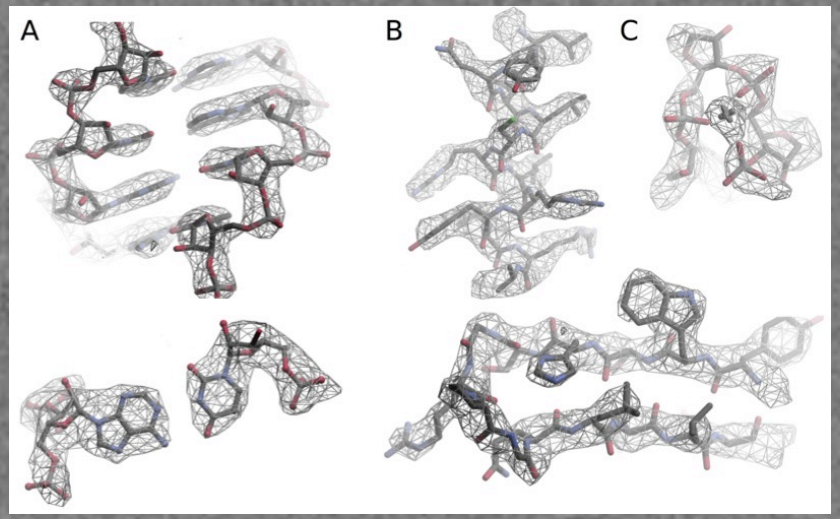
Methods for heterogeneity analysis

(with bias)



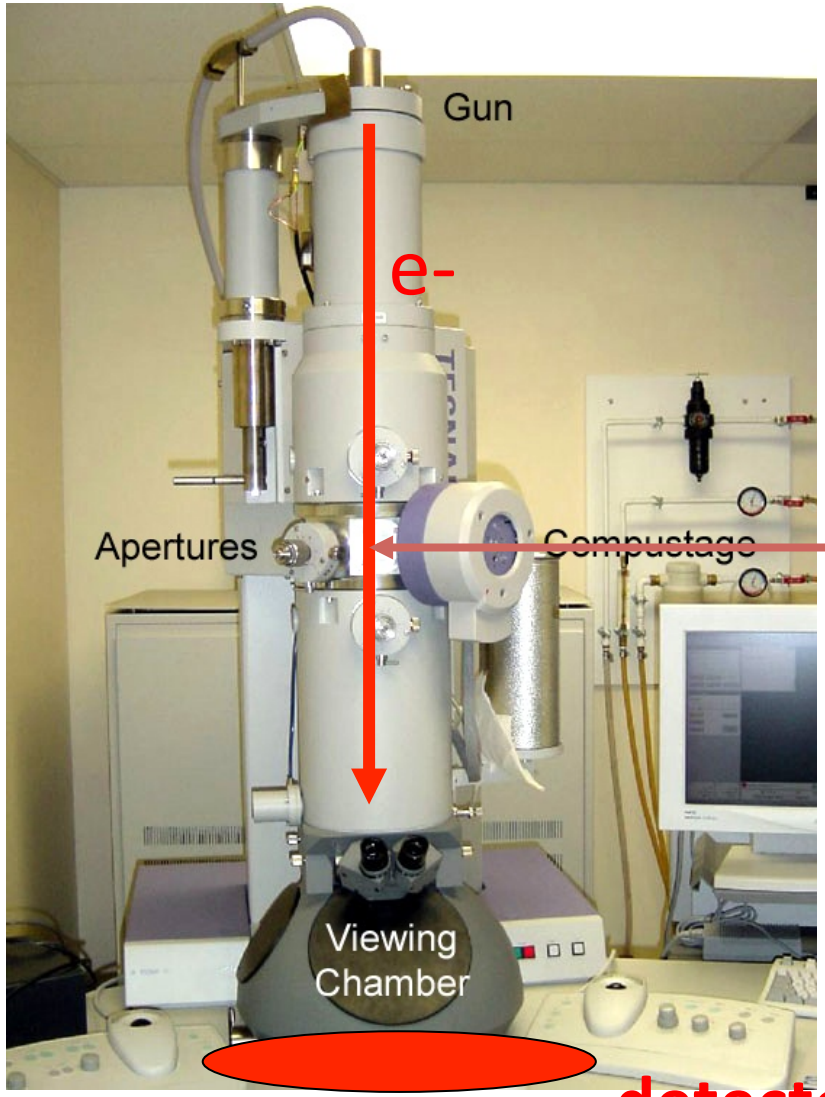
Sjors H.W. Scheres

EMAN workshop, Houston, October 2015

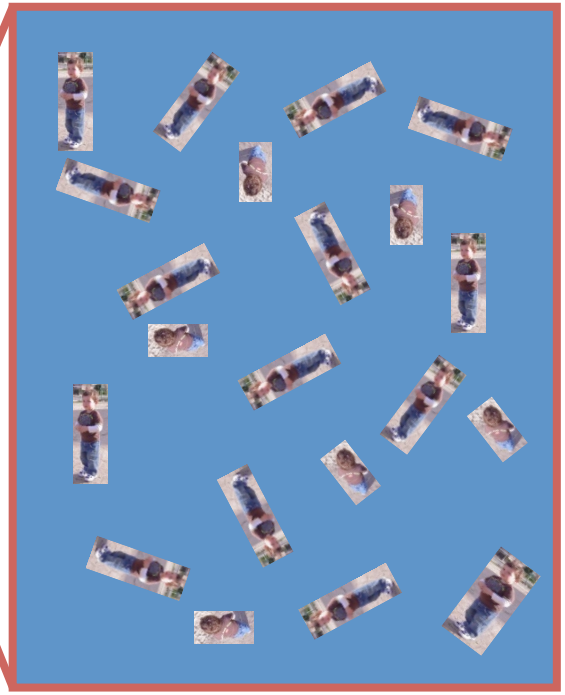
MRC Laboratory of Molecular Biology

An example “protein”

Experimental setup

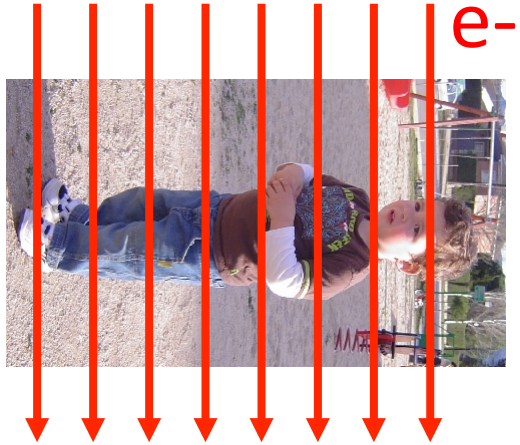


sample



Electron microscopy imaging

3D object



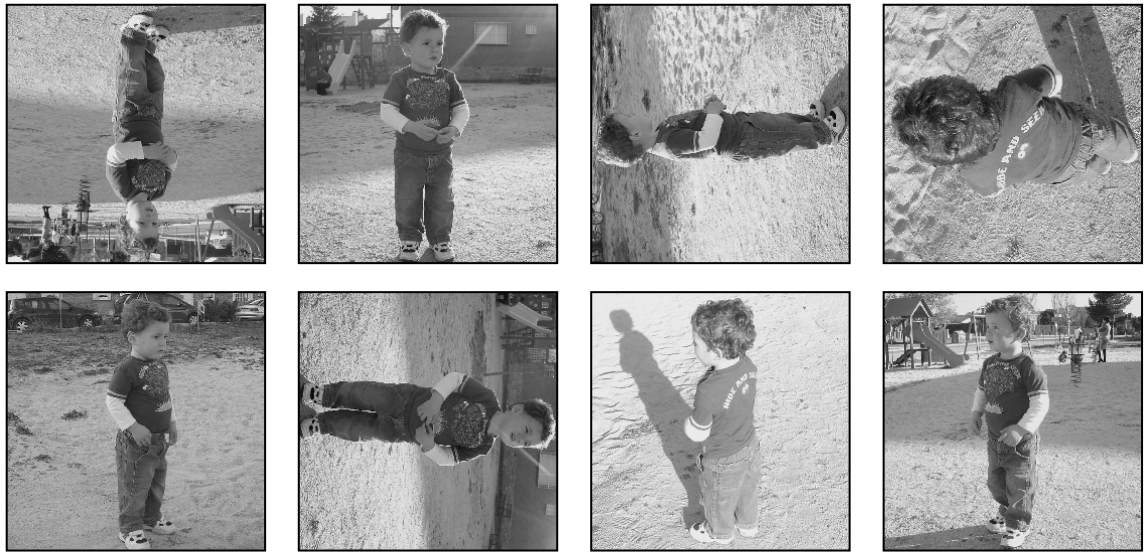
2D projection

We collect data in 2D,

but we want 3D info!

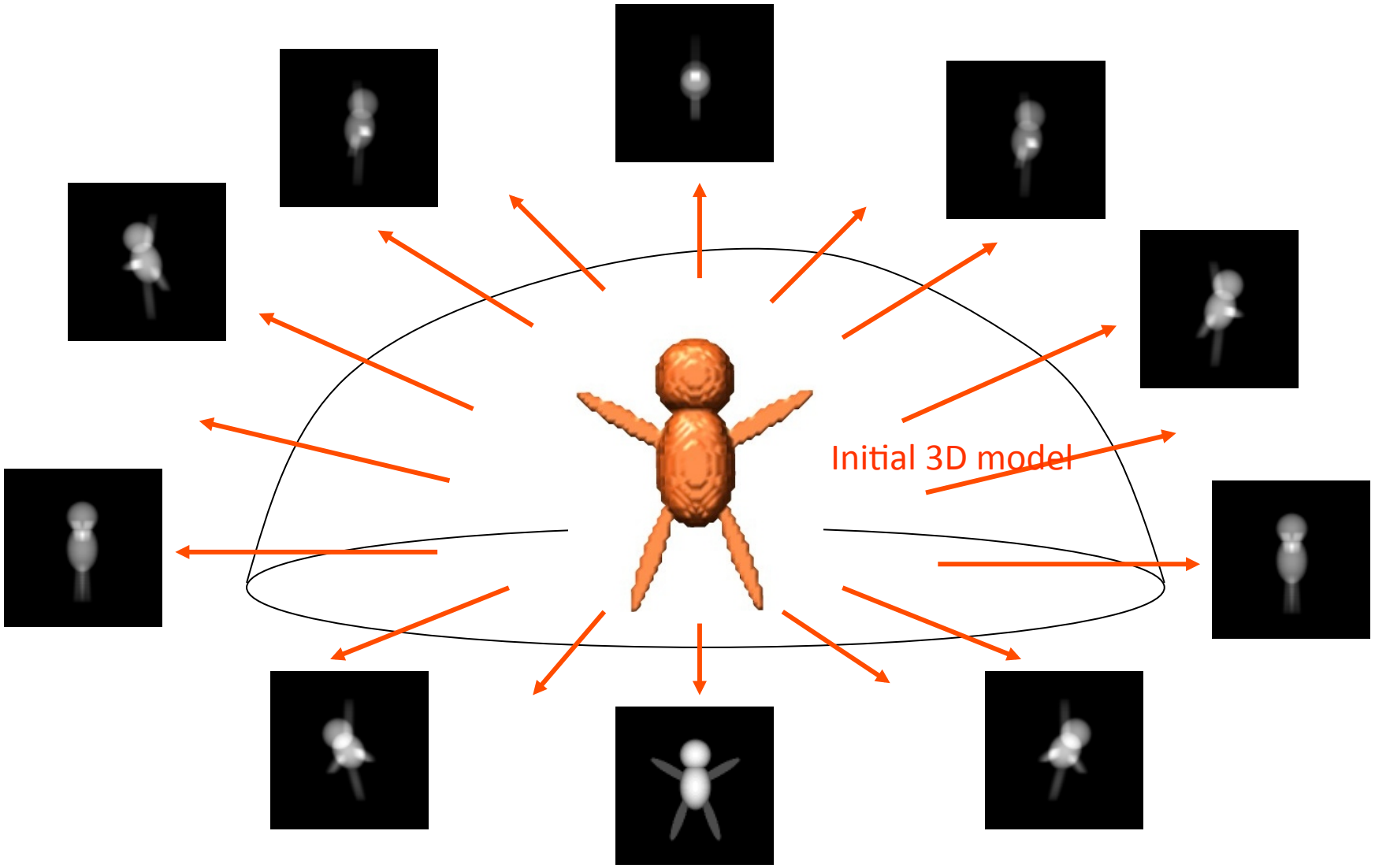
Single particle analysis

- Embedded in ice: many unknown orientations

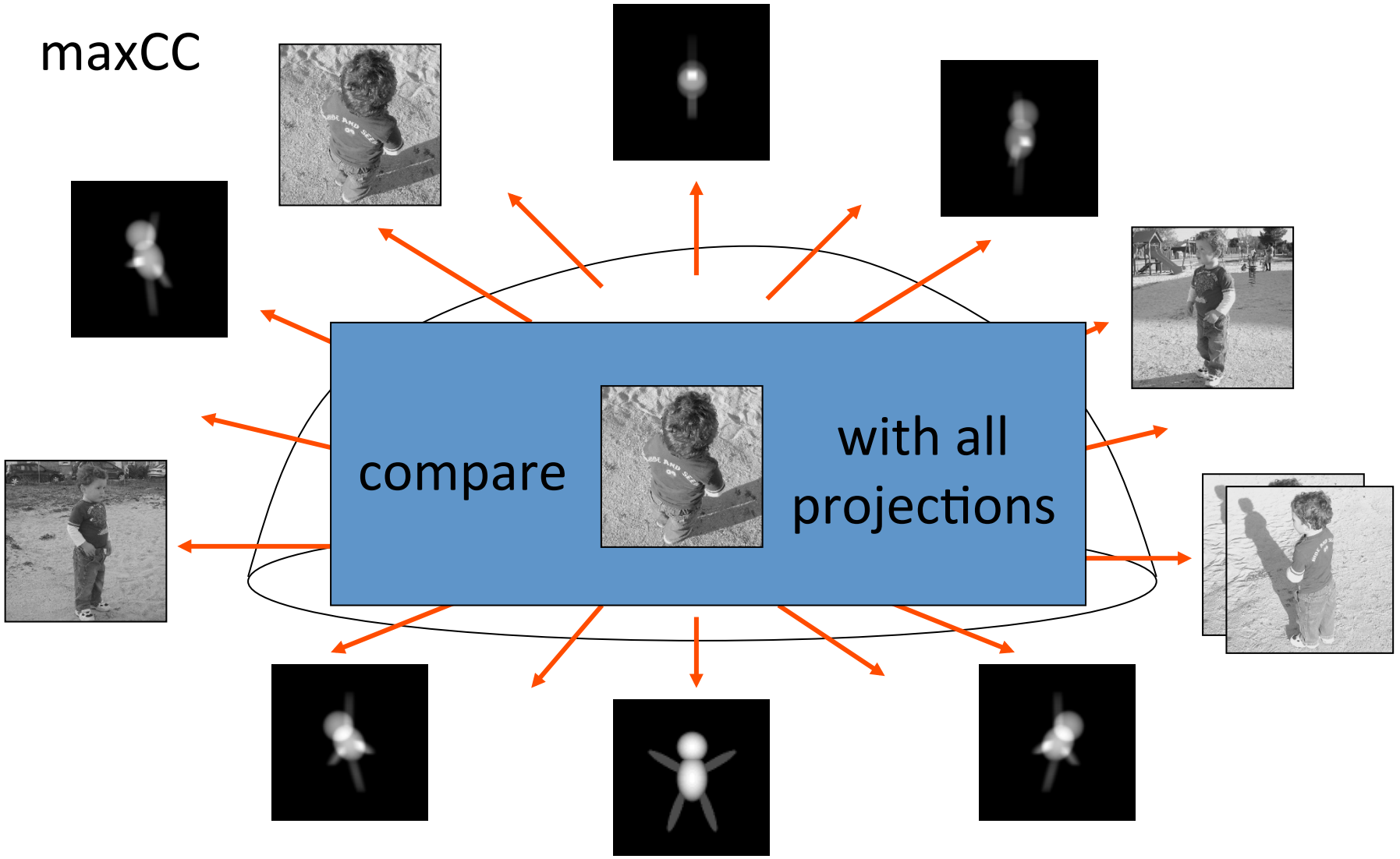


- Combine all 2D projections into a 3D reconstruction

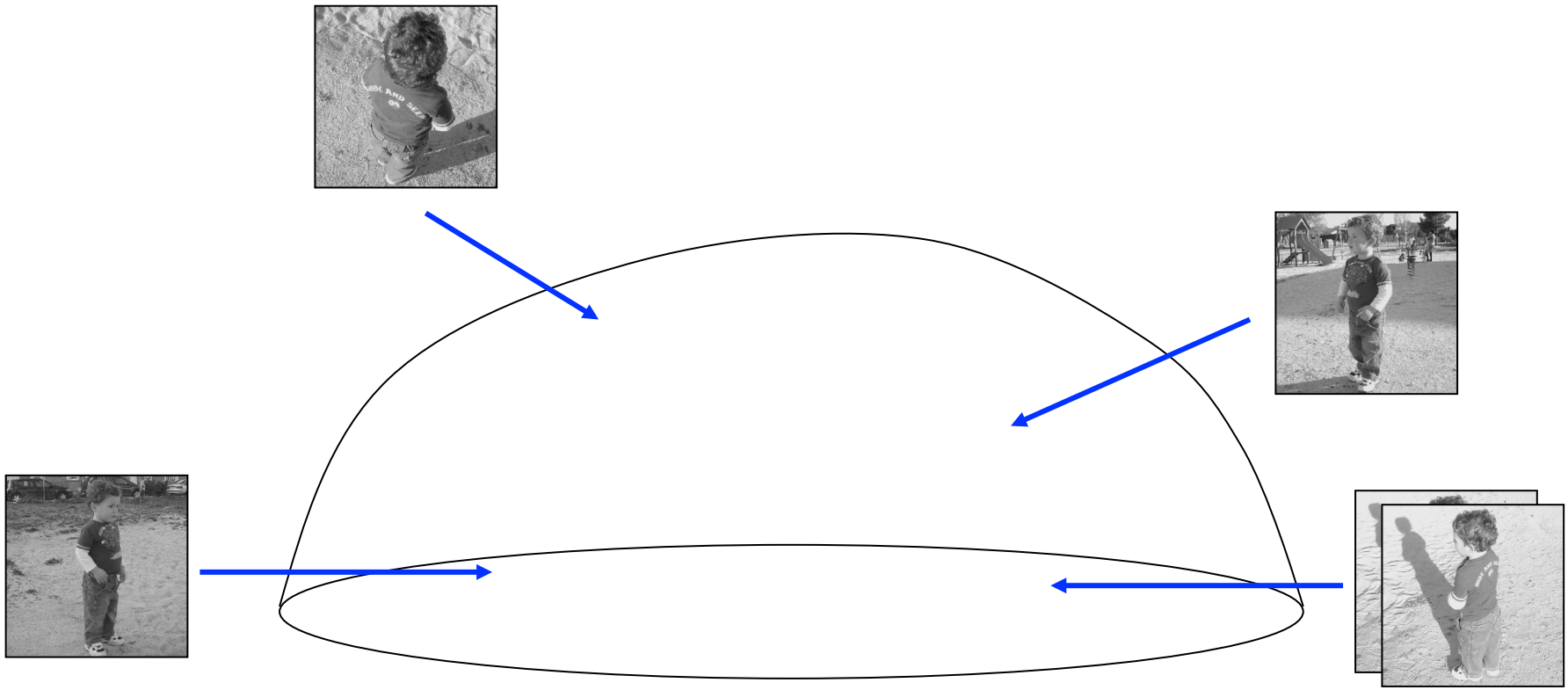
Projection matching



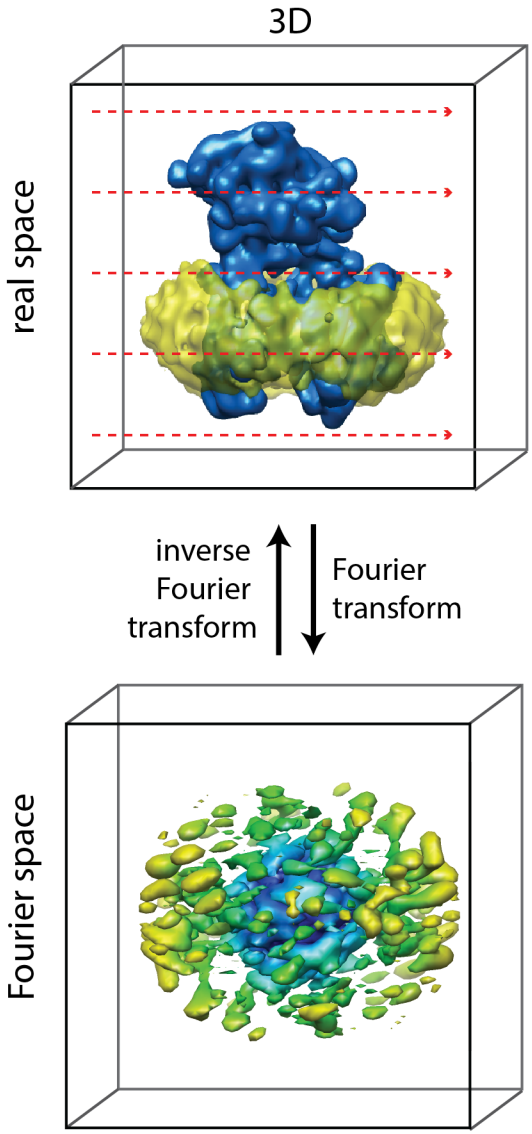
Projection matching



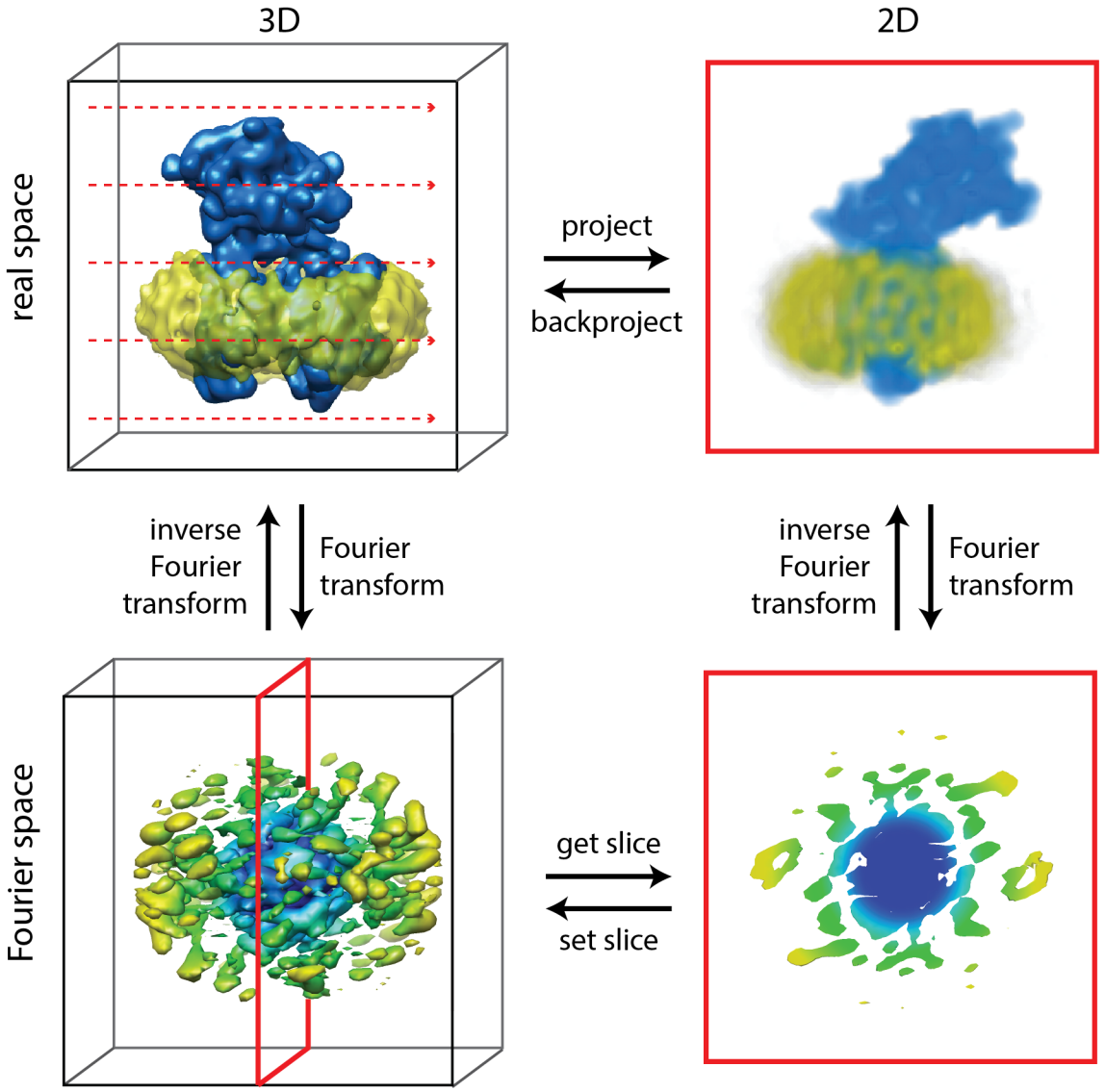
3D reconstruction



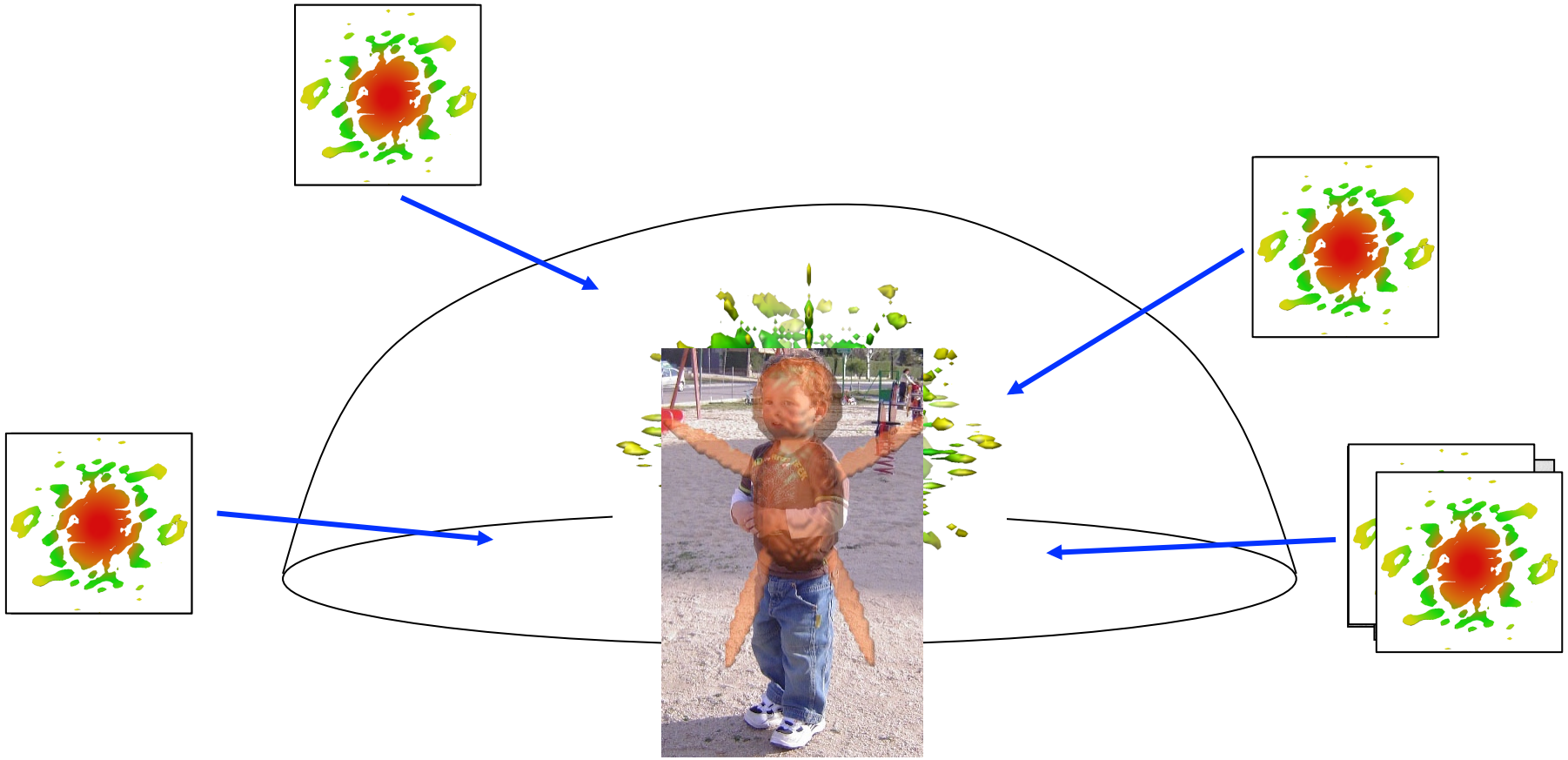
Projection slice theorem



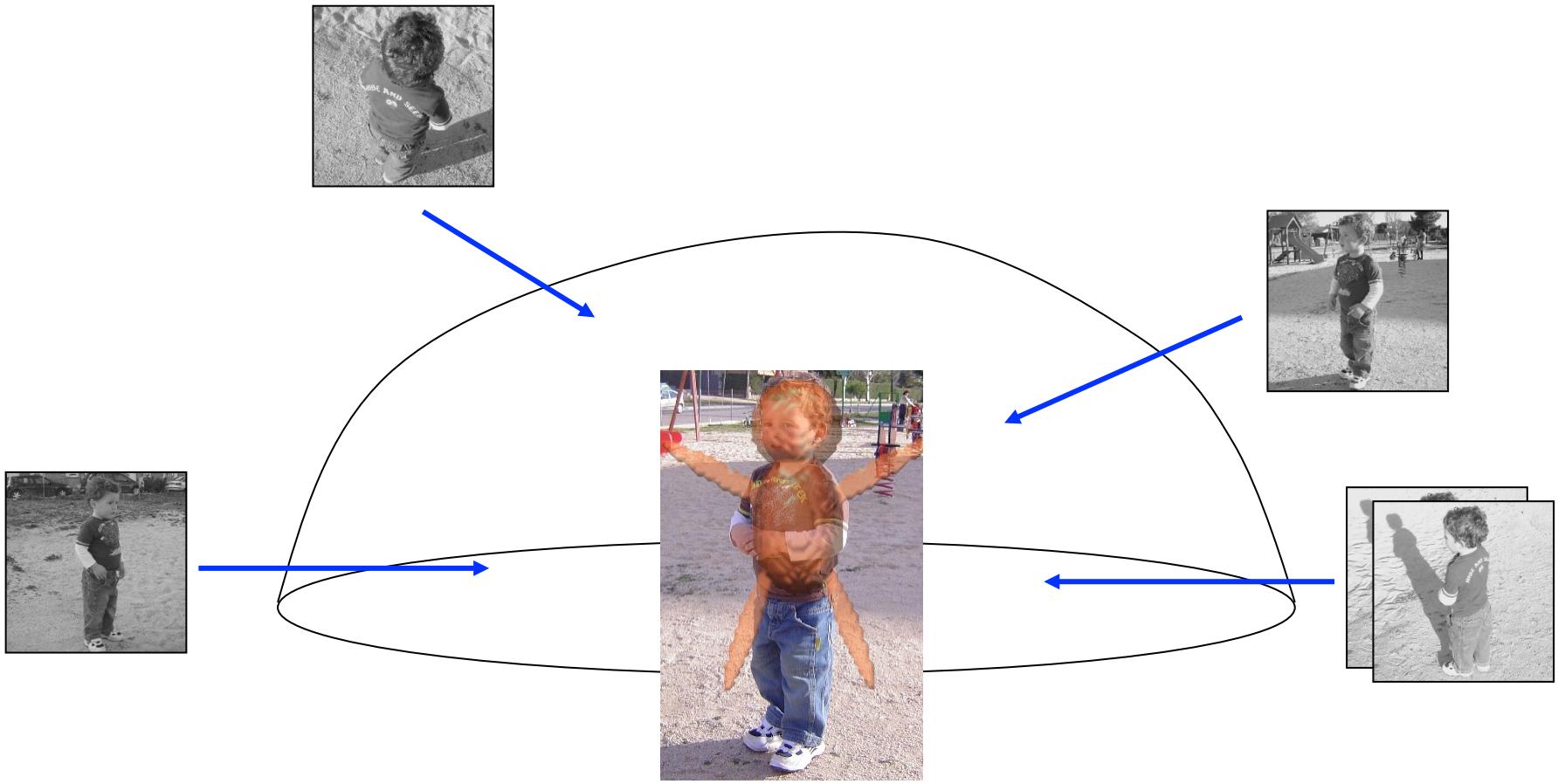
Projection slice theorem



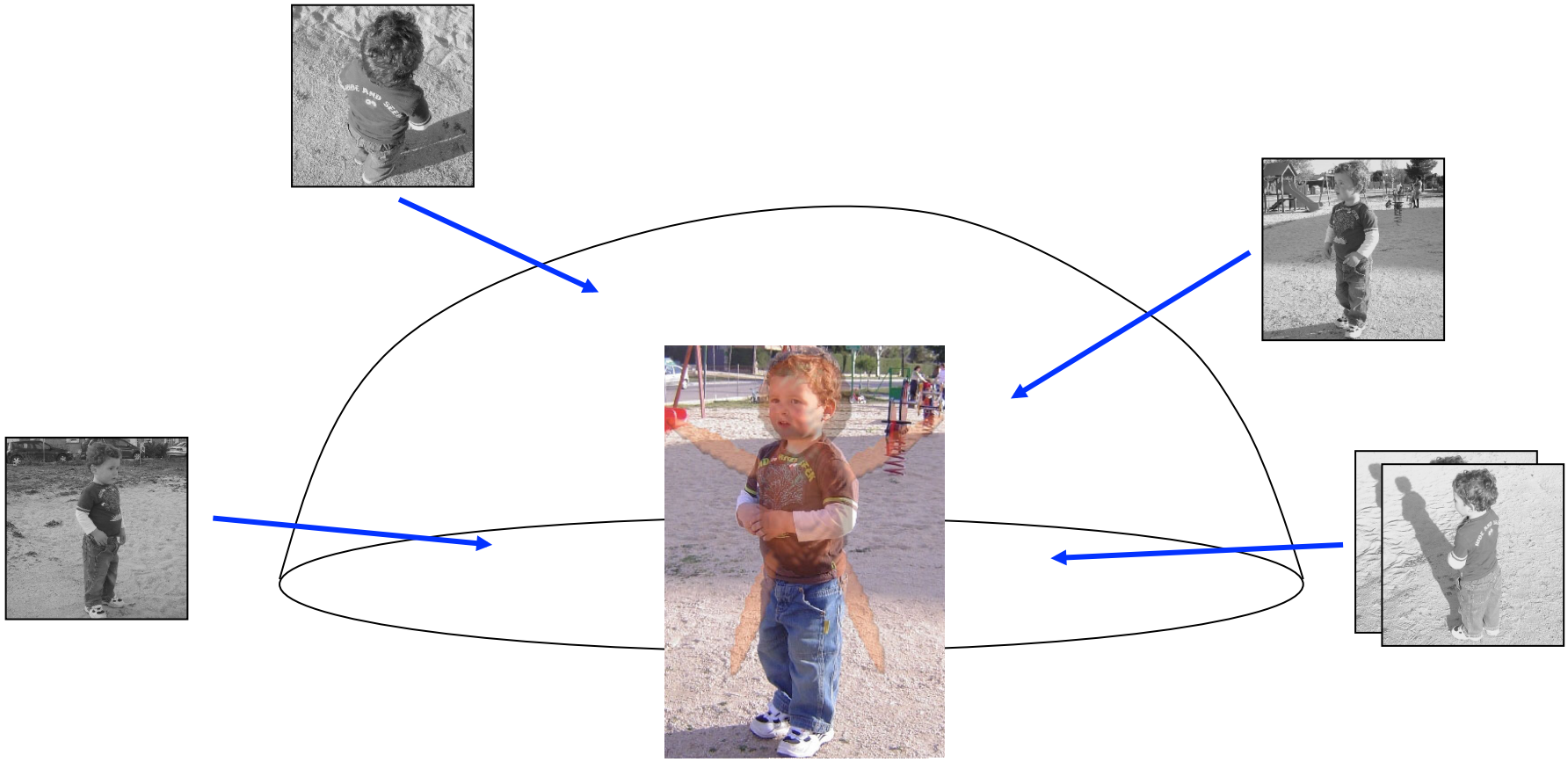
Iterative refinement



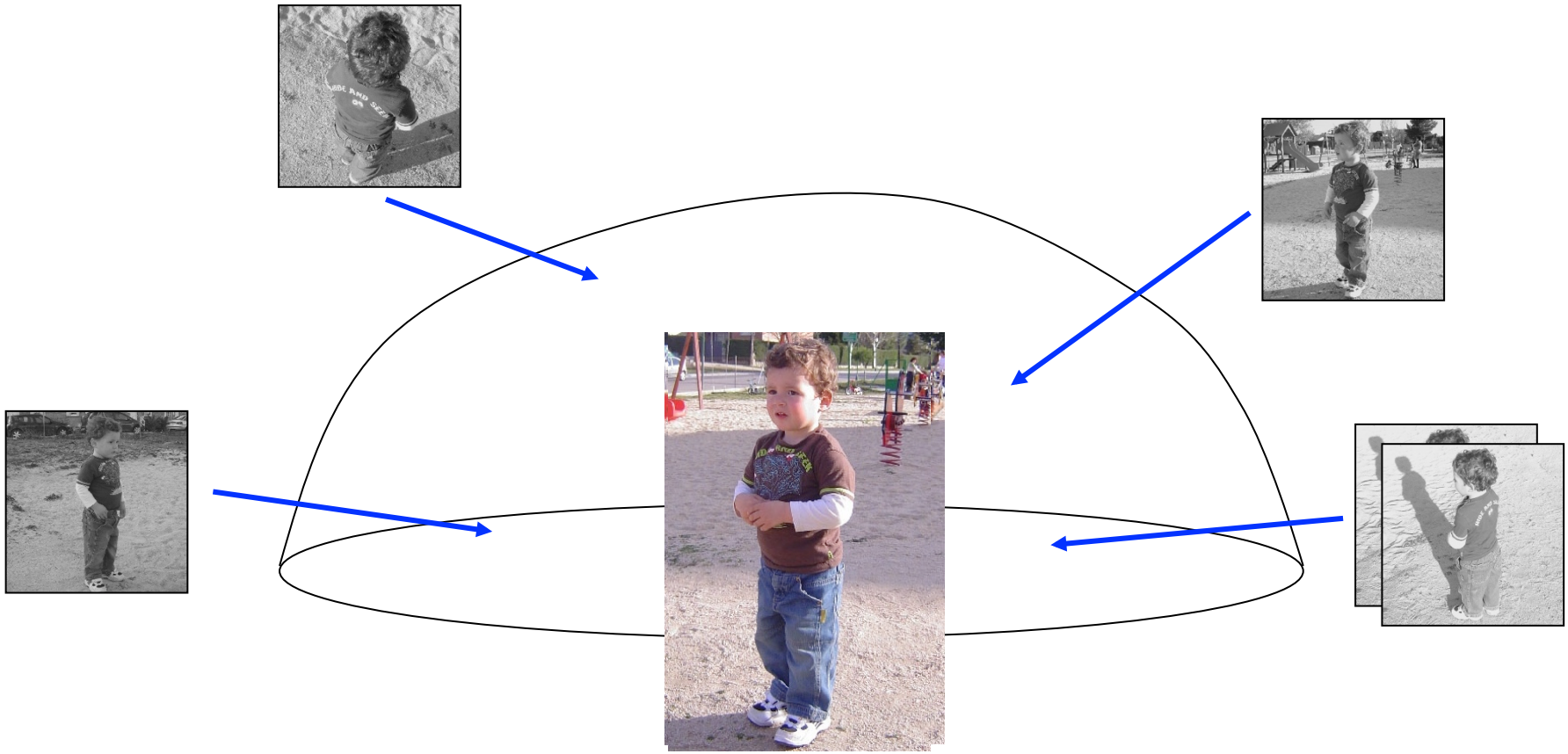
3D reconstruction



Iterative refinement

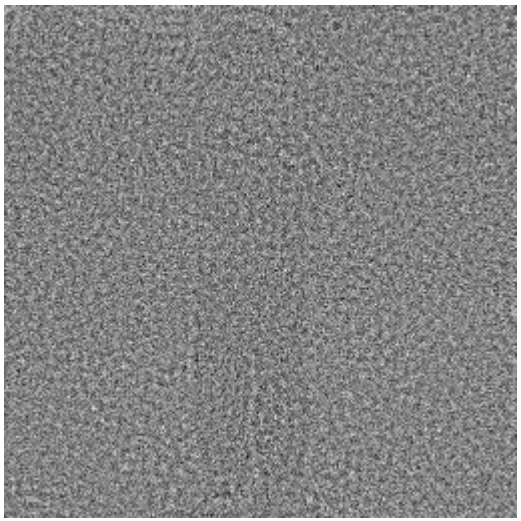


Iterative refinement



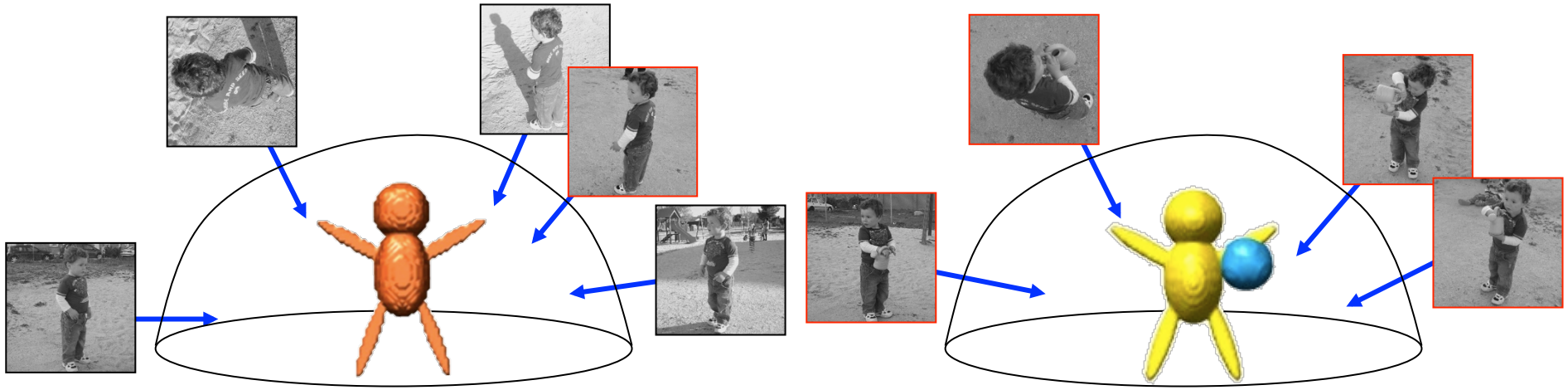
Further inconveniences

- Defocussing & microscope imperfections introduce artefacts (-> CTF correction)
- Low dose: large amounts of noise
- Structural heterogeneity!

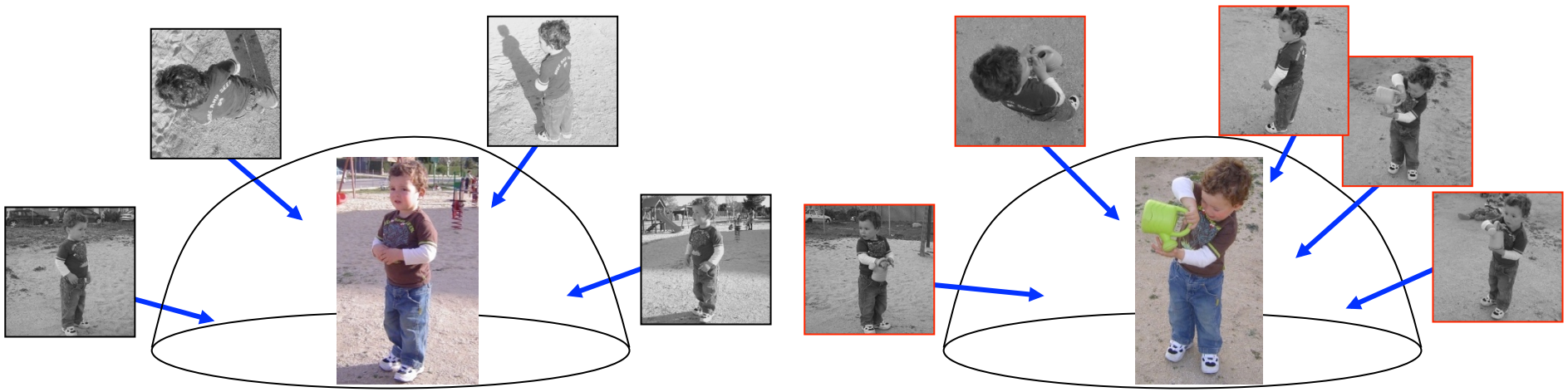


All samples are structurally heterogeneous!

Multi-reference refinement

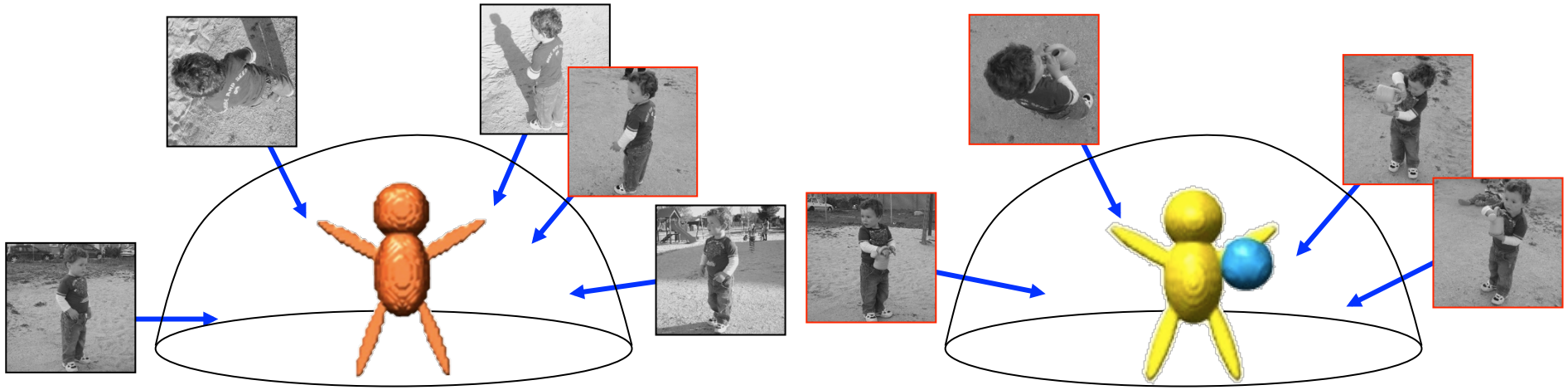


Multi-reference refinement



Supervised classification

(developed in the Frank lab)

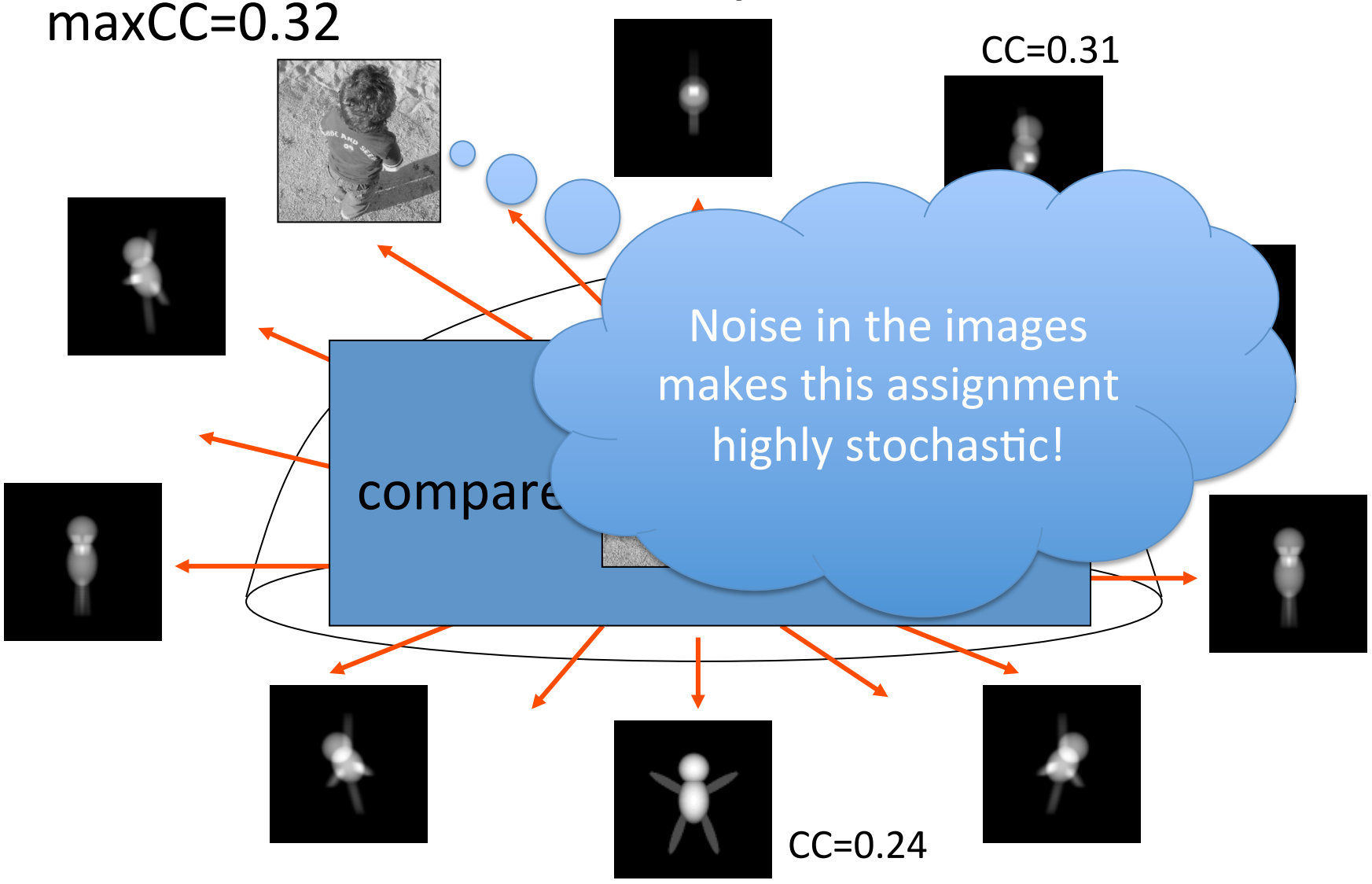


You kind-of need to know the answer already….

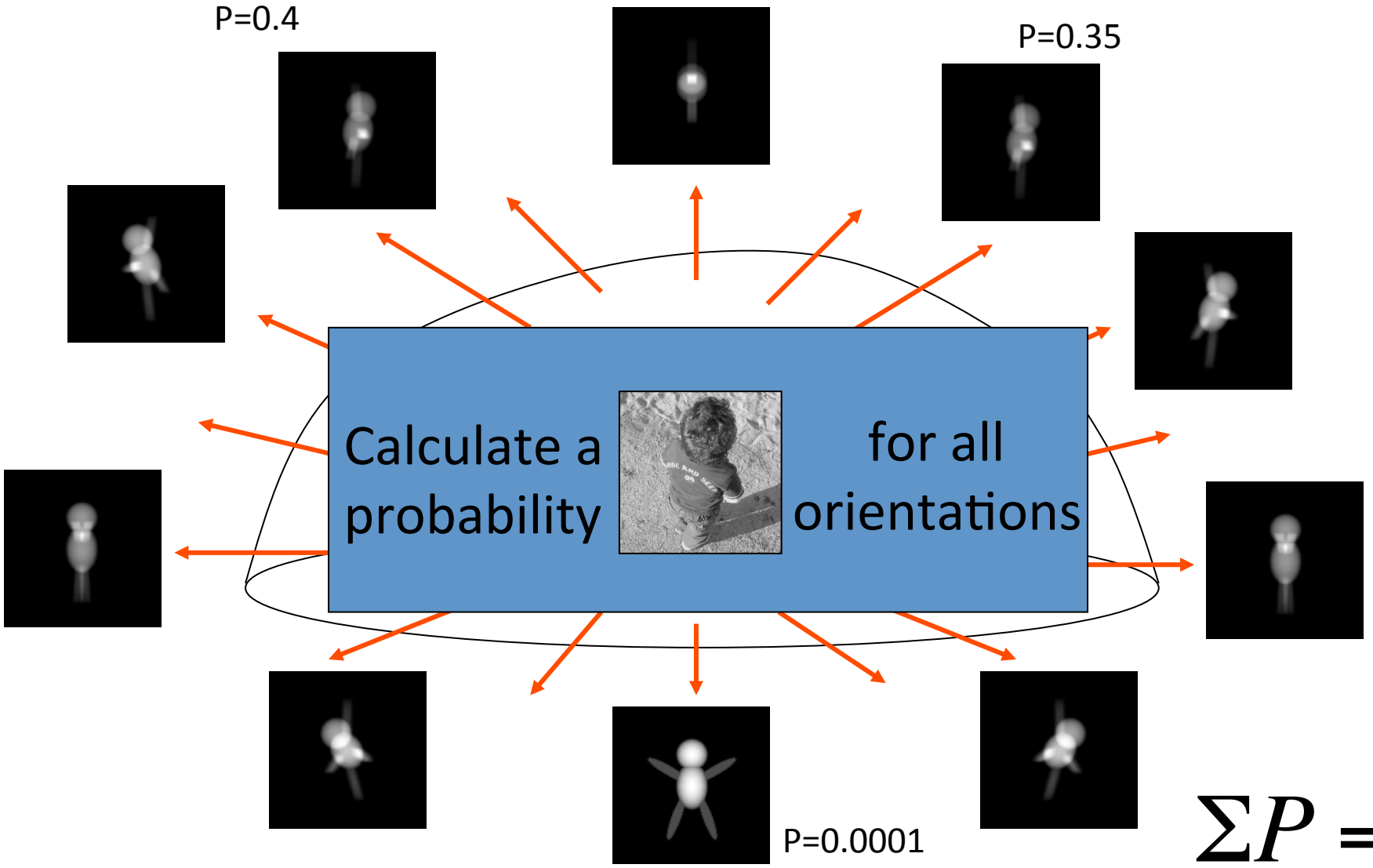
Maximum-likelihood approaches

- Marginalize over orientations & classes
  - Probability-weighted assignments
- First described by Fred Sigworth (JSB, 1998)
  - For 2D-alignment, single-reference
  - Real-space data model (white-noise model)
  - Matlab scripts
- Then extended for 2D & 3D classification (2005-2010)
  - XMIPP Scheres et al, JMB 2005; Nat Methods 2007;
- 3D ML-based classification without marginalizing over orientations
  - FREALIGN
    Lyumkis et al, JSB, 2013

Maximum cross-correlation (least-squares)

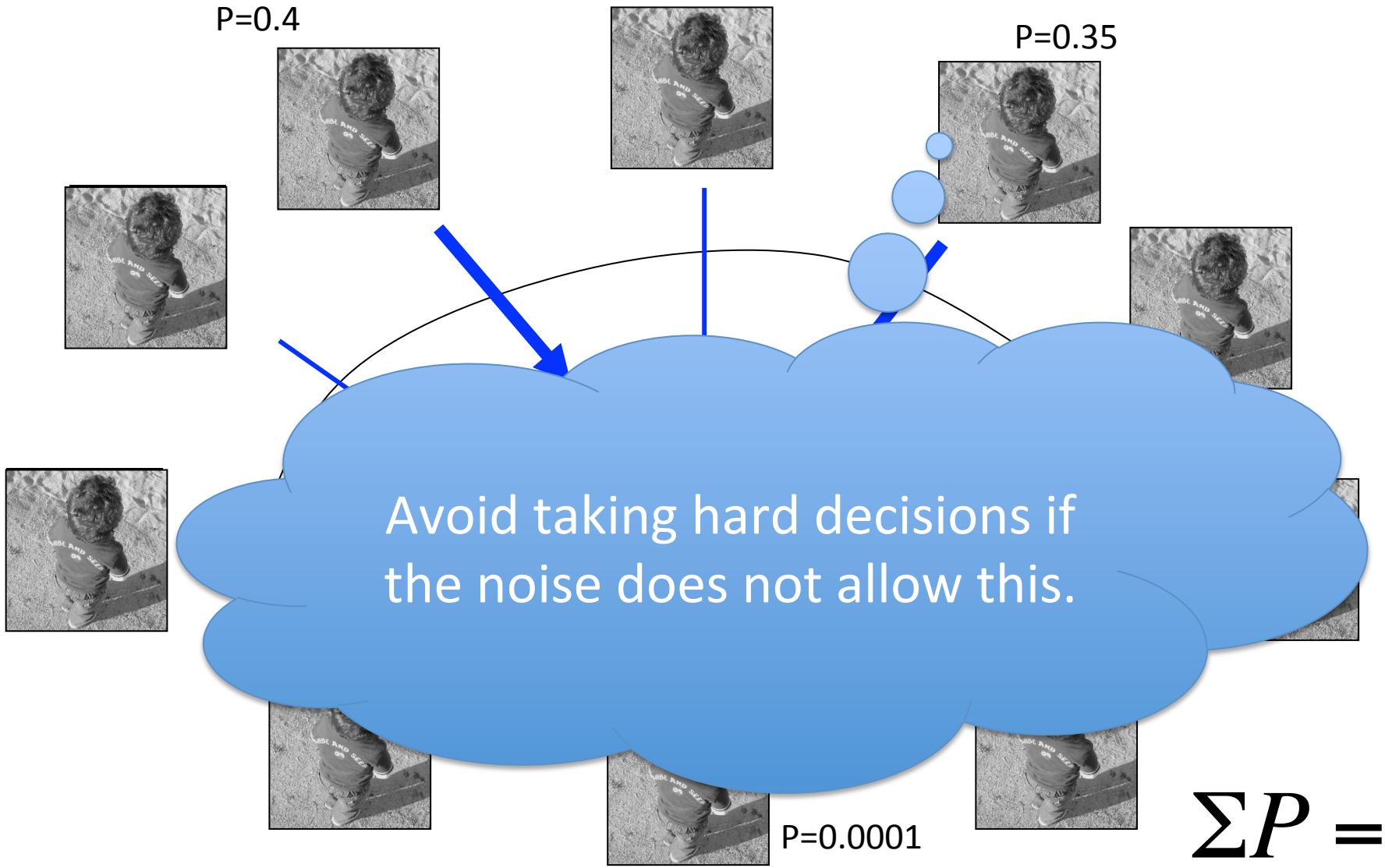


Maximum likelihood



$$\Sigma P = 1$$

Maximum likelihood



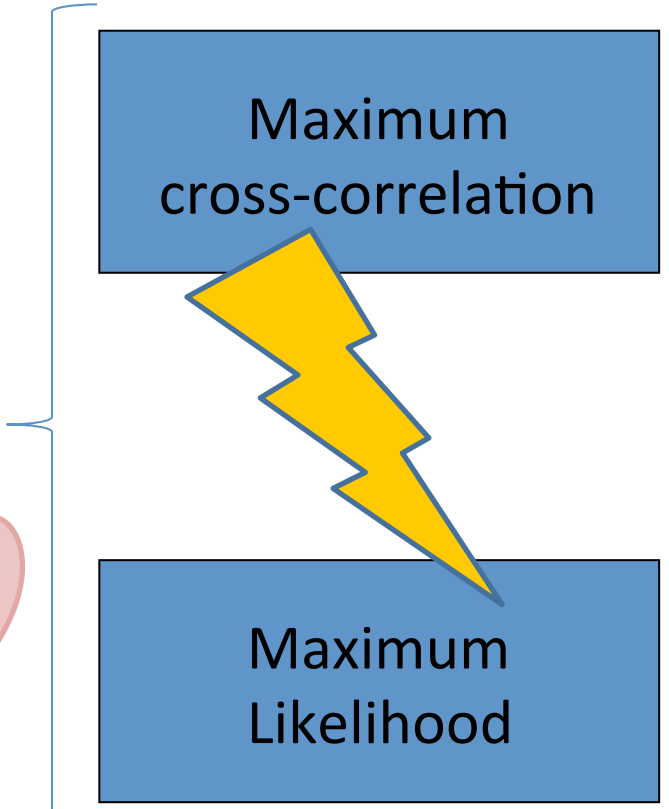
$$\Sigma P = 1$$

Incomplete data problems

- Option 1: add *Y* to the model

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

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



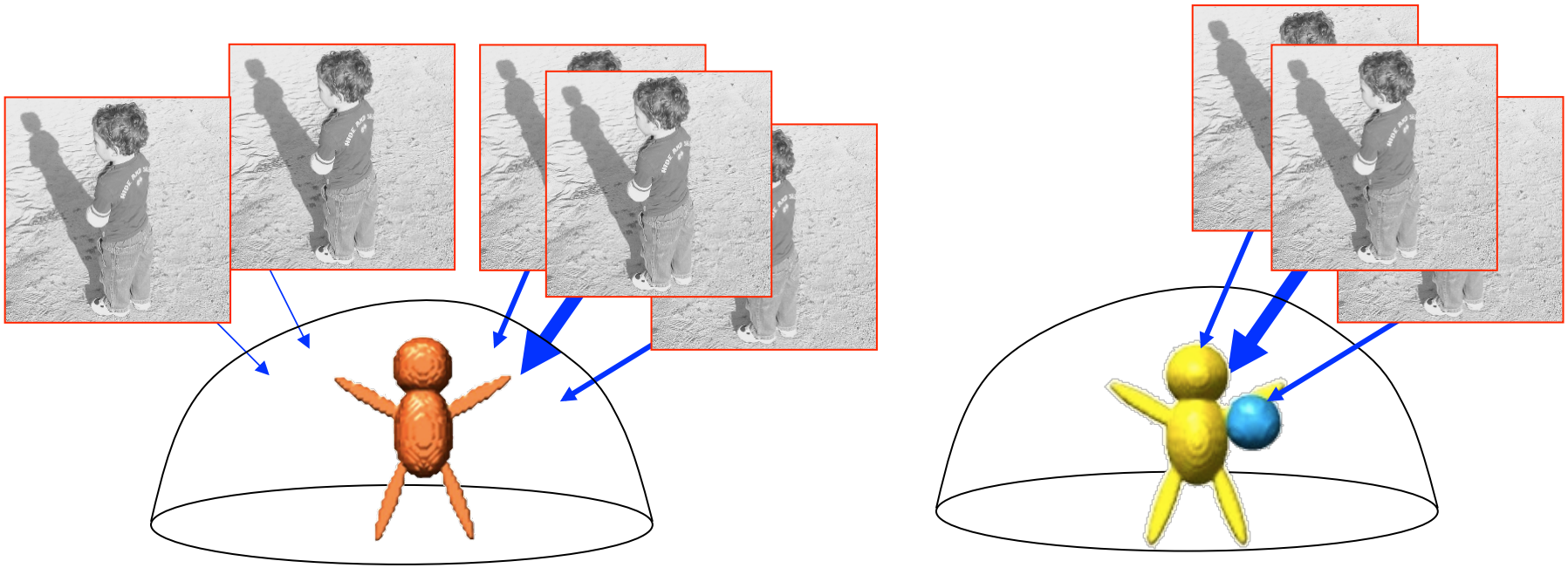
- Option 2: marginalize over *Y*

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

Probability of X, regardless Y

Read more? See *Methods in Enzymology*, **482** (2010)

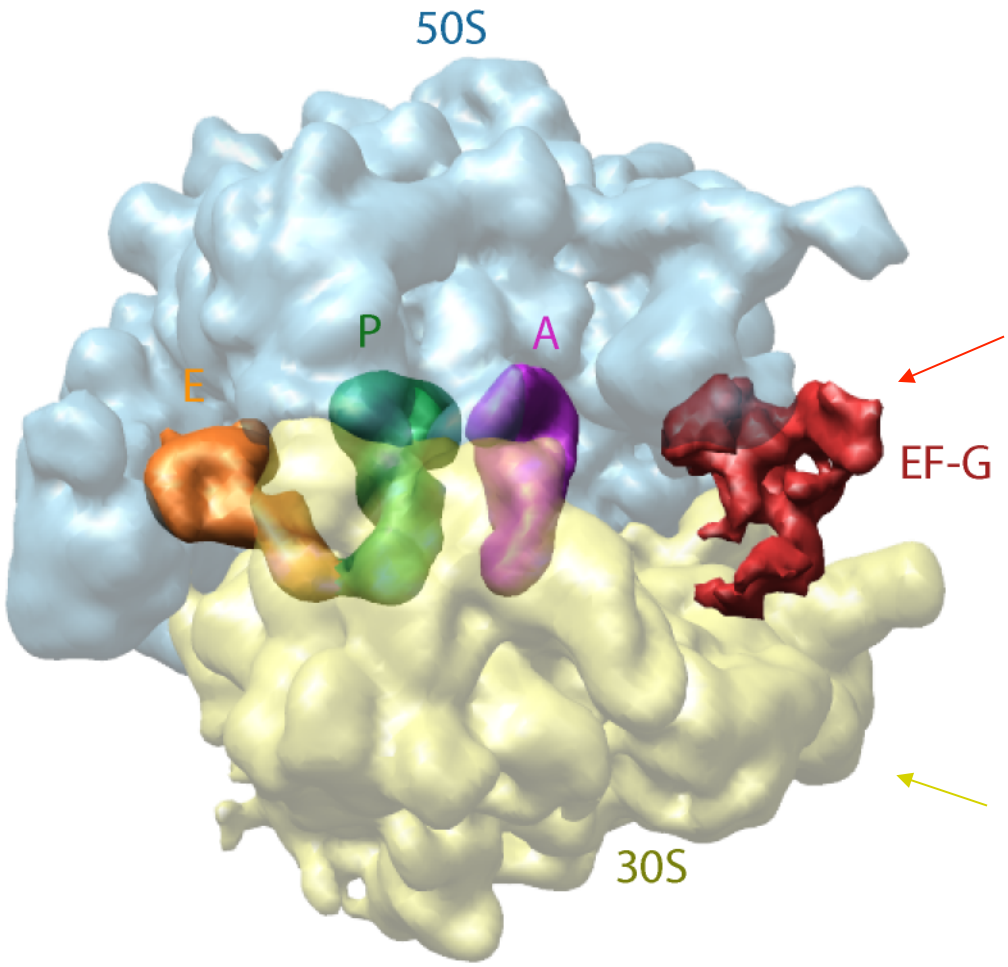
ML3D classification



Probability-weighted angular & class assignments

Prelim. ribosome reconstruction

91,114 particles; 9.9 Å resolution



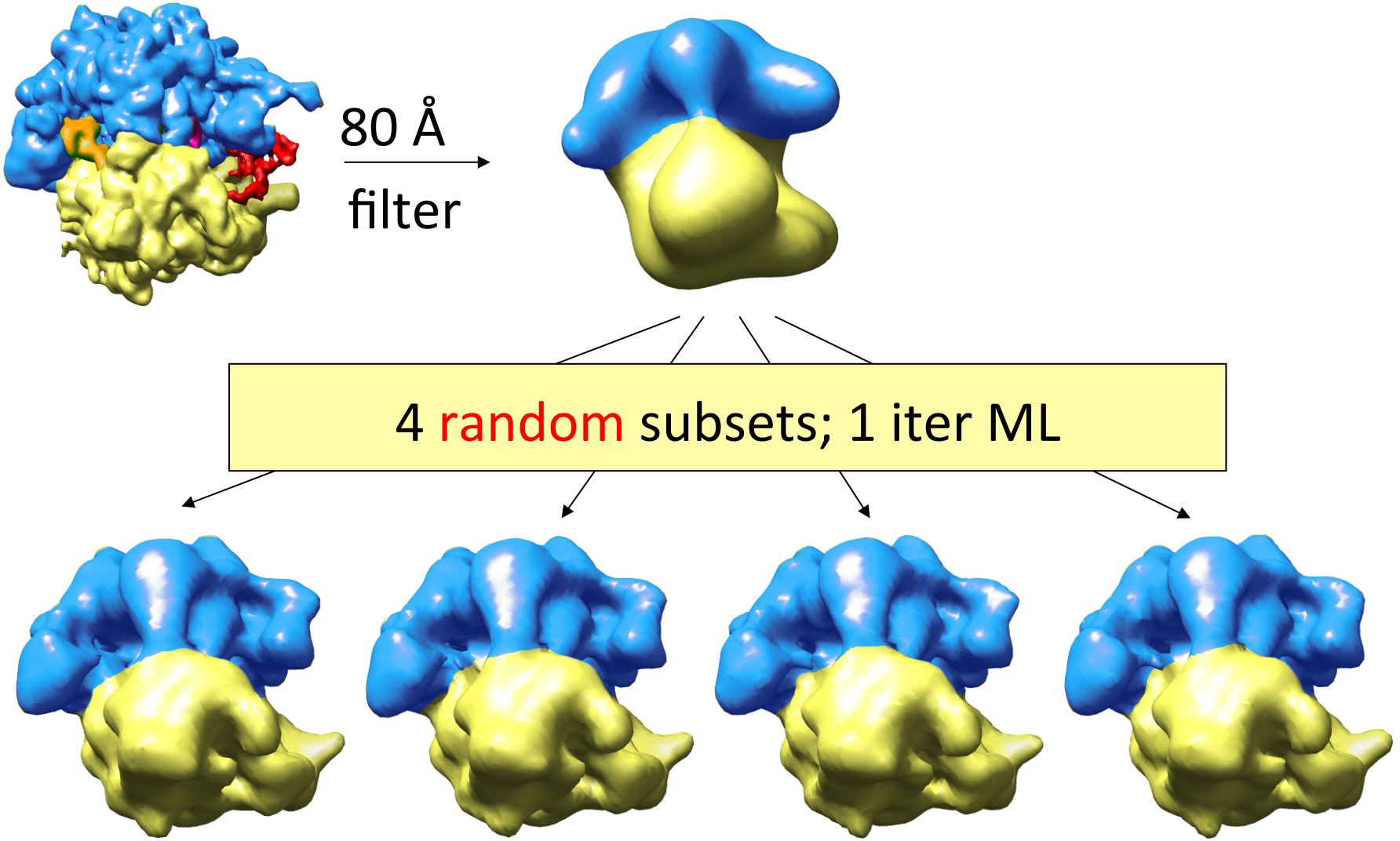
fragmented

(depicted at a lower threshold)

blurred

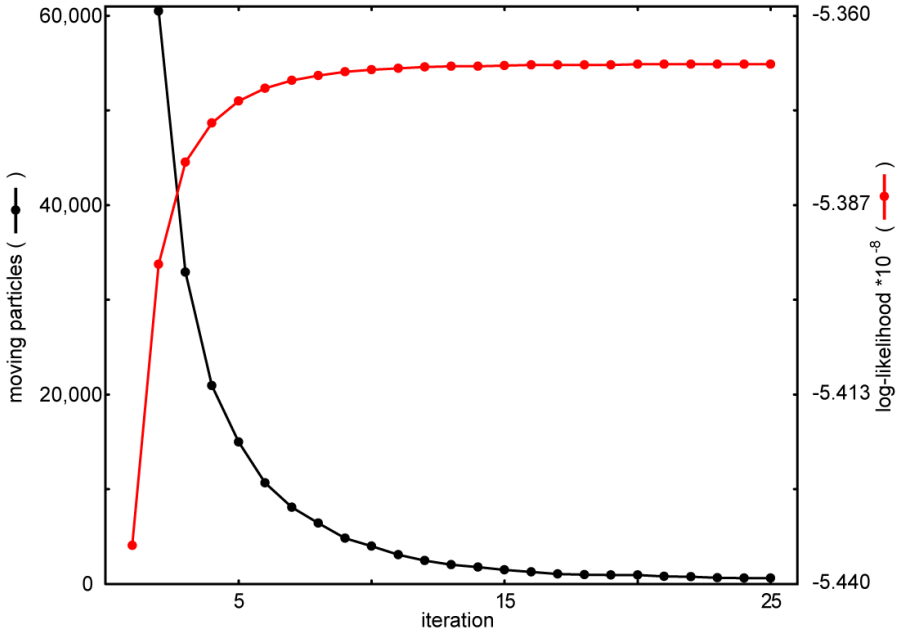
In collaboration with Haixiao Gao & Joachim Frank

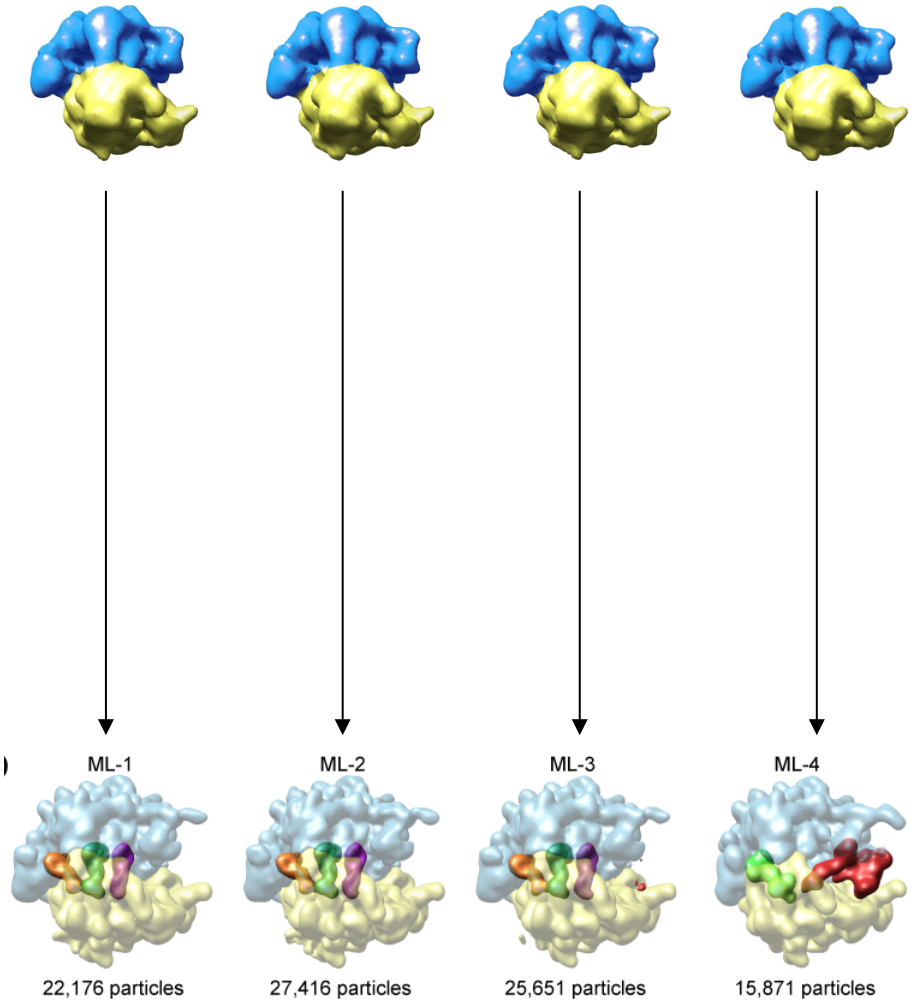
Seed generation



ML3D-classification

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





Scheres et al, Nat Methods, 2007

Regularised likelihood approach

- Data model in Fourier-space
  - Colored (correlated) noise
  - CTF-correction
- Marginalize over orientations & classes
  - Probability-weighted assignments
- Regularization term
  - Penalize high-frequency components
  - Elegant derivation of 3D Wiener filter
  - Iteratively learn power of signal and noise from the data
  - No user-expertise required to optimally filter data/map
  - Objectivity
- RELION

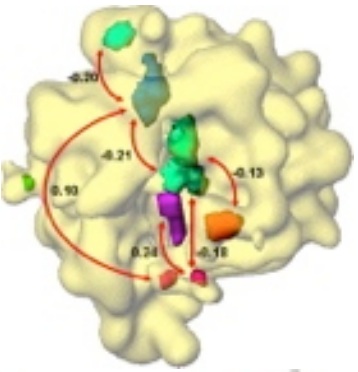
  Scheres, JMB 2012; JSB 2012

Other 3D classification tools (I)

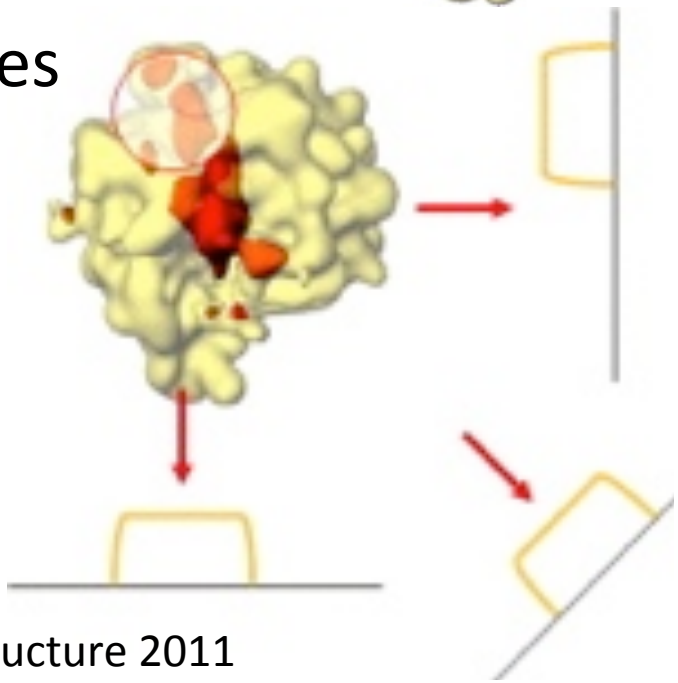
- Non-ML multi-reference refinement
  - IMAGIC/SPIDER Van Heel / Frank labs
  - EMAN2 (new similarity measures, alternate 2D/3D)
    Tang et al, JSB 2012; Ludtke et al, JSB 1999
  - SIMPLE (stochastic hill-climbing)
    Elmlund&Elmlund, JSB 2012
- Multi-variate statistical analysis
  - IMAGIC/SPIDER
    Elad et al, JSB 2008

Other 3D classification tools (II)

- Boot-strapping & 3D (co-)variance map
  - Detect and quantify heterogeneity!



- Focused classification
  - Mask out relevant areas in images

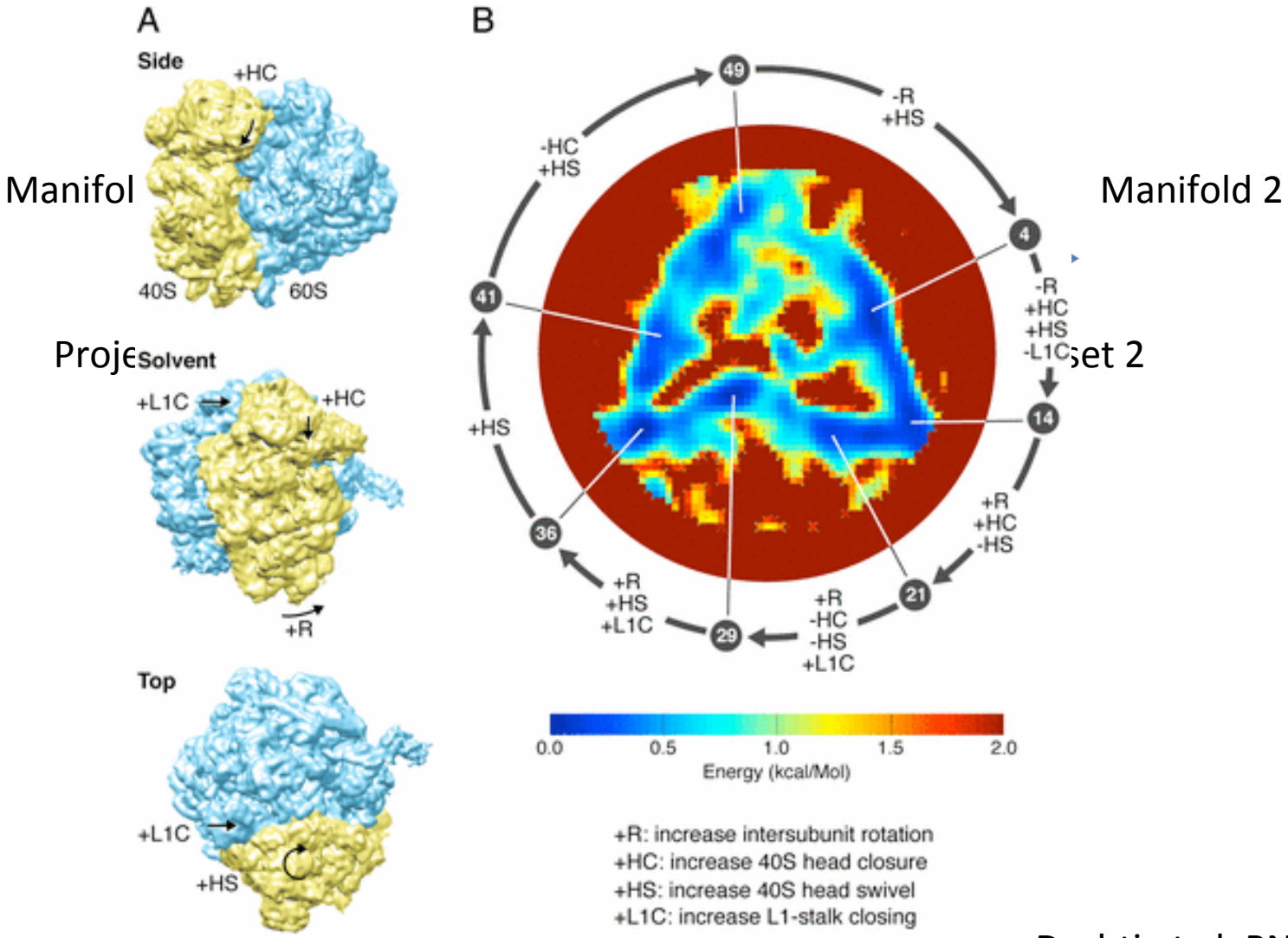


- MSA of bootstrapped maps
  - More generally applicable
  - Pawel: SPARX

Penczek, JSB 2006; Zhang et al, Structure 2008; Penzek et al Structure 2011

Classification of a continuum of states, and mapping of the energy landscape

Joachim Frank (Columbia), Peter Schwander and Abbas Ourmazd (U. of Wisconsin)



Manifol

Manifold 2

Proje

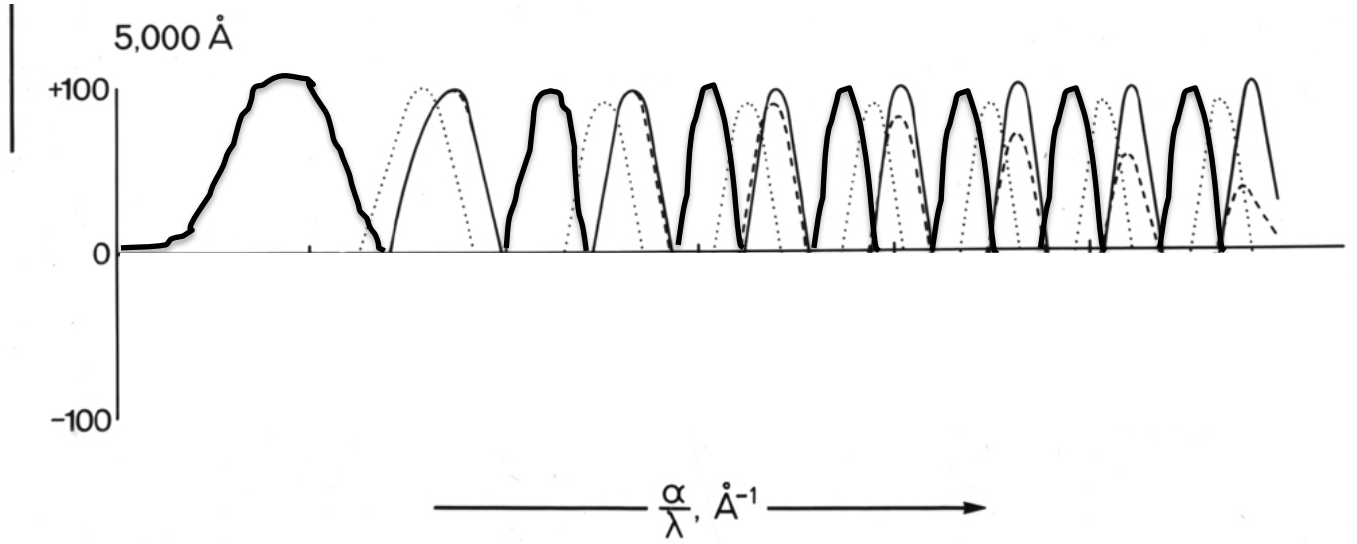
set 2

Dashti et al, PNAS, 2014

Many variations/applications

Possible in different software packages

Phase flipping



- Easy to do
- Reasonably effective
- Problems in classification?

(3D) Wiener filter

Optimal linear filter

$$V = \frac{\sum_{i=1}^{N} \mathbf{P}_{\varphi}^{\mathrm{T}} \frac{\mathrm{CTF}_i}{\sigma_i^2} X_i}{\sum_{i=1}^{N} \mathbf{P}_{\varphi}^{\mathrm{T}} \frac{\mathrm{CTF}_i^2}{\sigma_i^2} + \frac{1}{\tau^2}}$$

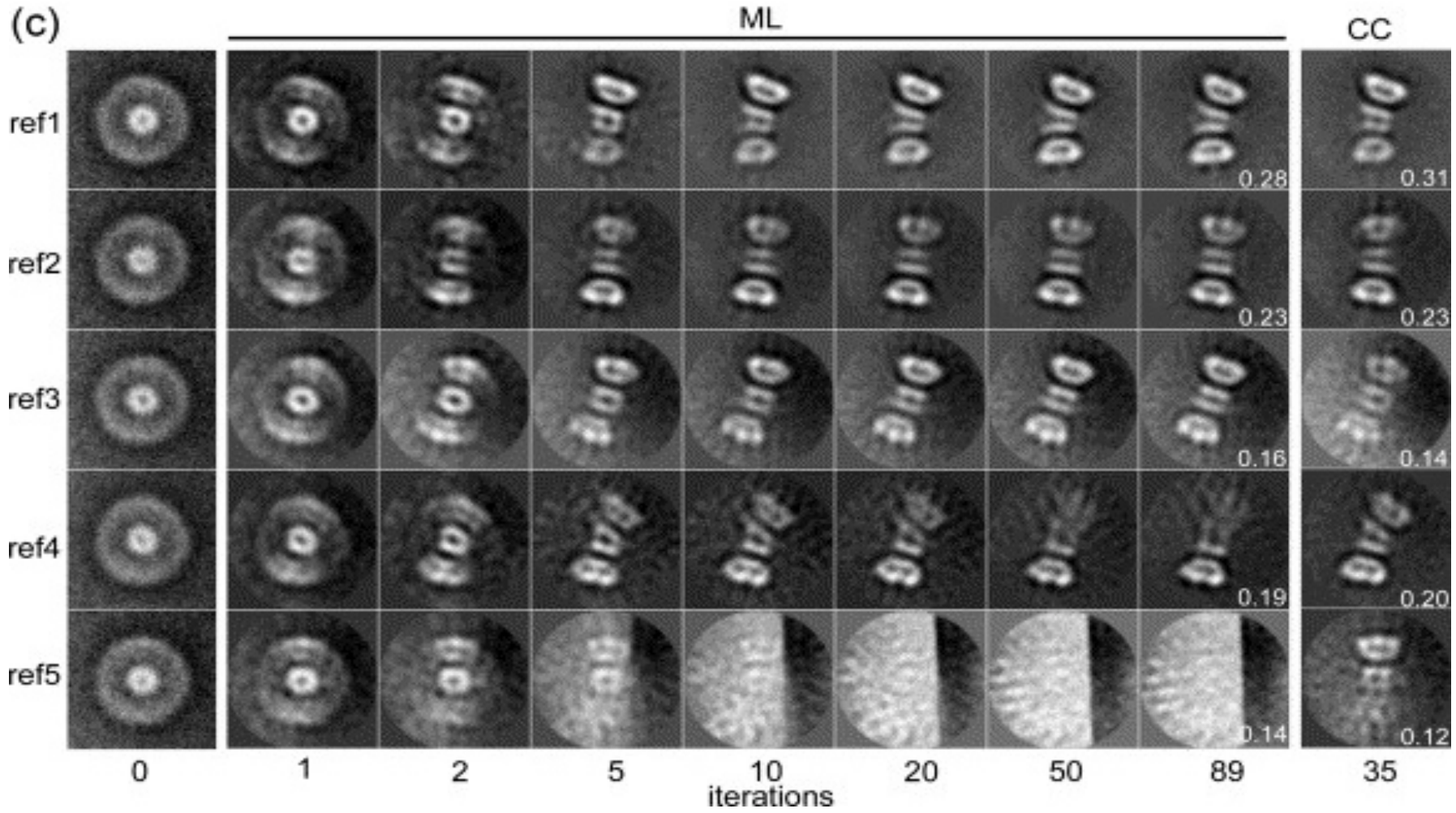
- σ^2: noise power
- τ^2: signal power

- Low-pass filters & corrects for CTF
- τ^2/σ^2 is often approximated as a constant
  => low-pass filter effect is lost
- You cannot pre-Wiener filter your data!

2D classification

- Multi-reference 2D refinement/alignment
  - RELION, XMIPP, EMAN2, SPARX (ISAC), SPIDER, IMAGIC
- MSA/PCA
  - SPIDER, IMAGIC, XMIPP, EMAN/SPARX?

Reference-free 2D classification



Scheres et al (2005) J.Mol.Biol.

2D classification

- We ALWAYS do 2D class averaging to tidy up the data set
  - Use at least ~100 particles/class for cryo-EM
  - Fewer for negative stain
- Often:
  - Large, high-resolution classes with nice particles
  - Small, low-resolution classes with crap
- Delete bad classes (and possibly repeat)

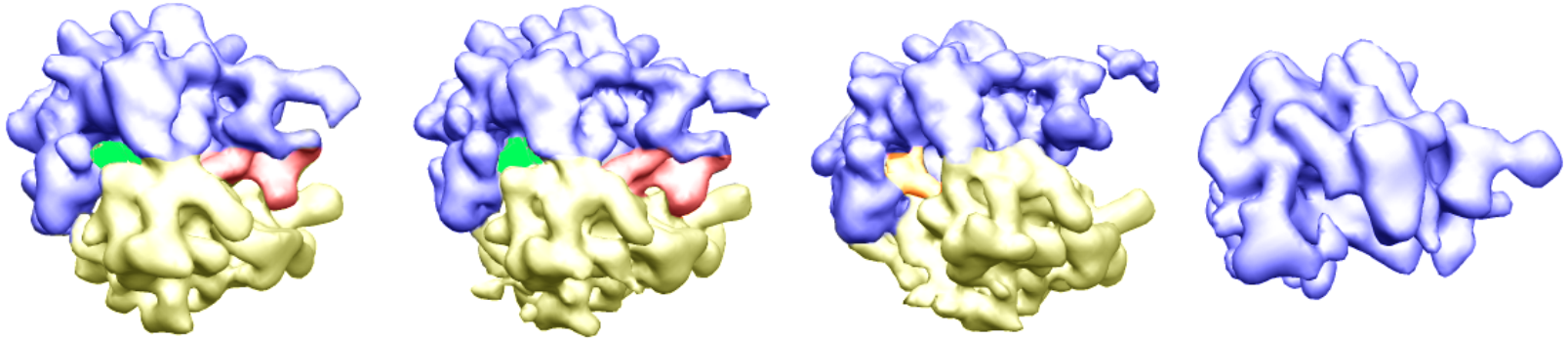
3D classification

- We ALMOST ALWAYS do 3D classification
  - Almost all samples are heterogeneous!
  - Use at least ~3,000 particles/class for cryo-EM
  - Computational cost often limits to 4-10 classes.

- Main scenarios:
  - 7.5° angular sampling; exhaustive angular searches
  - Finer angular sampling (e.g 0.9° or 1.8°); local searches around angles from 3D single-reference refinement
  - NEW: keep angles fixed and only classify (within a mask)
    - good for presence/absence of small factor

Classify structural variability

- Standard data set from the Frank lab
  - 10,000 70S ribosomes (50% +EFG; 50% -EFG)
  - MAP-refinement K=4

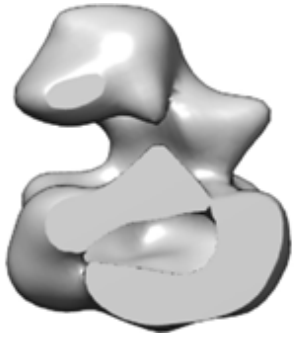


24%	28%	42%	6%
26Å	19Å	19Å	30Å

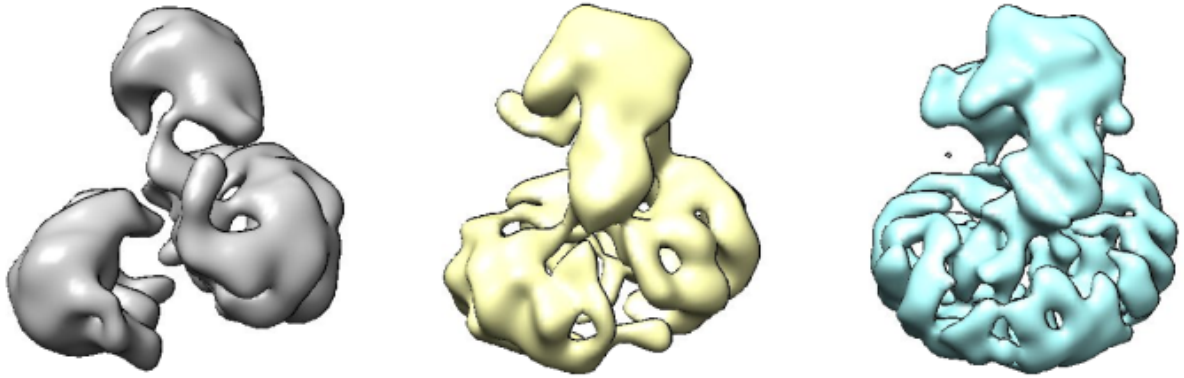
Scheres, JMB, 2012

Data cleaning

- One/few good classes
- Discard bad classes

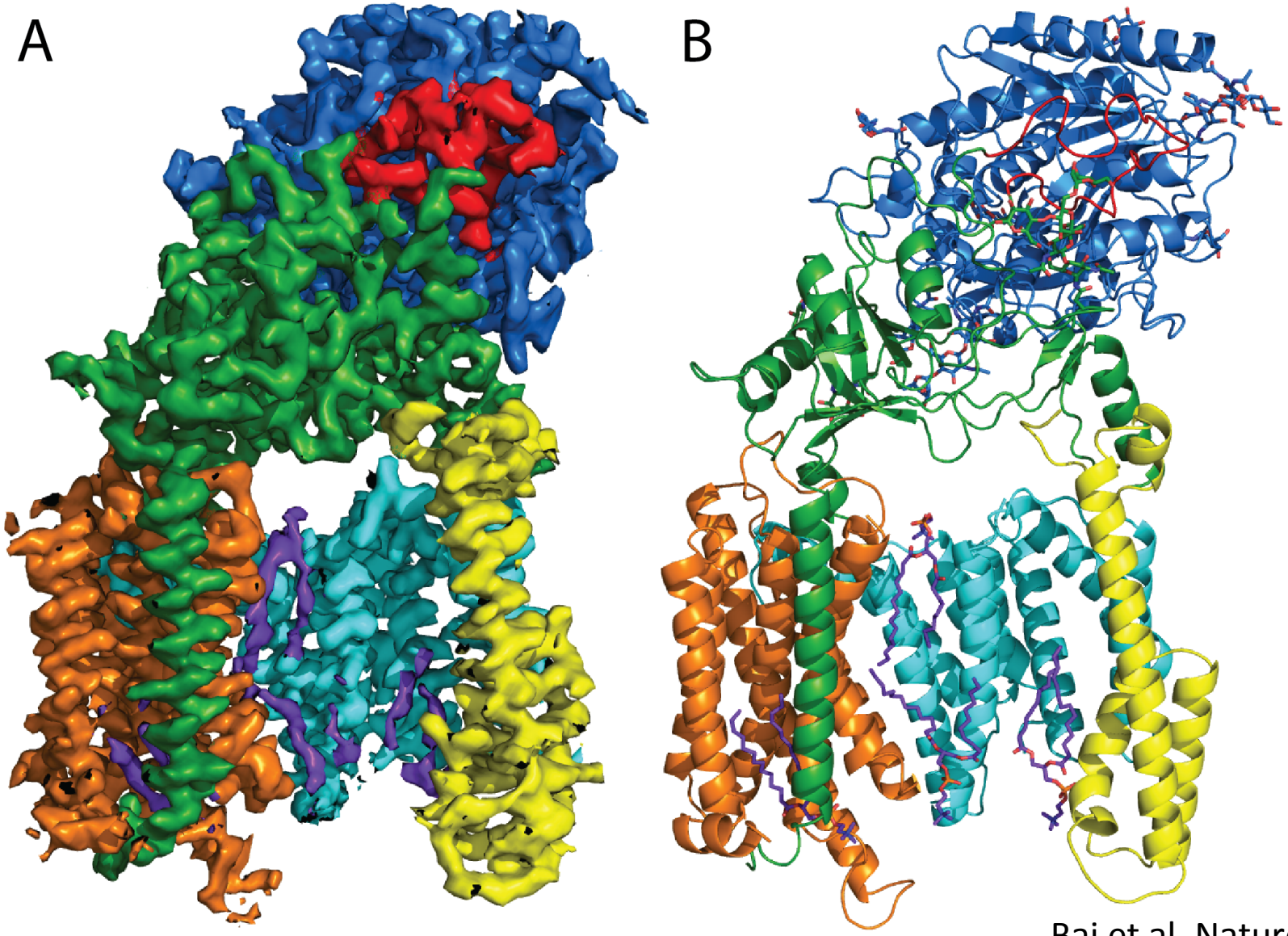


γ-secretase



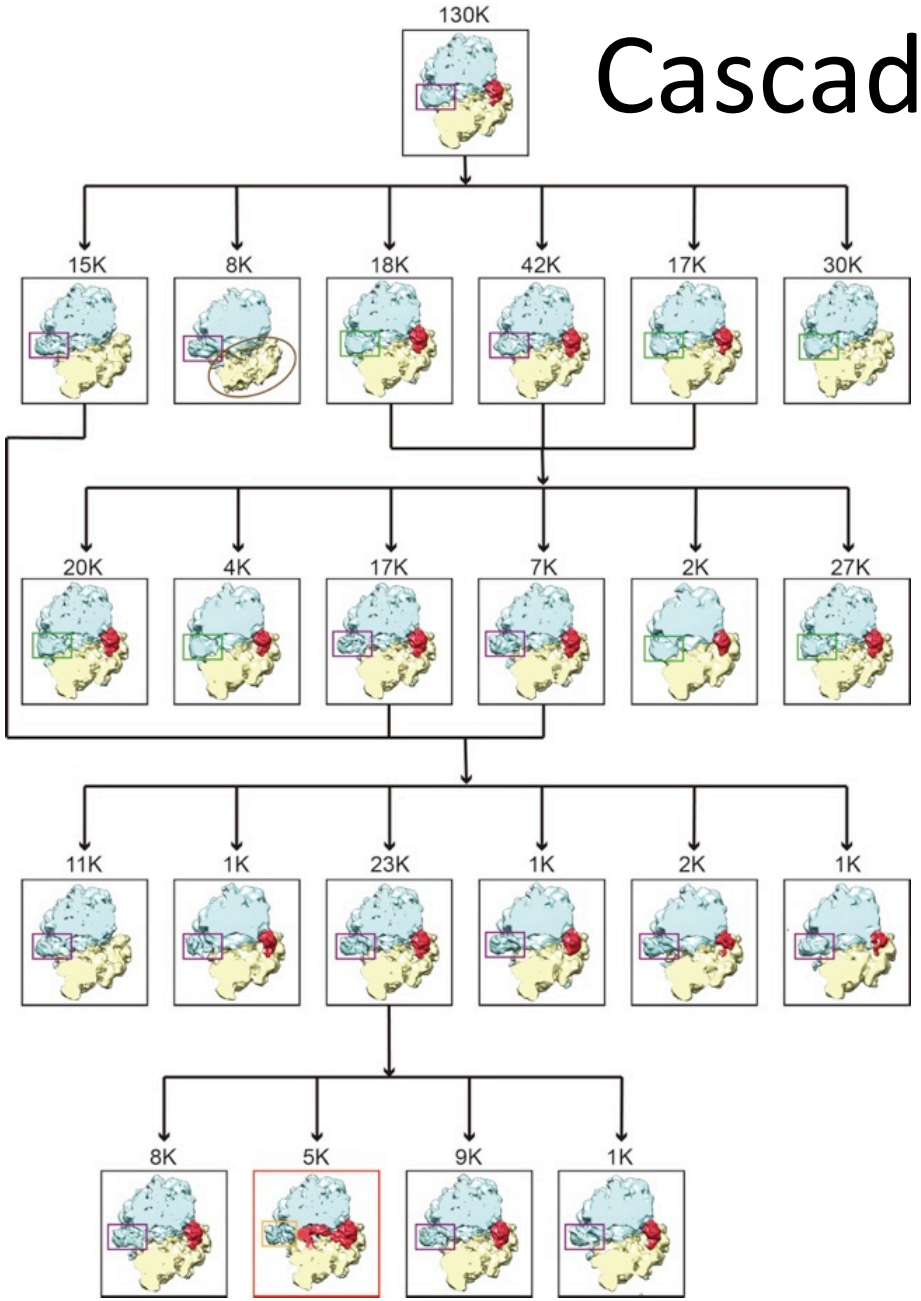
Lu et al, Nature, 2014

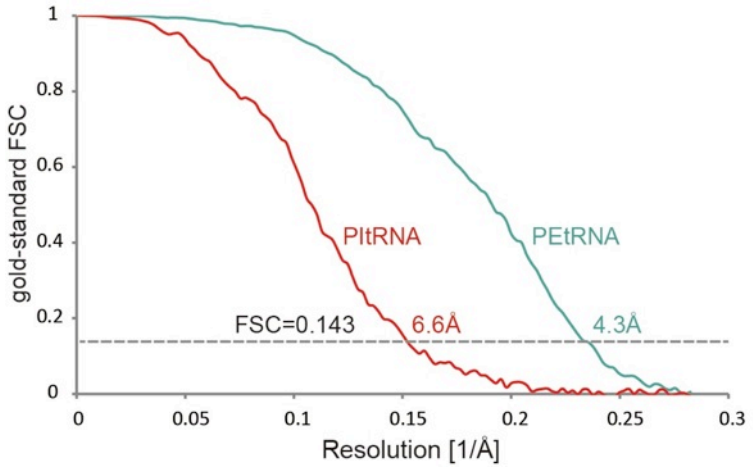
3.4 Å map, ~130 kDa ordered mass

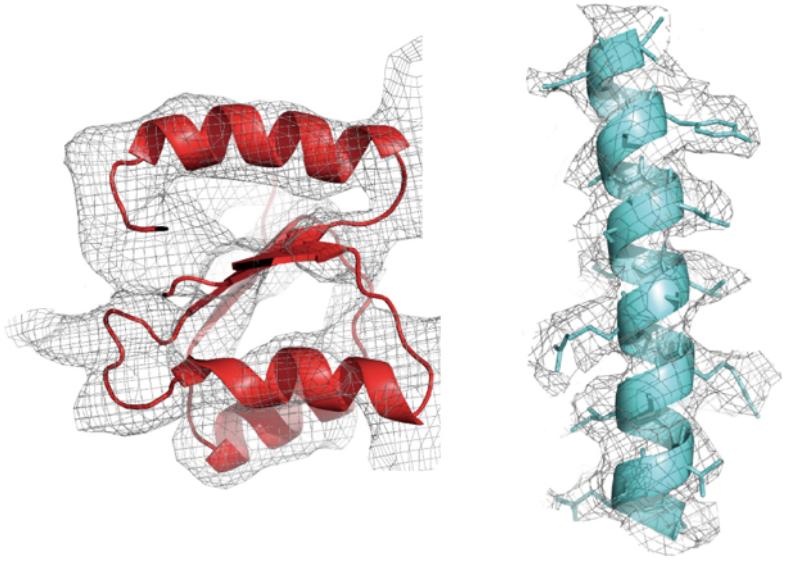


Bai et al, Nature 2015

Cascaded classification



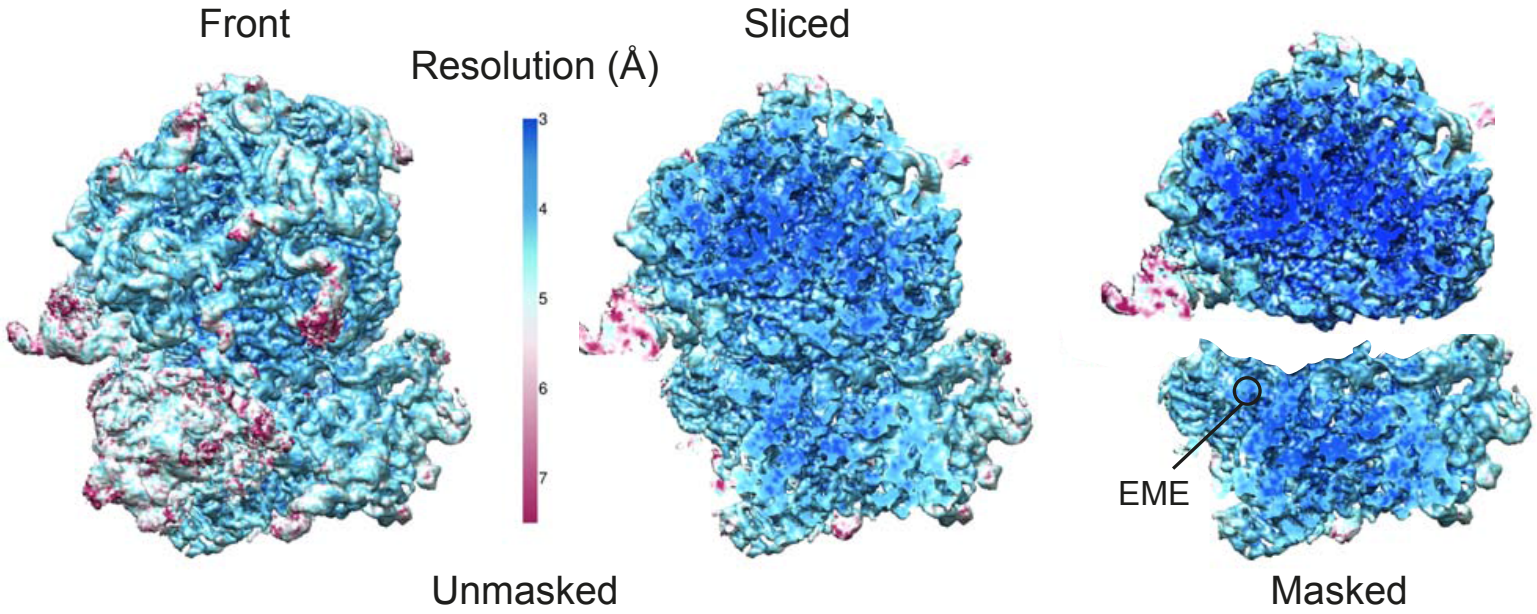




Fernandez et al, Science, 2013

Continuous heterogeneity: Masked refinements

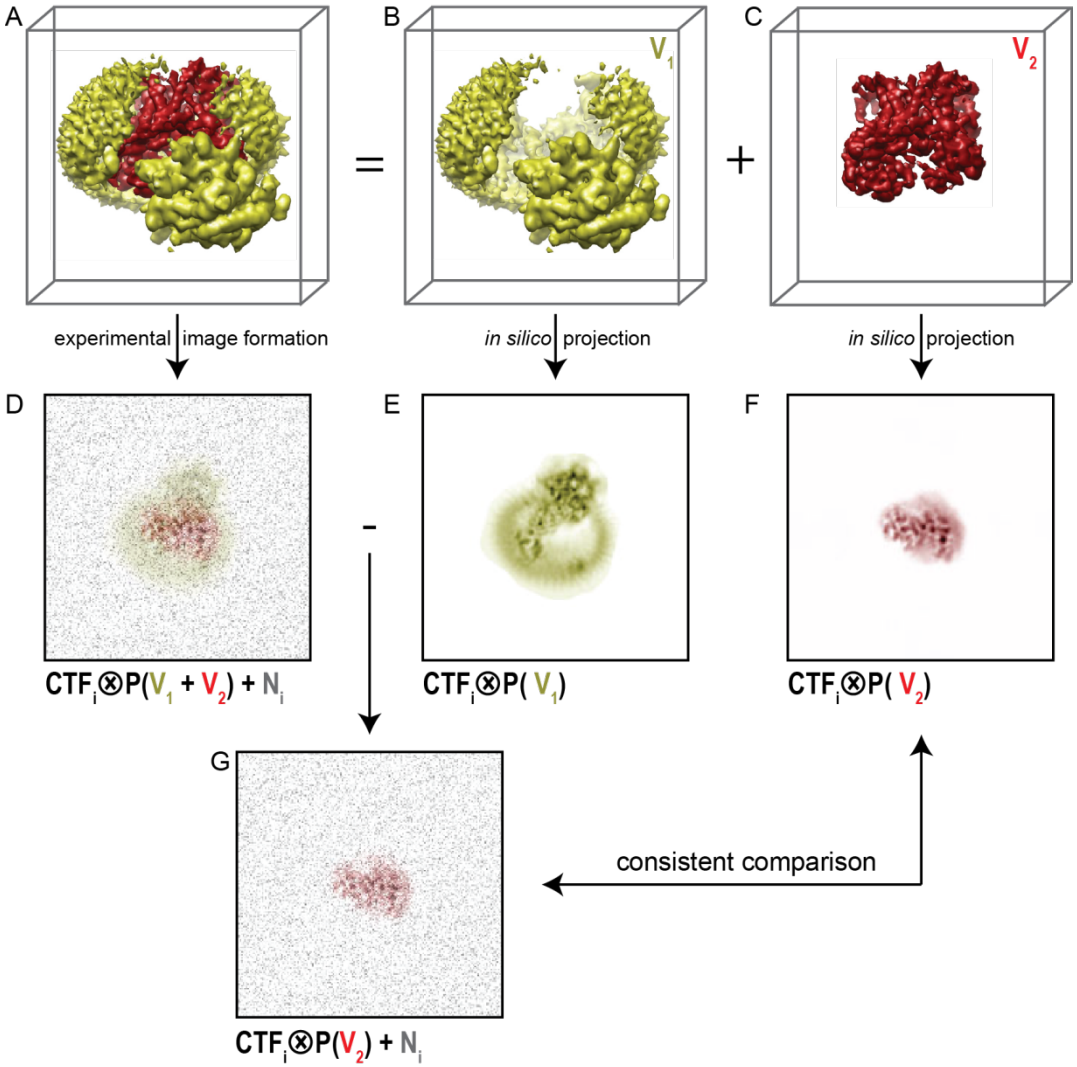
- Mask out volume of interest in reference at every step of 3D-(single-reference) refinement



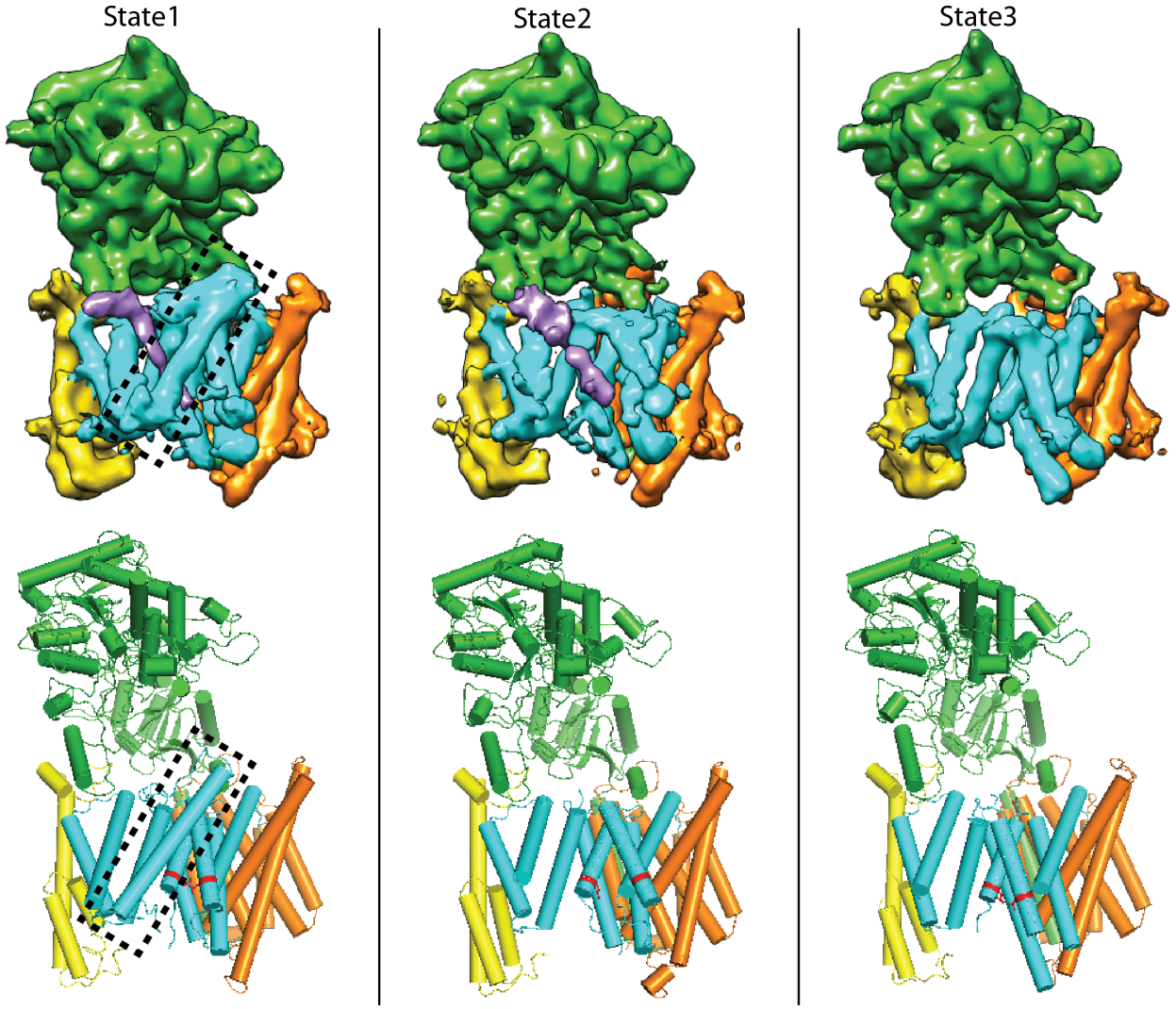
Plasmodium falciparum ribosome (+emetine)

Wong et al, eLife, 2014

Masked classification + signal subtraction

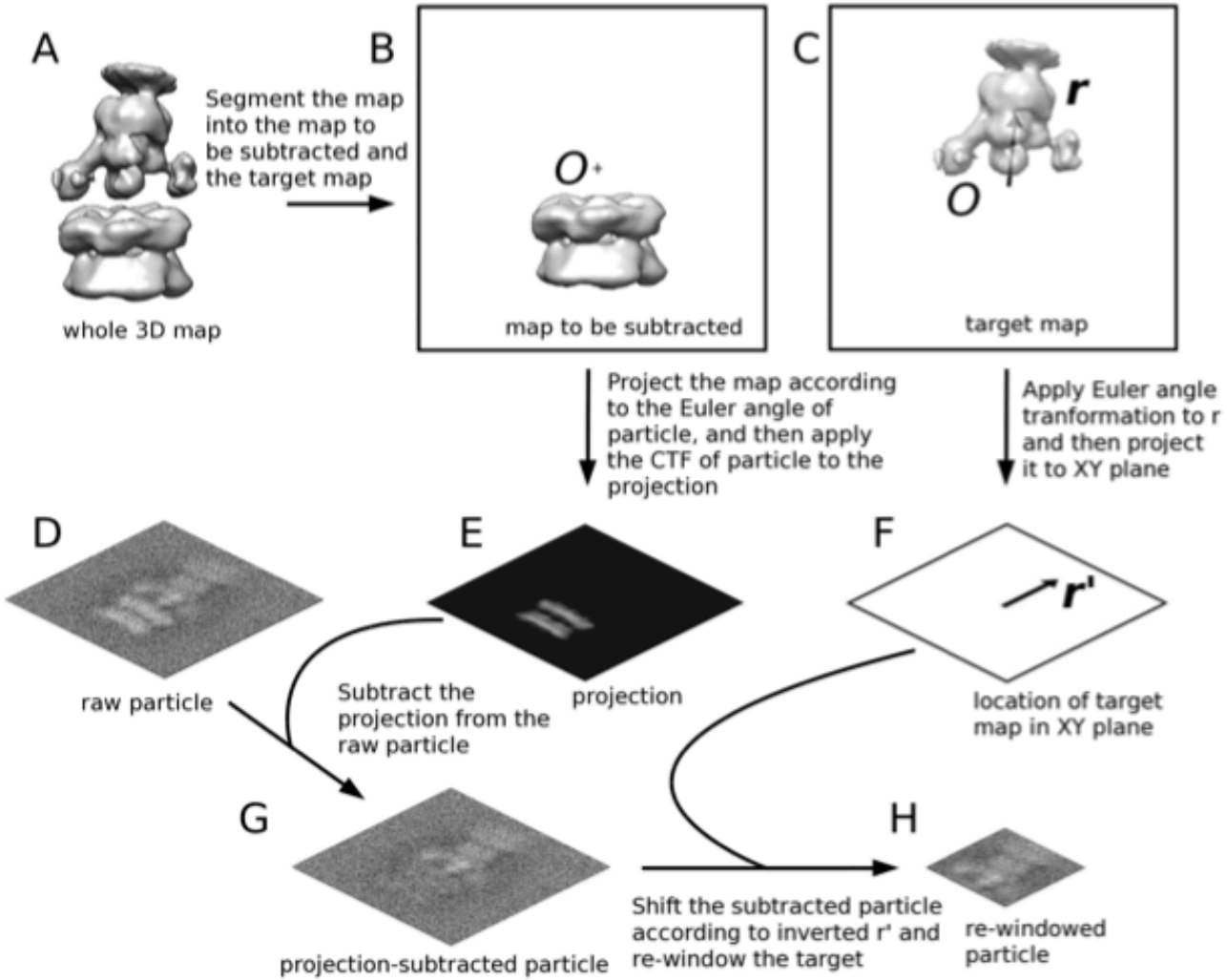


Conformational heterogeneity



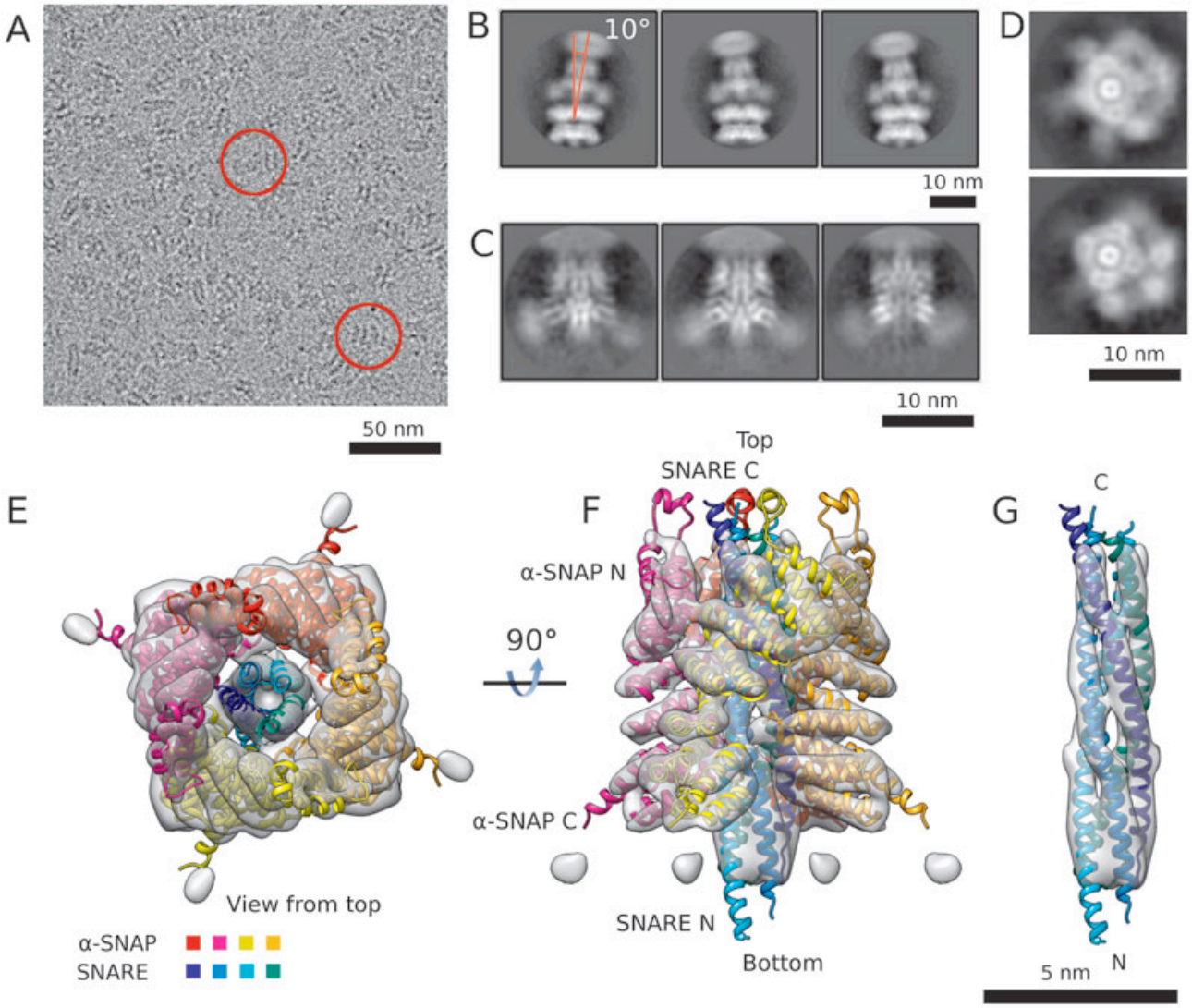
Bai et al, bioRxiv, 2015

Independent development

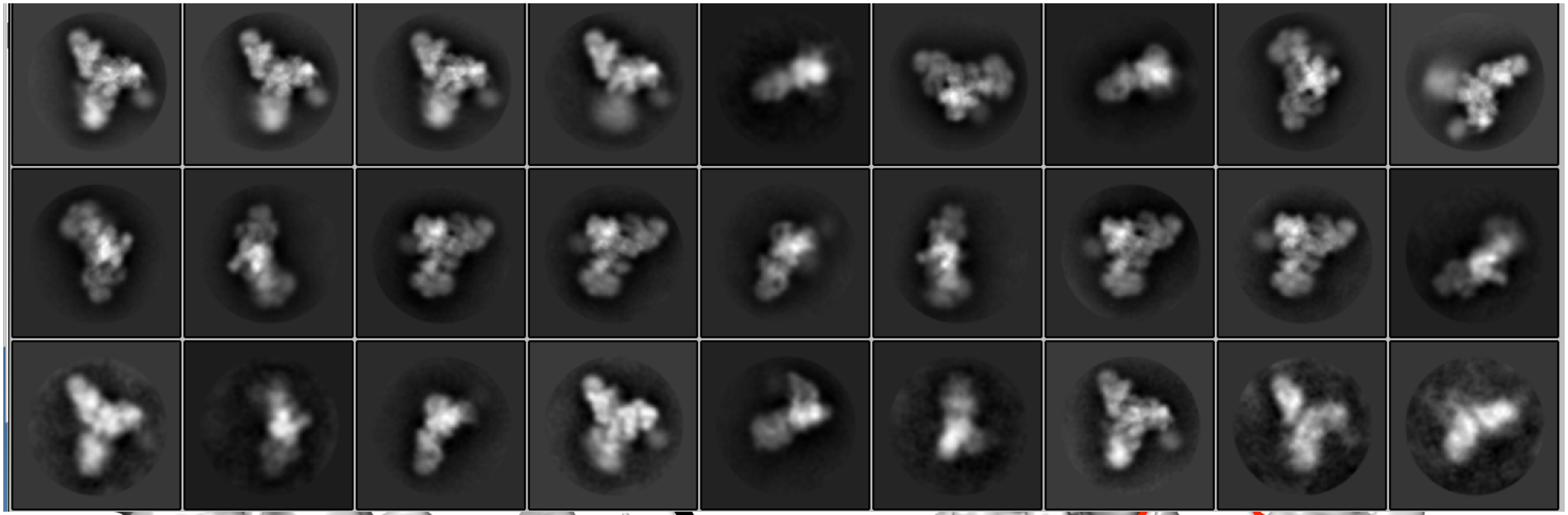


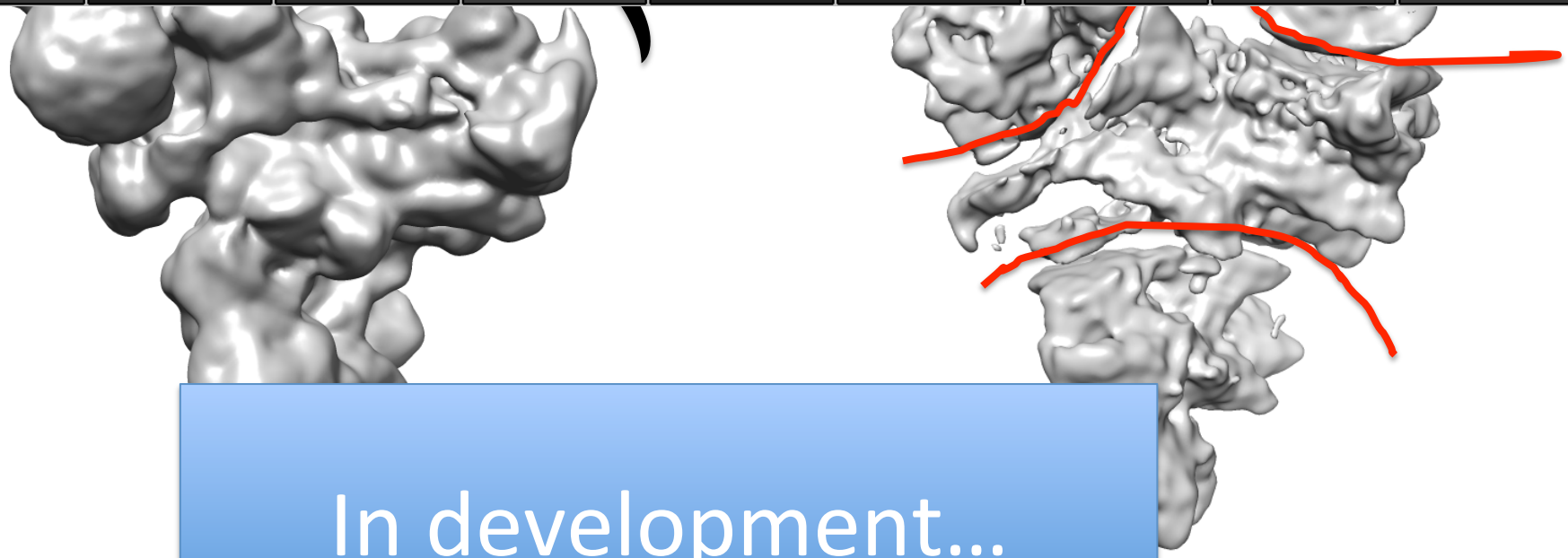
Zhou, …, Hongwei Wang, Senfang Sui
Cell Research, Apr 2015

SNAP-SNARE



Zhou, …, Hongwei Wang, Senfang Sui
Cell Research, Apr 2015

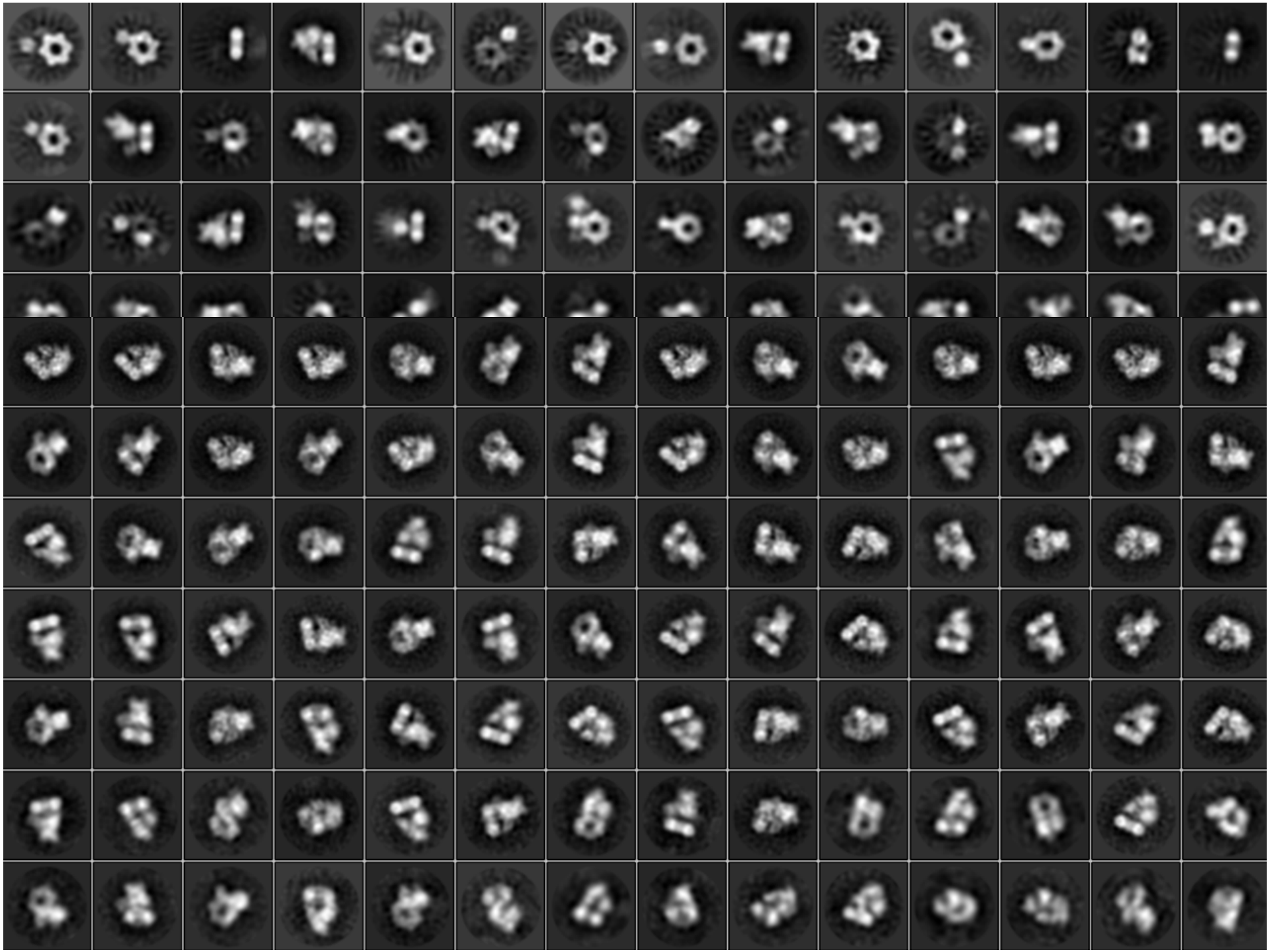




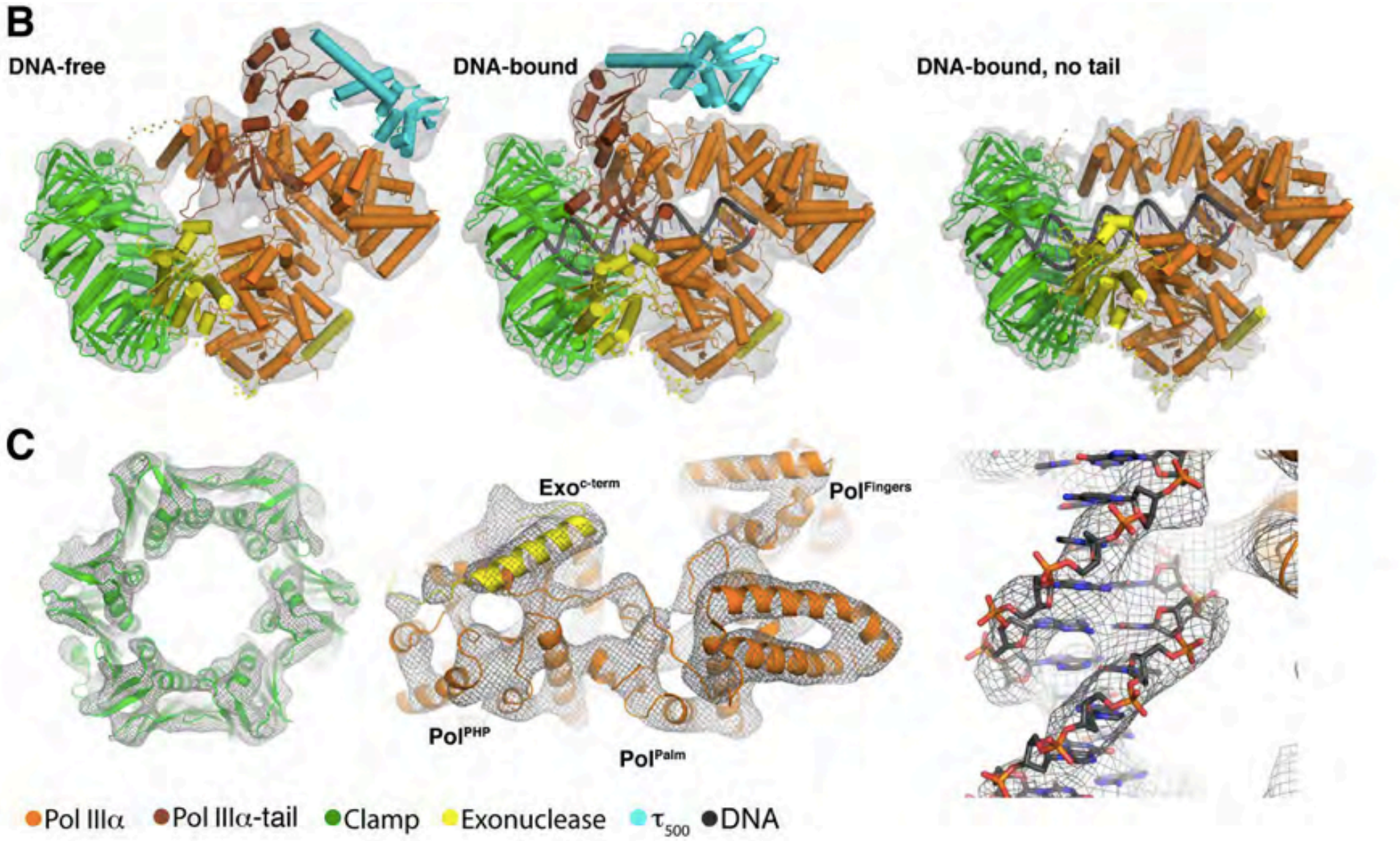
In development…

Nguyen et al, Nature, 2015

Some mistakes to avoid…



Replication complex



Leiro et al, eLife, advanced online

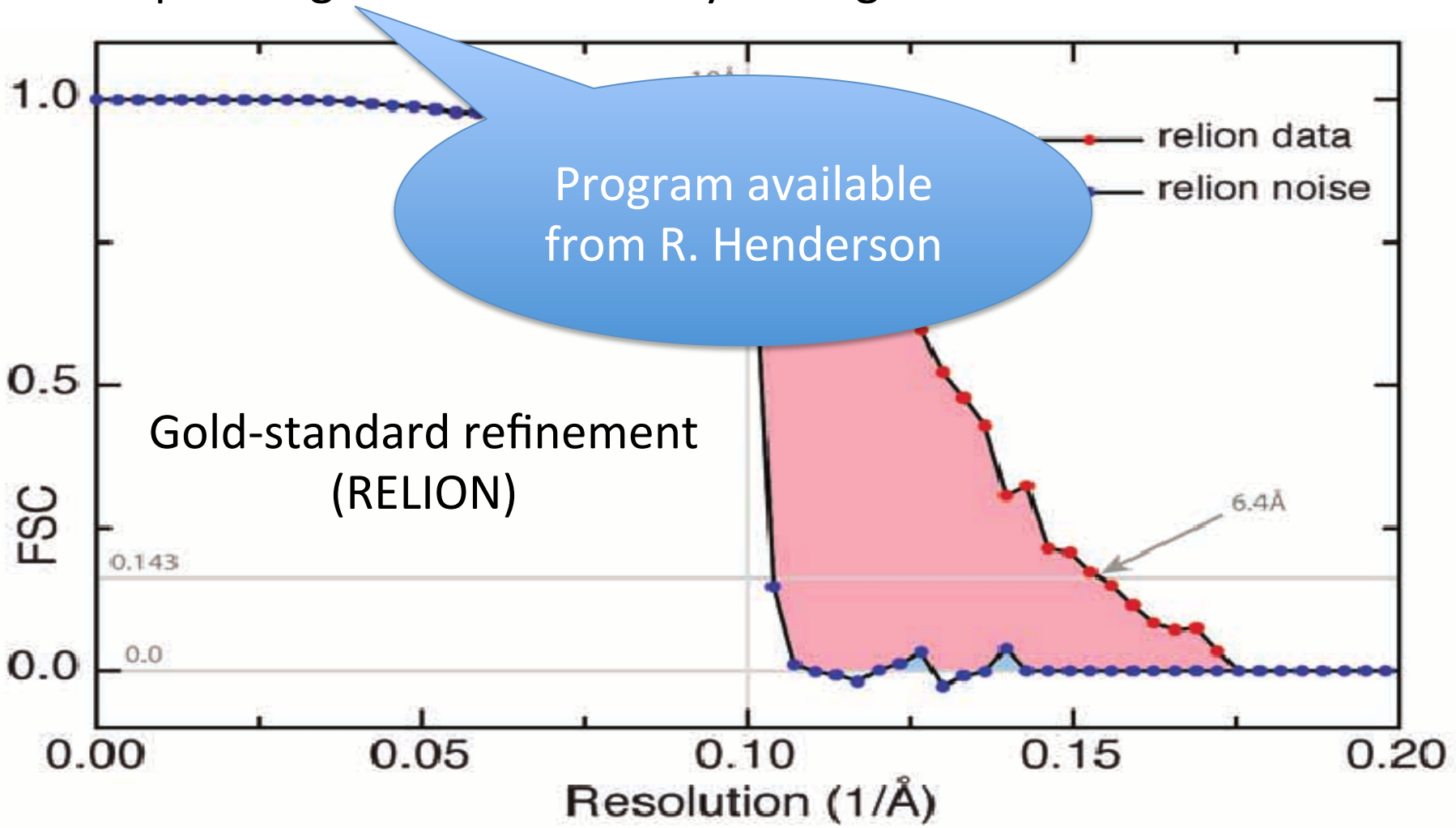
Overfitting

- Always use gold-standard refinement OR limited resolution refinement

- Some new algorithm?
    - Test high-resolution noise substitution

High-resolution noise-substitution

- Replace signal in the data beyond a given resolution *d* with noise



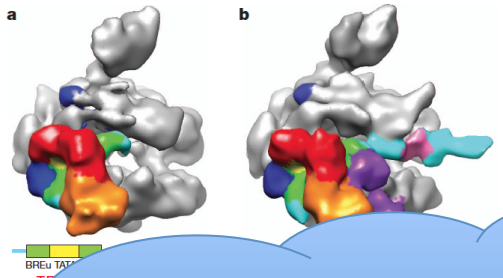
Get stuck with a wrong initial model

No program is guaranteed to find the global minimum…

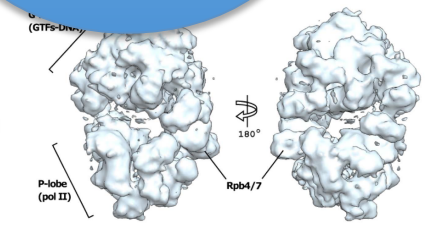
Human RNA polymerase II PIC

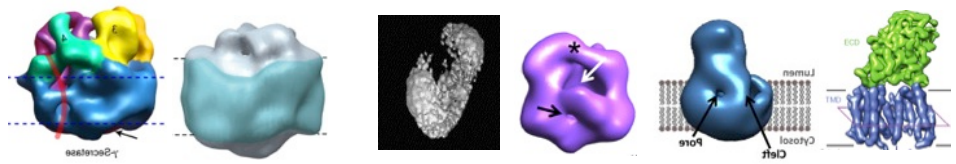
He et al & Nogales, Nature (2013)

As resolutions improve, this will be ever less of a problem.

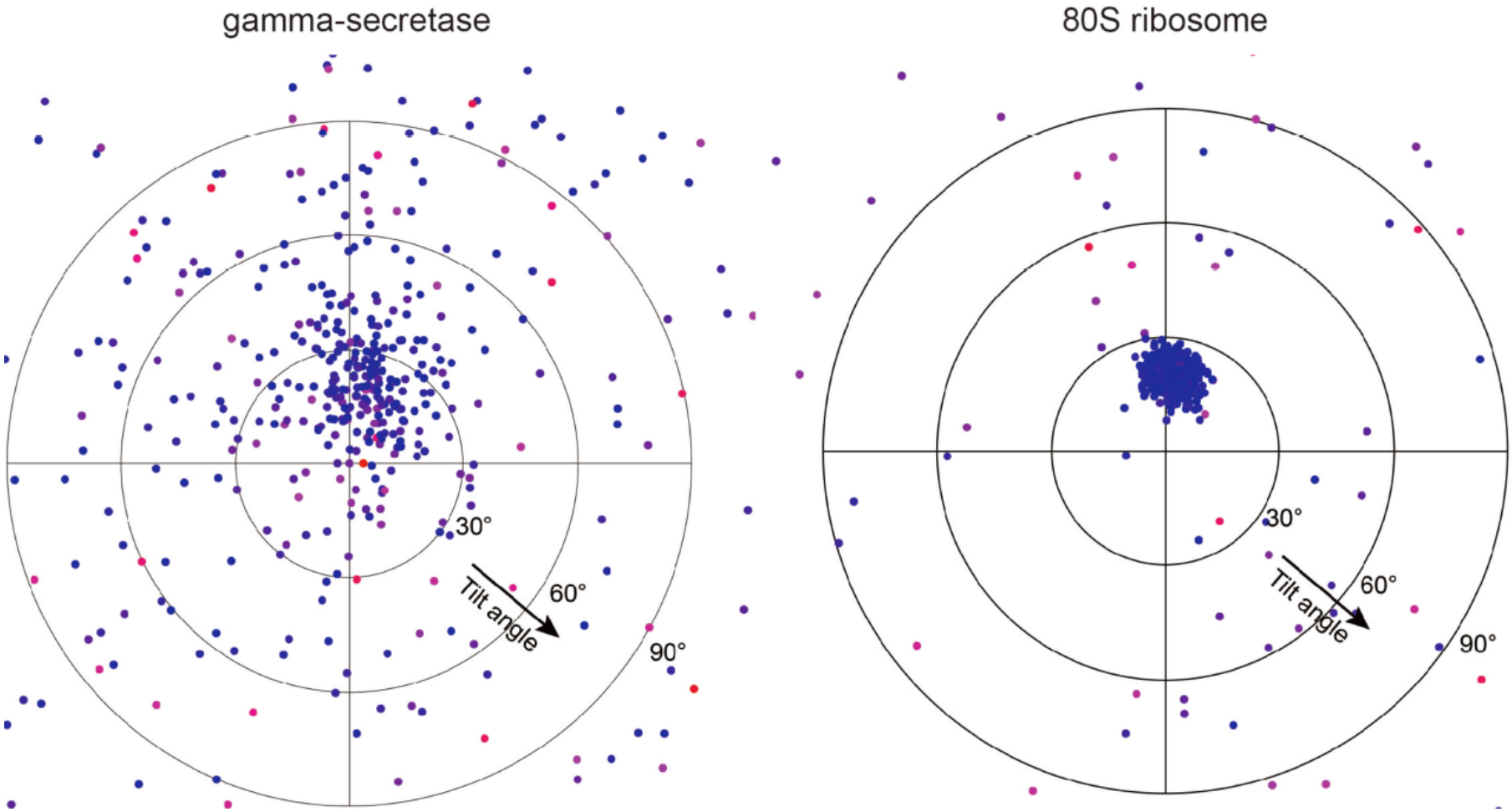


Should we stop publishing blobs?





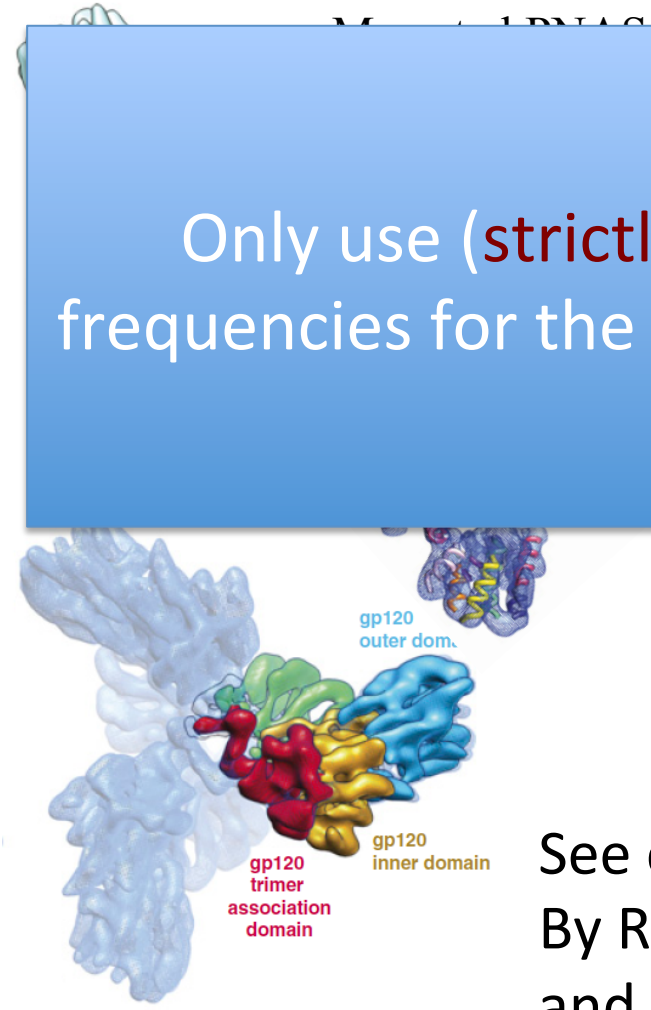
Tilt-pair validation

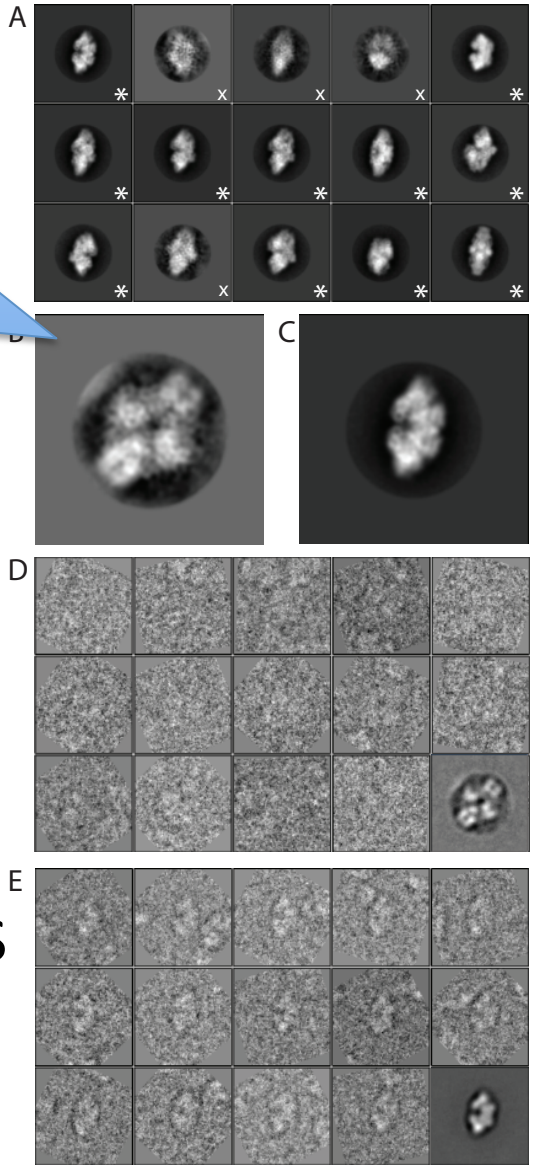


(like in RELION-1.3)

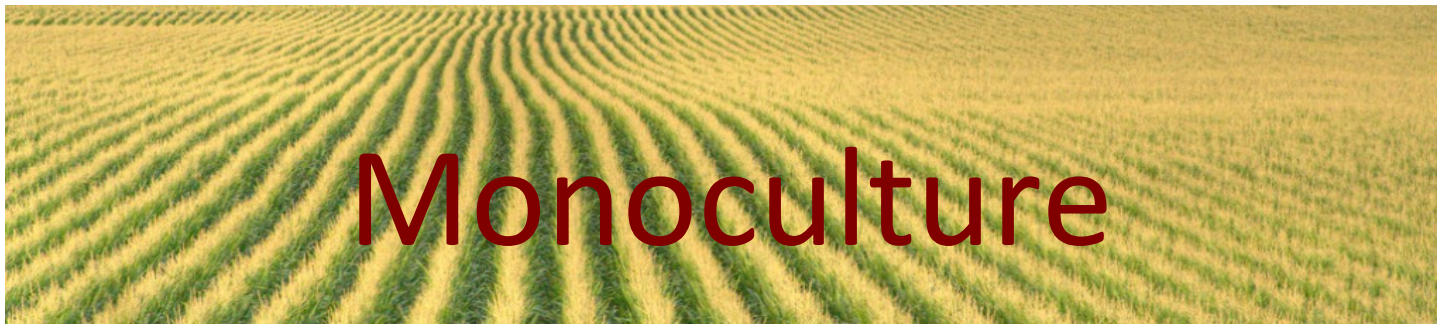
Template-based auto-picking

Only use (strictly) low-frequencies for the templates!





See comments in PNAS
By Richard Henderson
and Marin van Heel

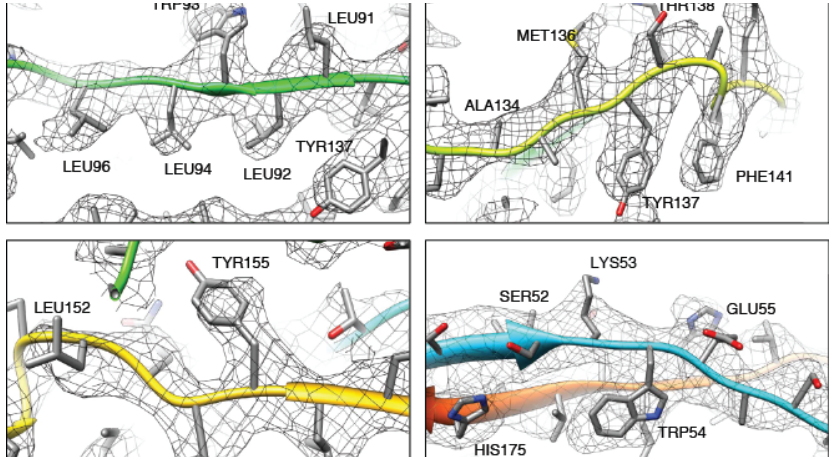


Microscopes: FEI, Jeol, Zeiss, …

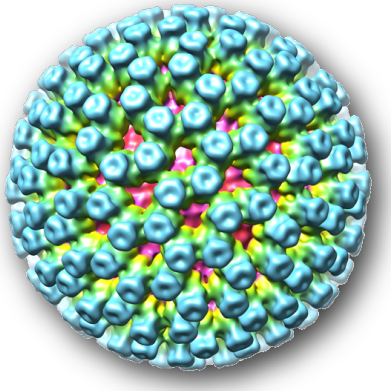
Detectors: K2, Falcon, DE, TVIPS, …

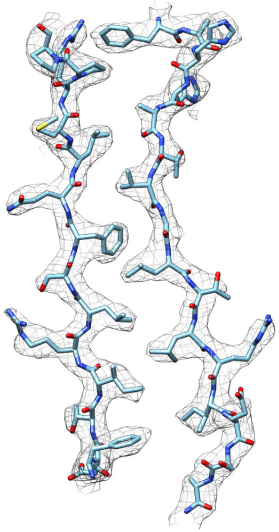
Software: SPIDER, IMAGIC, EMAN, SPARX, XMIPP, BSOFT, FREALIGN, RELION, …

Wang et al (2014) Nat Comm.



JEOL3200, DE-12,
EMAN (3.8 Å)

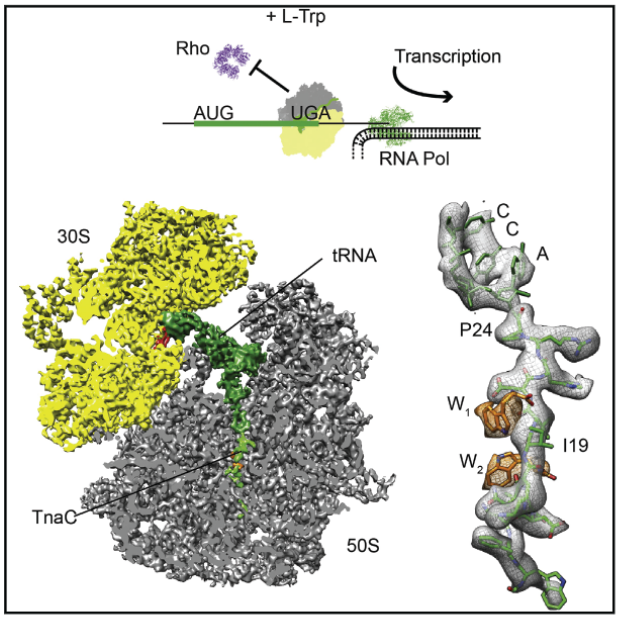




Cell Reports

Molecular Basis for the Ribosome Functioning as an L-Tryptophan Sensor

Graphical Abstract



Authors

Lukas Bischoff, Otto Berninghausen, Roland Beckmann

Correspondence

beckmann@lmb.uni-muenchen.de

In Brief

Bischoff et al. now present a cryoelectron microscopy reconstruction of a TnaC stalled ribosome, revealing two L-Trp molecules in the ribosomal exit tunnel. As a result, the peptidyl transferase center adopts a distinct conformation that precludes productive accommodation of release factor 2.

Titan Krios, Falcon-II,
SPIDER (3.8 Å)

Tim Grant & Niko Grigorieff, eLife 2015

Titan Krios, K2, FREALIGN (2.6 Å)

Conclusions

- Image processing will continue to drive this field forward
  - A variety of software solutions will be most efficient
- New hardware will continue to have huge impacts
  - Better SNRs: distinction between smaller differences
- Making good samples remains crucial!
  - Good classification algorithms are no excuse for bad samples…
- Structural heterogeneity can be an opportunity!
  - If addressed adequately

Thanks!

LMB EM-course 2014

Daily in the MPLT from 9:30-10:30am

Mon May 12: Tony Crowther
Course introduction with a historical perspective

Tue May 13: Sjors Scheres
Image formation, Fourier analysis, CTF theory

Wed May 14: Chris Russo
Microscopy physics and optics

Thu May 15: Lori Passmore
sample preparation

Fri May 16: Paula da Fonseca
Initial data analysis

Mon May 19: Sjors Scheres
Image refinement in 2D and 3D

Tue May 20: Tanmay Bharat
Tomography and sub-tomogram averaging

Wed May 21: Richard Henderson
Map validation

Thu May 22: David Barford & Alan Brown
Low- and high-resolution modeling

Thu May 22: Shaoxia Chen, Christos Savva & others
(11am-12pm) Local setup and training & 2 example applications

Enquiries: scheres@mrc-lmb.cam.ac.uk

Lecture PDFs and professionally edited videos available on:

ftp://ftp.mrc-lmb.cam.ac.uk/pub/scheres/EM-course