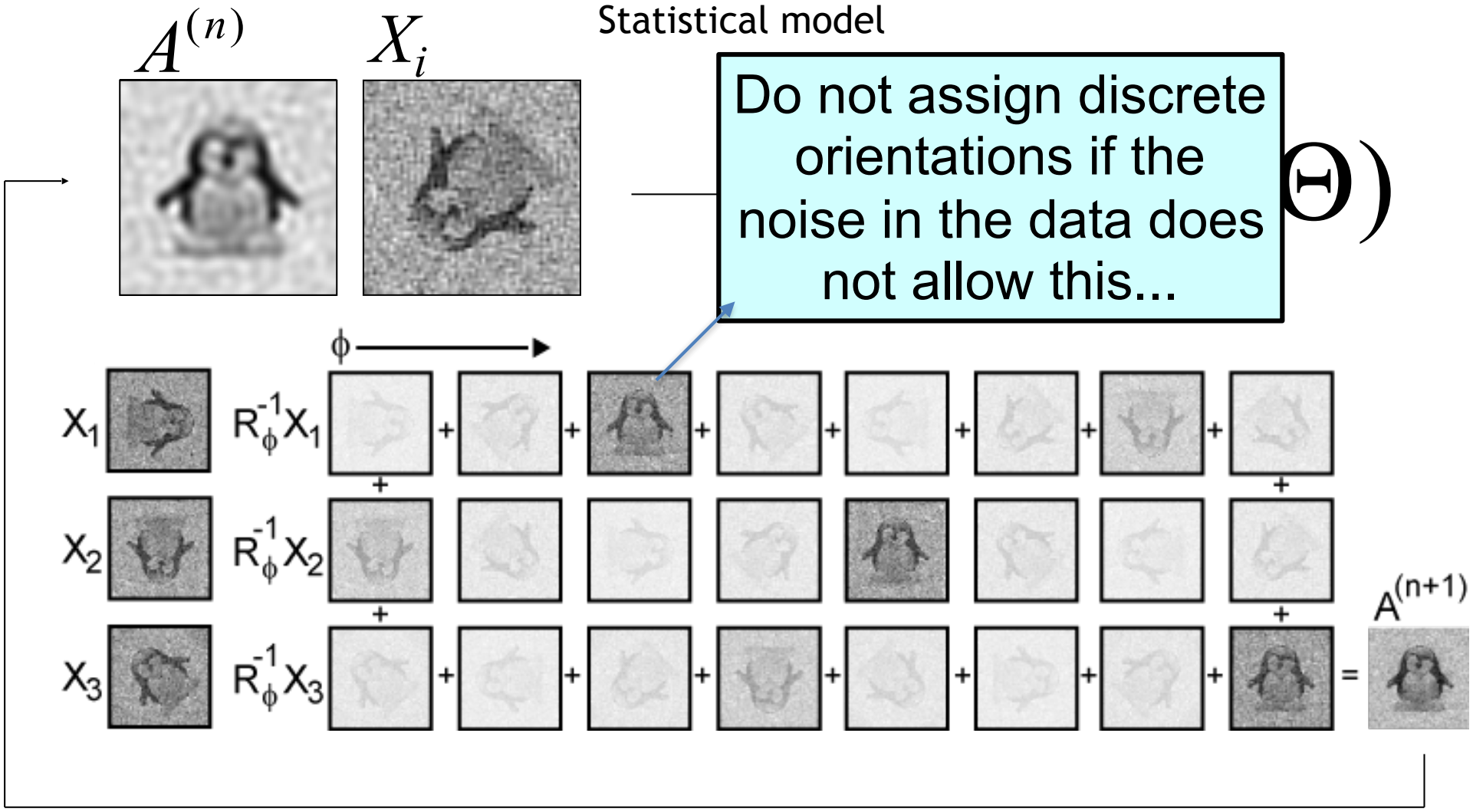


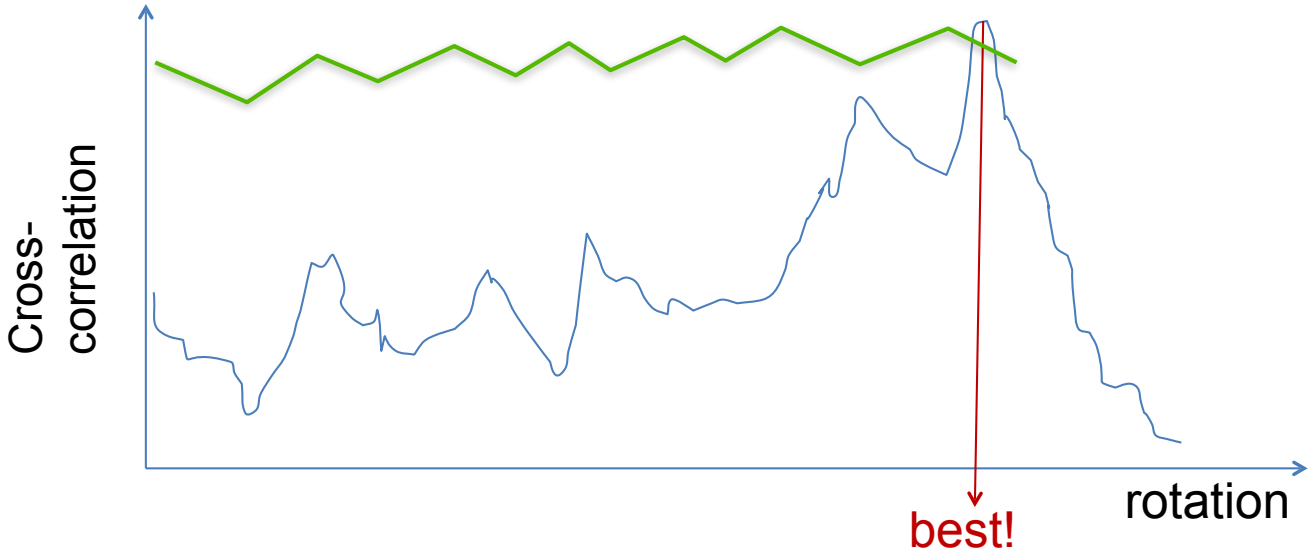
Statistical model



$$P(X_i \mid \phi, \Theta)$$

Maximum likelihood

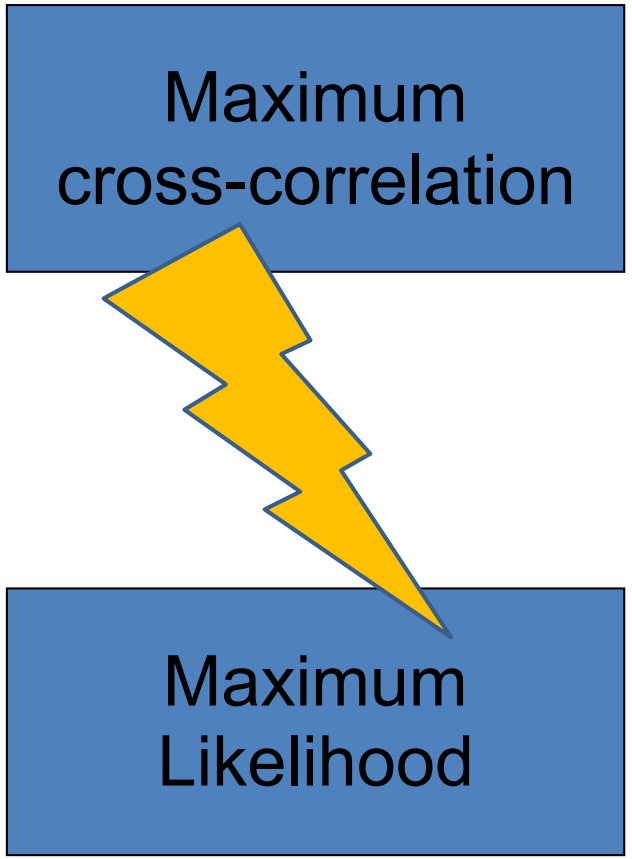



Cross-
correlation
rotation
best!

Incomplete data problems

- Option 1: add *Y* to the model

$$L(Y,\Theta) = P(X \mid Y,\Theta)$$



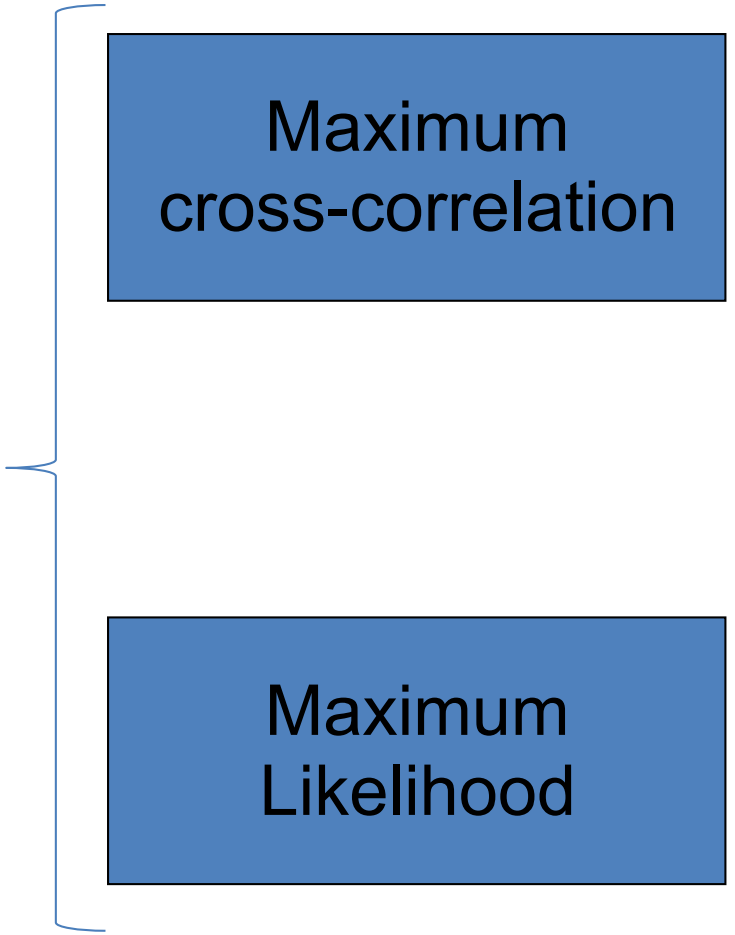
- Option 2: marginalize over *Y*

$$L(\Theta) = P(X \mid \Theta) = \int_Y P(X \mid Y,\Theta) P(Y \mid \Theta) d\phi$$

Probability of X,
regardless Y

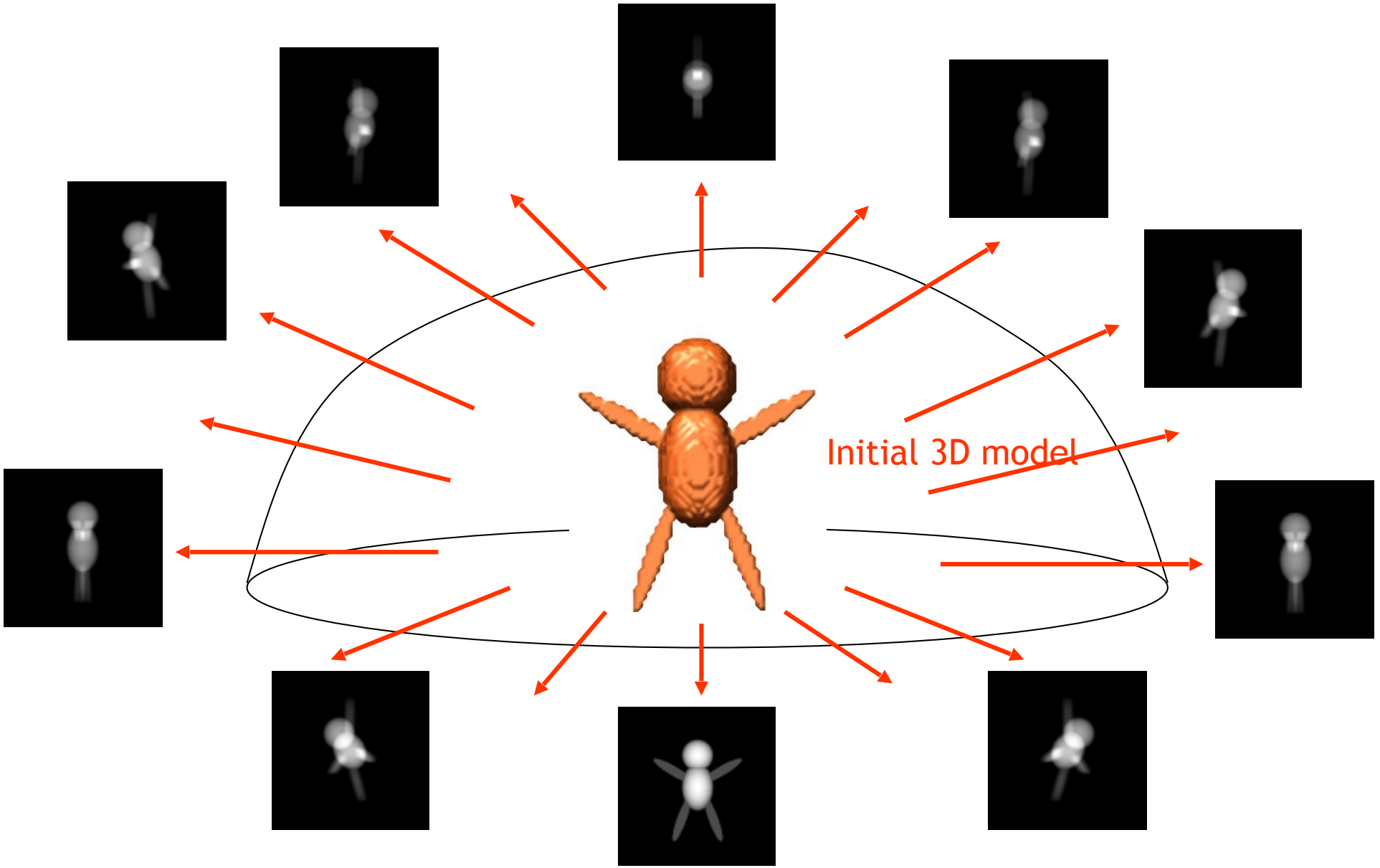
Incomplete data problems

In the limit of **noiseless data** the
Two techniques are equivalent!

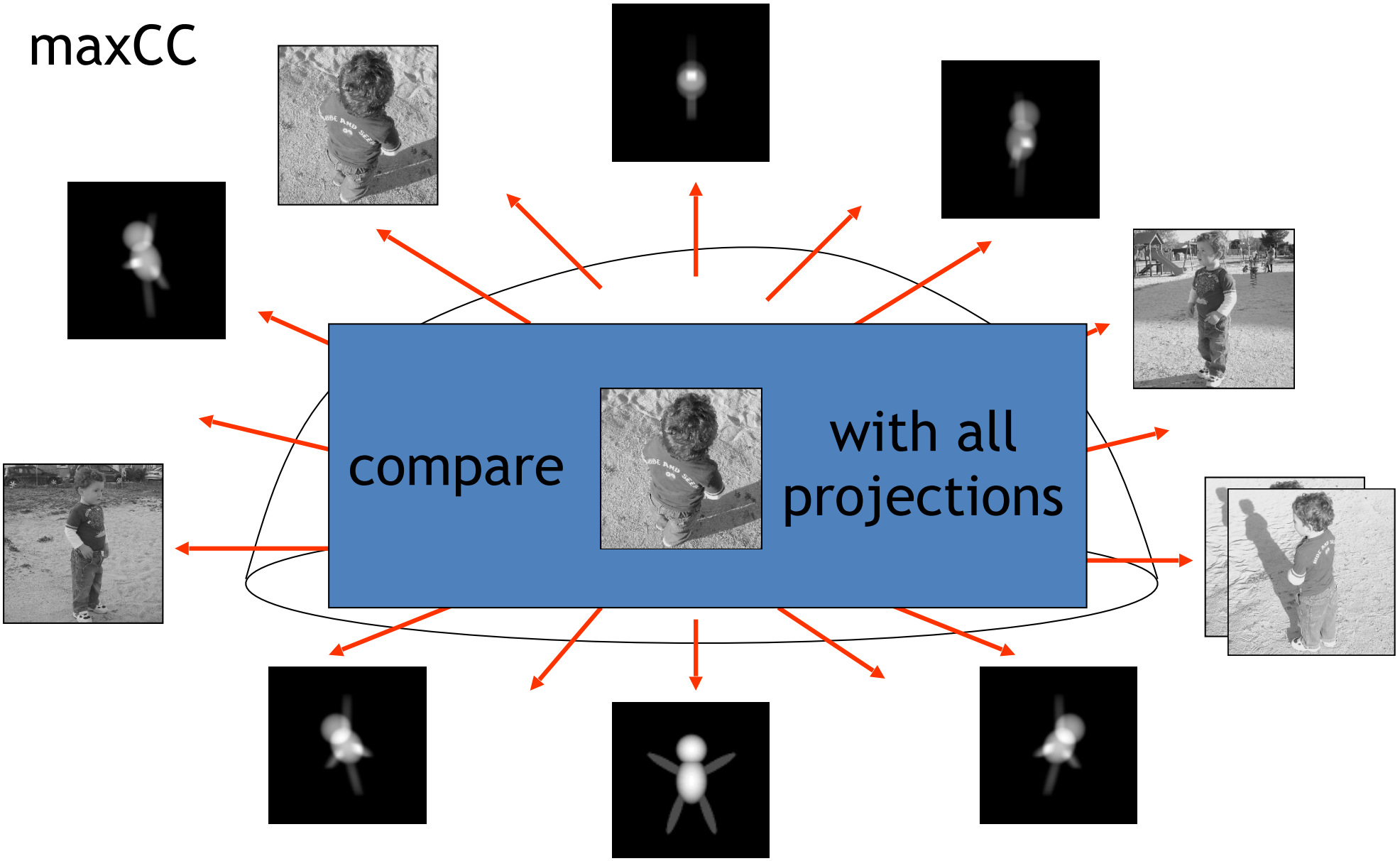


Scheres et al *Methods in Enzymology*, **482** (2010)

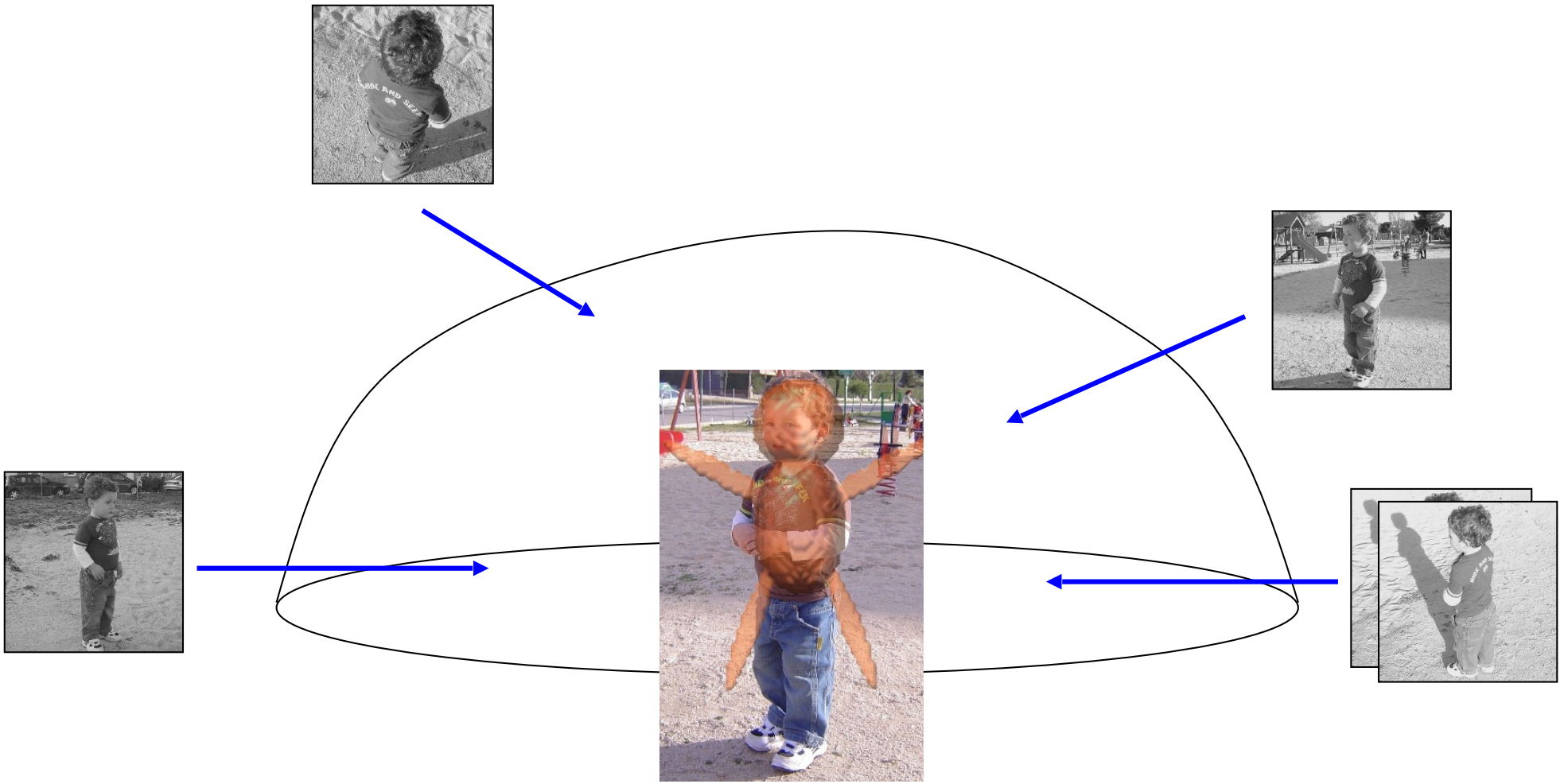
Projection matching



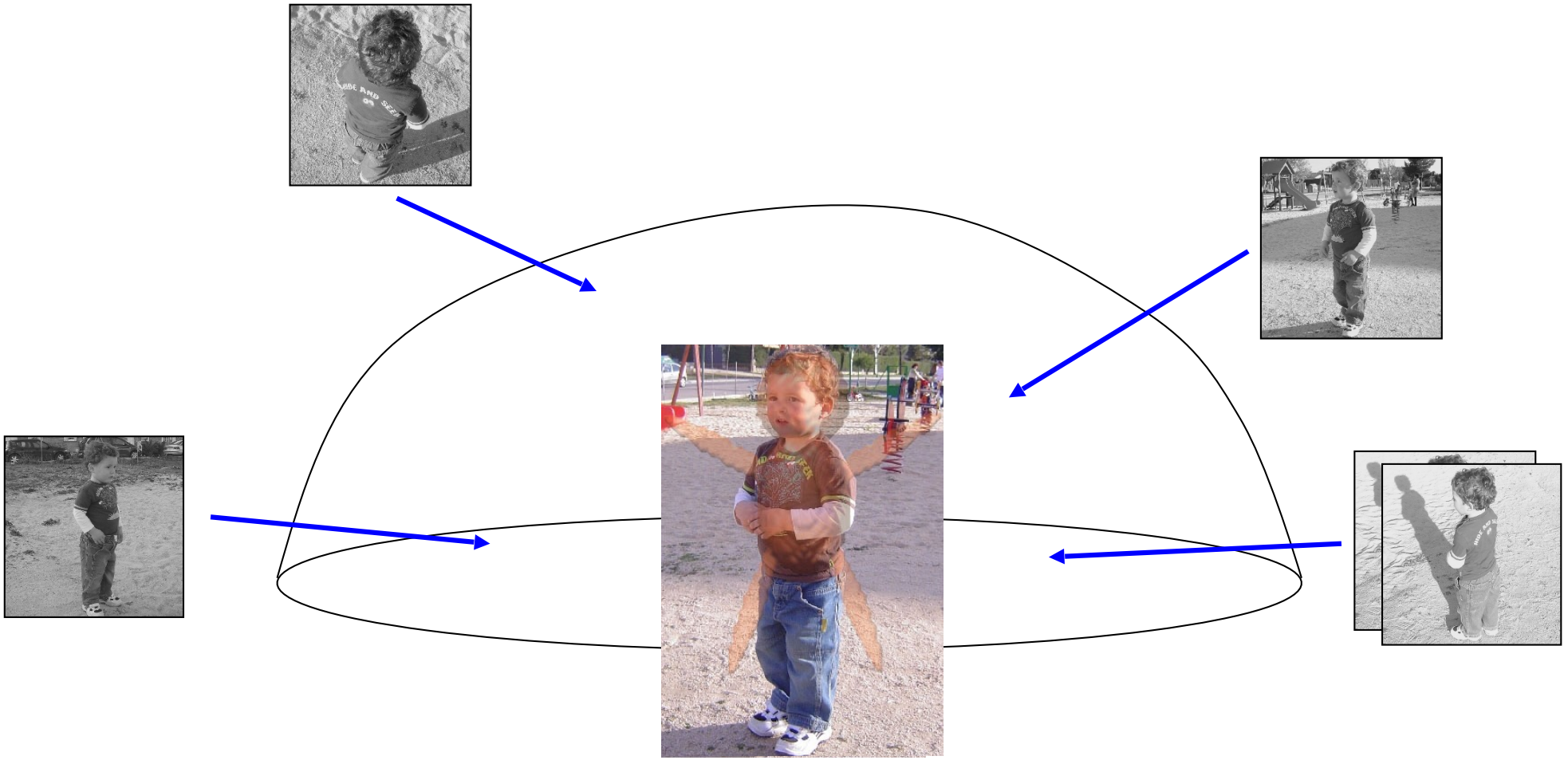
Projection matching



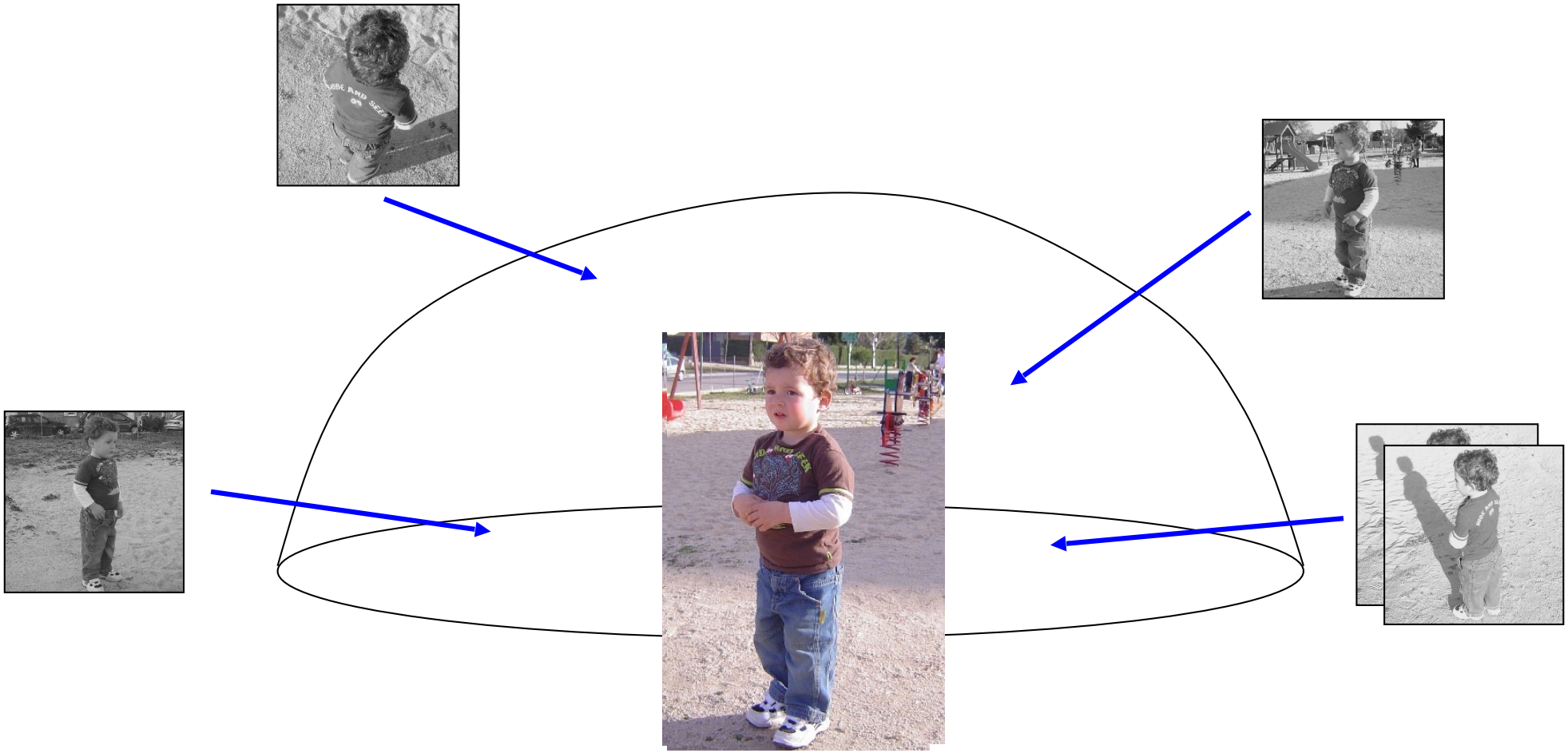
3D reconstruction



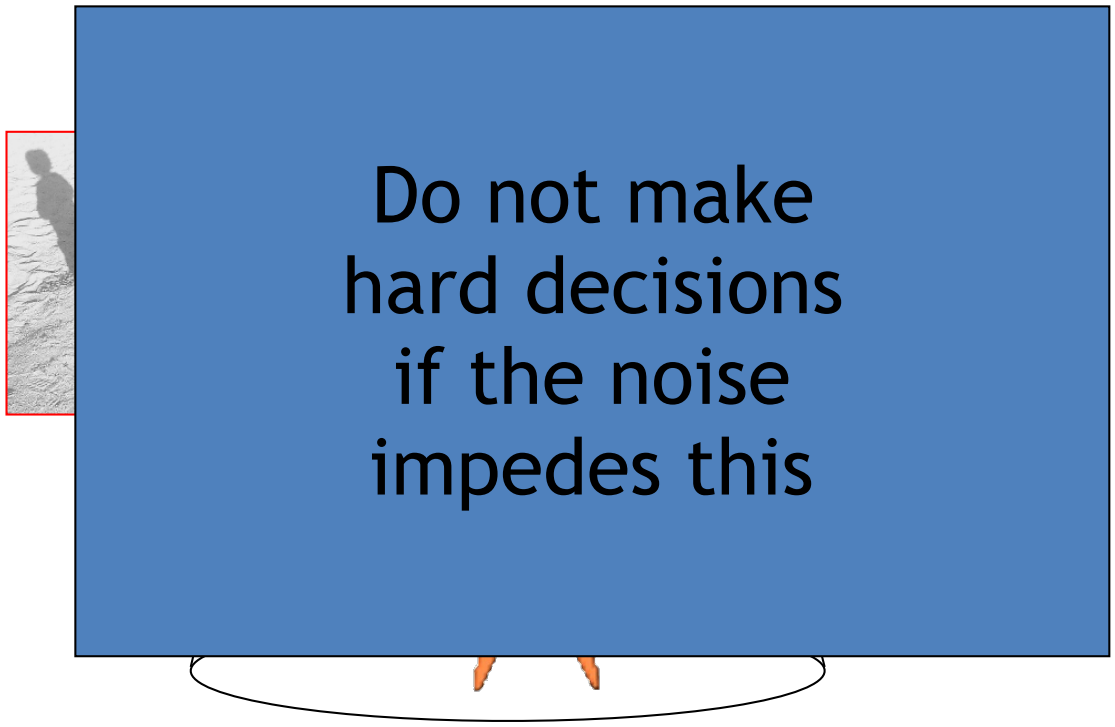
3D reconstruction
(Iterative refinement)



Iterative refinement

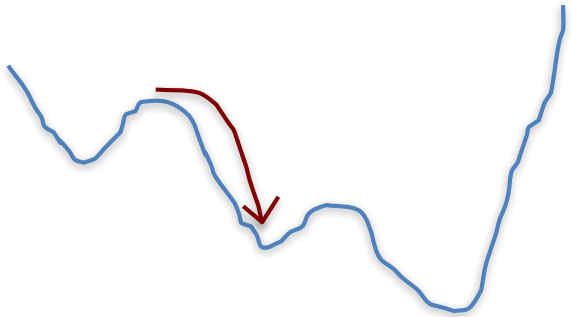


3D ML refinement



“Probability-weighted angular assignment”

Initial model

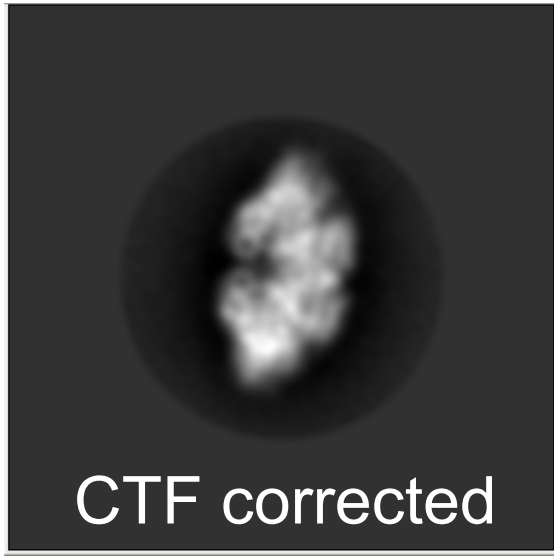


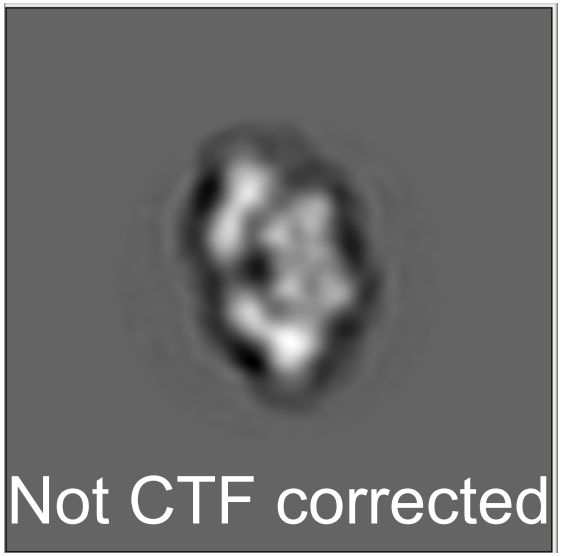
- Local optimizer!
  - Gets stuck in nearest local minimum
  - (as most other approaches in the field)
- Need an initial model (see next lecture)
  - Start from random angles will often NOT work
  - (unlike the 2D case)
- Bad model in -> bad model out!!!
  - Model bias
  - Bad common-lines models are notoriously difficult
- Stochasticity
  - Randomly perturb optimisation (á la Simulated Annealing in X-ray)
  - Can ***potentially*** reach global minimum
  - Do refine from random blobs
  - SIMPLE (Hans Elmlund), EMAN2 (Steve Ludtke)

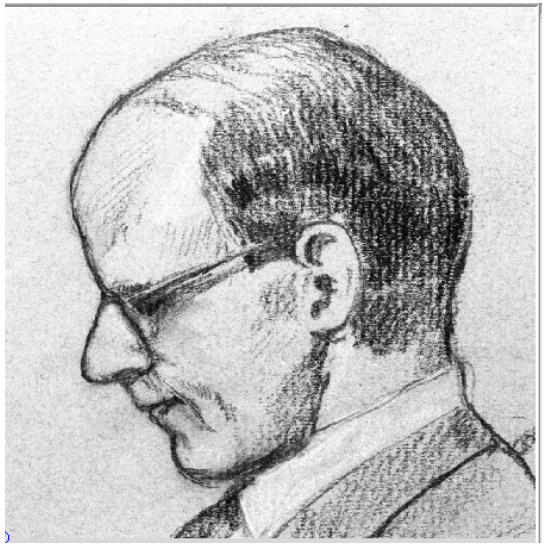
Fourier-space formulation

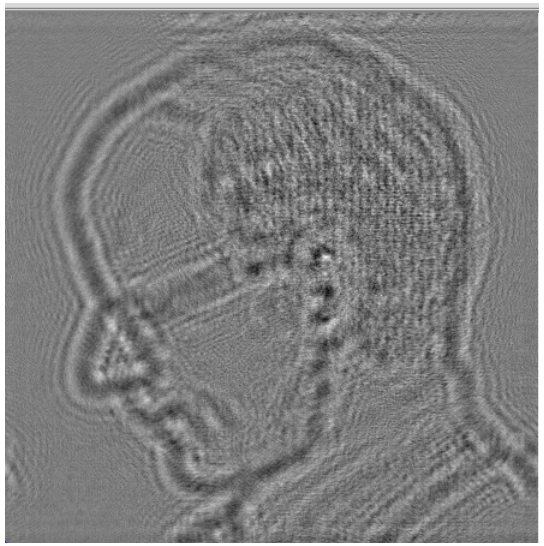
CTF correction

CTF-correction

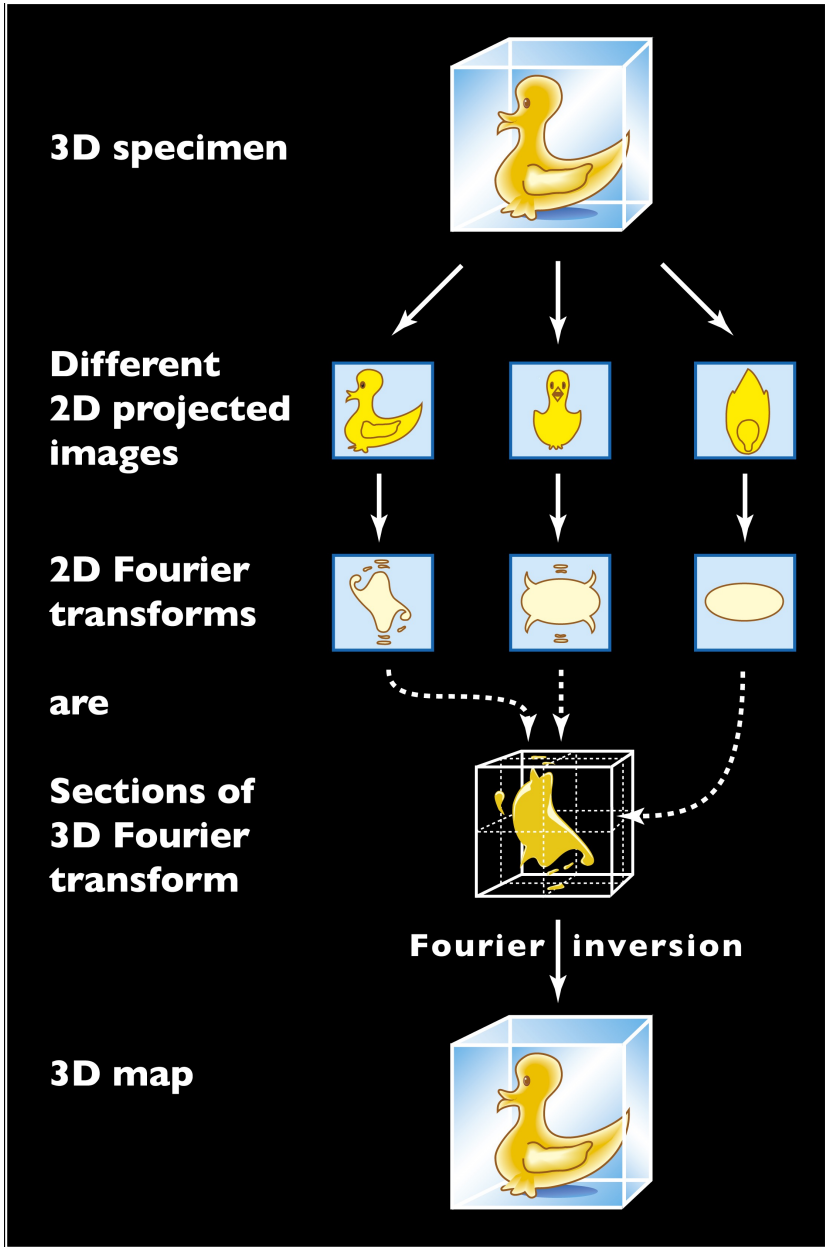








Projection slice theorem



Data model

- Real-space

$$X_i = \mathrm{CTF}_i \otimes \mathbf{P}_\varphi V_k + N_i$$

- Convolute w/ CTF
- $\mathbf{P}_\varphi$ implements integrals
- N_i describes white noise

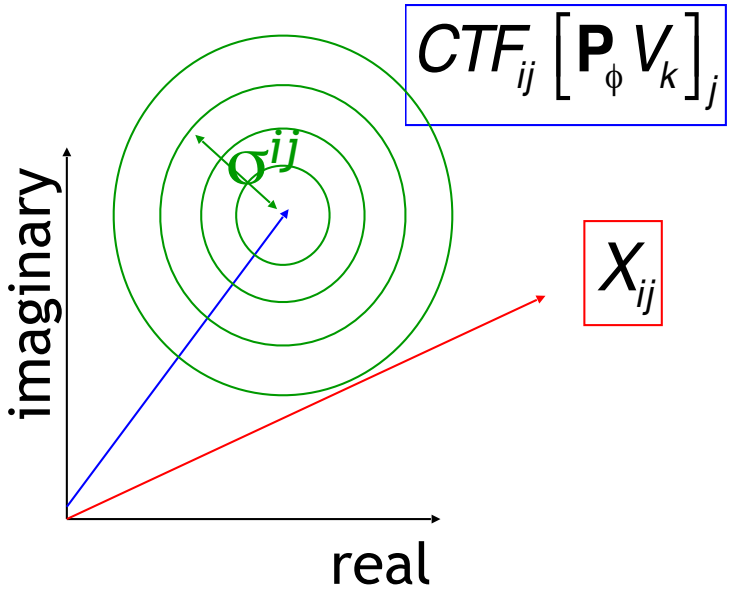
- Fourier space

$$X_i = \mathrm{CTF}_i \mathbf{P}_\varphi V_k + N_i$$

- Multiply w/ CTF
- $\mathbf{P}_\varphi$ takes a slice
- N_i describes coloured noise

Coloured noise model

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



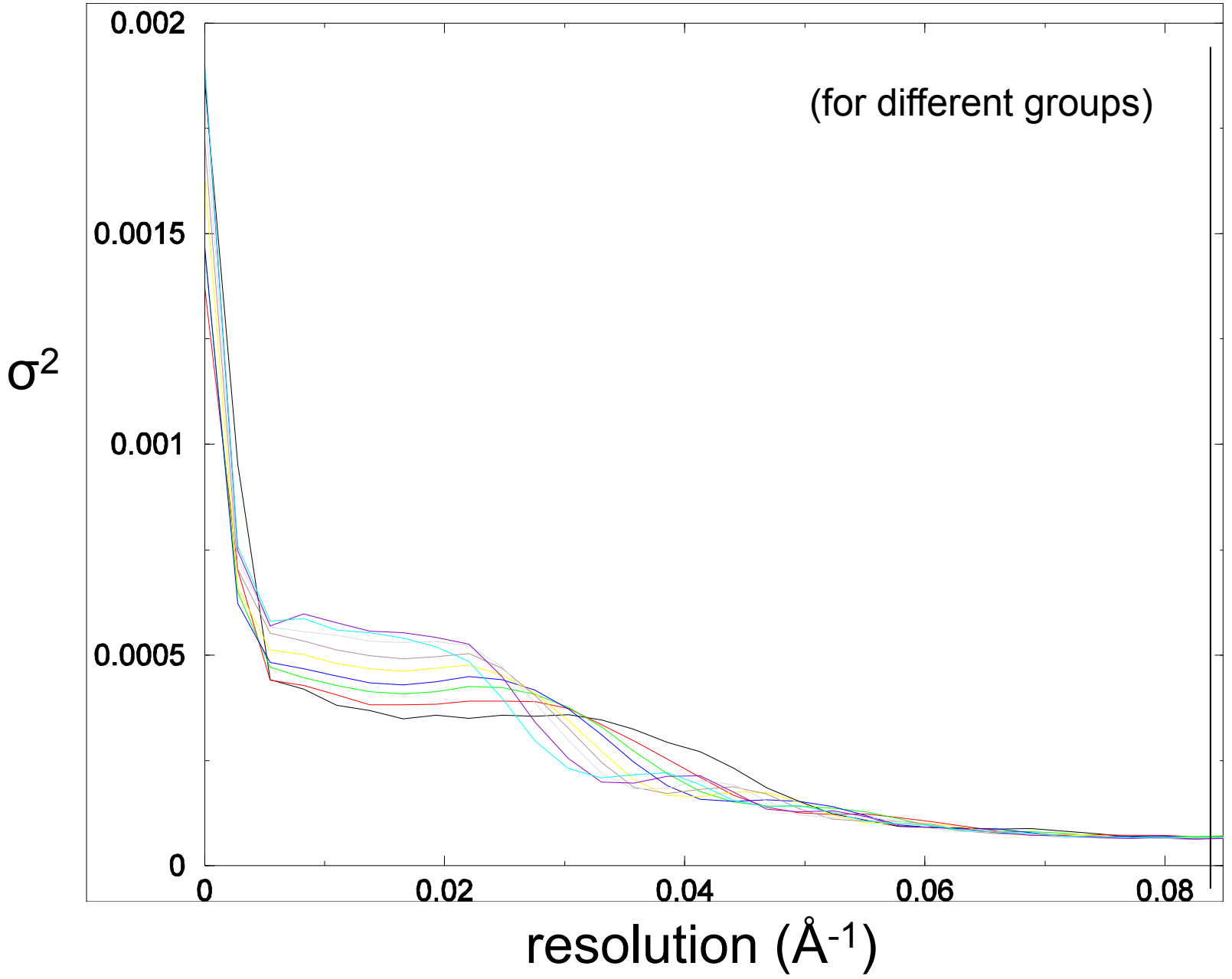
Assuming independence of noise between all Fourier terms:

$$P(X_i \mid k,\phi,\Theta) = \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}^2} \exp\left(\frac{\left|CTF_{ij}\left[\mathbf{P}_\phi V_k\right]_j - X_{ij}\right|^2}{-2\sigma_{ij}^2}\right)$$

resolution-dependent noise model!

Scheres *et al.* (2007) *Structure*

Coloured noise!!



Ill-posedness

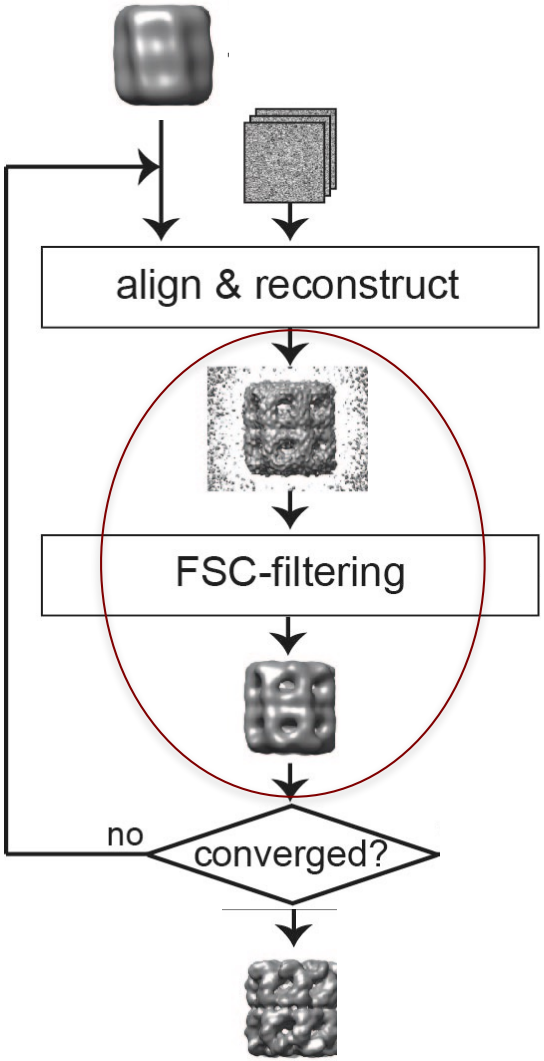
- The experimental data alone is not enough to determine a unique solution!
- There are many noisy reconstructions that describe the data equally well…
- Danger of incorrect interpretation…

- By incorporating external information, a different problem may be solved for which a unique solution does exist!

- Regularization

- Conventional regularization approaches
  - Wiener filtering
  - Low-pass filtering

Ad-hoc regularisation



Low-pass filter the map
at every iteration

Many different ways and implementations…

A Bayesian view on regularization

$$P(\Theta \mid X) = \frac{P(X \mid \Theta)P(\Theta)}{P(X)}$$

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

Regularised likelihood optimisation - RELION

Likelihood

- Assume noise is Gaussian and independent
  - in Fourier space
  - with spectral power $\sigma^2(\upsilon)$: *coloured noise*

$$P(X_i \mid k, \phi, \Theta) = \prod_{j=1}^{J} \frac{1}{2\pi\sigma_{ij}} \exp\left(\frac{\left\| X_{ij} - \mathrm{CTF}_{ij} (\mathbf{P}_\phi V_k)_j \right\|^2}{-2\sigma_{ij}^2} \right)$$

Prior

- Assume signal is Gaussian and independent
  - in Fourier space
  - Limited power $\tau^2(u)$: *smoothness in real space!*

$$P(\Theta) = \prod_l \frac{1}{2\pi\tau_{kl}} \exp\left\{ \frac{\|V_{kl}\|^2}{-2\tau_{kl}^2} \right\}$$

Expectation maximization

$$V^{(n+1)} = \frac{\sum_{i=1}^{N} \int_{\phi} \Gamma_{i\phi}^{(n)} \mathbf{P}_{\phi}^{\mathrm{T}} \frac{\mathrm{CTF}_i}{\sigma_i^{2(n)}} X_i d\phi}{\sum_{i=1}^{N} \int_{\phi} \Gamma_{i\phi}^{(n)} \mathbf{P}_{\phi}^{\mathrm{T}} \frac{\mathrm{CTF}_i^2}{\sigma_i^{2(n)}} d\phi + \frac{1}{\tau^{2(n)}}}$$

→ Wiener (optimal) filter for CTF-corrected 3D reconstruction / 2D class averages

$$\sigma_i^{2(n+1)} = \frac{1}{2} \int_{\phi} \Gamma_{i\phi}^{(n)} \left\| X_i - \mathrm{CTF}_i \mathbf{P}_{\phi} V^{(n)} \right\|^2 d\phi$$

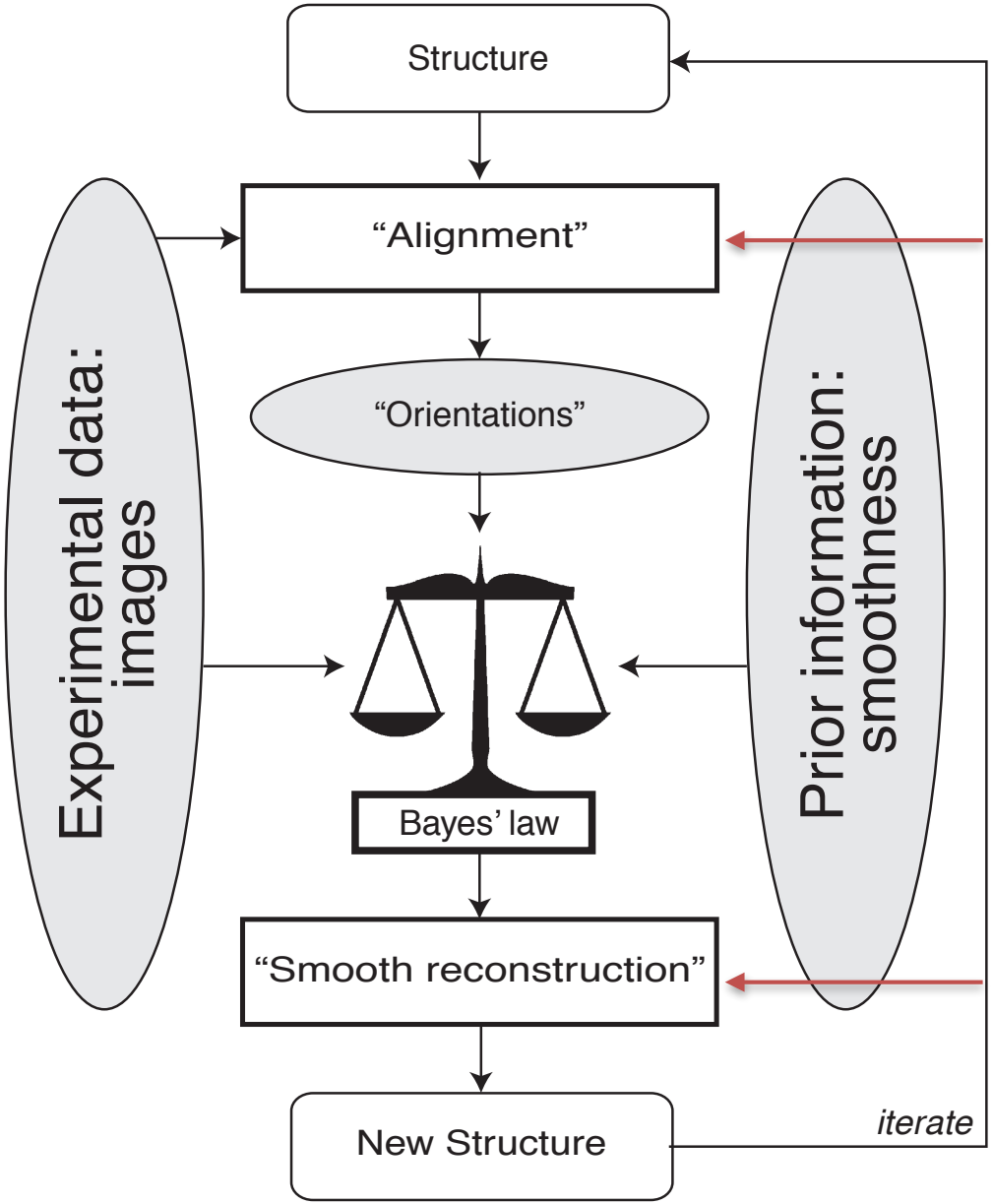
→ Estimate resolution-dependent power of noise from the data

$$\tau^{2(n+1)} = \frac{1}{2} \left\| V^{(n)} \right\|^2$$

→ Estimate resolution-dependent power of signal from the data

$$\Gamma_{i\phi}^{(n)} = \frac{P(X_i | \phi, \Theta^{(n)}) P(\phi | \Theta^{(n)})}{\int_{\phi'} P(X_i | \phi', \Theta^{(n)}) P(\phi' | \Theta^{(n)}) d\phi'}$$

Scheres 2012



Expectation

$$\Gamma_{i\phi}^{(n)}$$

Maximization

Scheres 2012

- 3D-EM refinement
  - Inverse problem: needs iterating
  - Incomplete problem: needs marginalizing
  - Ill-posed problem: needs regularizing

- Regularised likelihood approach
  - Does all 3 things in optimizing a single function!
  - 2D classes, 3D classes or 3D refinement
  - “Learns” optimal parameters from the data
  - Few *ad-hoc* parameters to tune by the user

Software -EM

EMAN - Electron Micrograph Analysis, multipurpose

FREALIGN - only refinement and reconstruction (ML in classification)

I**MAGIC** - multipurpose, ab-initio

SPIDER - multipurpose

SPARX - multipurpose and robust clustering

Xmipp - multipurpose, uses ML in many steps

RELION - classification, refinement and reconstruction, uses ML

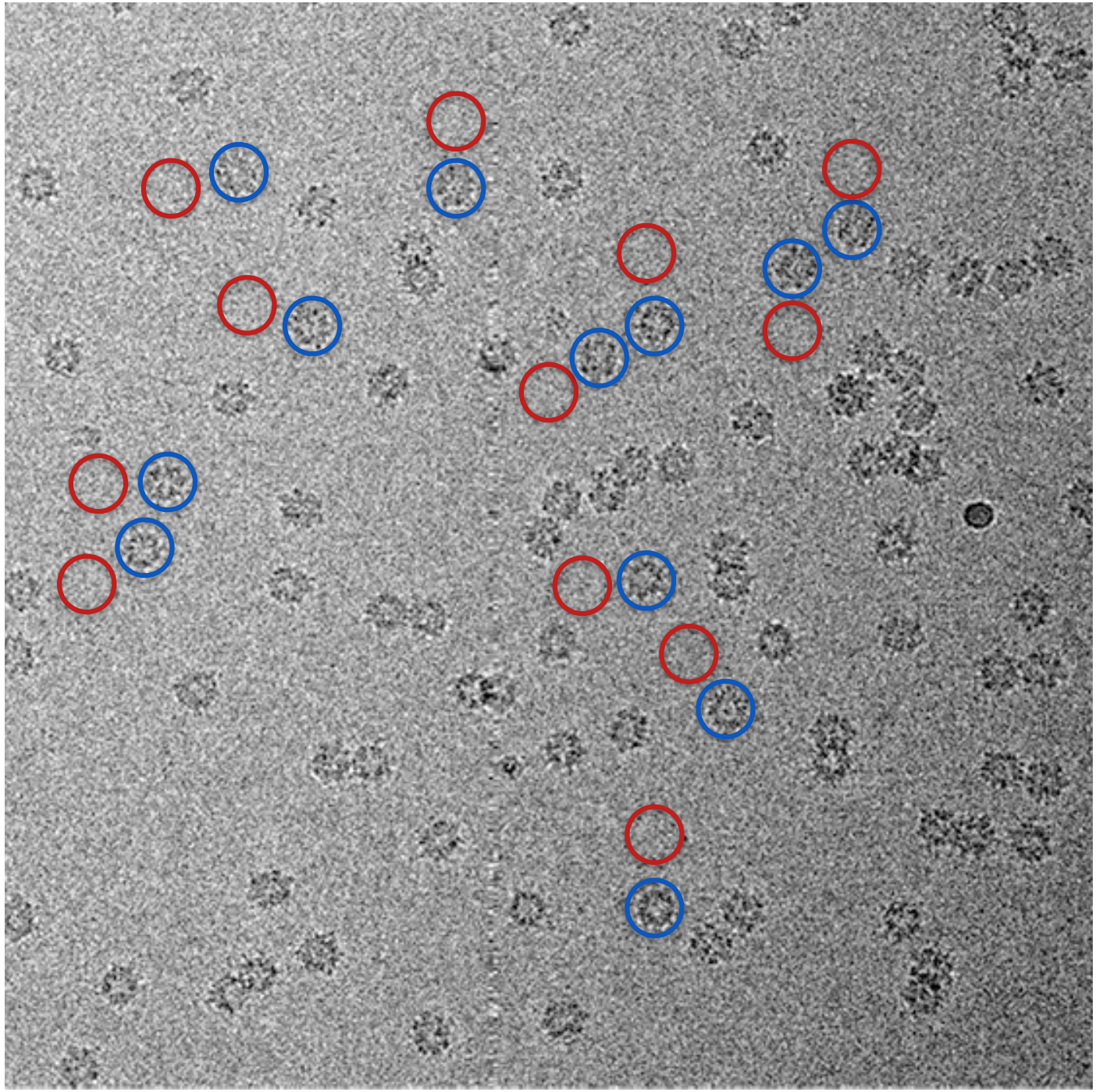
SIMPLE/PRIME - multipurpose, ab-initio

Bsoft - multipurpose

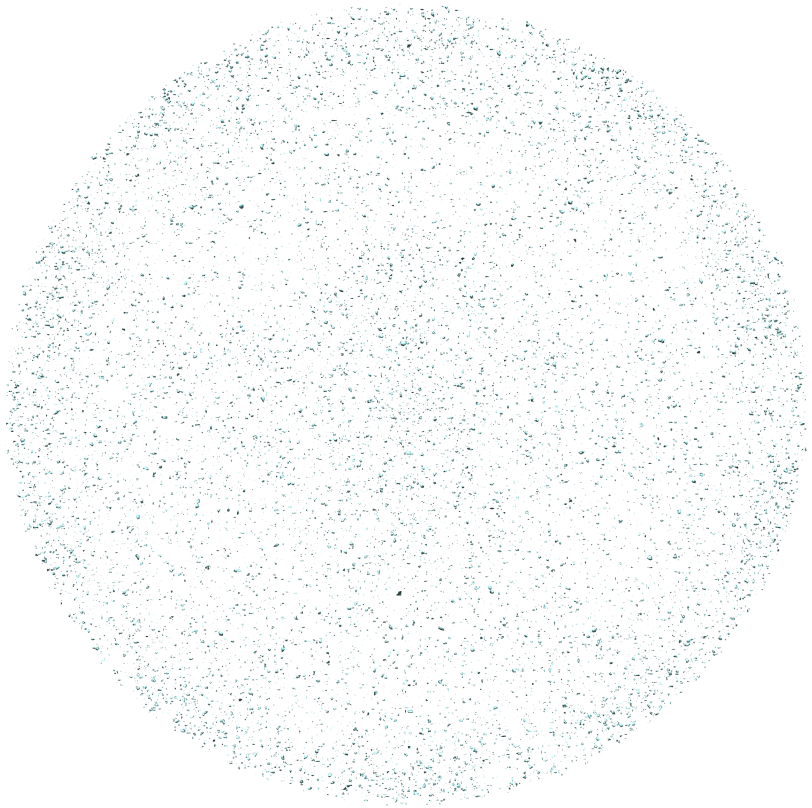
3D refinement: practical problems

- overfitting
- is my refinement improving ?
- how many particles do I need to get ‘X’ Å resolution?
- resolution doesn’t improve despite adding more particles
- orientation distribution
- validation

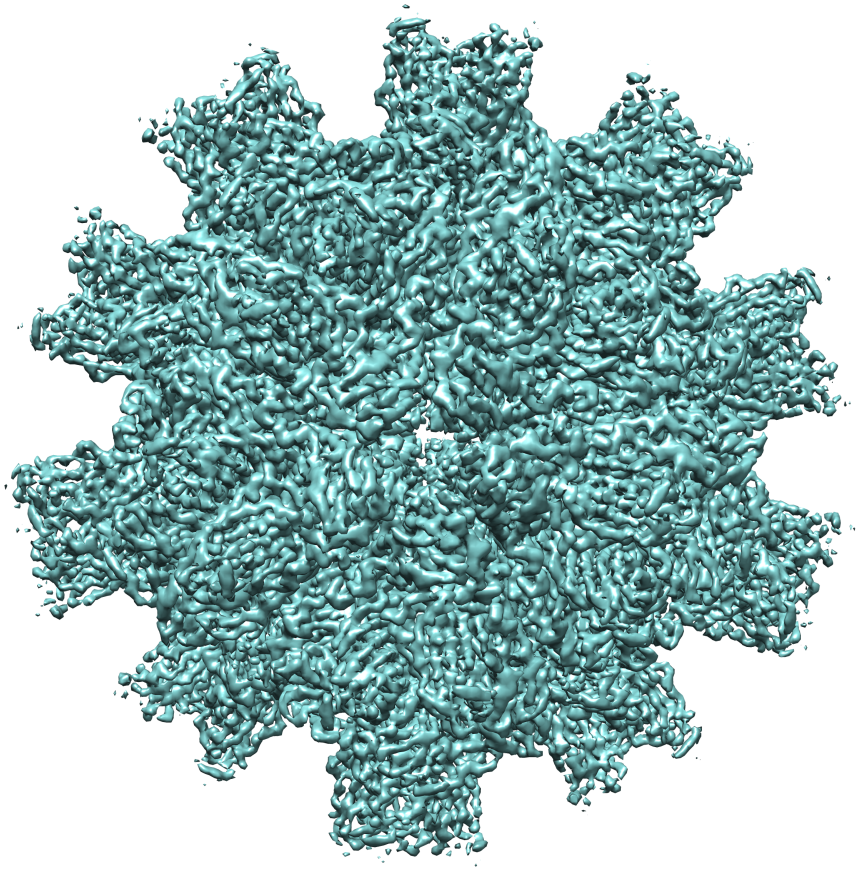
Overfitting: refinement of noise



Overfitting: refinement of noise

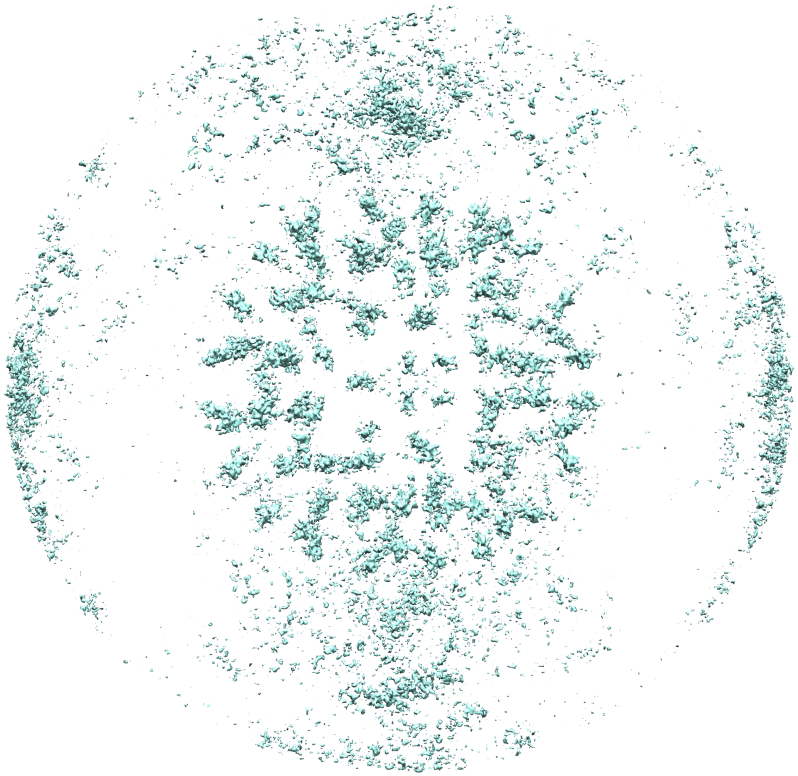


Map from noise
(red circles)

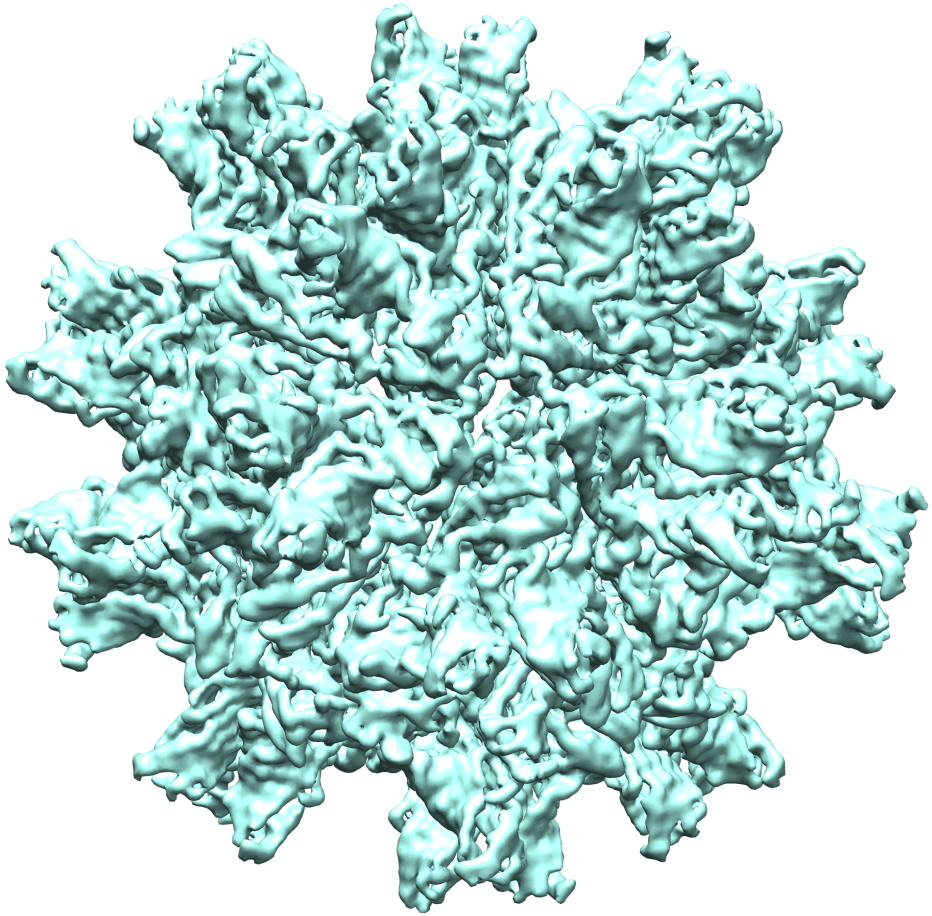


Map from real particles
(blue circles)

Overfitting: refinement of noise



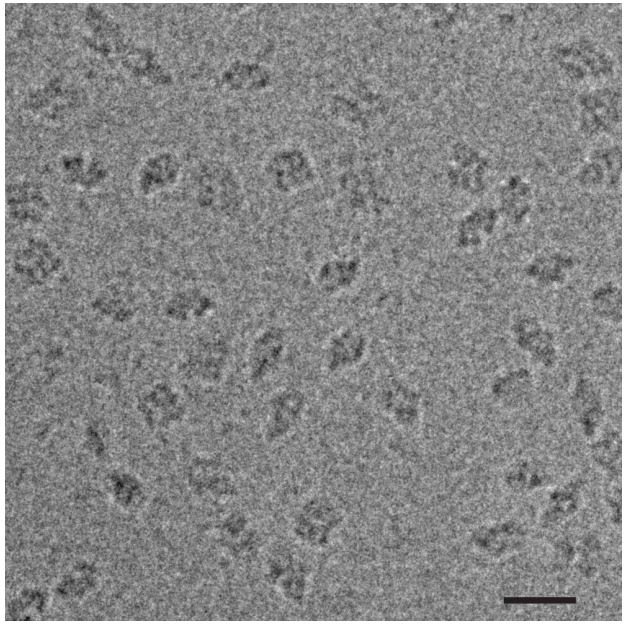
Map from noise
(red circles)

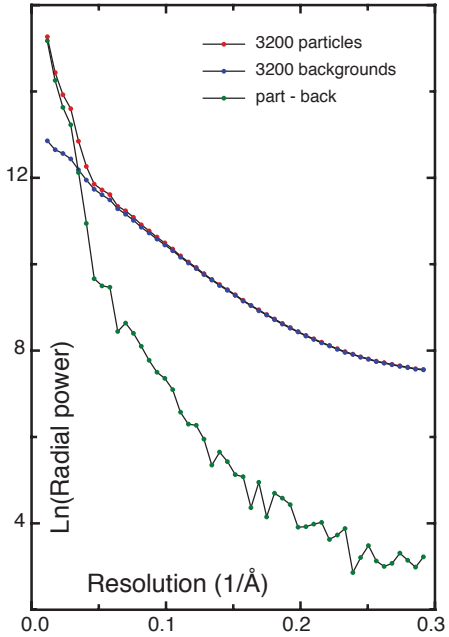


Map from real particles
(blue circles)

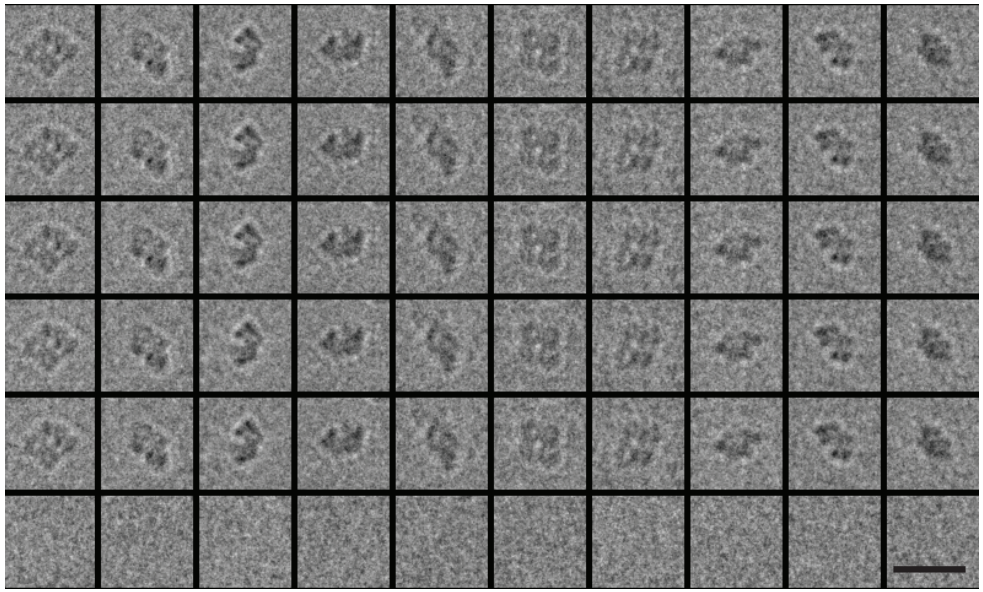
High-resolution noise substitution/phase randomisation

micrograph





power spectra
particles (red)
background (blue)
signal (green)



(c) raw particles

(d) random > 1/17 $Å^{-1}$

(e) noise > 1/17 $Å^{-1}$

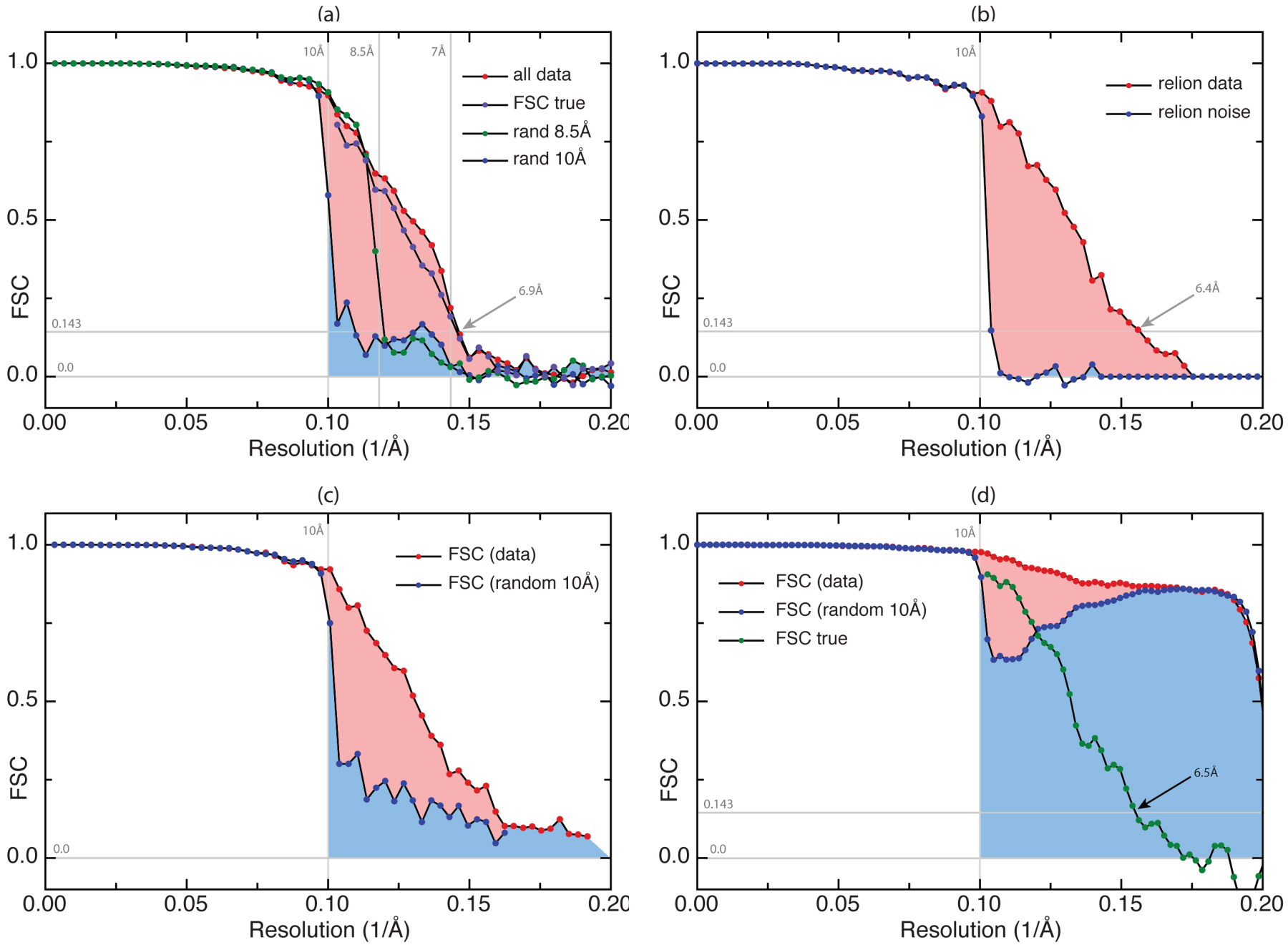
(f) random > 1/10 $Å^{-1}$

(g) noise > 1/10 $Å^{-1}$

(h) noise regions

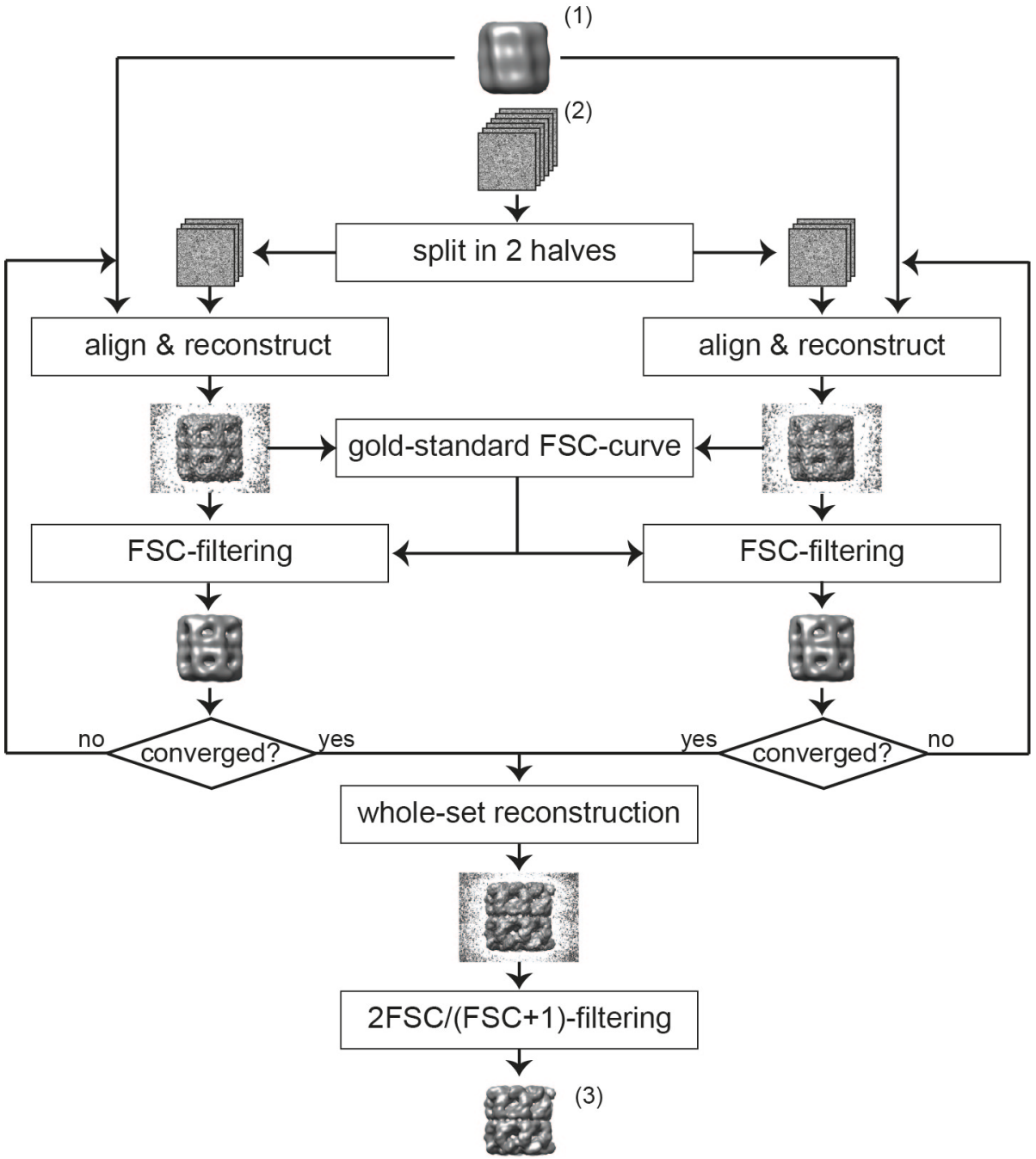
Chen et al., 2013

High-resolution noise substitution/phase randomisation



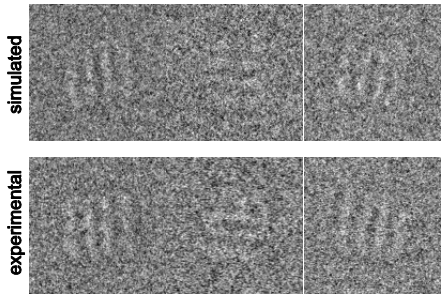
Chen et al., 2013

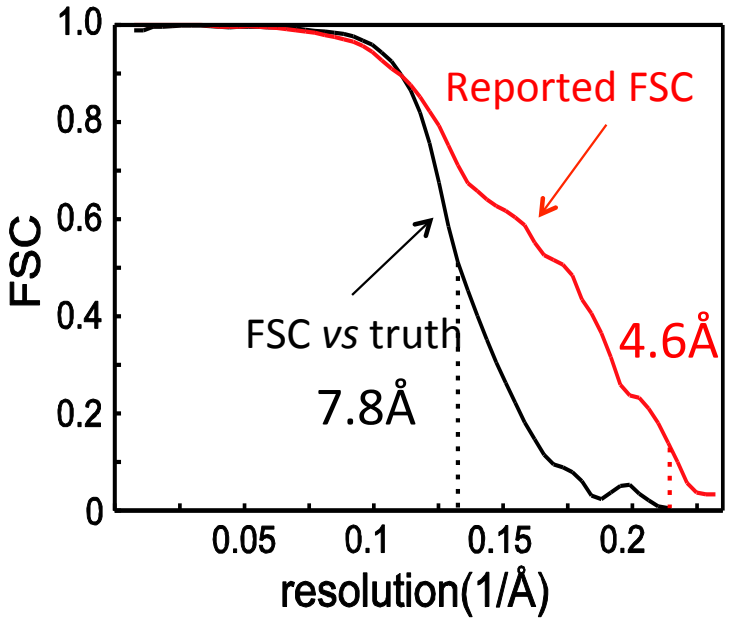
Gold-standard refinement



The pitfalls of undetected overfitting

- 20,000 simulated GroEL particles
- Conventional projection matching





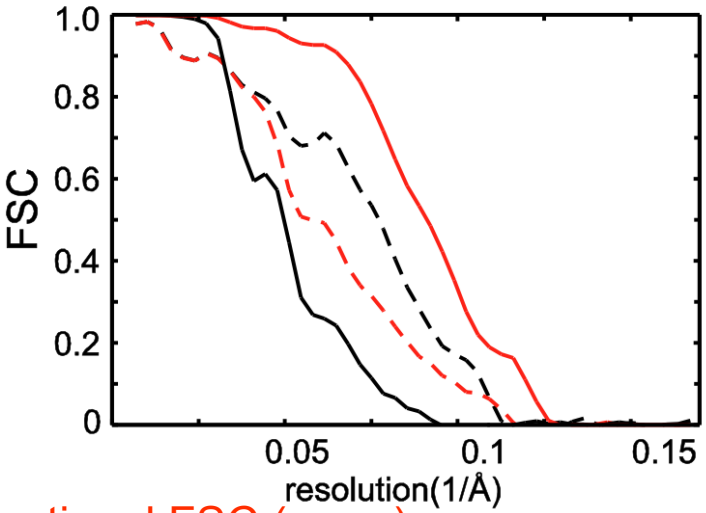
FSC between map and (perfect) model at FSC = 0.5

FSC between two independent half data sets at FSC = 0.143

Scheres and Chen 2012

Rosenthal and Henderson 2003

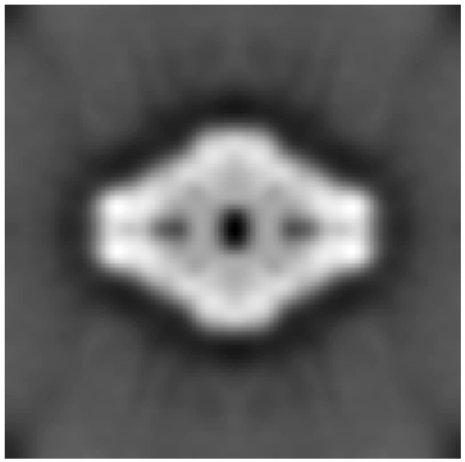
ß-galactosidase



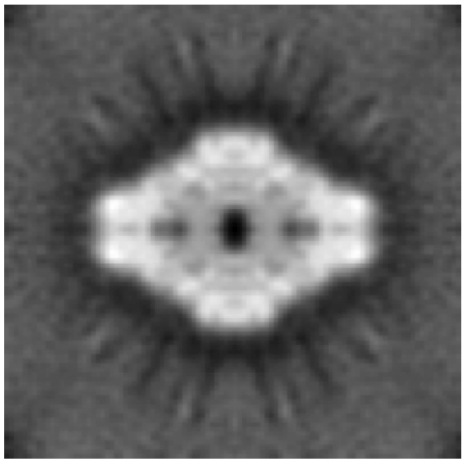
Conventional FSC (———-)
gold-standard FSC (——)

map vs model FSC (- - - -)
map vs model FSC (- - - -)

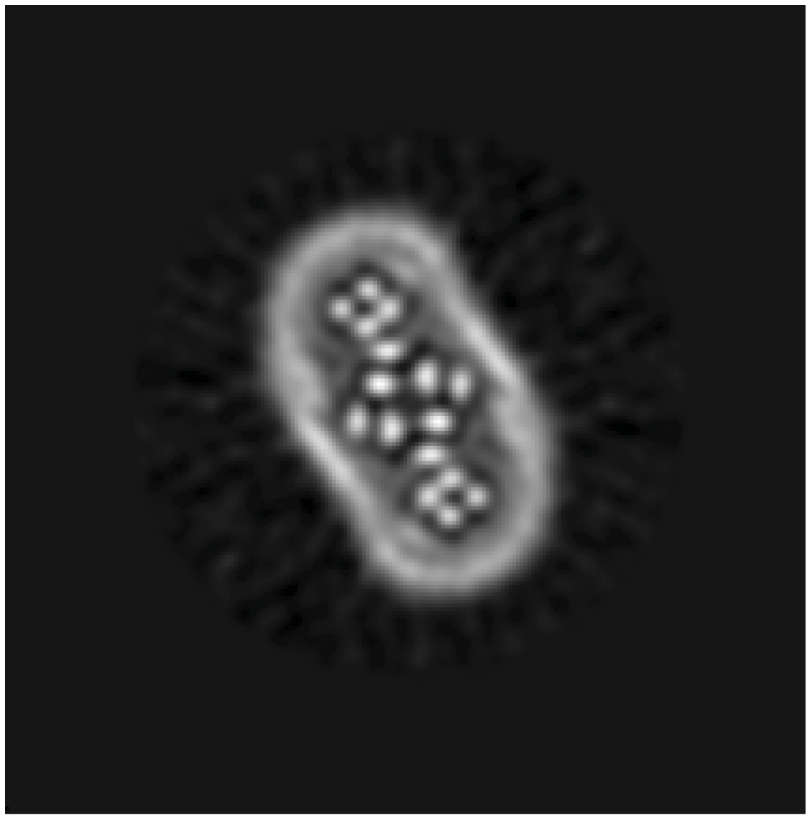
gold-standard

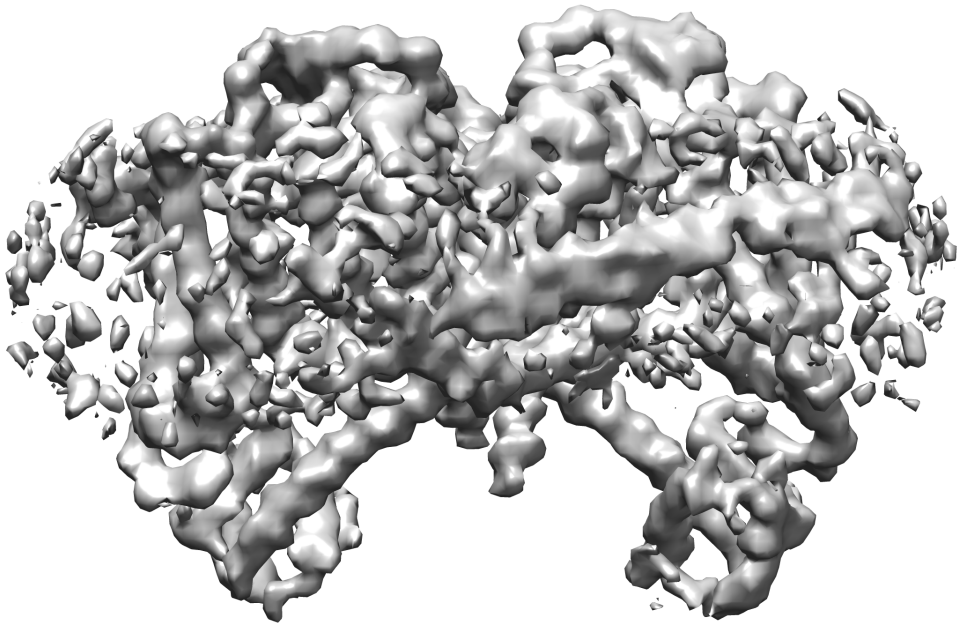


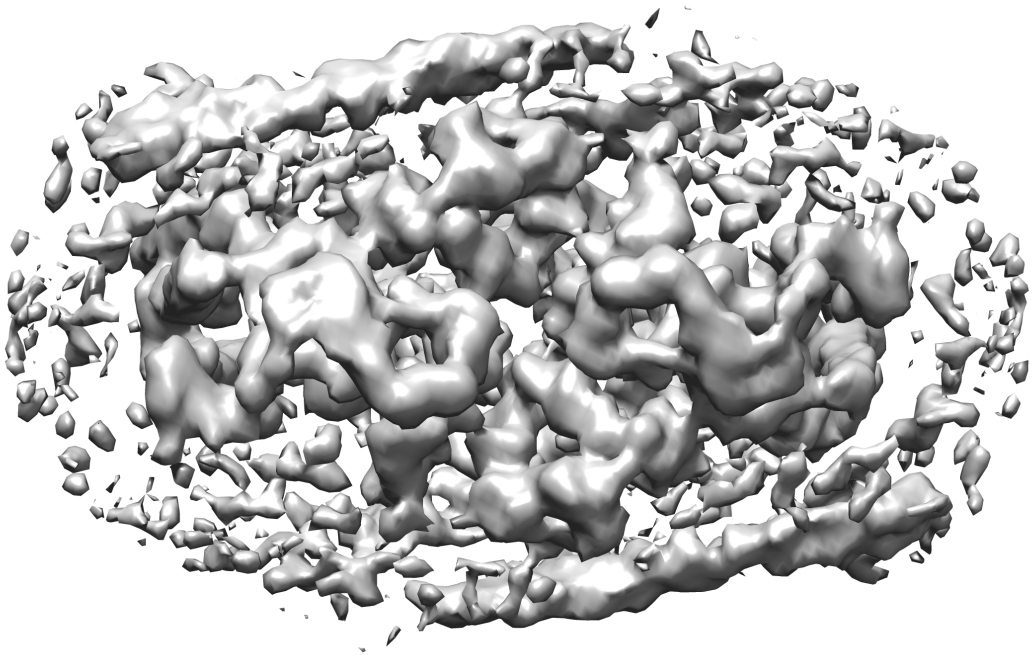
conventional



Overfitting within the half-maps

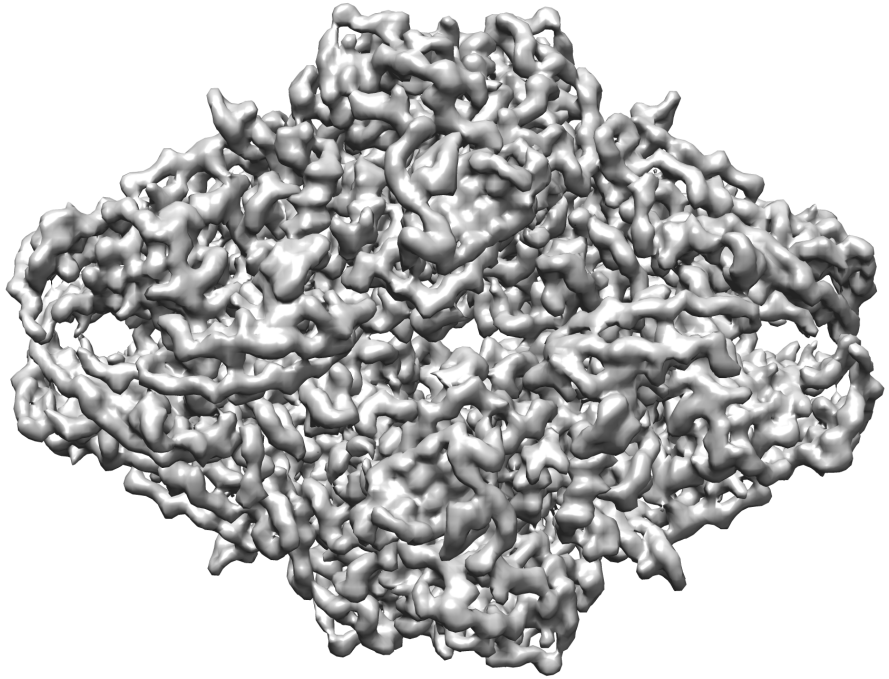




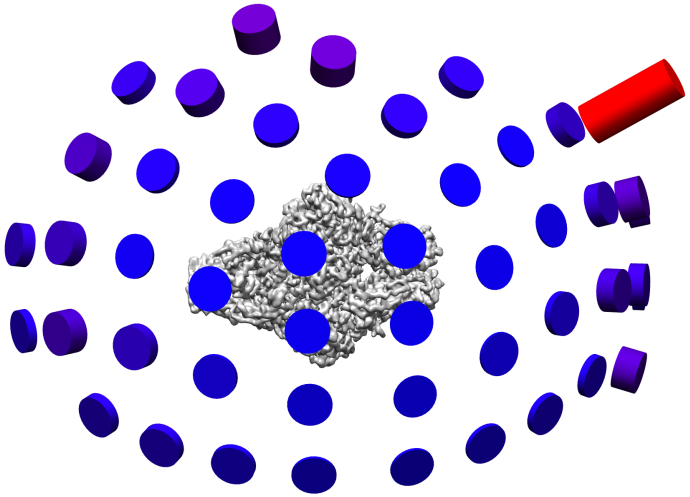


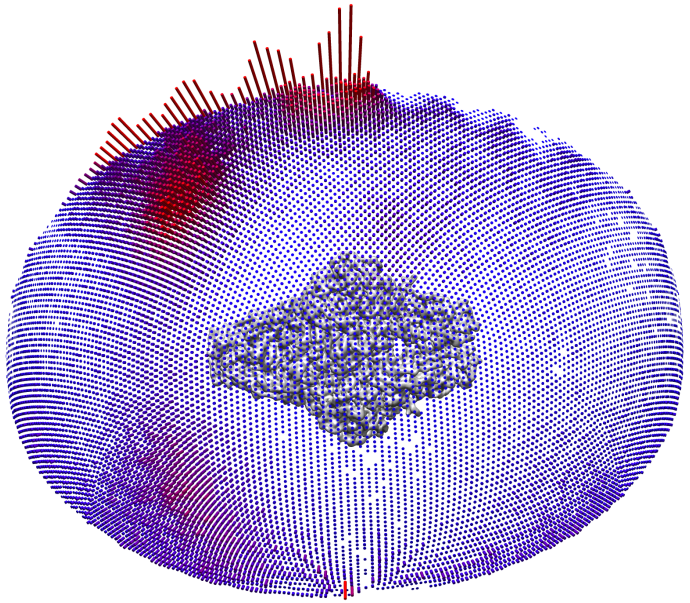
Refinement: Improvement of map

Orientation accuracy -999 deg
50 Å
start

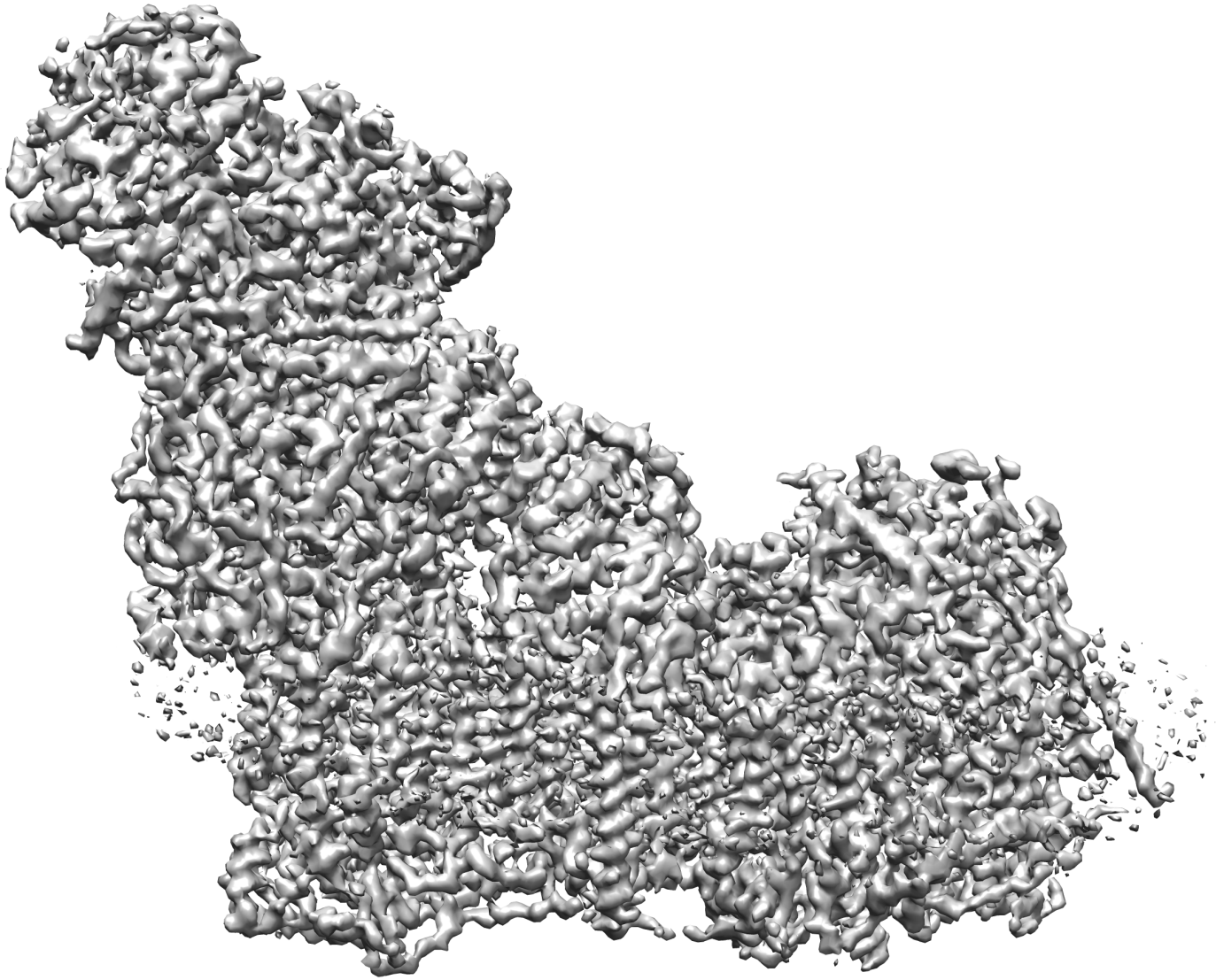


Orientation accuracy, 0.9 deg
4.5 Å
final





Resolution: features that are visible

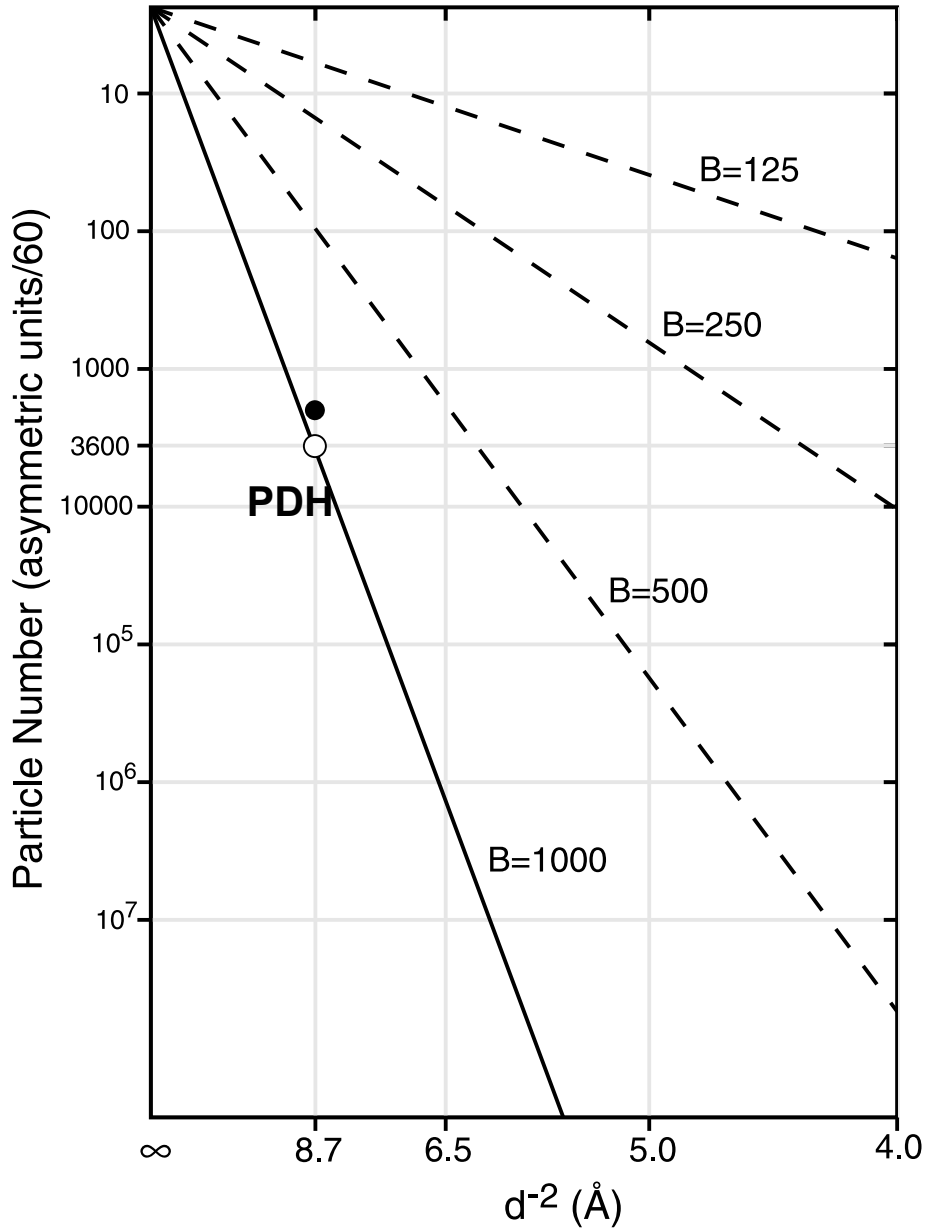


4.2 Å

No. of particles - B-factors

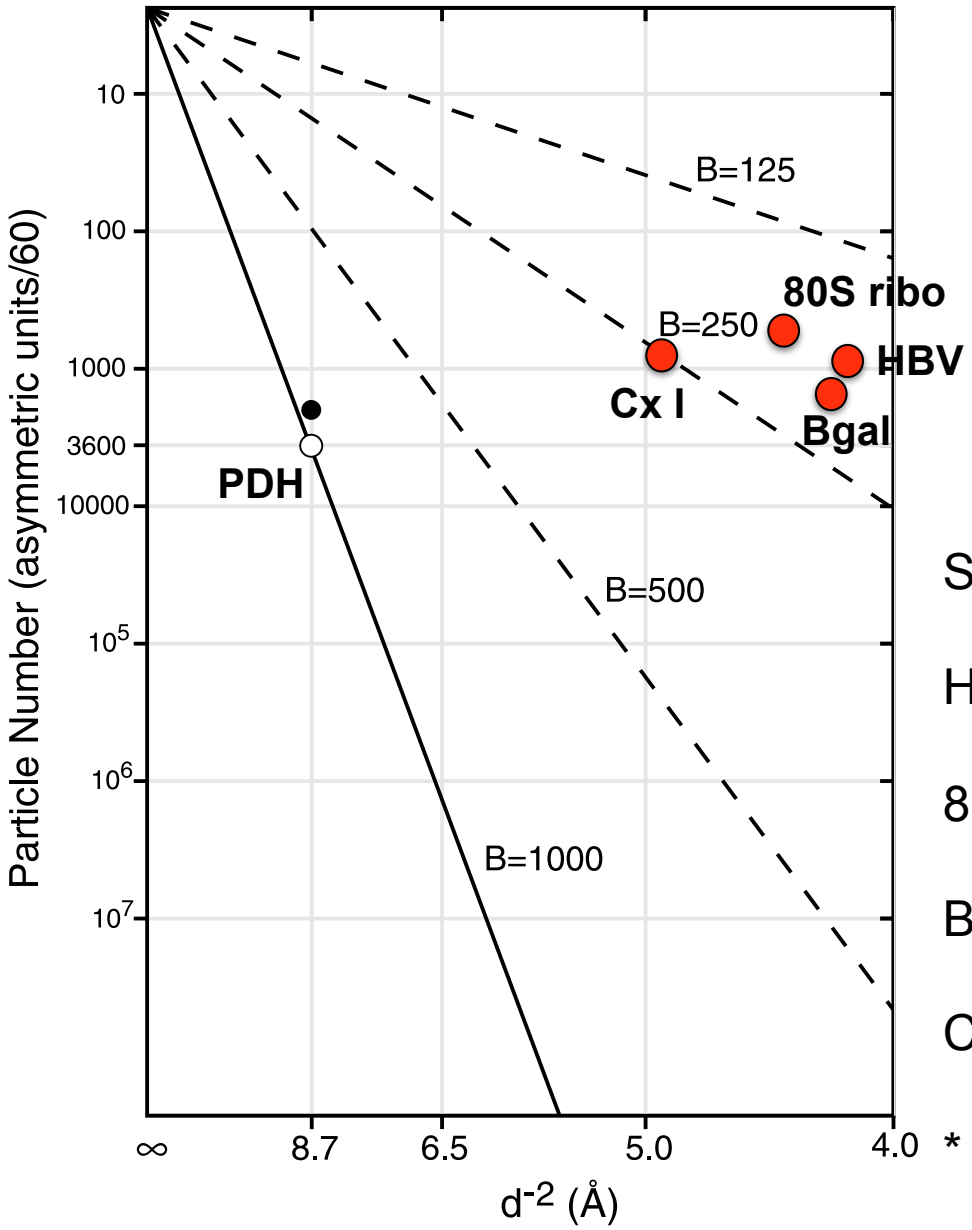
Btotal = Bimage + Bcomputational

$$N_{part} = (1/N_{asym})[(\langle S\rangle^2/\langle N\rangle^2)30\pi/N_e s_e d]e^{B/2d^2}$$



Rosenthal and Henderson 2003

No. of particles - B-factors



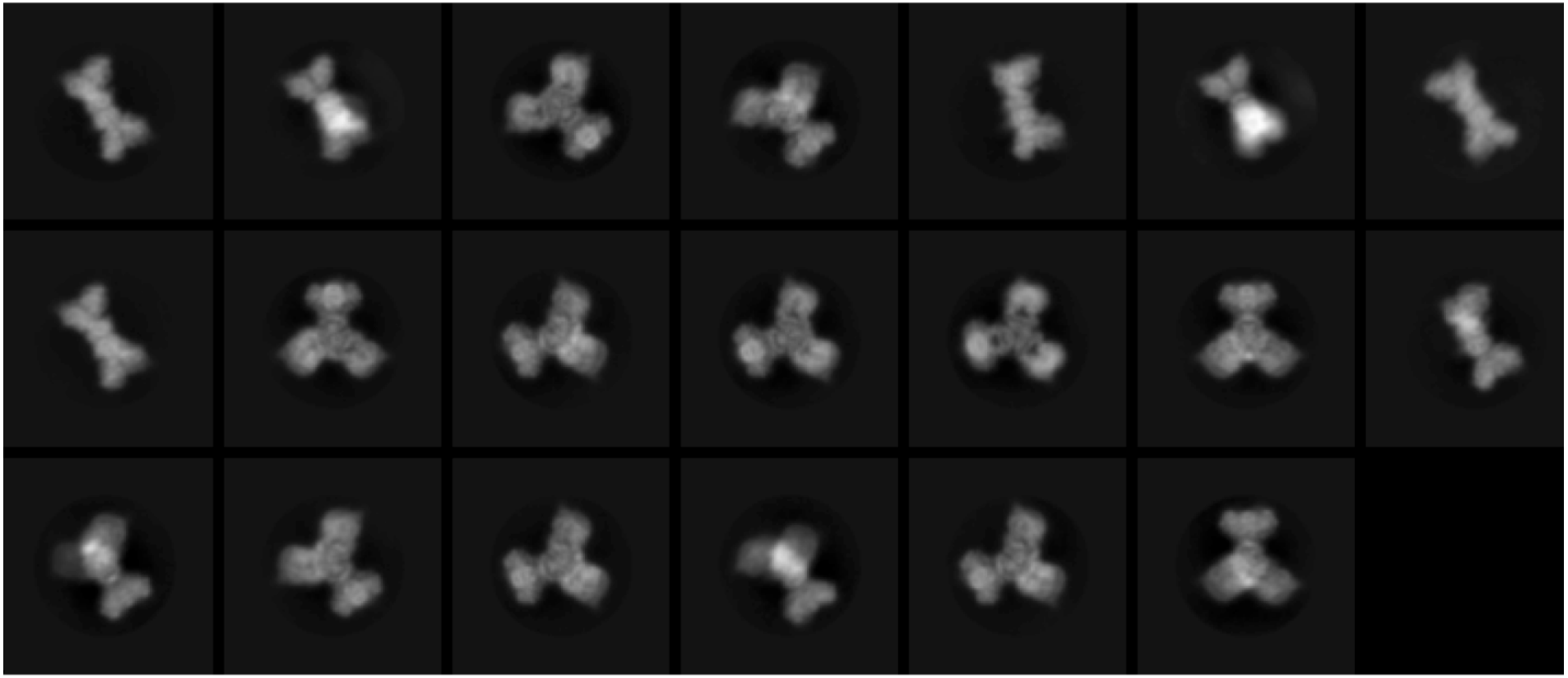
Structure	# particles	# asym units	Resolution (Å)
HBV	1078	64680	4.2
80S ribo*	35813	35813	4.5
Beta Gal	25101	100404	4.2
ComplexI	25373	25373	4.9

* - Bai et al 2012

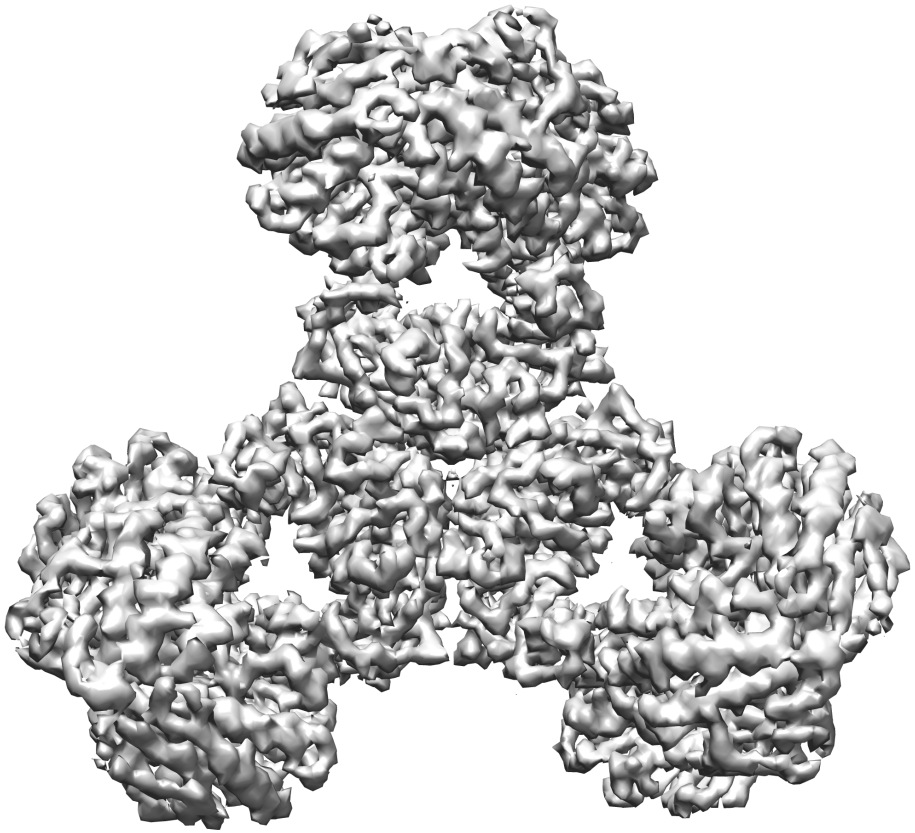
Rosenthal and Henderson 2003

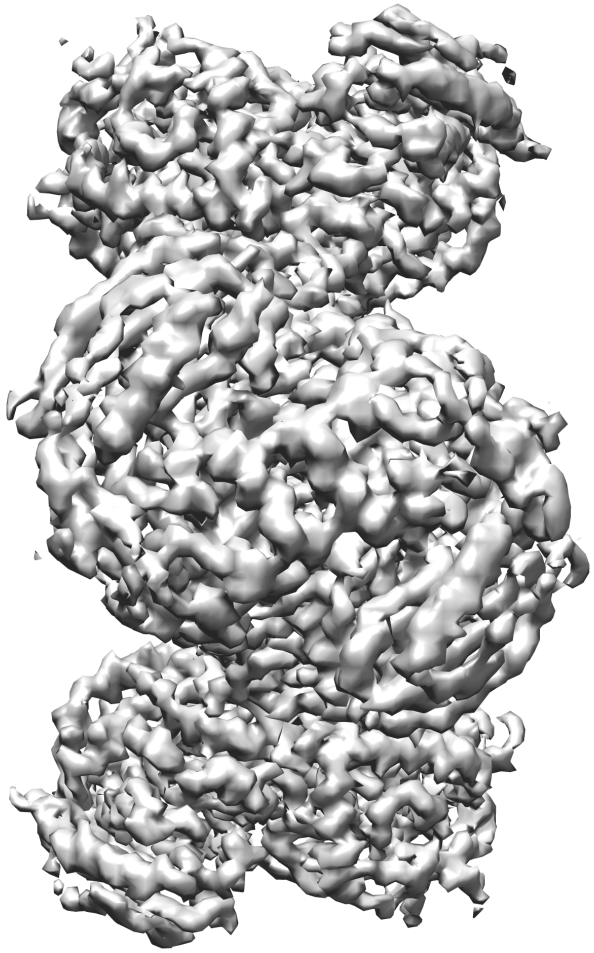
Orientation distribution

Paa Z - a bifunctional enzyme

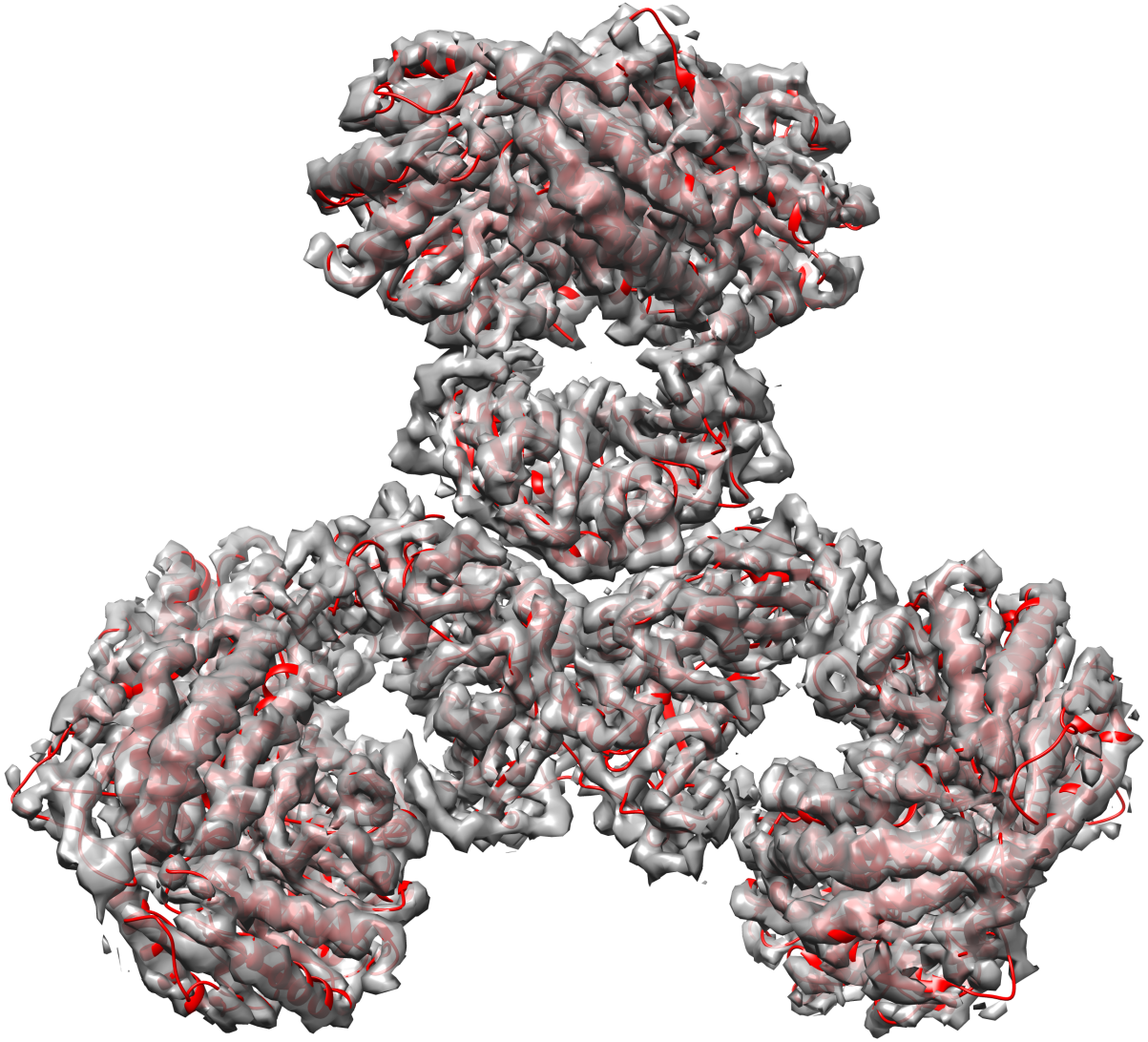


Paa Z

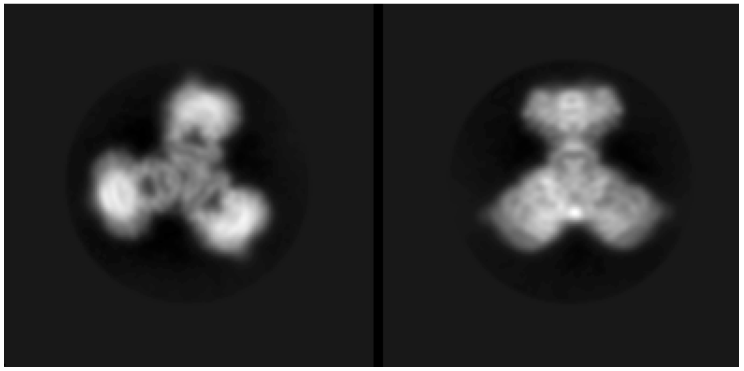


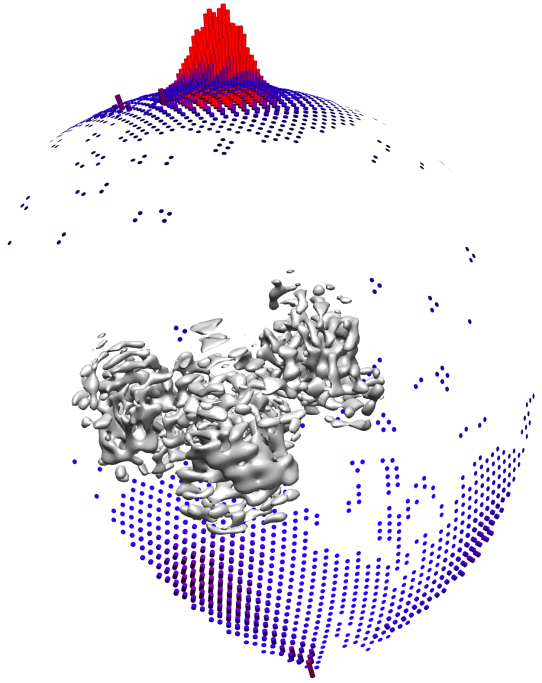


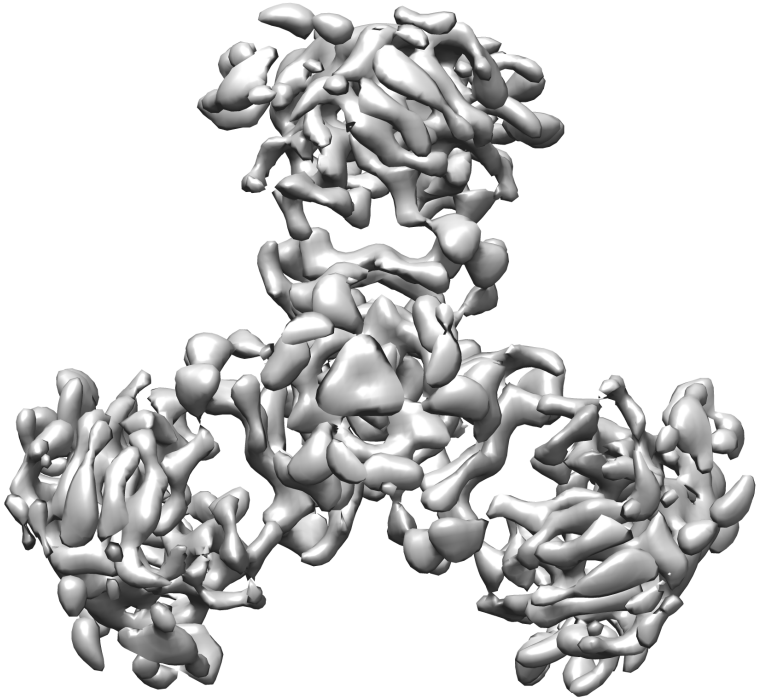
4.2 Å, 66783 orientation accuracy - 1.3

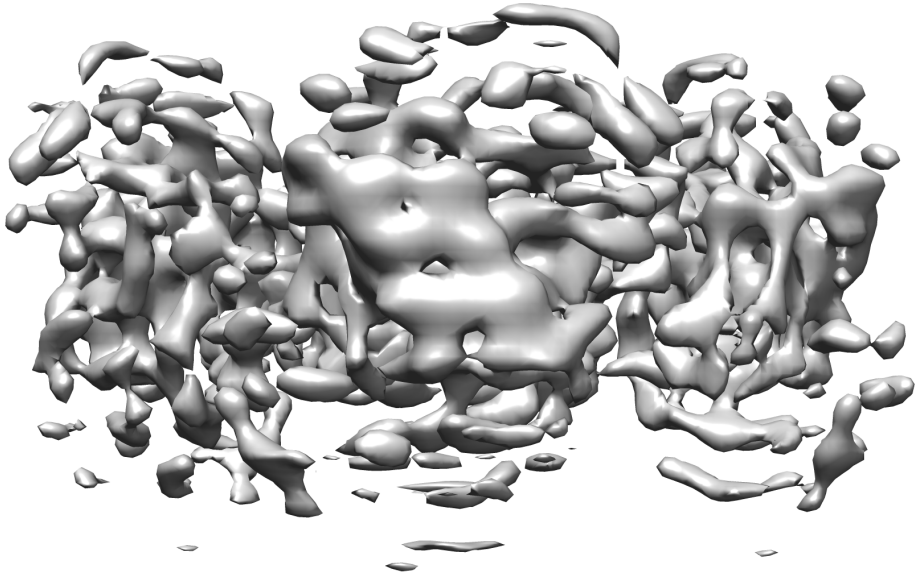


Paa Z - map, with only three-fold view



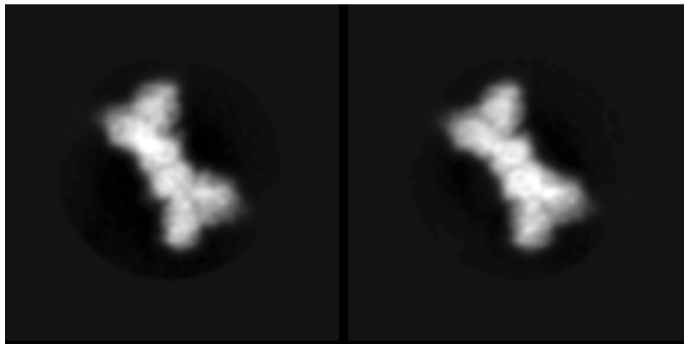


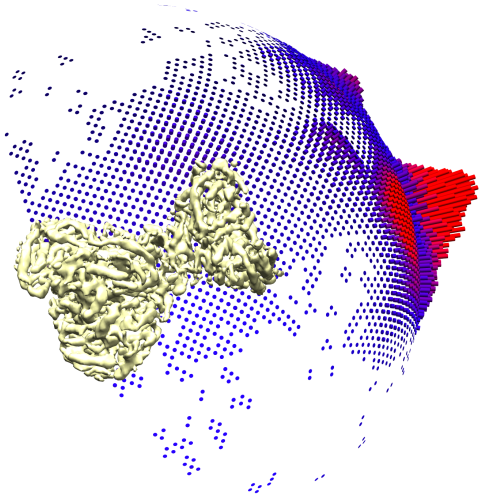


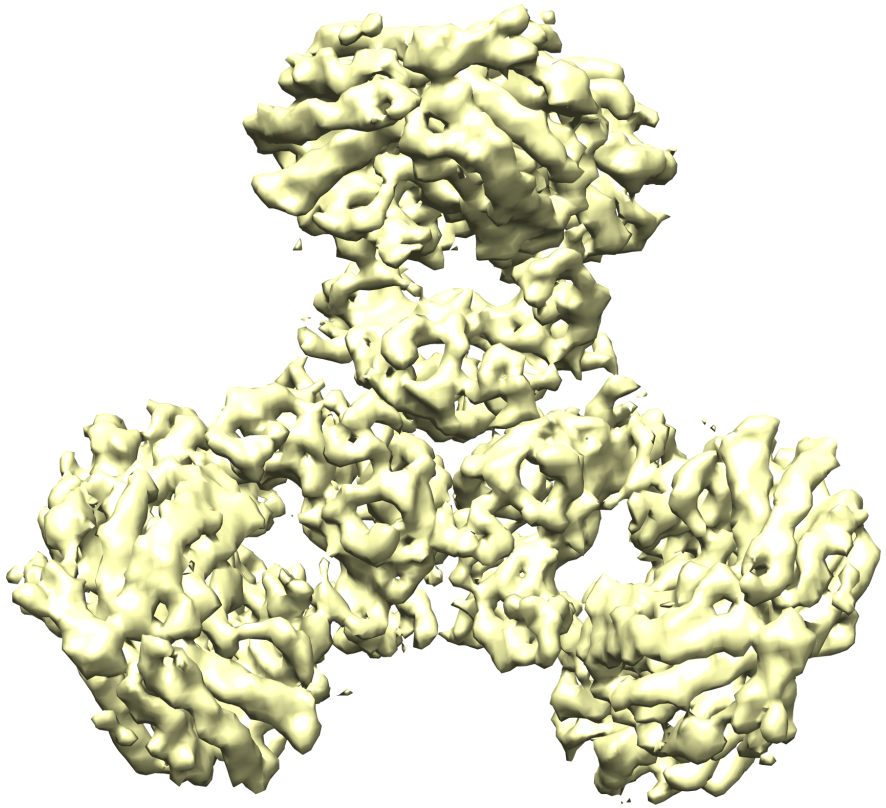


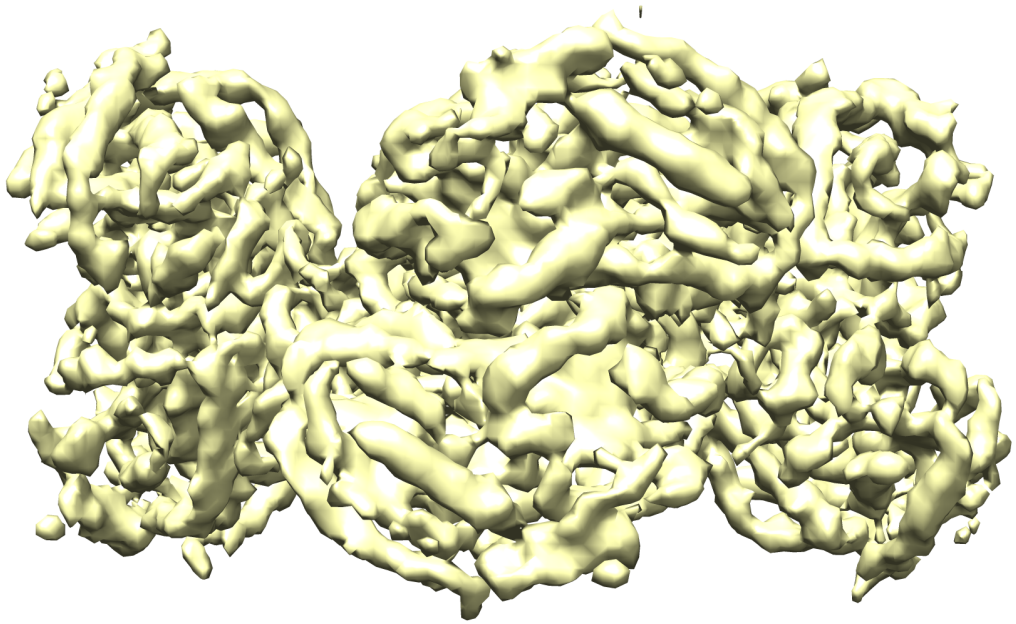
~9.7 Å, 36753, orientation accuracy - 2.48

Paa Z - map, with only side view



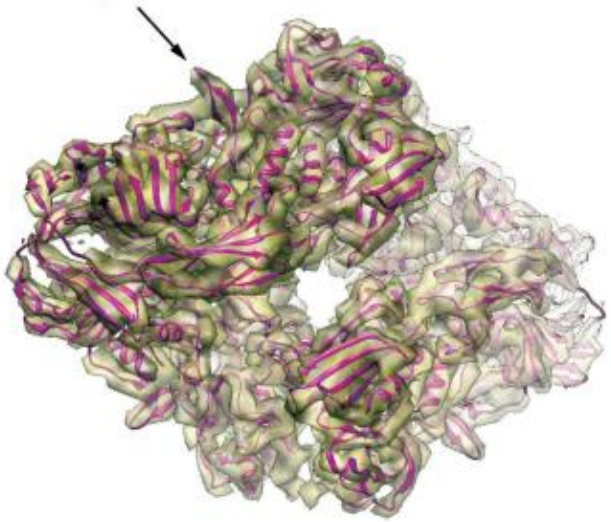


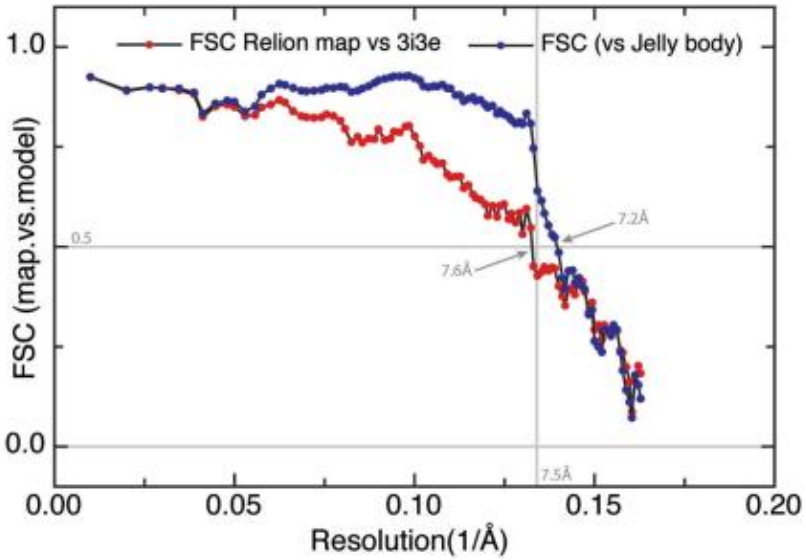


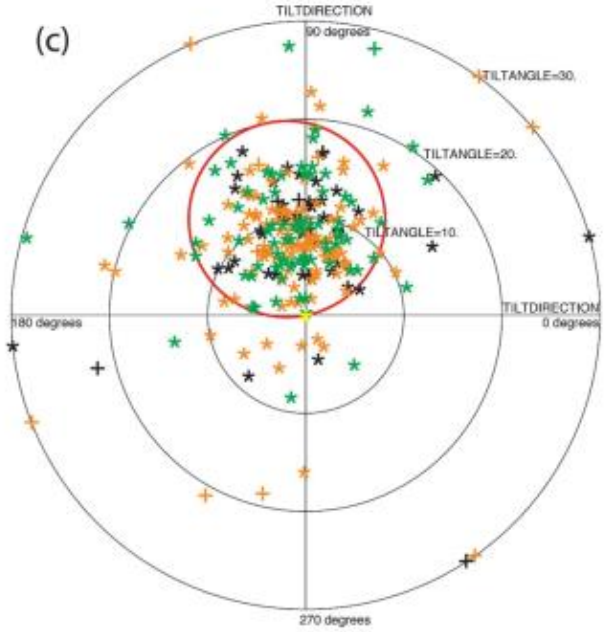


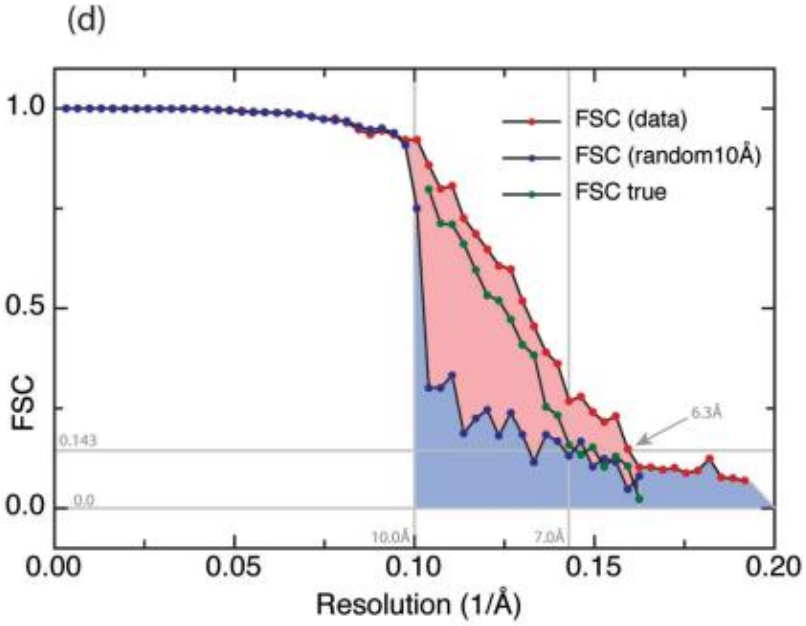
5.1 Å, 30050, orientation accuracy - 1.6

Validation tests for an EM map









In practice - some thoughts when starting to process data

- Perform 2D class averaging before 3D reconstruction!
  - Powerful means of cleaning data set
  - Quality of classes -> quality reconstruction! (when using direct detectors high resolution structural features - alpha helices are already observed)
  - Play with number of classes
    - Cryo-EM: 100-200 particles/class
    - Negative stain: <50 particles/class
- Check for 3D heterogeneity!
  - Remember: proteins are machines, highly dynamic …
  - Again: play with number of classes (some biochemical knowledge can be used to determine the classes but also to interpret if the classes are biologically meaningful)
- Smaller, high-quality data sets are often better than large and dirty ones
  - Cheaper computationally
  - Cleaner reconstructions
  - final B-factor of map is a useful measure for the quality of the data

Further Reading

- Penczek, Fundamentals of Three-Dimensional Reconstruction from Projections, *Methods in Enzymology*, , **482** (2010) p 1
- Penczek, Image restoration in cryo-electron microscopy, *Methods in Enzymology*, , **482** (2010) p 35
- Sigworth, Doerschuk, Carazo & Scheres, An Introduction to Maximum-Likelihood Methods in Cryo-EM, *Methods in Enzymology*, **482** (2010) p 263
- Scheres, Classification of Structural Heterogeneity by Maximum-Likelihood Methods, *Methods in Enzymology*, **482** (2010) p 295
- Rosenthal & Henderson, Optimal determination of particle orientation, Absolute Hand, and Contrast Loss, J.Mol.Biol., **333** (2003), p 721
- Scheres, A Bayesian View on Cryo-EM structure determination, J.Mol.Biol., **415** (2012) p 406
- Scheres, RELION: Implementation of a Bayesian approach to cryo-EM structure determination, J.Stru.Biol., **180** (2012) p 519