Processing:

Introduction and new approaches

Sjors H.W. Scheres

NRAMM cryo-EM workshop, San Diego, November 2014

MRC Laboratory of Molecular Biology

Introduction and new approaches

Lots of hard work in early image processing developments (Joachim, Marin, Michael, Pawel, …)

A comprehensive overview of
last few years that have e

Topics to be covered include:

- 3D reconstruction
- image restoration techniques
- how to deal with heterogeneous populations.

- What are the hot topics in processing?
- What are the major mathematical approaches and available software?
- What are the success stories and the failures?
- Where are the greatest challenges right now and how are we approaching these?
- Do we need completely new algorithms or just incremental improvements on the current ones?
- Mistakes to avoid!

Introduction and new approaches

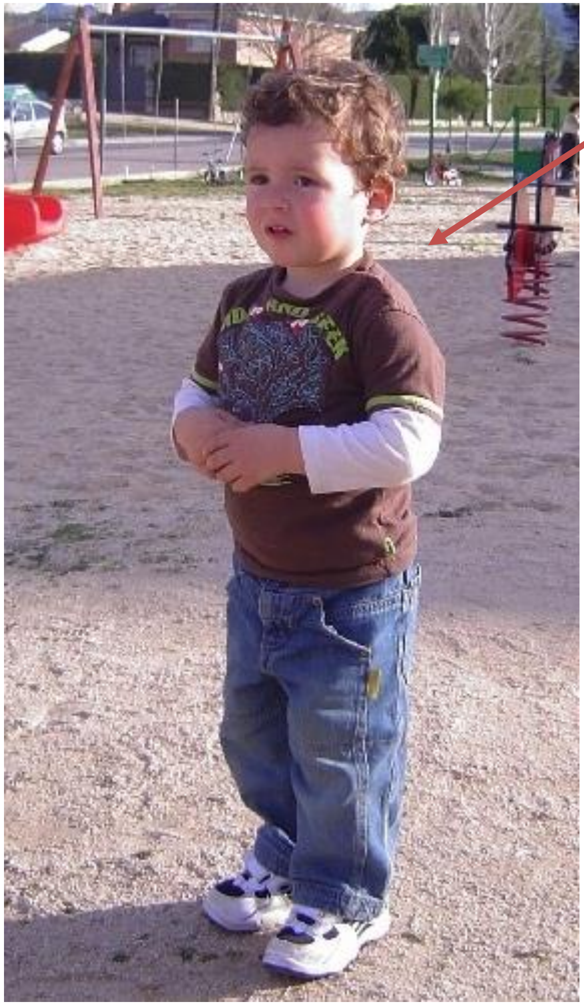
A comprehensive overview of the major advances that have taken place in the last few years that have enabled maps to achieve "atomic" resolution.

Topics to be covered include:

- 3D reconstruction
- image restoration techniques
- how to deal with heterogeneous populations.

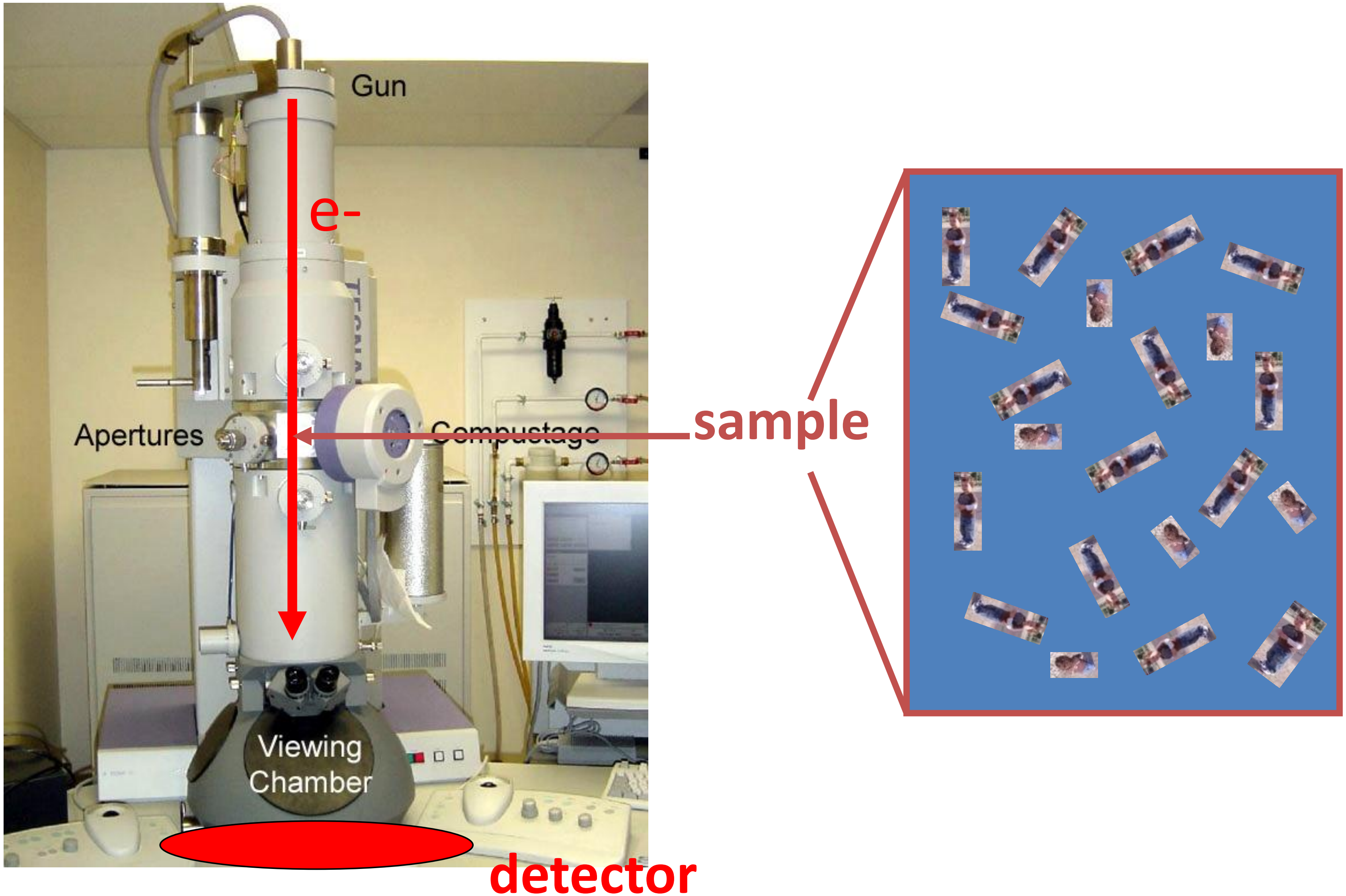
- What are the hot topics in processing?
- What are the major mathematical approaches and available software?
- What are the success stories and the failures?
- Where are the greatest challenges right now and how are we approaching these?
- Do we need completely new algorithms or just incremental improvements on the current ones?
- Mistakes to avoid!

An example “protein”



Jan

Experimental setup



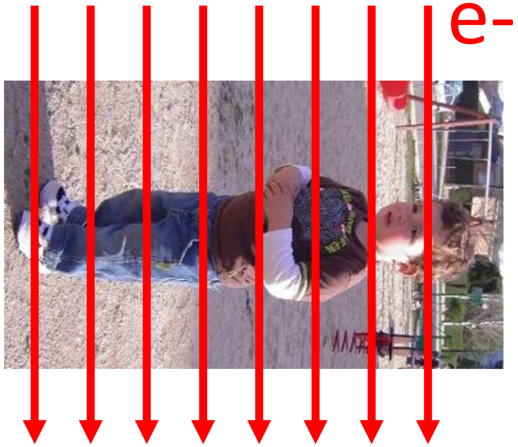
e-

sample

detector

Electron microscopy imaging

3D object



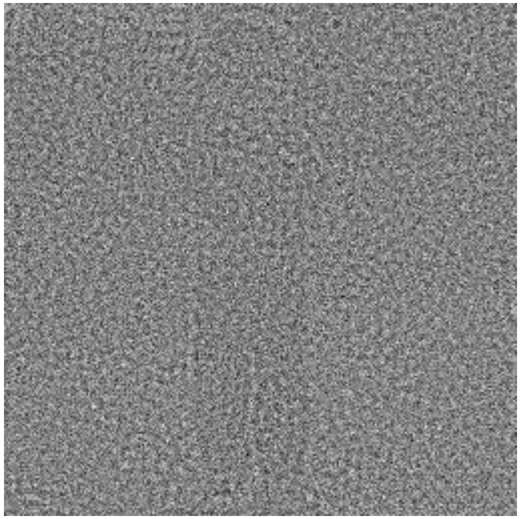
2D projection

We collect data in 2D,

but we want 3D info!

Further inconveniences

- Defocussing & microscope imperfections introduce artefacts
- Low dose: large amounts of noise

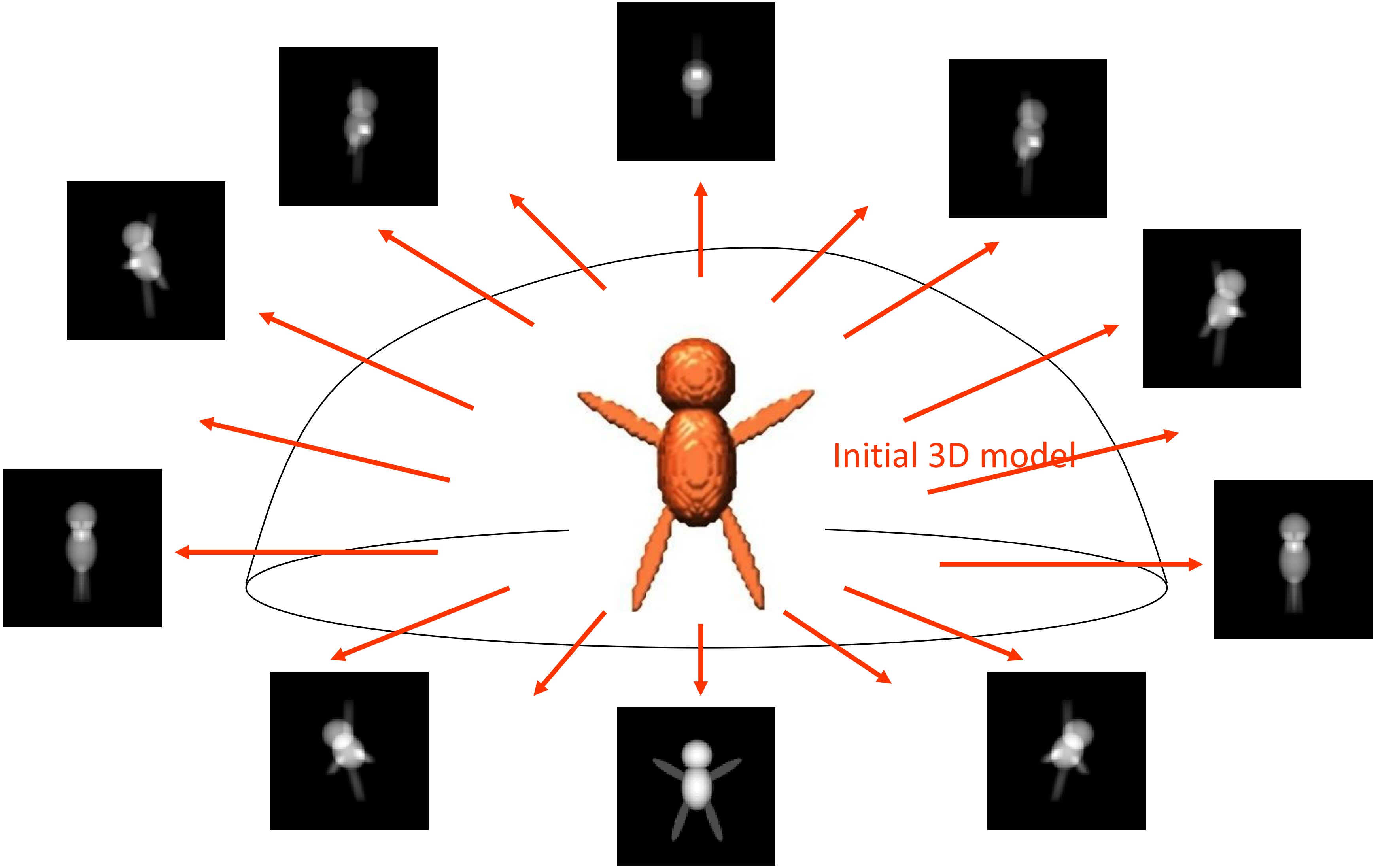


Single particle analysis

- Embedded in ice: many unknown orientations

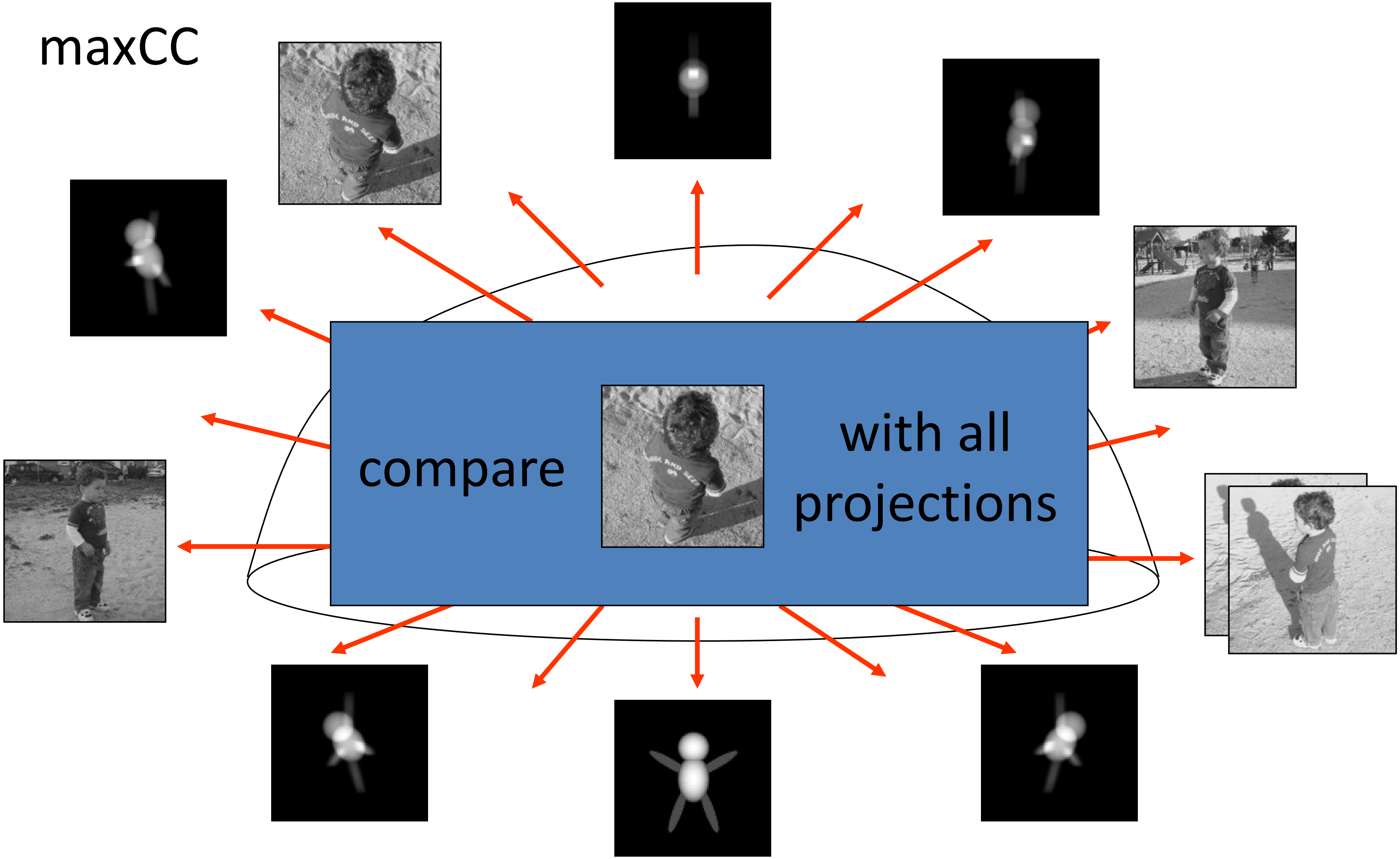
- Combine all 2D projections into a 3D reconstruction

Projection matching



Projection matching

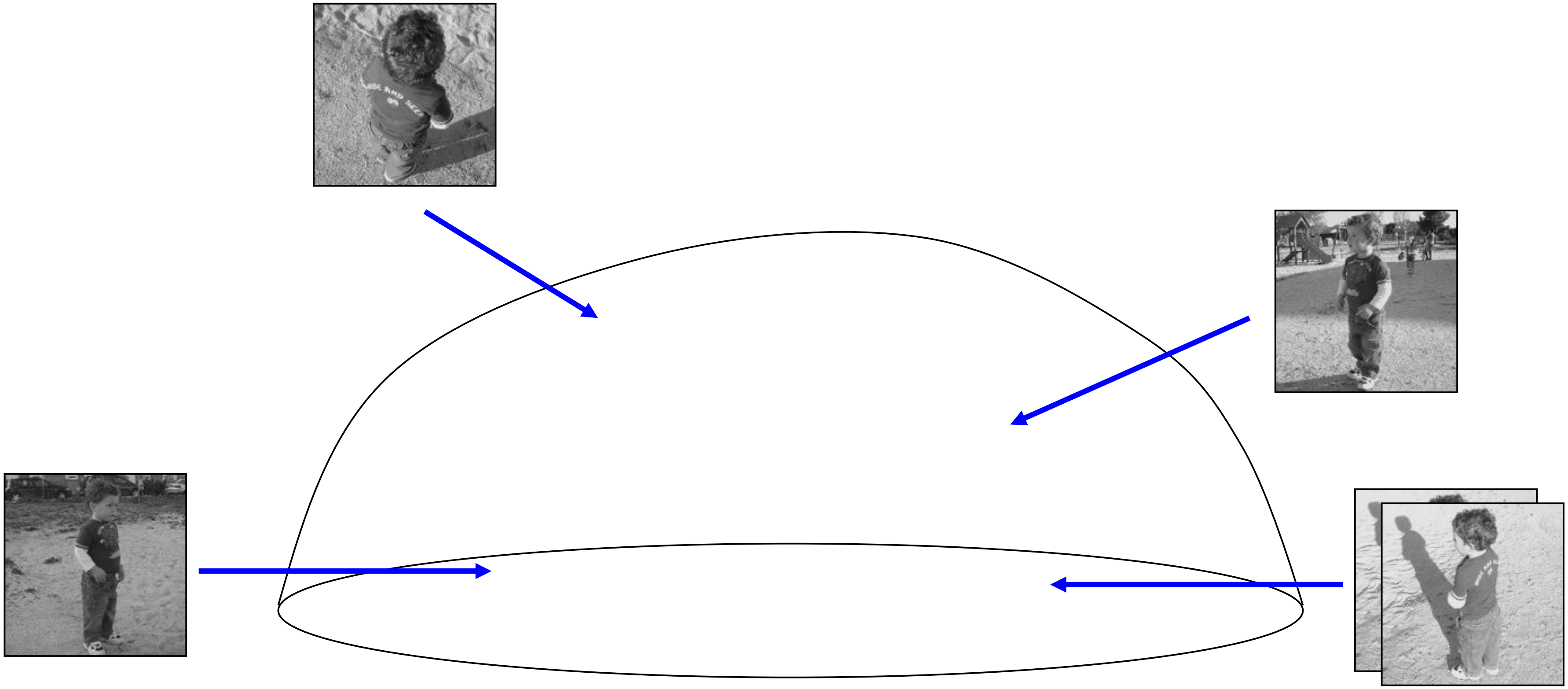
maxCC



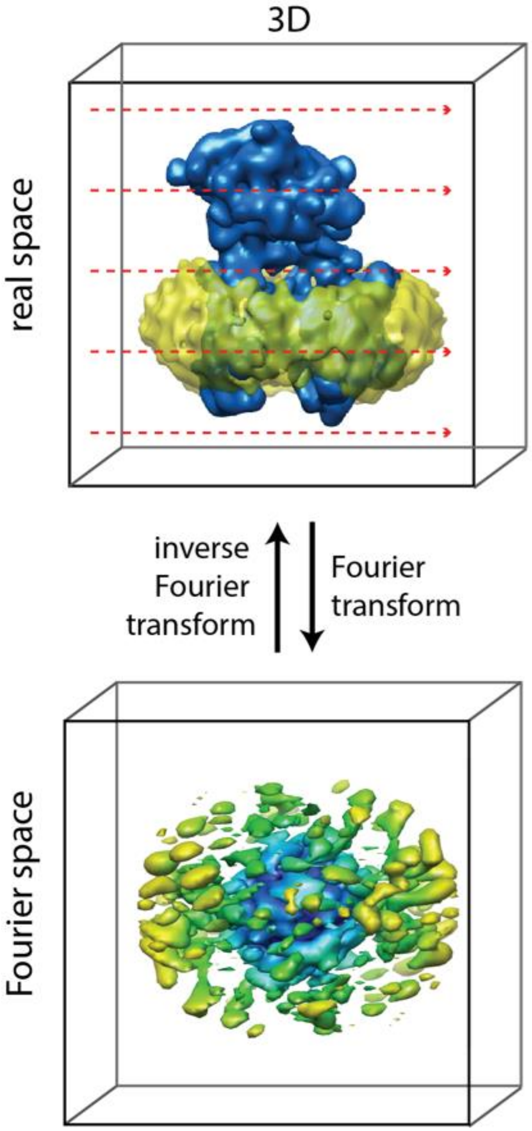
compare

with all projections

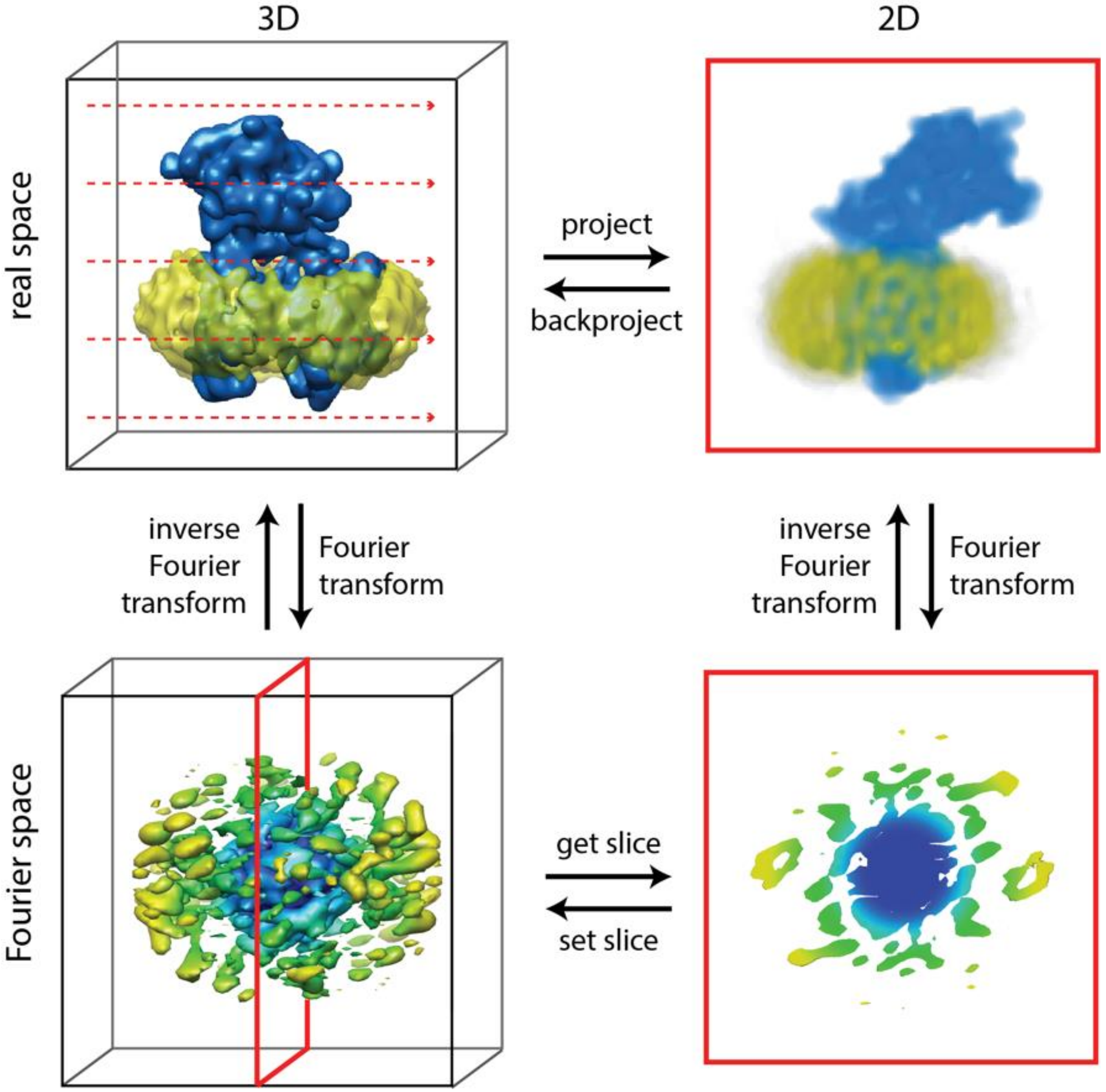
3D reconstruction



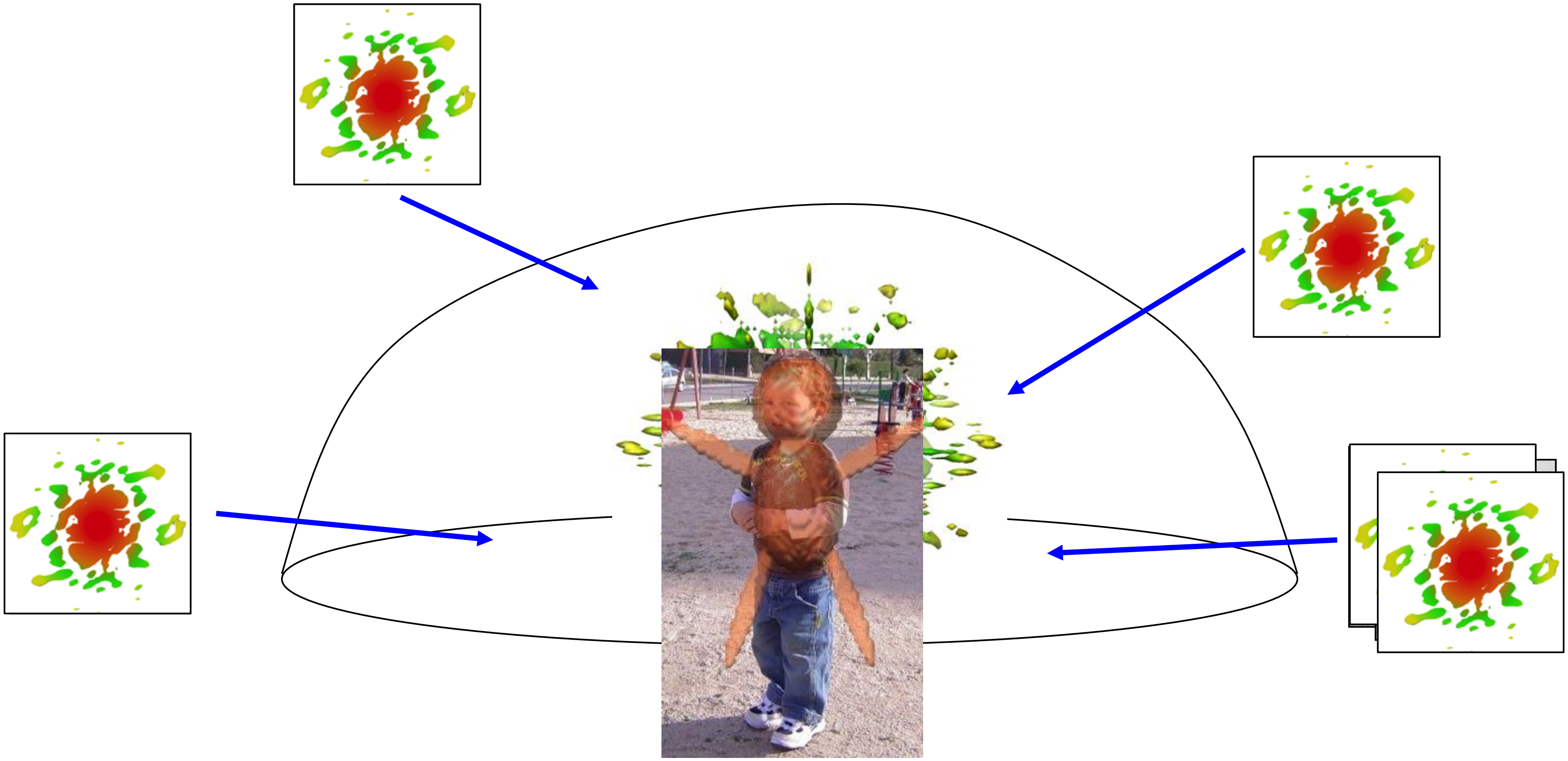
Projection slice theorem



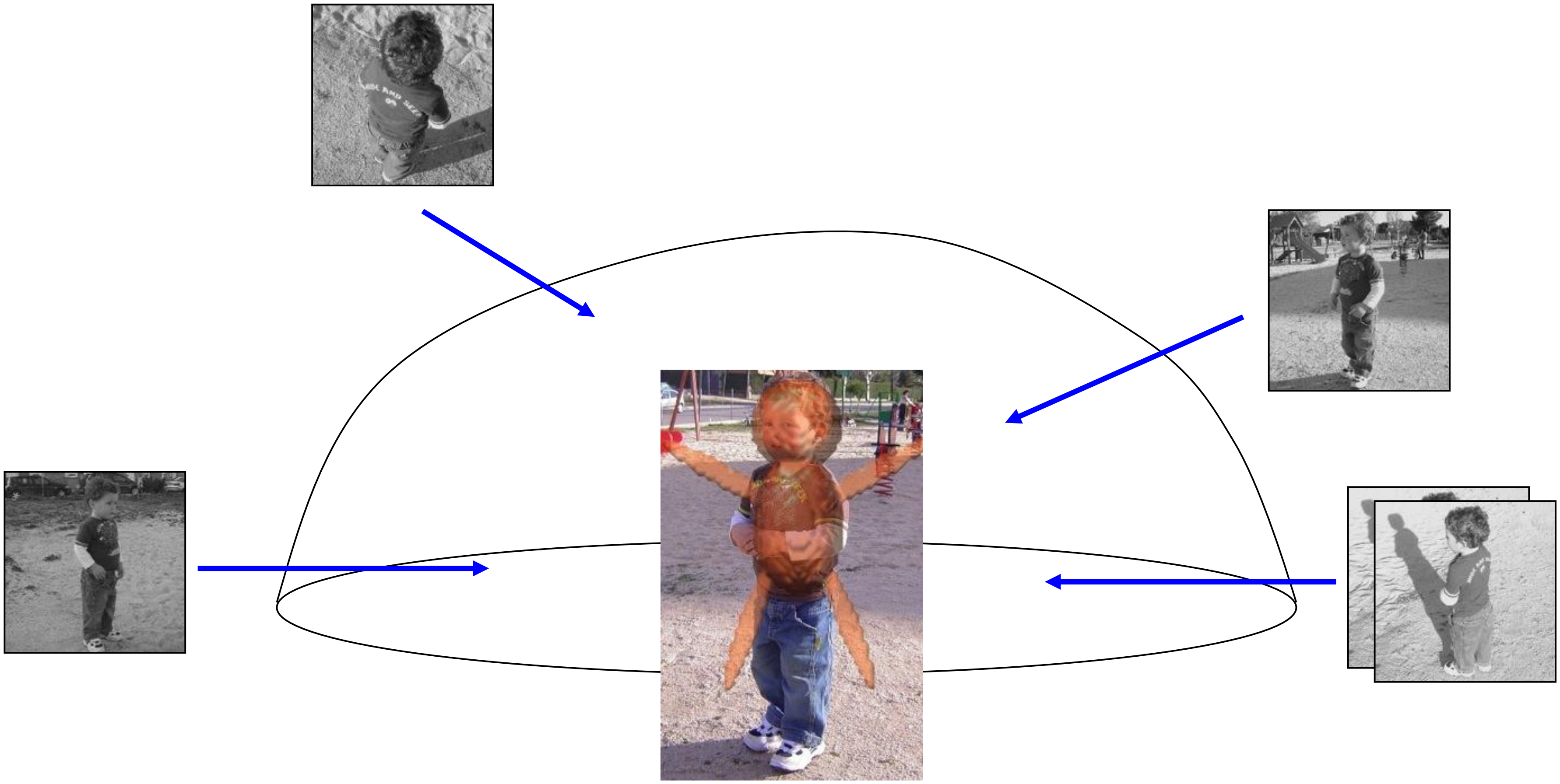
Projection slice theorem



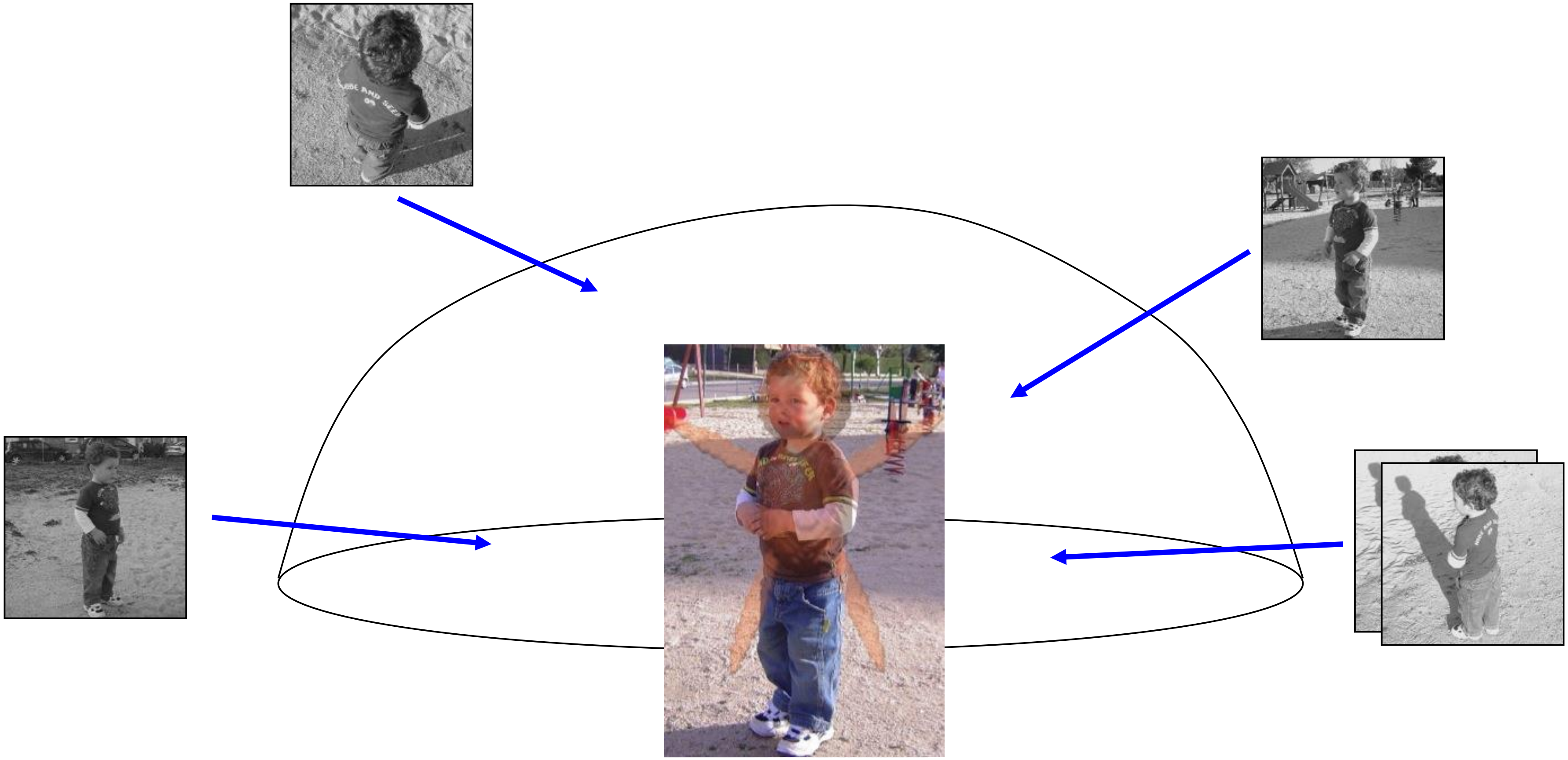
Iterative refinement



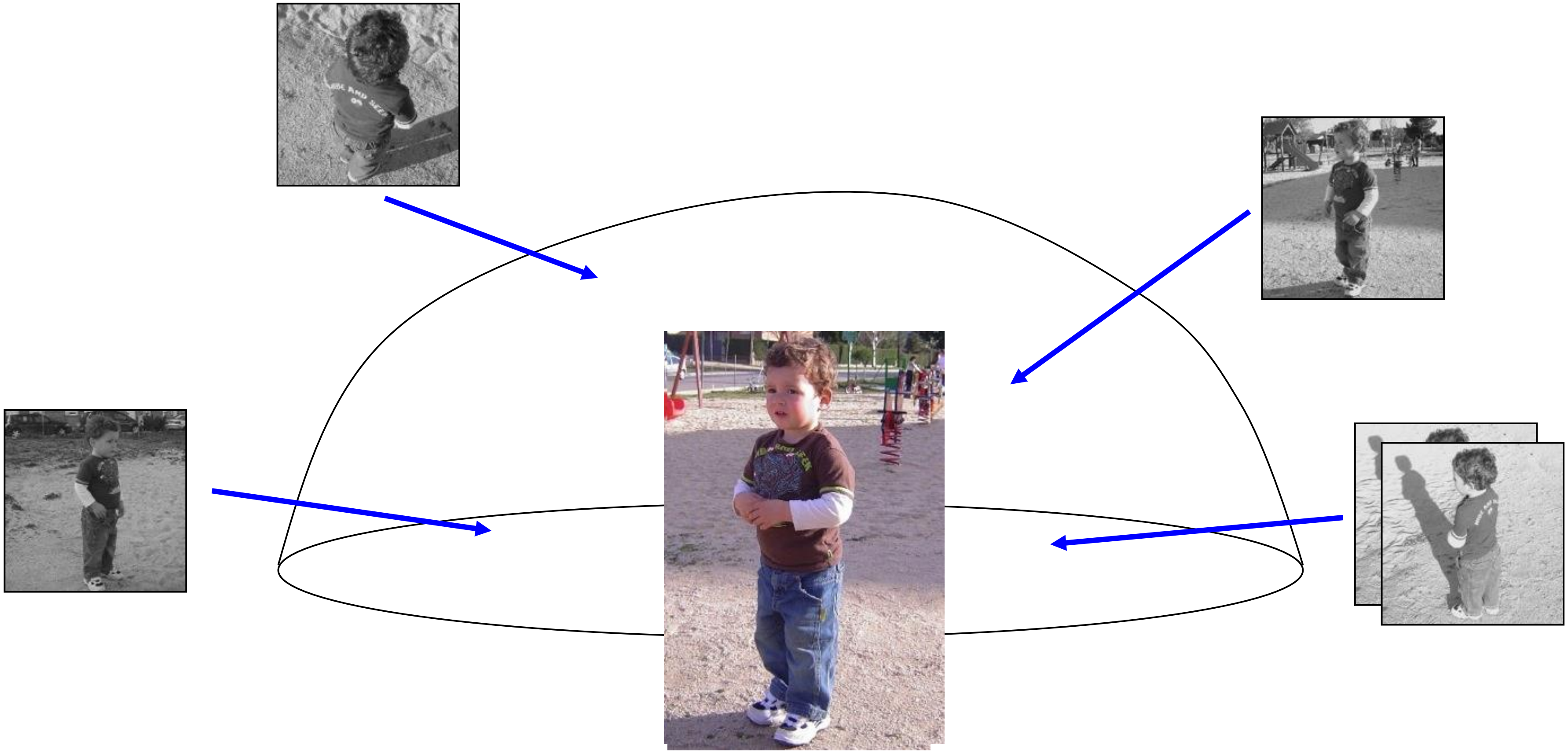
3D reconstruction



Iterative refinement



Iterative refinement



Introduction and new approaches

A comprehensive overview of the major advances that have taken place in the last few years that have enabled maps to achieve "atomic" resolution.

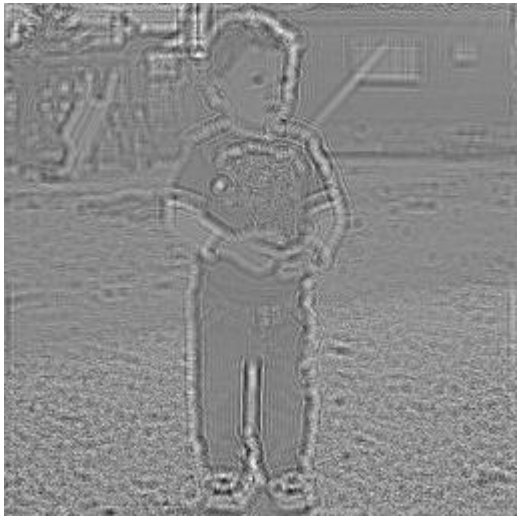
Topics to be covered include:

- 3D reconstruction
- image restoration techniques
- how to deal with heterogeneous populations.

- What are the hot topics in processing?
- What are the major mathematical approaches and available software?
- What are the success stories and the failures?
- Where are the greatest challenges right now and how are we approaching these?
- Do we need completely new algorithms or just incremental improvements on the current ones?
- Mistakes to avoid!

Further inconveniences

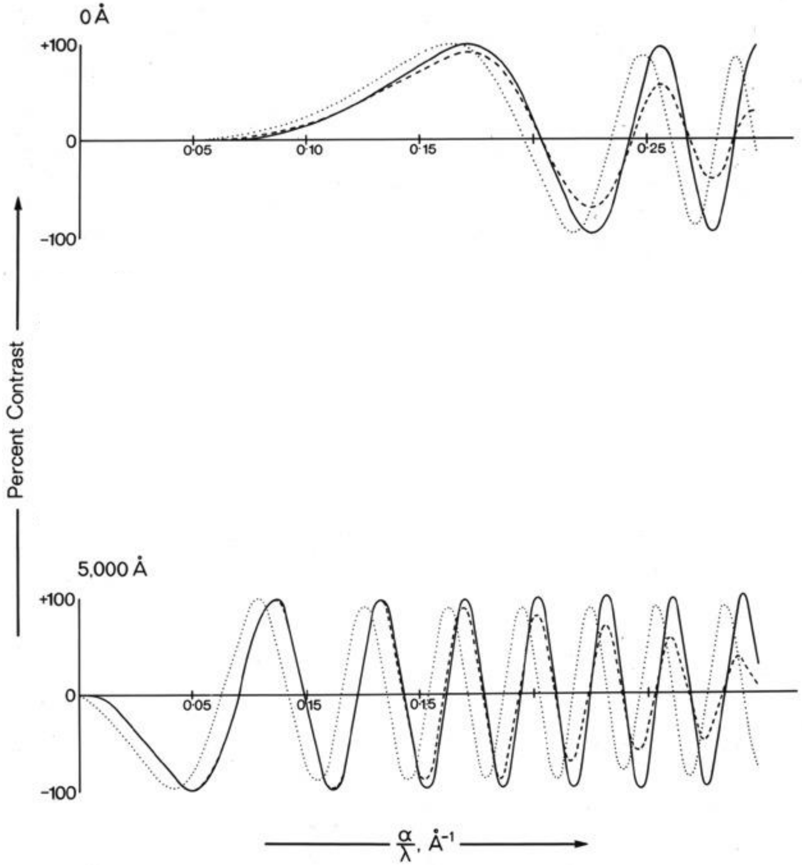
- Defocussing & microscope imperfections introduce artefacts



Phil. Trans. Roy. Soc. Lond. B. **261**, 105–118 (1971) [105]
Printed in Great Britain

Measurement and compensation of defocusing and aberrations by Fourier processing of electron micrographs

By H. P. Erickson and A. Klug, F.R.S.
Medical Research Council Laboratory of Molecular Biology, Cambridge



Data model

- **Real-space**

$$X_i = \mathrm{CTF}_i \text{ Ä } \mathbf{P}_j V_k + N_i$$

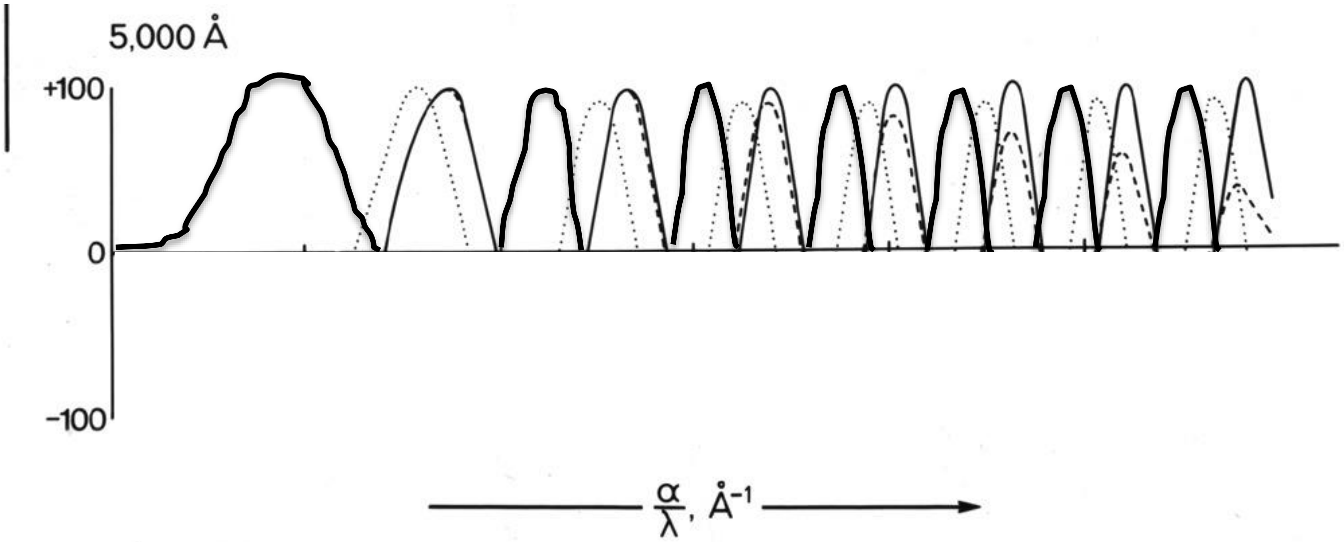
- Convolute w/ CTF
- $\mathbf{P}_\phi$ implements integrals

- **Fourier space**

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

- Multiply w/ CTF
- $\mathbf{P}_\phi$ takes a slice

Phase flipping



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

(3D) Wiener filter

Optimal linear filter

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

- 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!

Introduction and new approaches

A comprehensive overview of the major advances that have taken place in the last few years that have enabled maps to achieve "atomic" resolution.

Topics to be covered include:

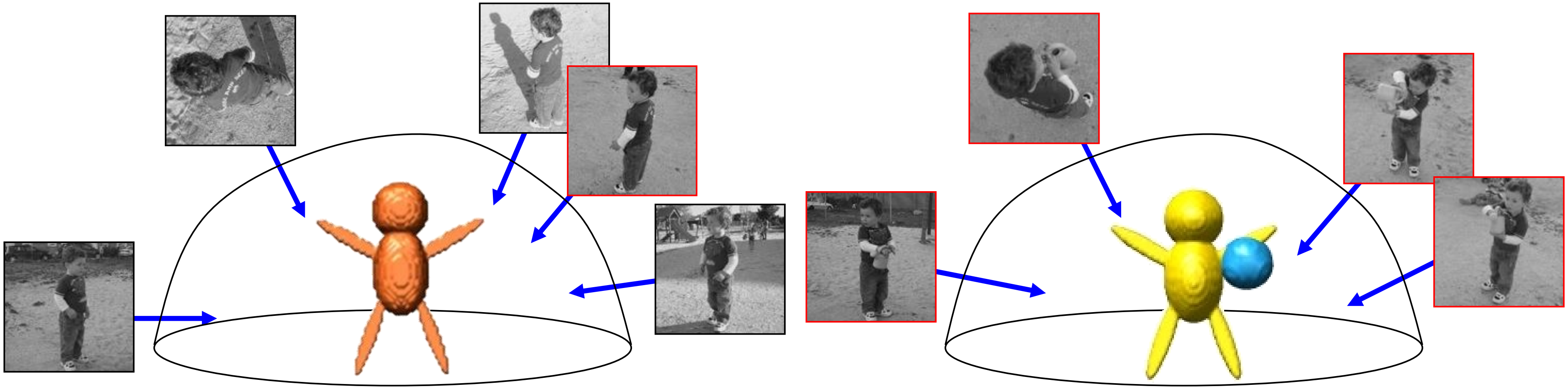
- 3D reconstruction
- image restoration techniques
- how to deal with heterogeneous populations.

- What are the hot topics in processing?
- What are the major mathematical approaches and available software?
- What are the success stories and the failures?
- Where are the greatest challenges right now and how are we approaching these?
- Do we need completely new algorithms or just incremental improvements on the current ones?
- Mistakes to avoid!

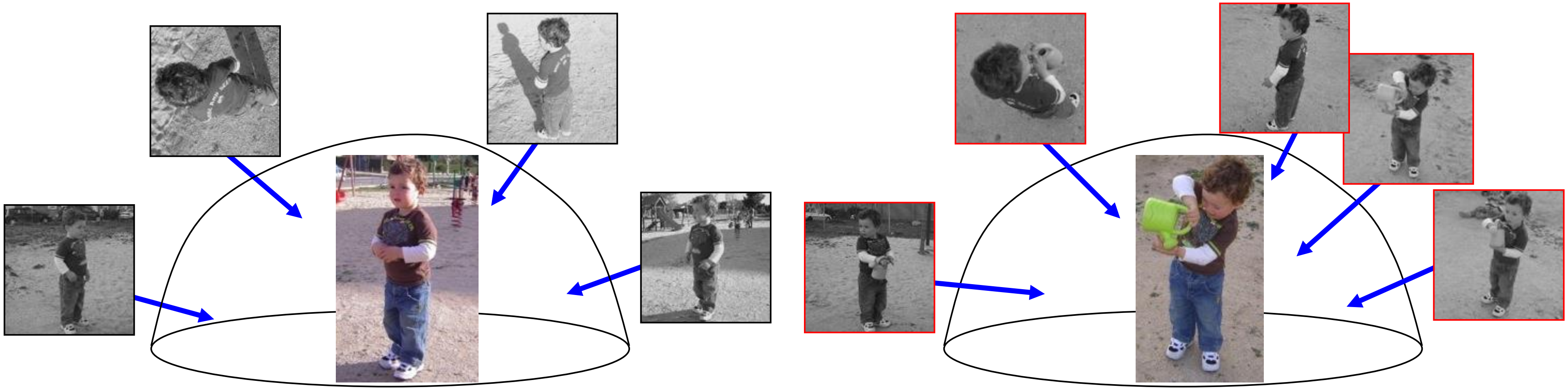
Structural heterogeneity

complex!

Multi-reference refinement



Multi-reference refinement



Introduction and new approaches

A comprehensive overview of the major advances that have taken place in the last few years that have enabled maps to achieve "atomic" resolution.

Topics to be covered include:

- 3D reconstruction
- image restoration techniques
- how to deal with heterogeneous populations.

- What are the hot topics in processing?
- What are the major mathematical approaches and available software?
- What are the success stories and the failures?
- Where are the greatest challenges right now and how are we approaching these?
- Do we need completely new algorithms or just incremental improvements on the current ones?
- Mistakes to avoid!

Hot topics?

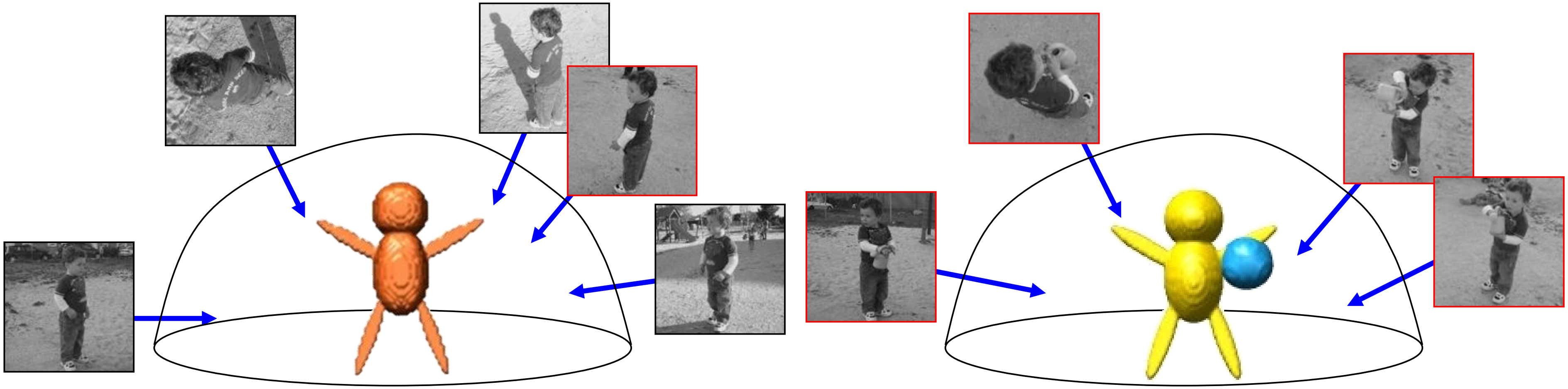
- Unsupervised (3D) classification
- High-resolution refinement
  - Prevention of overfitting
  - Movie-processing

Hot topics?

- Unsupervised (3D) classification

- High-resolution refinement
  - Prevention of overfitting
  - Movie-processing

Supervised classification



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

Maximum-likelihood approaches

- Marginalize over orientations & classes
  - Probability-weighted assignments

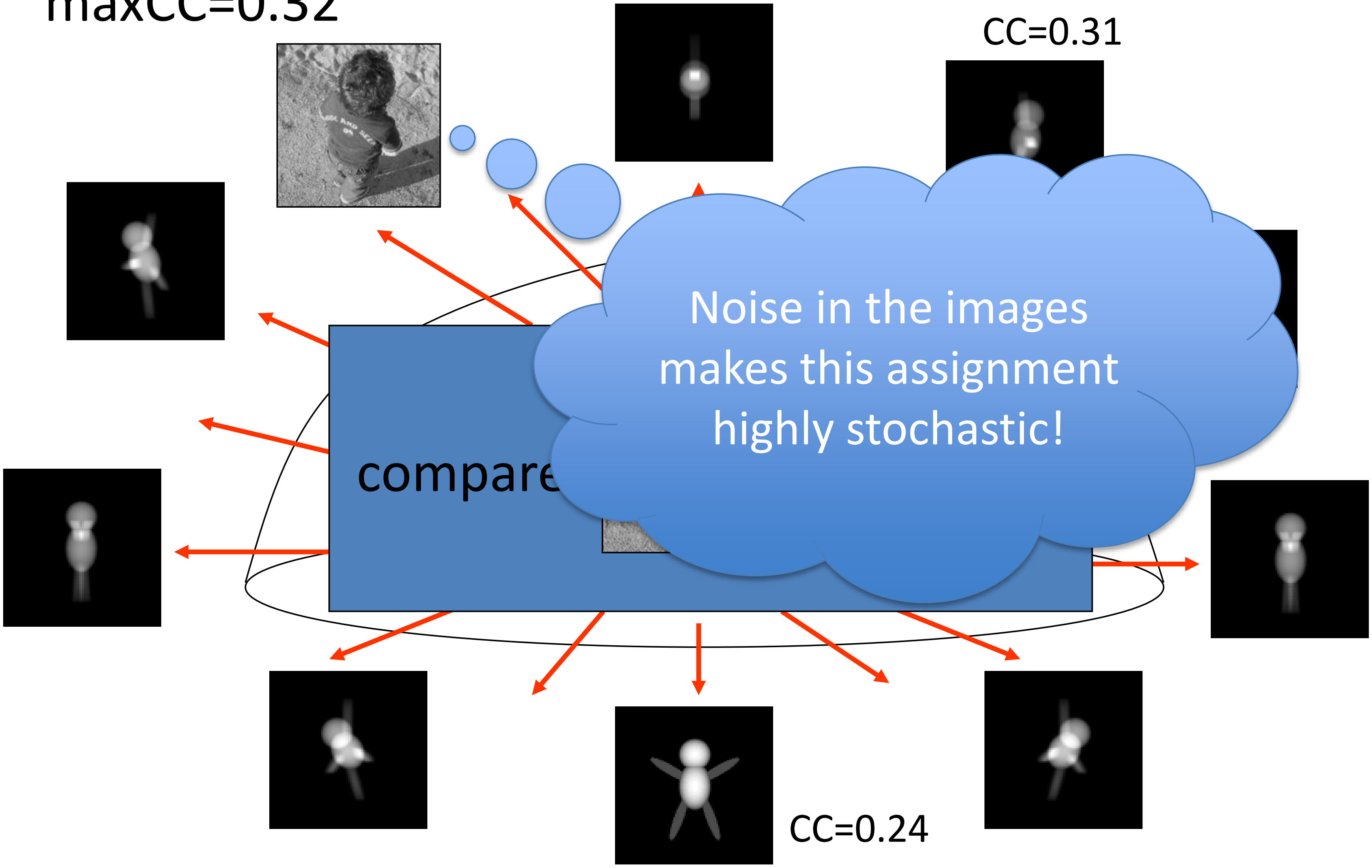
- First described by Fred Sigworth (JSB-1998)
  - For 2D-alignment, single-reference
  - Real-space data model (white-noise model)
  - Matlab scripts

- Then extended for 2D & 3D classification (2005-2010)
  - XMIPP

- 3D ML-based classification without marginalizing over orientations (Niko, 2013)
  - FREALIGN

Maximum cross-correlation (least-squares)

maxCC=0.32



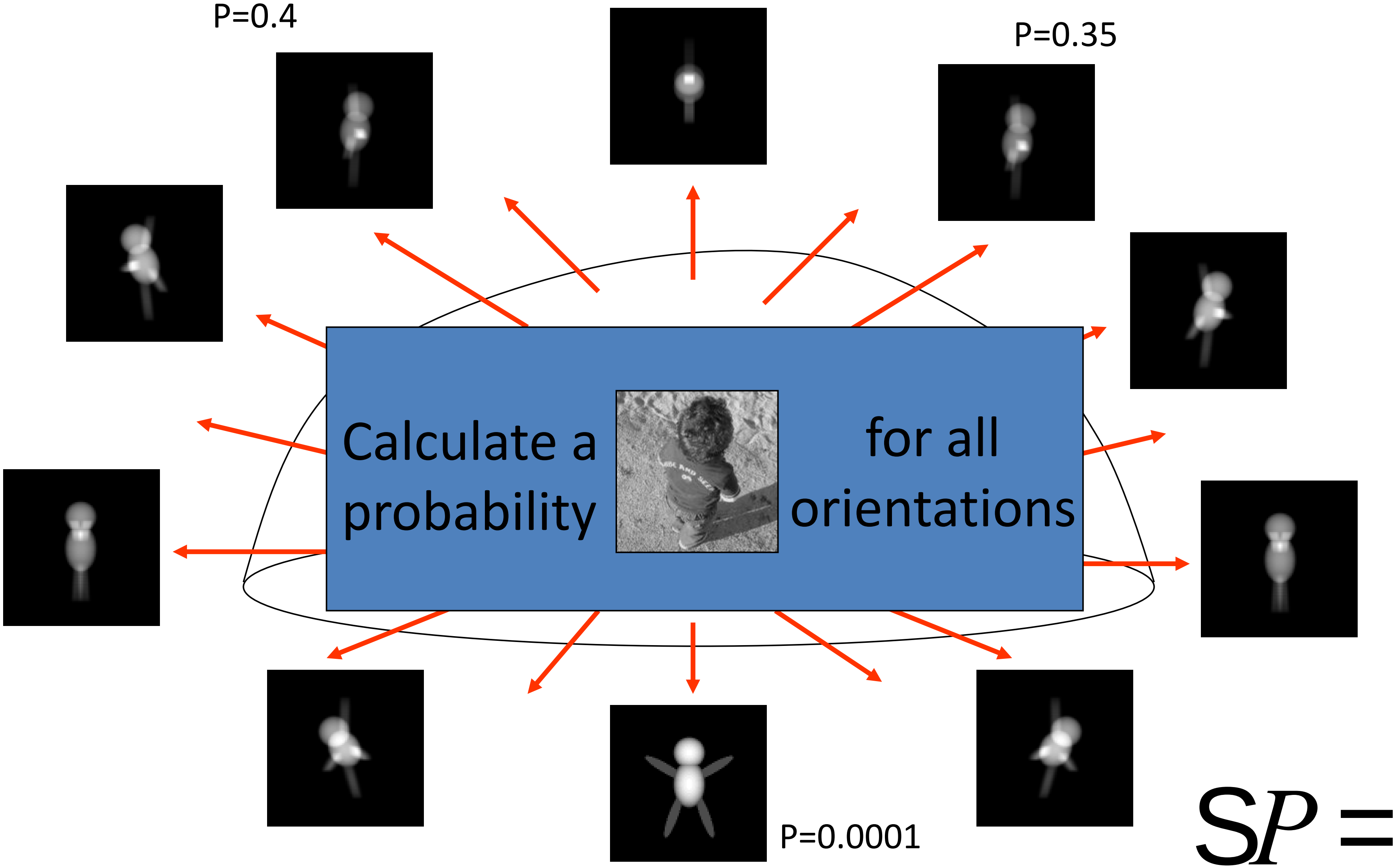
CC=0.31

compare

Noise in the images makes this assignment highly stochastic!

CC=0.24

Maximum likelihood



Calculate a probability for all orientations

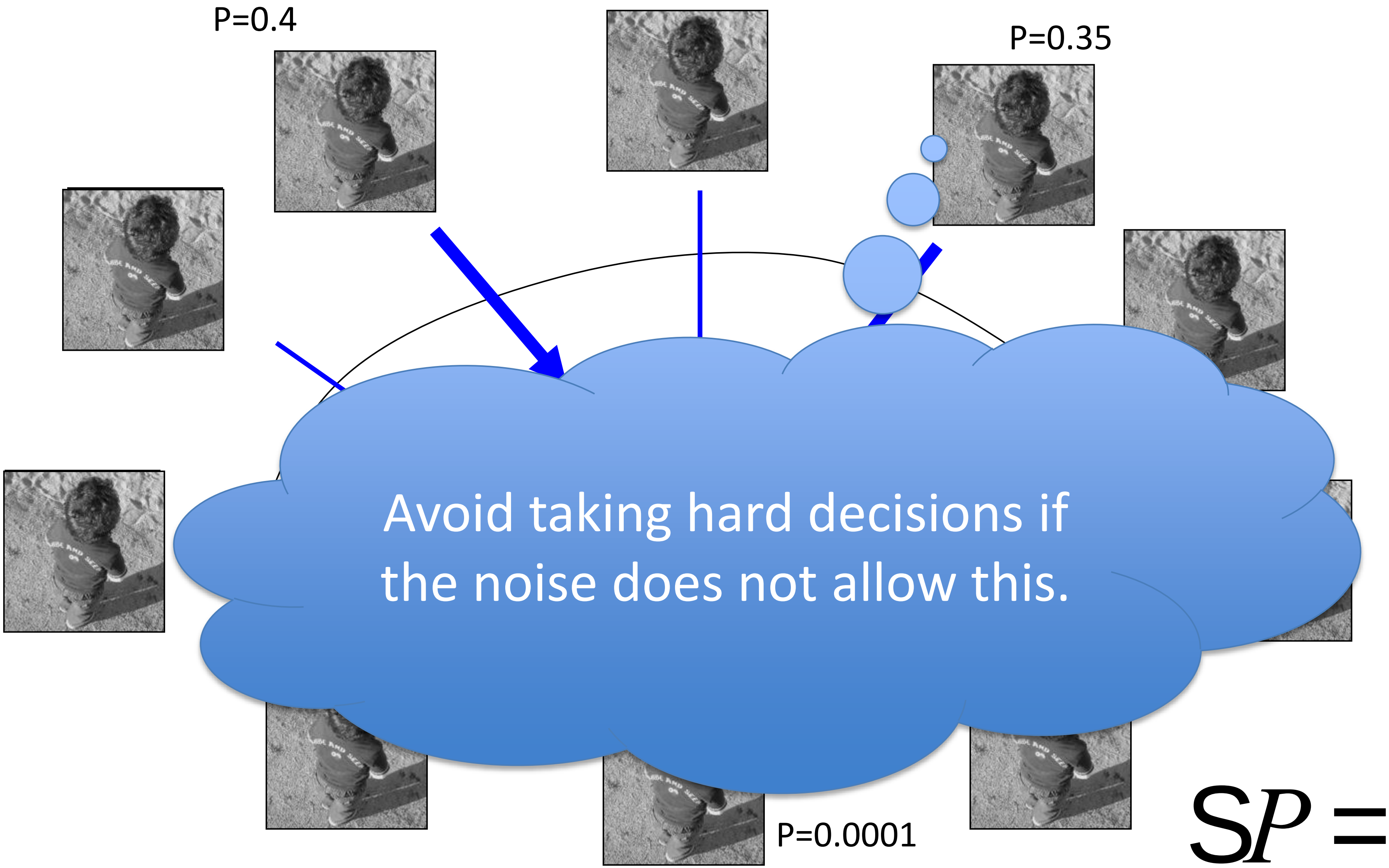
P=0.4

P=0.35

P=0.0001

$SP=1$

Maximum likelihood



P=0.4

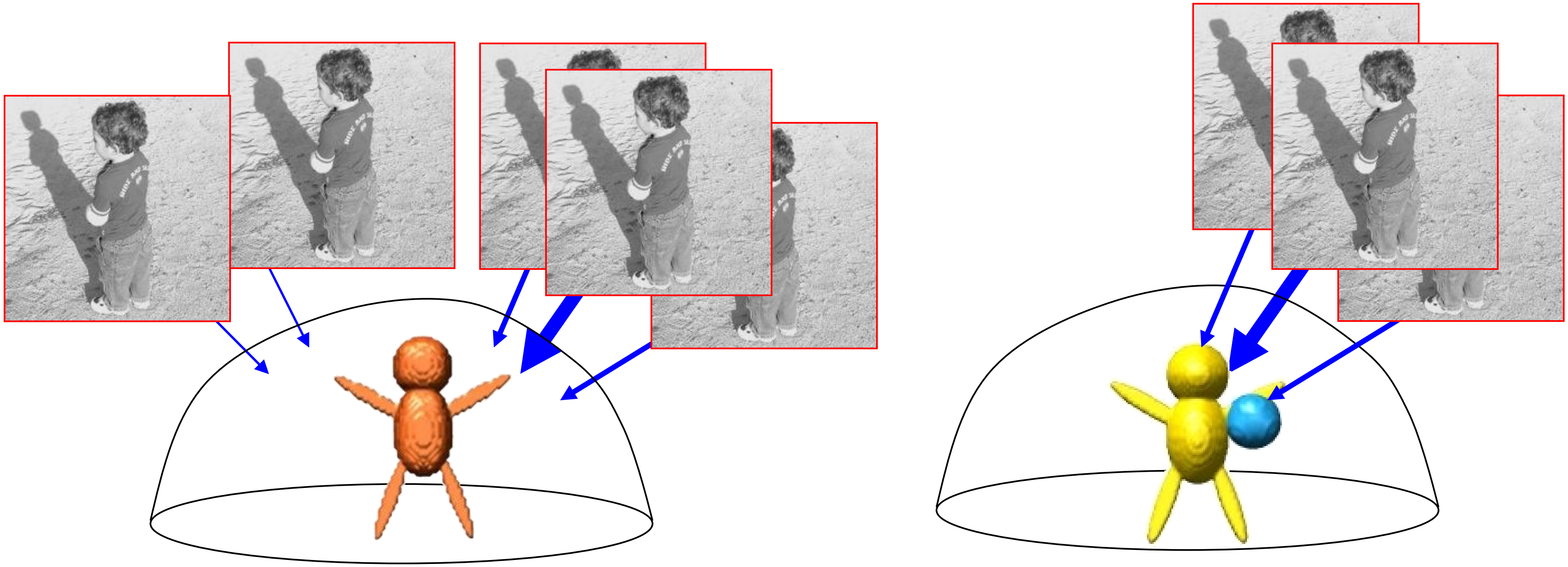
P=0.35

P=0.0001

Avoid taking hard decisions if the noise does not allow this.

$SP = 1$

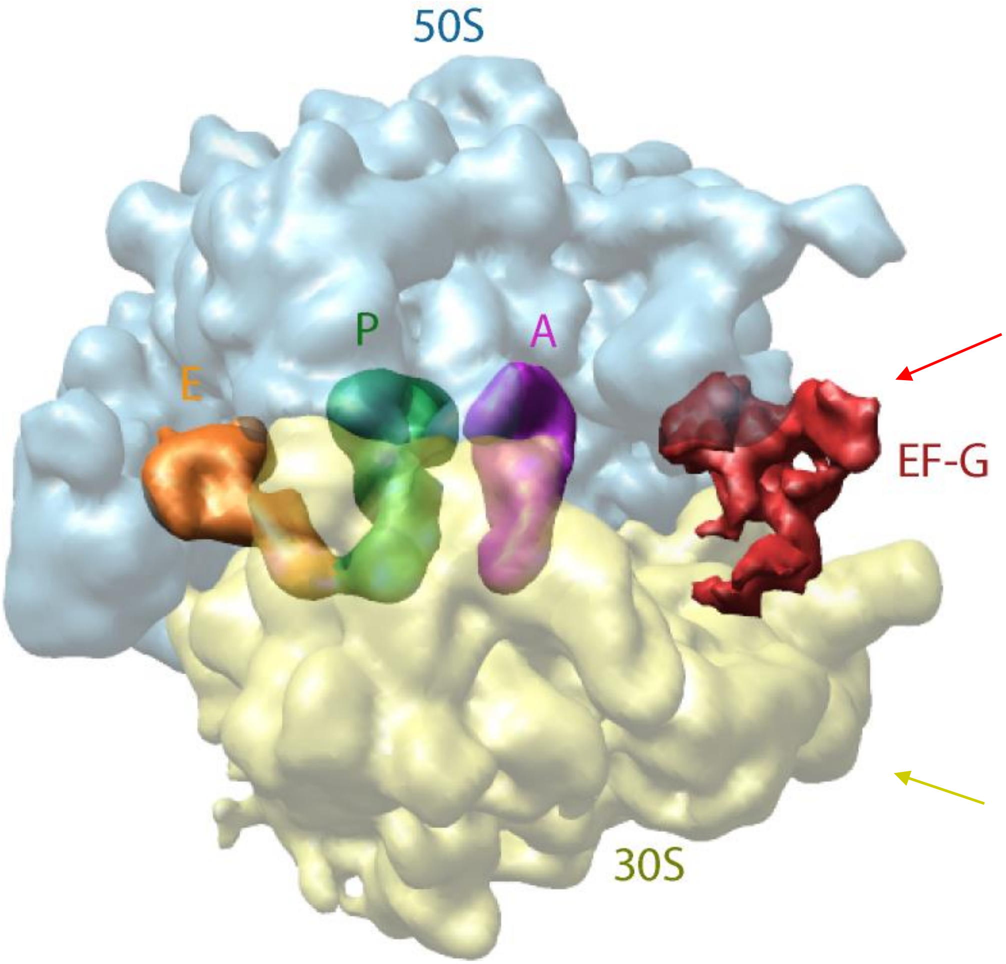
ML3D classification



Probability-weighted angular & class assignments

Prelim. ribosome reconstruction

91,114 particles; 9.9 Å resolution



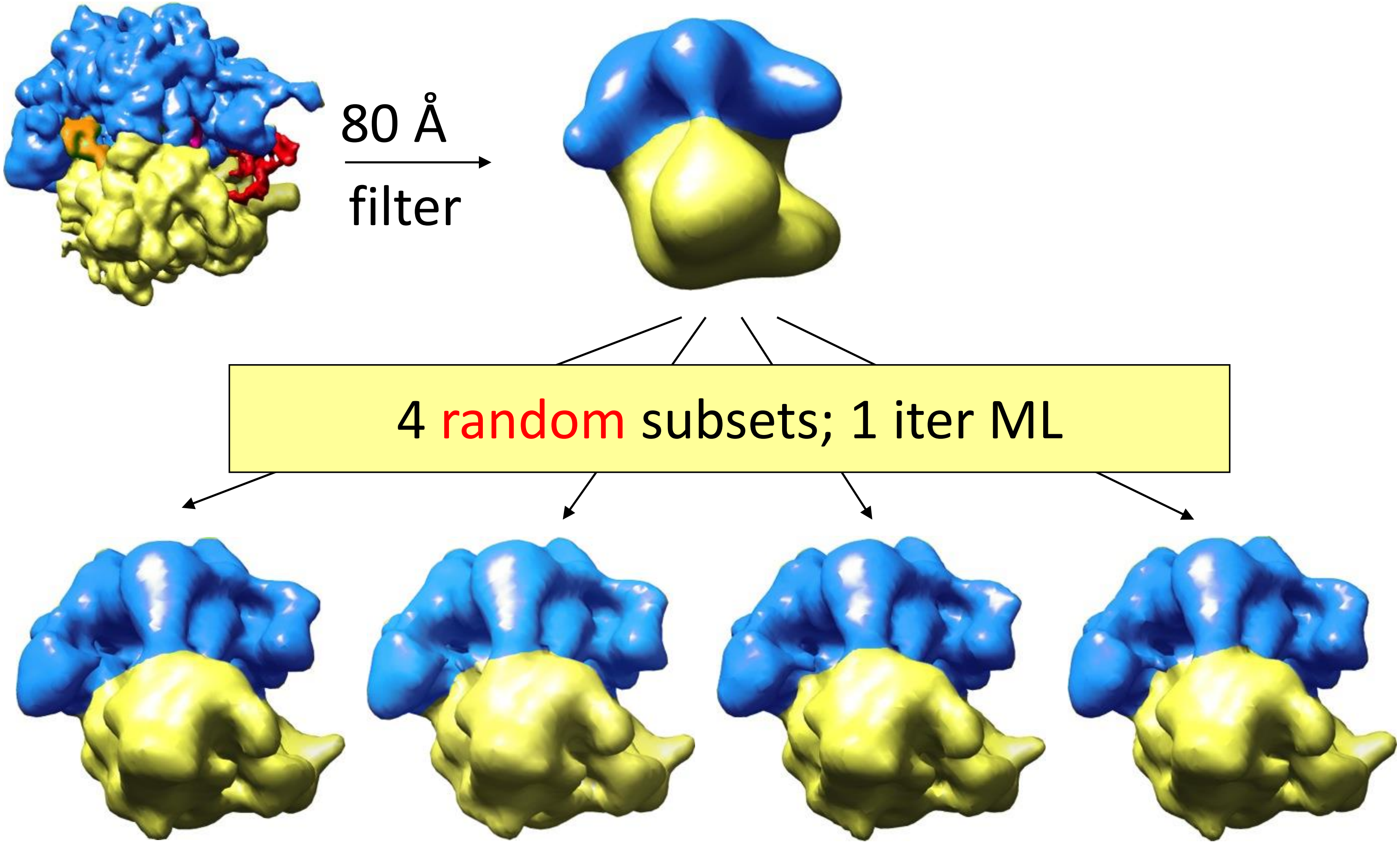
fragmented

(depicted at a lower threshold)

blurred

In collaboration with Haixiao Gao & Joachim Frank

Seed generation



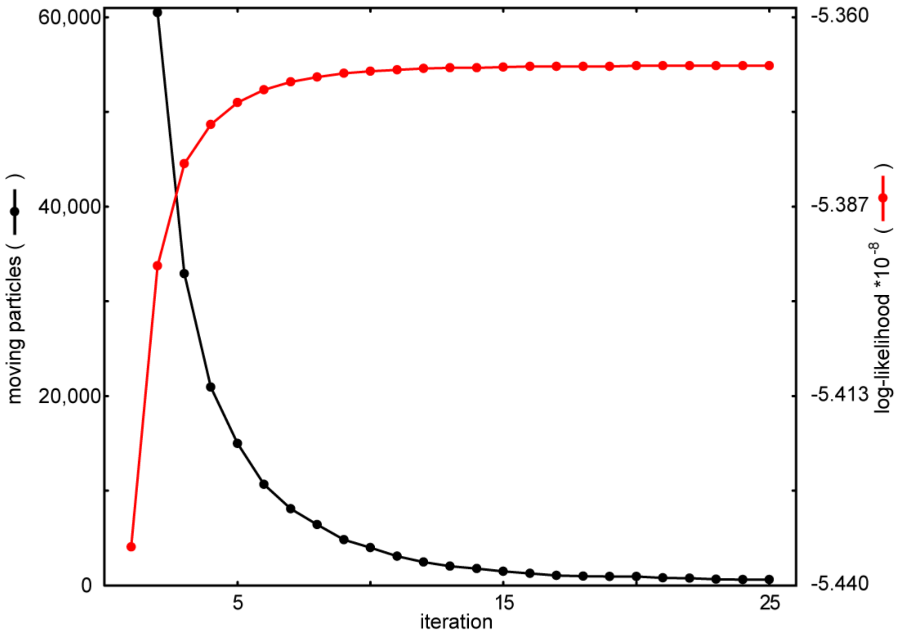
80 Å

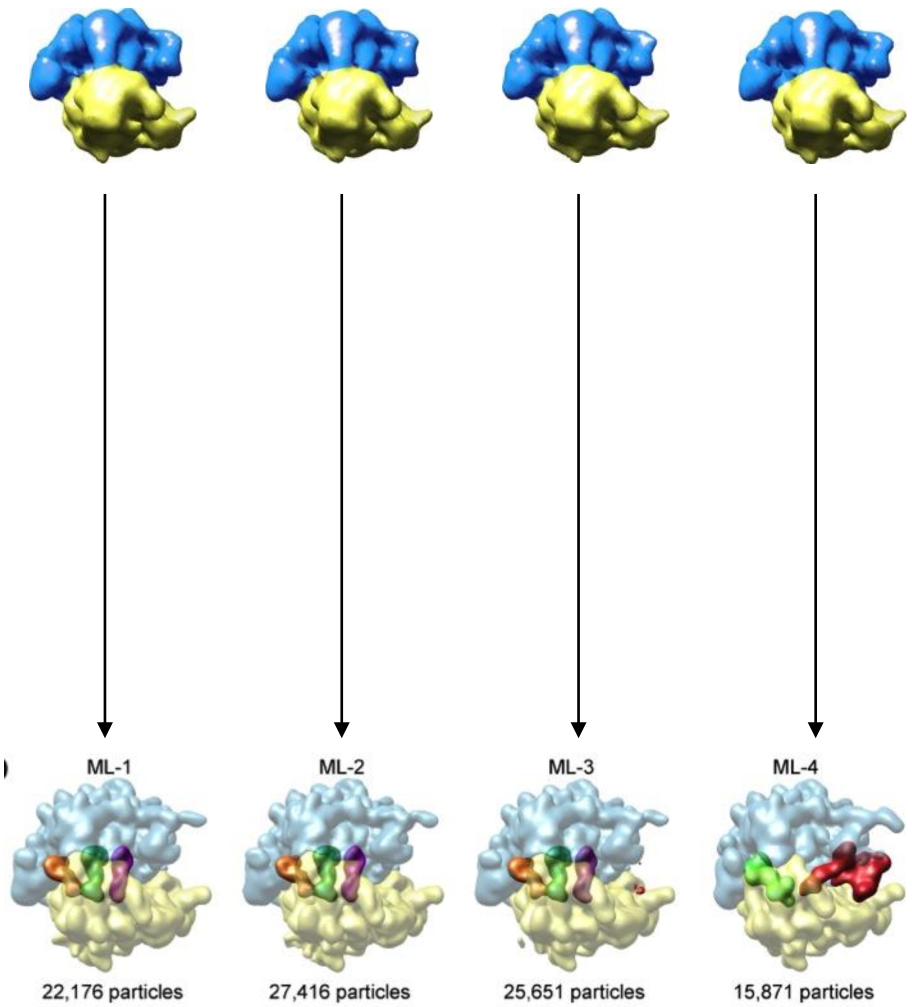
filter

4 random subsets; 1 iter ML

ML3D-classification

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





Regularised likelihood approach

(2012)

- Data model in Fourier-space
  - Colored (correlated) noise
  - CTF-correction

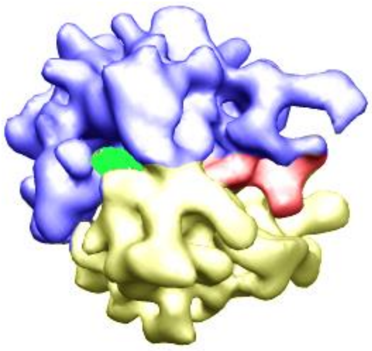
- Marginalize over orientations & classes
  - Probability-weighted assignments

- 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

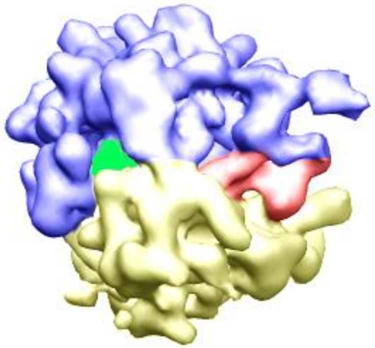
Classify structural variability

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



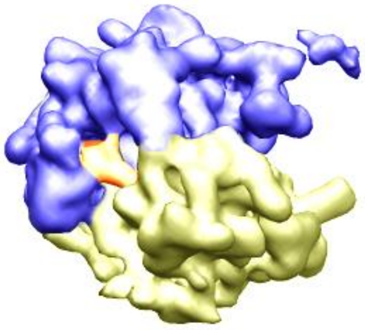
24%

26Å



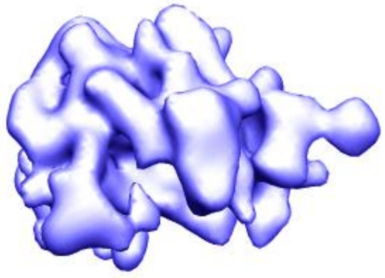
28%

19Å



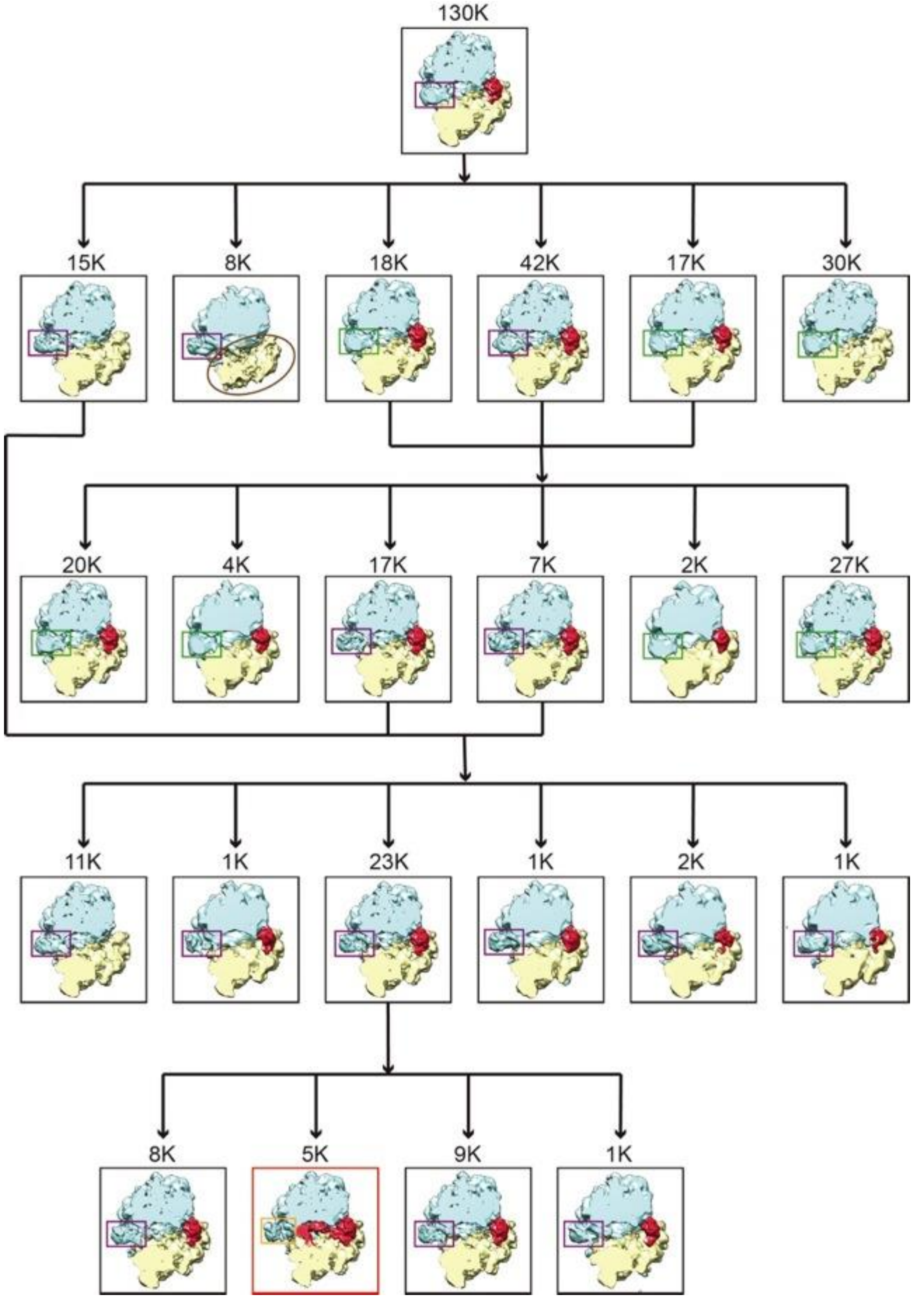
42%

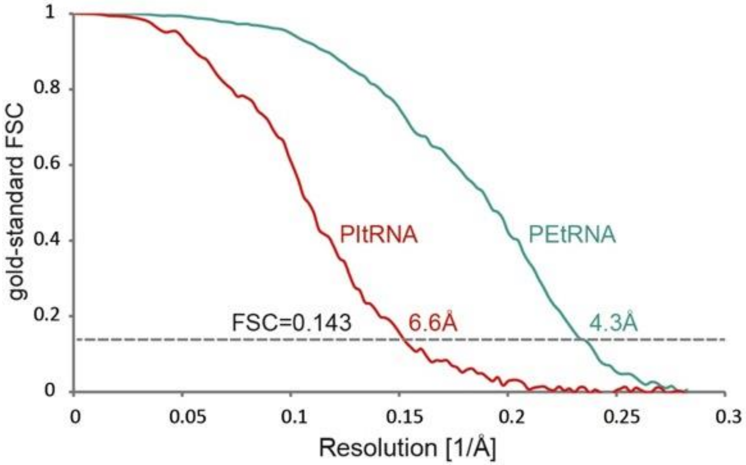
19Å

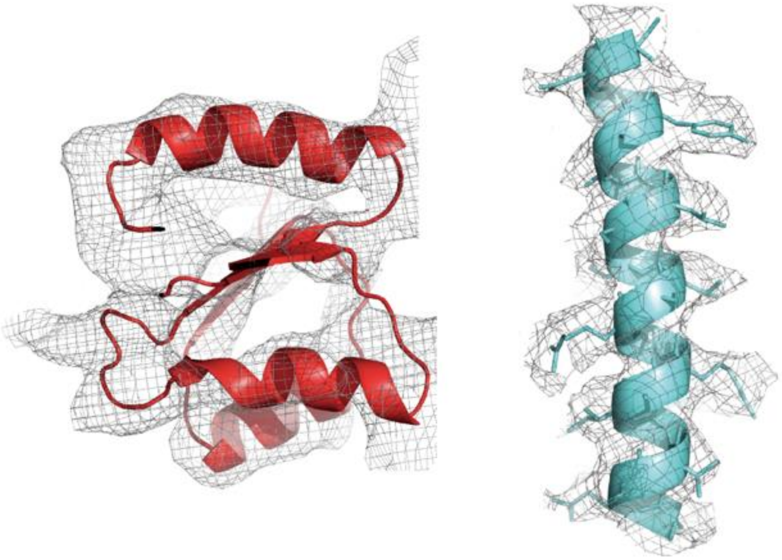


6%

30Å

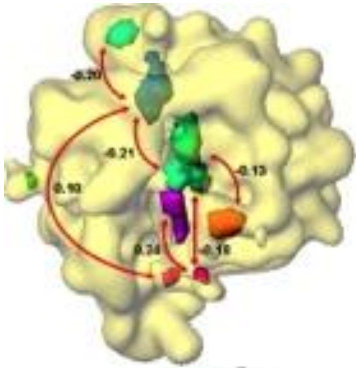

130K
15K
8K
18K
42K
17K
30K
20K
4K
17K
7K
2K
27K
11K
1K
23K
1K
2K
1K
8K
5K
9K
1K

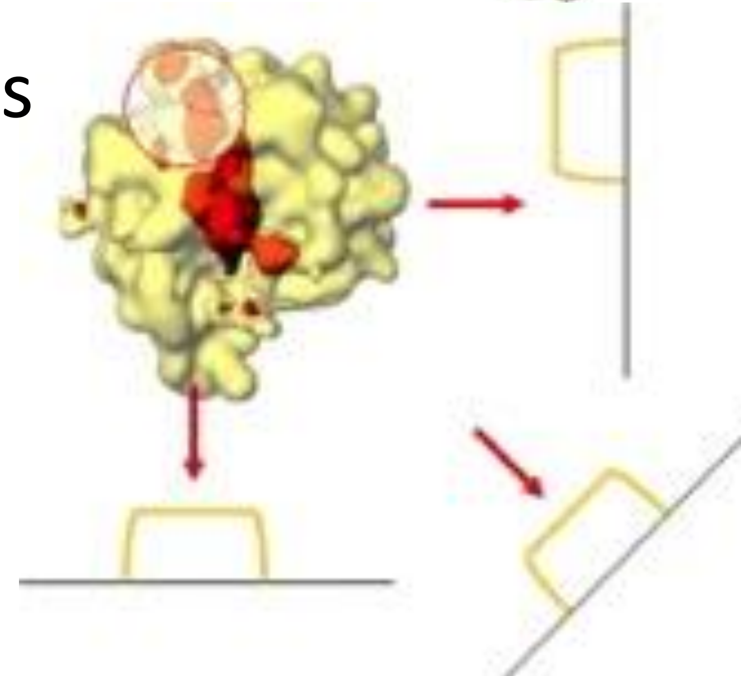

gold-standard FSC
1
0.8
0.6
0.4
0.2
0
0
0.05
0.1
0.15
0.2
0.25
0.3
Resolution [1/Å]
PItRNA
PEtRNA
FSC=0.143
6.6Å
4.3Å



Other 3D classification tools

- Non-ML multi-reference refinement
  - EMAN/IMAGIC/SPIDER/…
- Boot-strapping & 3D (co-)variance map
- Focussed classification & MSA of bootstrapped maps
  - SPARX

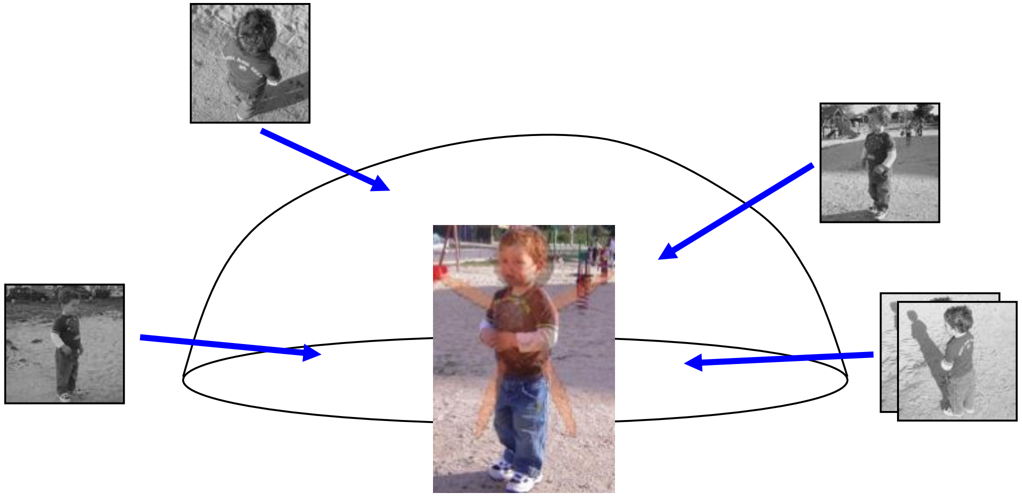




Hot topics?

- Unsupervised (3D) classification
- High-resolution refinement
    - Prevention of overfitting
    - Movie-processing

Overfitting



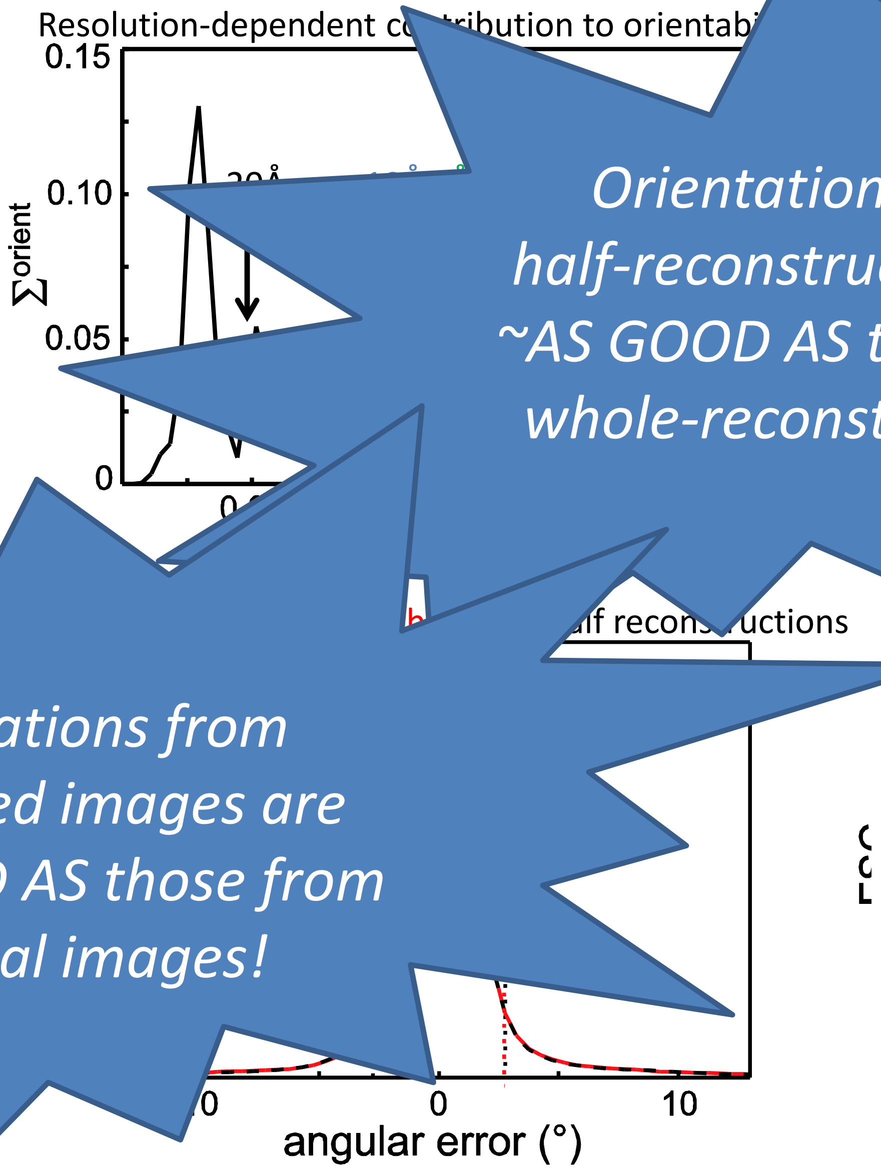
- Some noise slips into reconstruction
- Model bias reproduces that noise
- Iteration re-enforces the noise
- Over-estimated resolution & noisy maps

Prevention of overfitting

- Two main approaches
  - Limit resolution in your refinement
    - FREALIGN
    - (ANY)

  - Independently refine 2 independent data-halves
    - Gold-standard refinement / FSC (Steve Ludtke)
    - EMAN
    - RELION
    - (ANY)

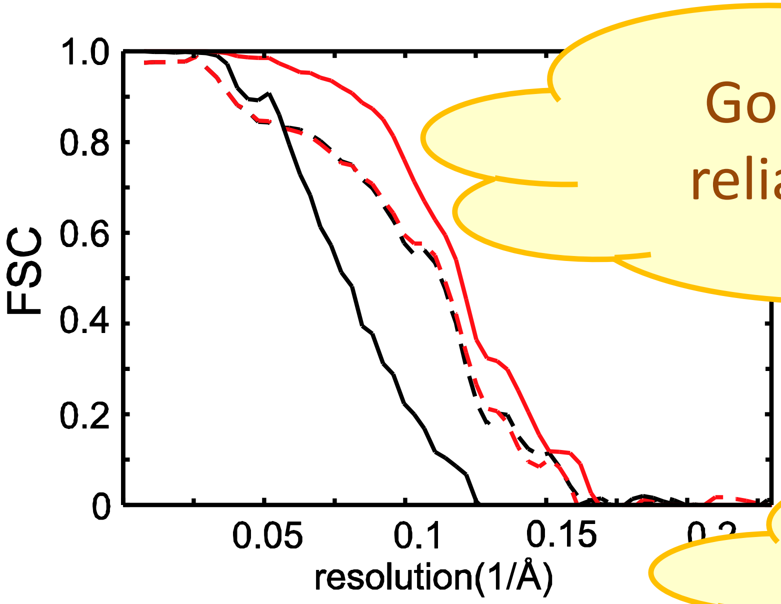
Only lower resolution data drive alignment



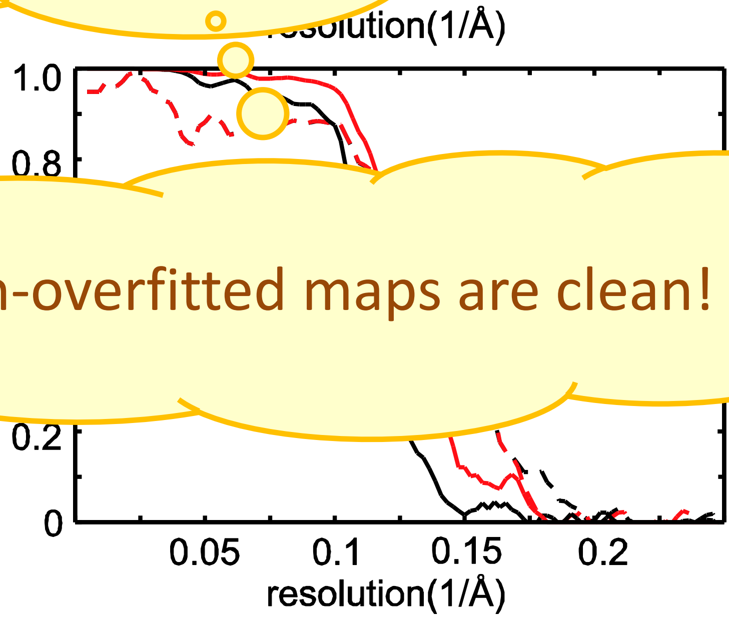
Orientations from half-reconstructions are ~AS GOOD AS those from whole-reconstructions!

Orientations from 8A-filtered images are ~AS GOOD AS those from original images!

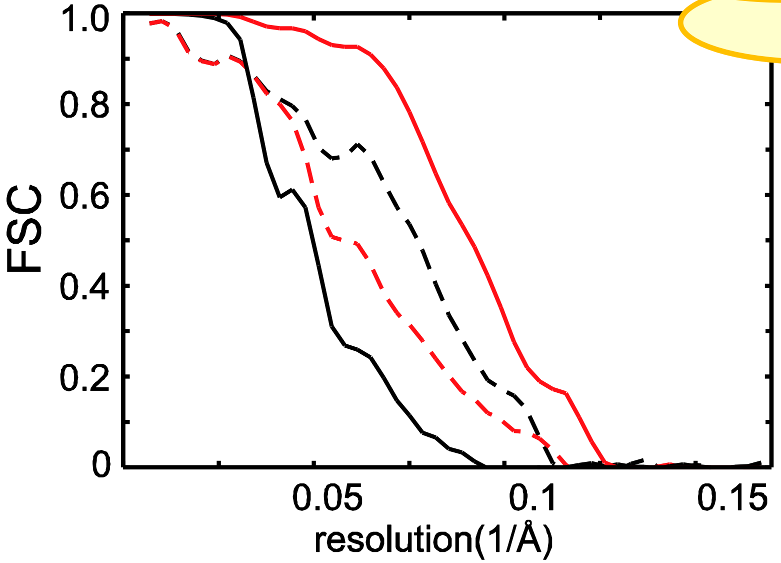
Resolution criteria...



Gold-standard FSC=0.143 is a reliable indicator of resolution



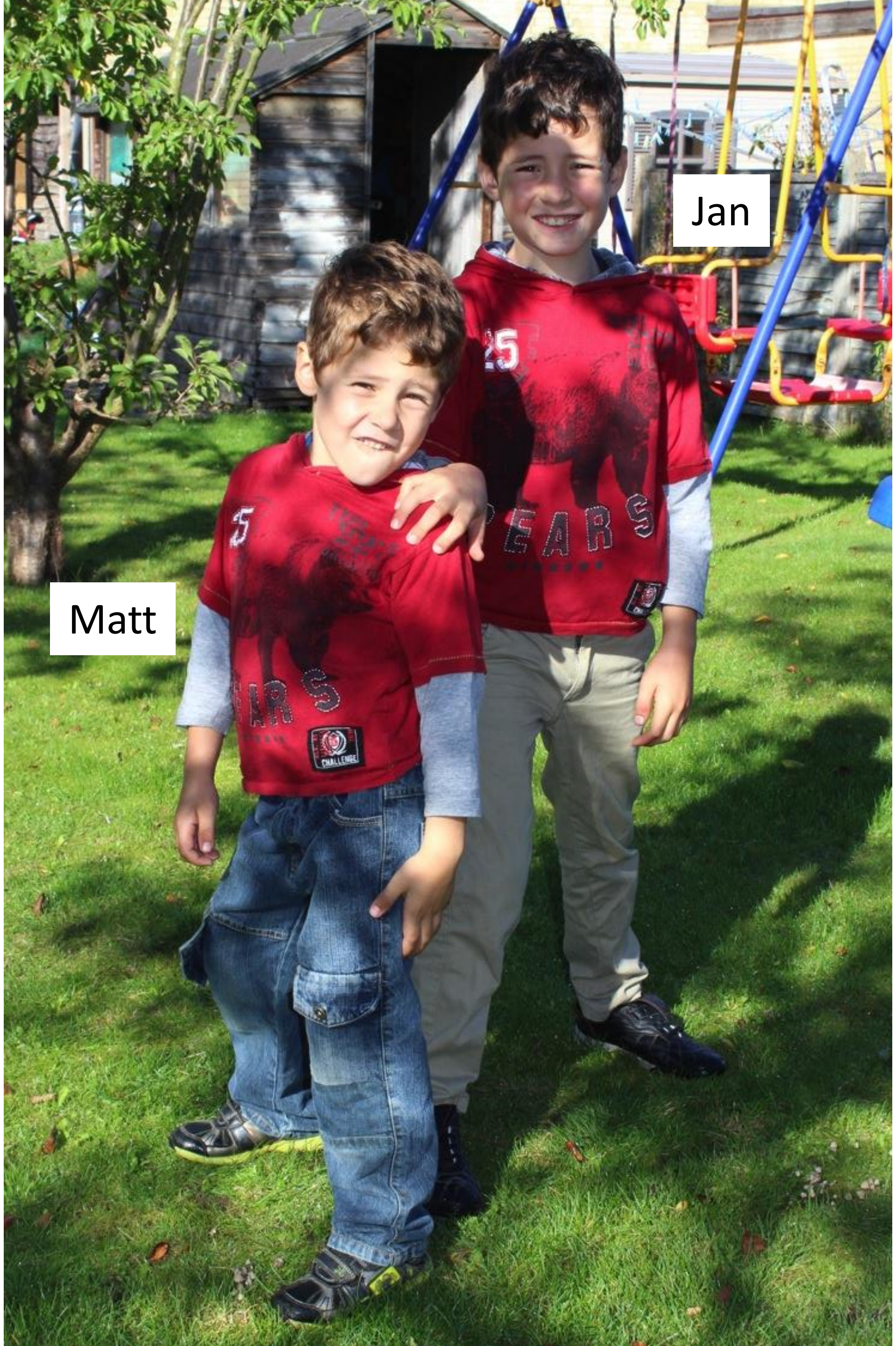
Non-overfitted maps are clean!



Hot topics?

- Unsupervised (3D) classification

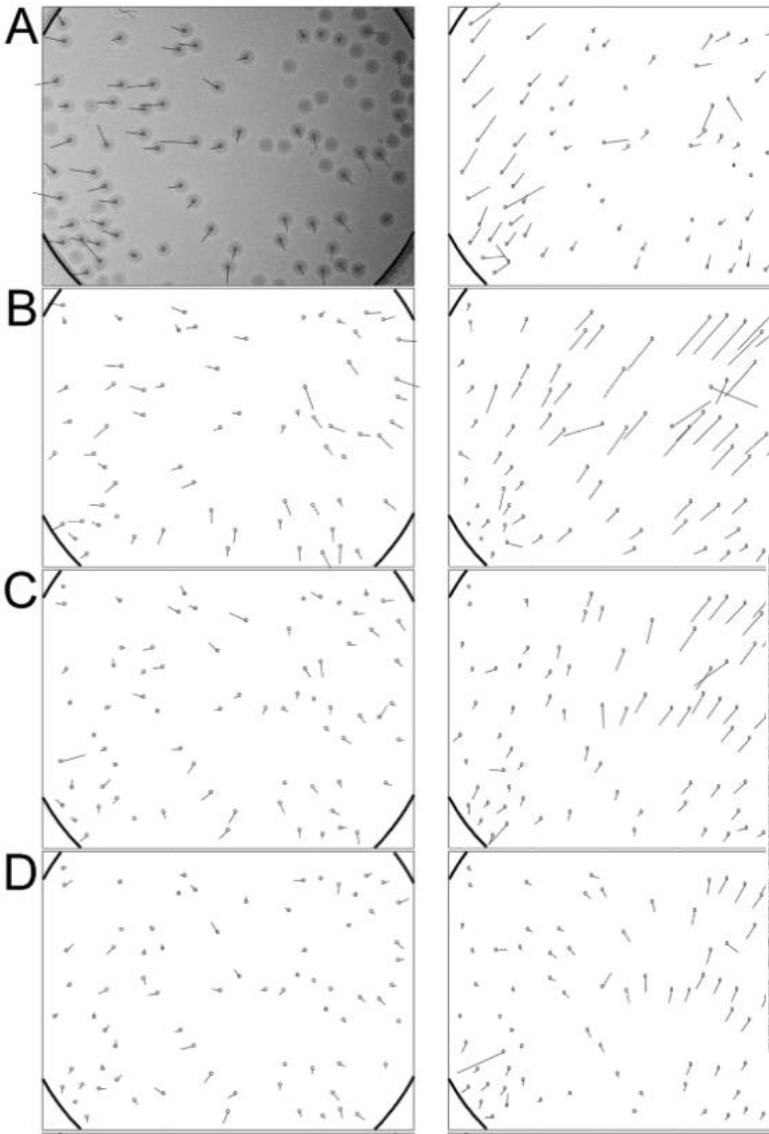
- High-resolution refinement
  - Prevention of overfitting
  - Movie-processing


Jan
Matt

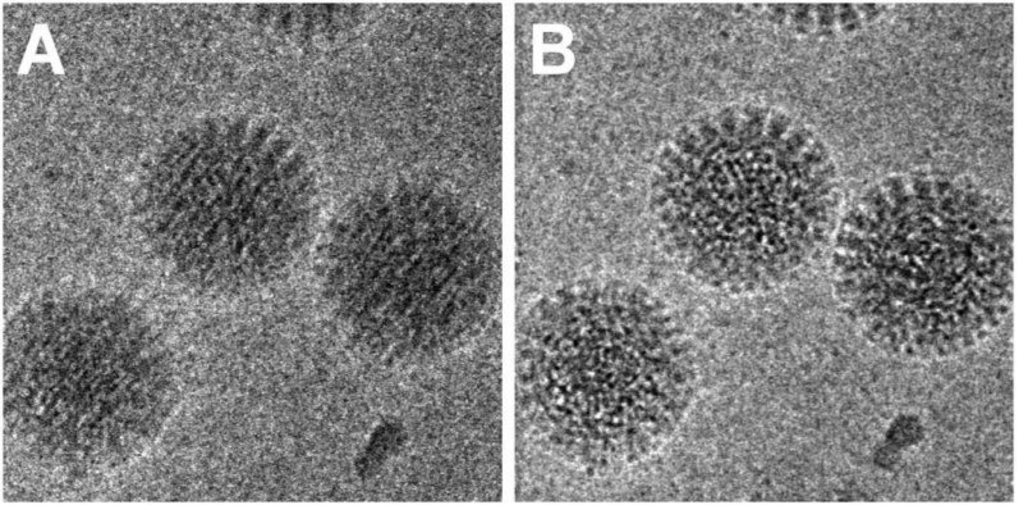
Take pictures of moving objects

Take fast snapshots
of moving objects

Motion-correction



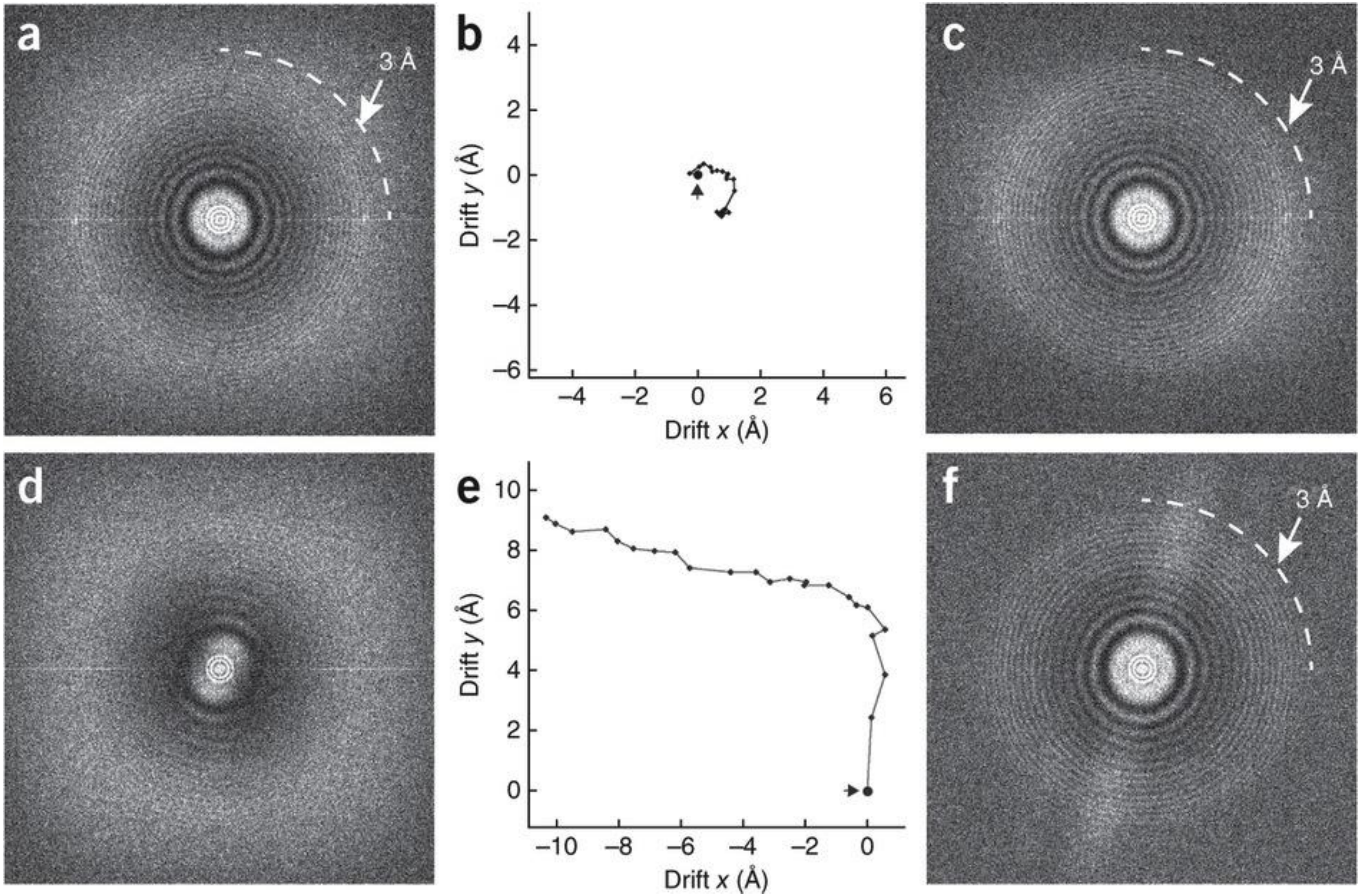
Brilot, … , Potter, Carragher, … , Grigorieff (2012) *J.Struct.Biol.*



Movie-processing programs

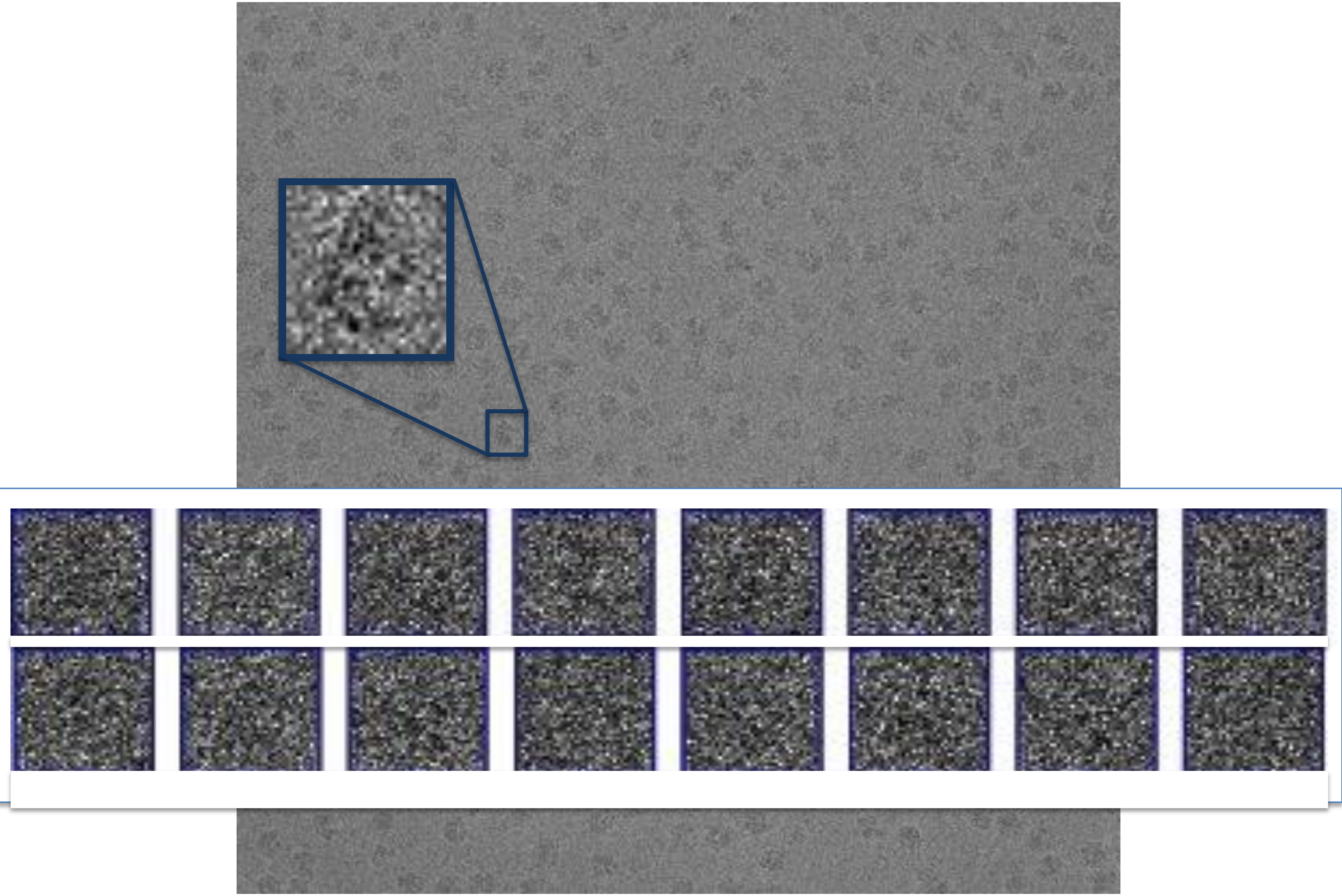
- Two main approaches
  - Per-micrograph
    - MOTIONCORR
      (Xueming Li, …, Yifan Cheng (2013) *Nat Meth.*)

MOTIONCORR

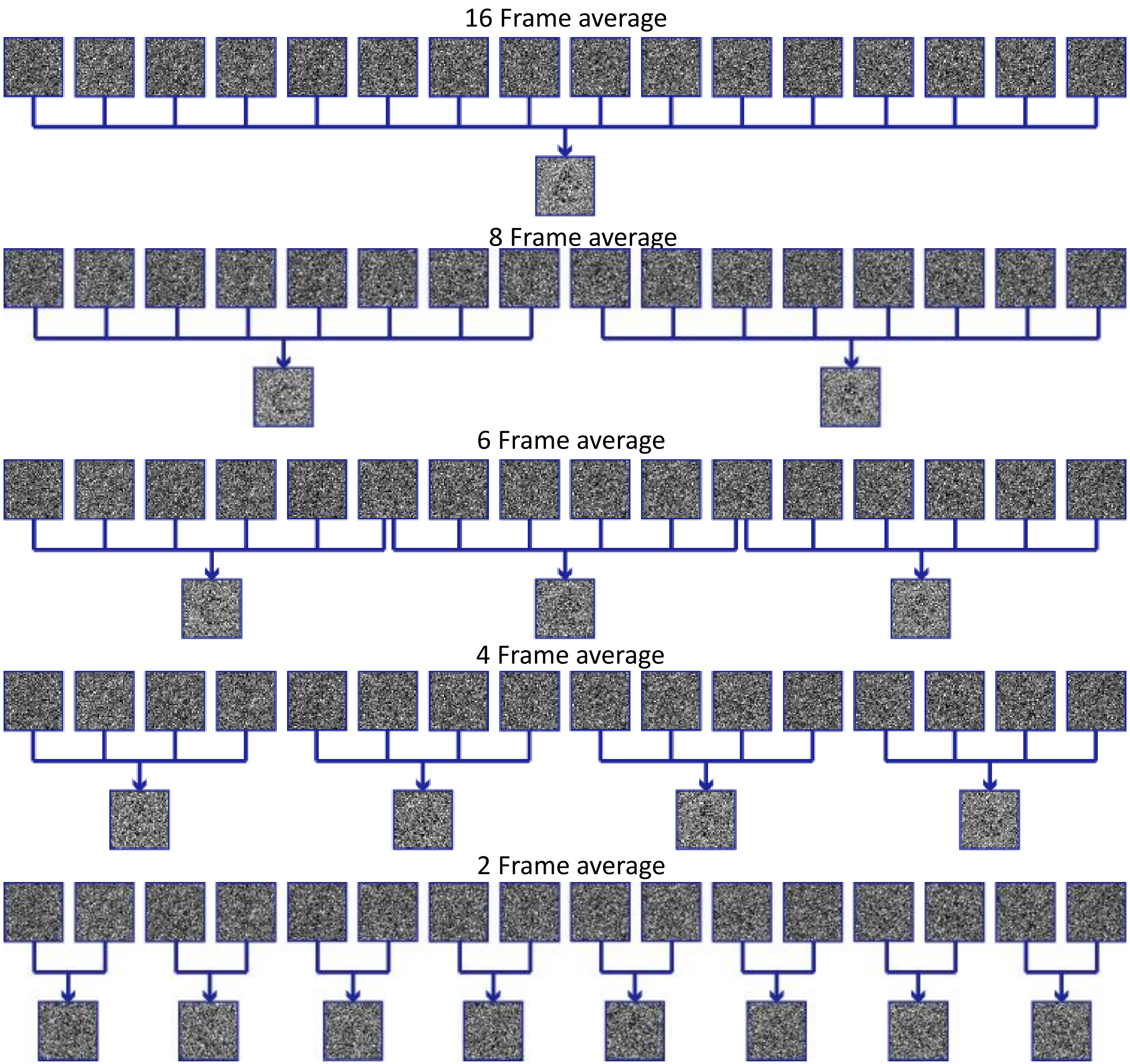

a
3 Å
b
Drift y (Å)
4
2
0
−2
−4
−6
−4
−2
0
2
4
6
Drift x (Å)
c
3 Å
d
e
Drift y (Å)
10
8
6
4
2
0
−10
−8
−6
−4
−2
0
Drift x (Å)
f
3 Å

Movie-processing programs

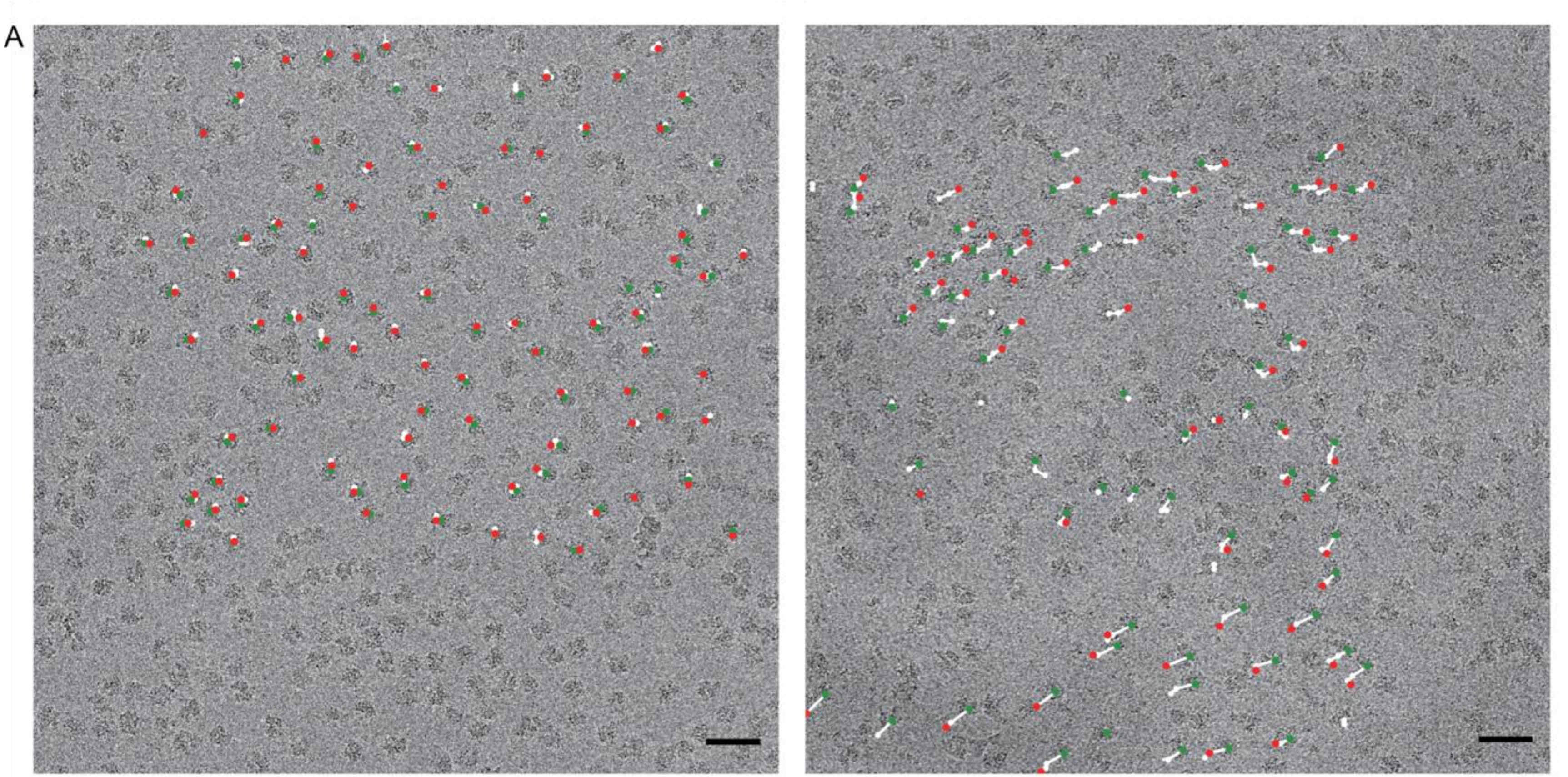
- Two types of approaches
  - Per-micrograph
    - MOTIONCORR (Xueming Li, …, Yifan Cheng (2013) *Nat Meth.*)
  - Per-particle
    - RELION
    - FREALIGN (?)
    - Align_lmbfgs (John Rubinstein)
    - Direct Electron (Ben Bammes)



Movie-processing (Bai *et al*, eLife 2013)



Beam-induced movements

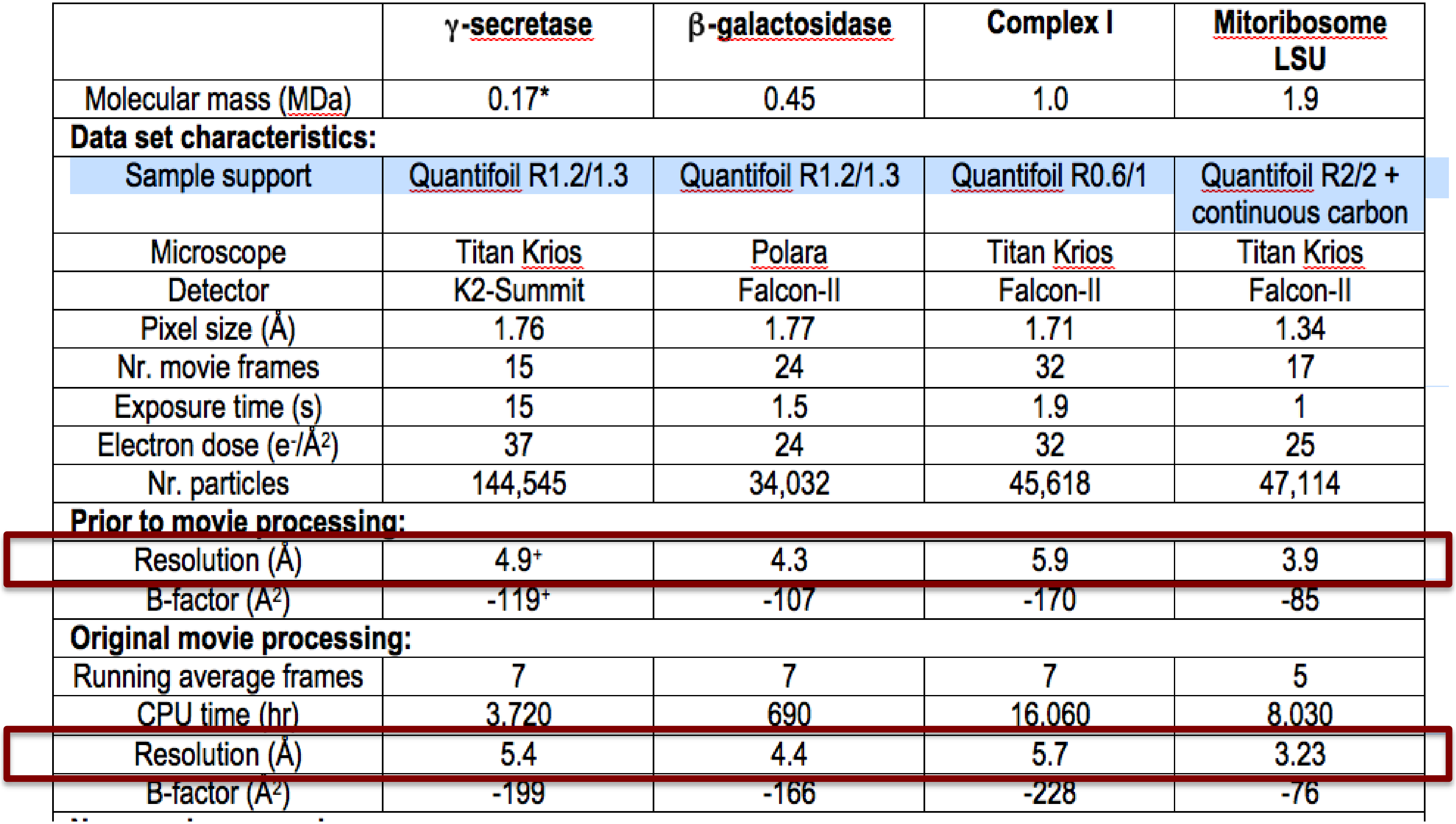


The 2013 approach

- Worked great for large particles (>1 MDa)
  - Ribosomes, viruses, etc

- Smaller particles: too noisy to follow beam-induced motions in several movie-frames

The 2013 approach

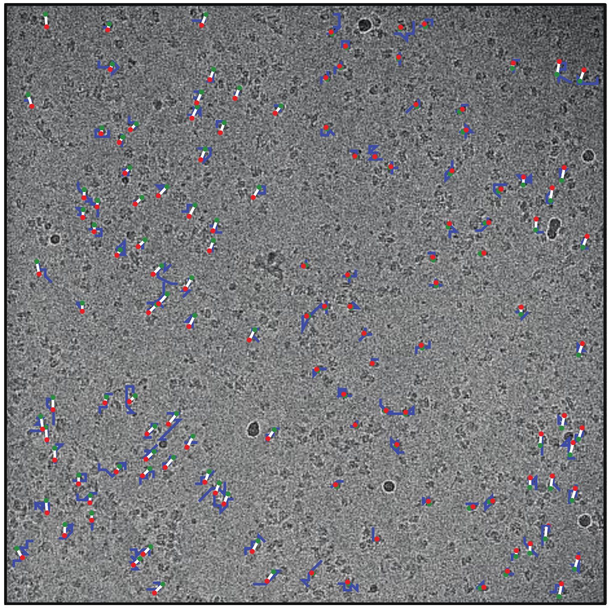


	γ-secretase	β-galactosidase	Complex I	Mitoribosome LSU
Molecular mass (MDa)	0.17*	0.45	1.0	1.9
Data set characteristics:				
Sample support	Quantifoil R1.2/1.3	Quantifoil R1.2/1.3	Quantifoil R0.6/1	Quantifoil R2/2 + continuous carbon
Microscope	Titan Krios	Polara	Titan Krios	Titan Krios
Detector	K2-Summit	Falcon-II	Falcon-II	Falcon-II
Pixel size (Å)	1.76	1.77	1.71	1.34
Nr. movie frames	15	24	32	17
Exposure time (s)	15	1.5	1.9	1
Electron dose ($e^-/Å^2$)	37	24	32	25
Nr. particles	144,545	34,032	45,618	47,114
Prior to movie processing:				
Resolution (Å)	4.9[+]	4.3	5.9	3.9
B-factor ($Å^2$)	-119[+]	-107	-170	-85
Original movie processing:				
Running average frames	7	7	7	5
CPU time (hr)	3,720	690	16,060	8,030
Resolution (Å)	5.4	4.4	5.7	3.23
B-factor ($Å^2$)	-199	-166	-228	-76

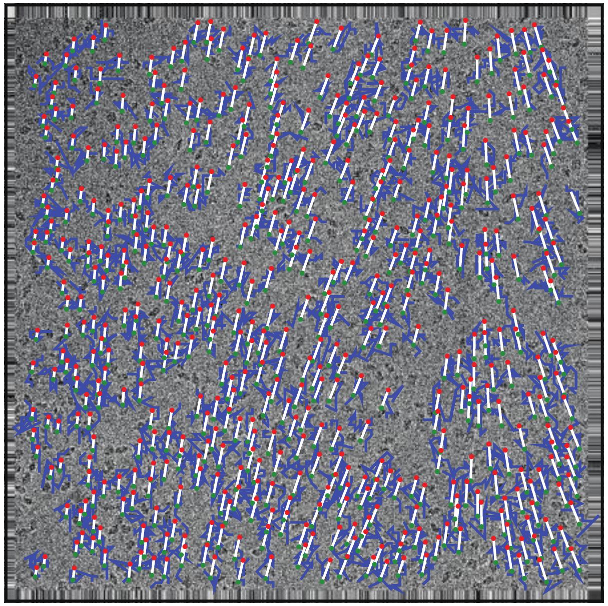
The new approach (eLife 2014) – part I

- Fit straight lines through beam-induced translations
- Weighted least-squares fits with neighboring particles on the micrograph
  - Neighbors often move in a similar direction
  - Weight: Gaussian of inter-particle distance
- Ignore rotations
  - Were small anyway (at limit of detectability)
  - Program becomes much (e.g. 6x) faster

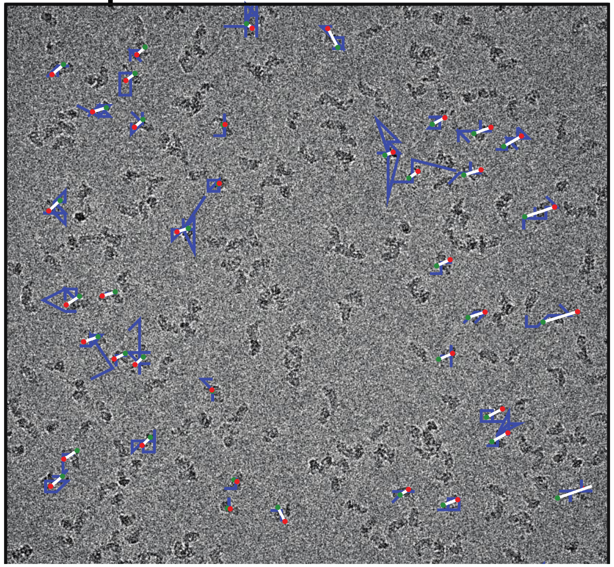
g-secretase (after UCSF scripts)



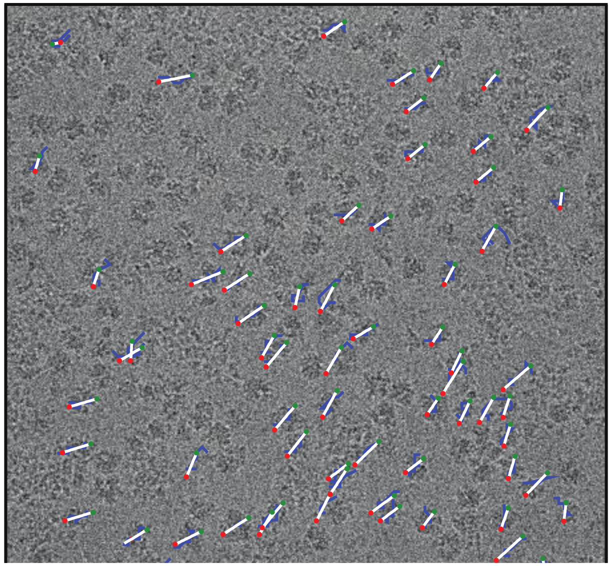
b-galactosidase



complex-I

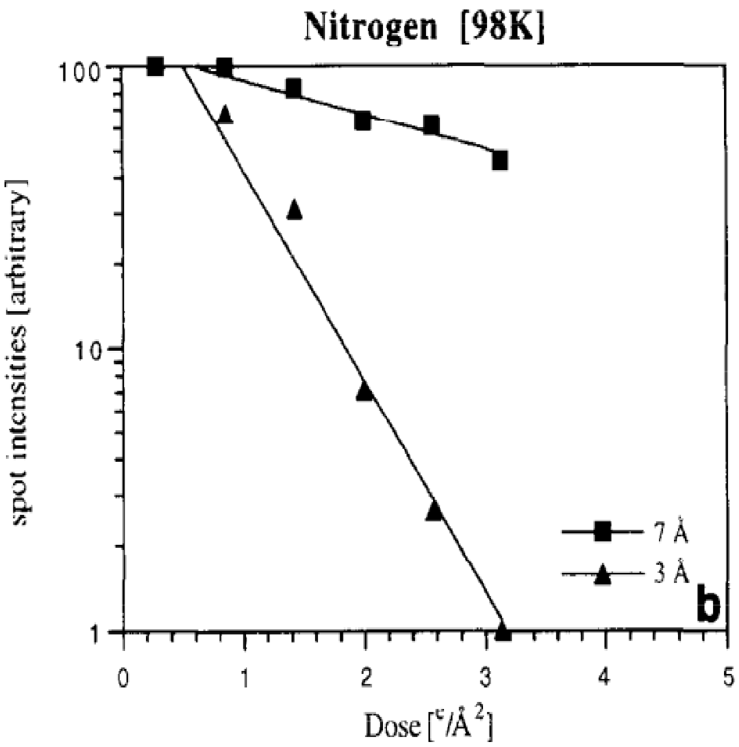


mitoribosome



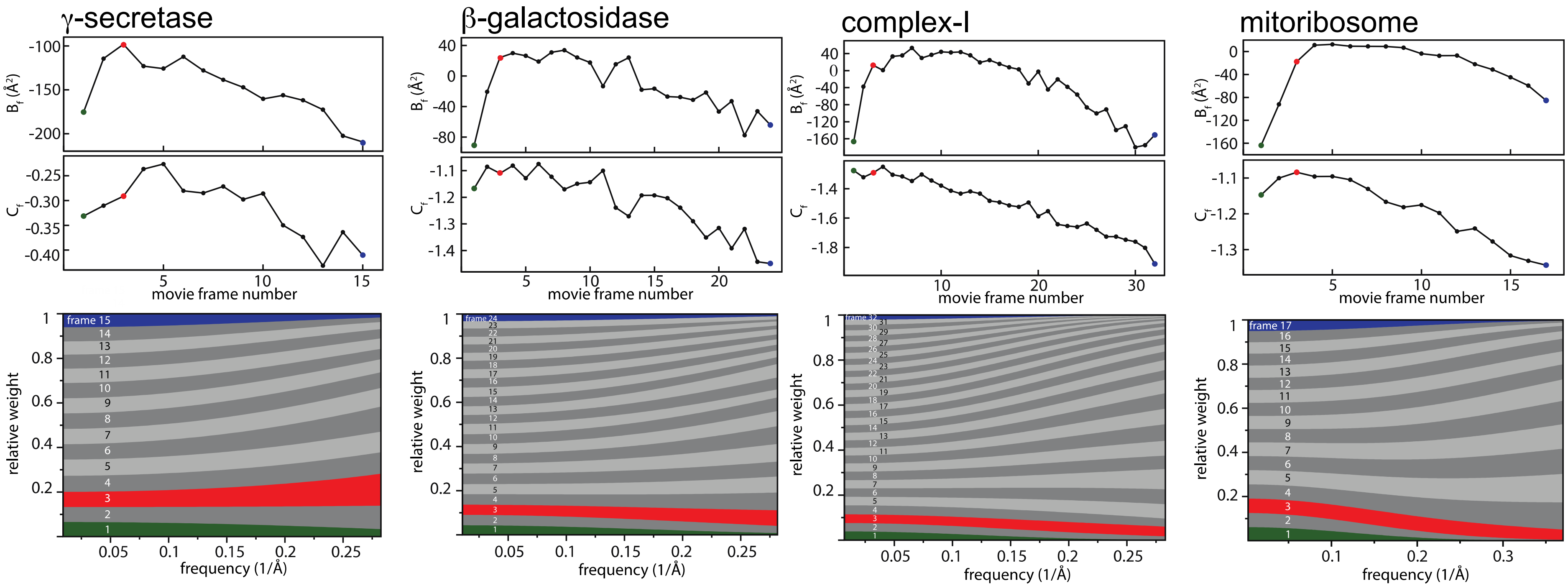
The new approach – part II

- Dose-dependent radiation-damage model
  - Higher-frequencies disappear at lower dose!
- Estimate B-factor for each movie frame



Stark et al, 1996

The new approach – part II



γ-secretase

β-galactosidase

complex-I

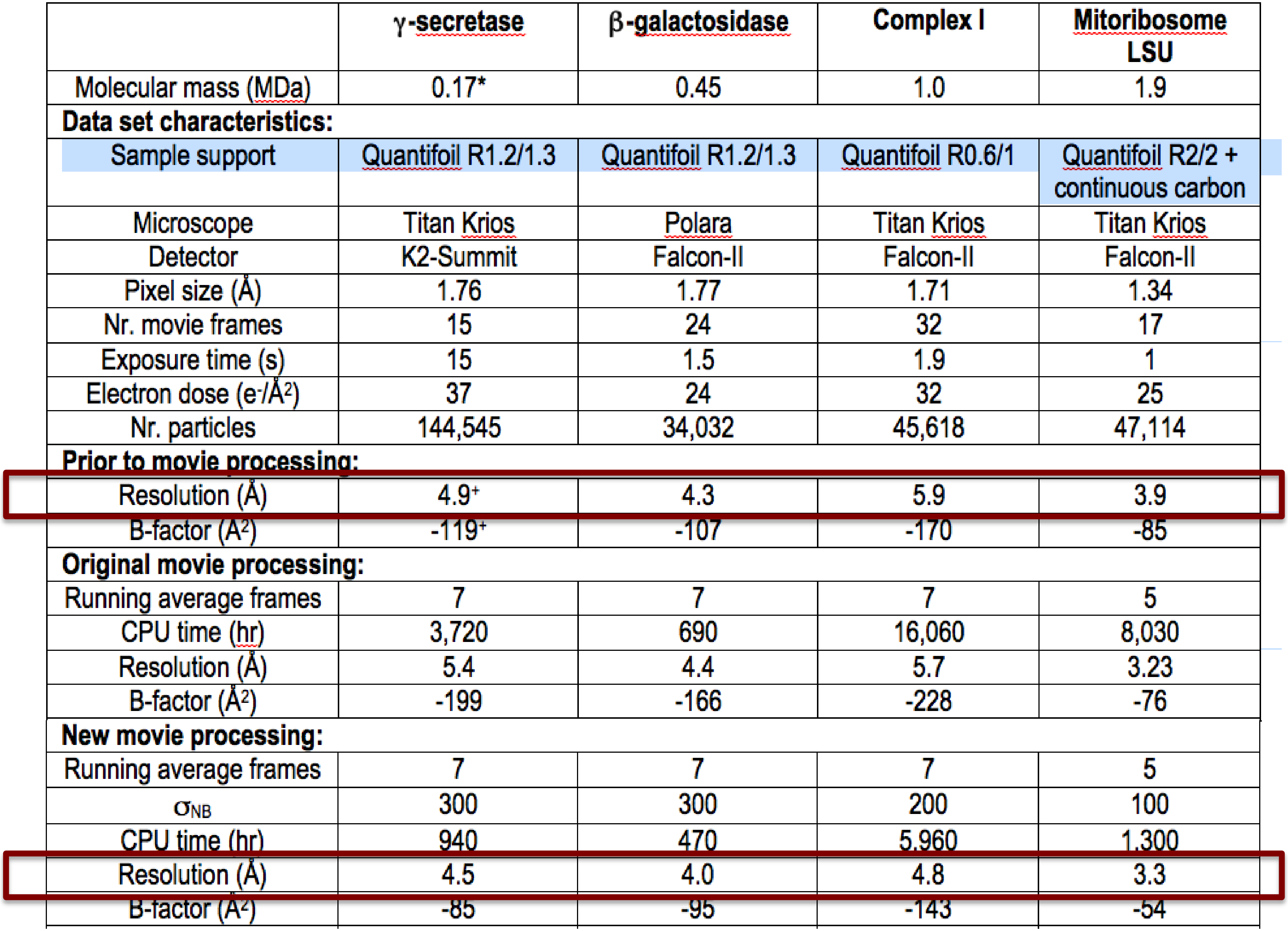
mitoribosome

The new approach – part III

- FOR EACH particle
  - Re-align movie-frames
  - Apply per-frame B-factor weighting
  - Average

- New set of "polished/shiny particles"
  - Increased SNRs

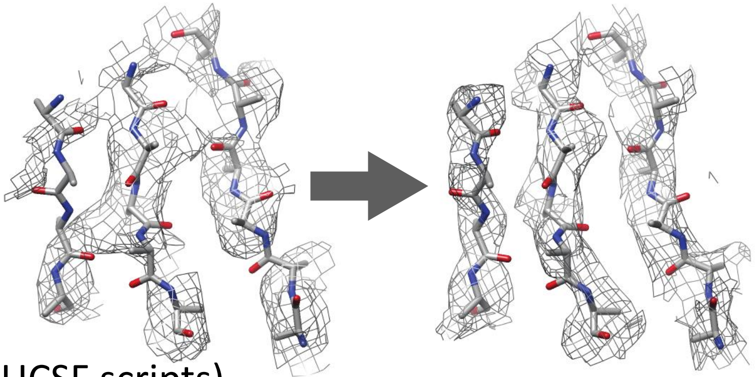
- Re-classify, re-refine



	γ-secretase	β-galactosidase	Complex I	Mitoribosome LSU
Molecular mass (MDa)	0.17*	0.45	1.0	1.9
Data set characteristics:				
Sample support	Quantifoil R1.2/1.3	Quantifoil R1.2/1.3	Quantifoil R0.6/1	Quantifoil R2/2 + continuous carbon
Microscope	Titan Krios	Polara	Titan Krios	Titan Krios
Detector	K2-Summit	Falcon-II	Falcon-II	Falcon-II
Pixel size (Å)	1.76	1.77	1.71	1.34
Nr. movie frames	15	24	32	17
Exposure time (s)	15	1.5	1.9	1
Electron dose ($e^-/Å^2$)	37	24	32	25
Nr. particles	144,545	34,032	45,618	47,114
Prior to movie processing:				
Resolution (Å)	4.9^+	4.3	5.9	3.9
B-factor ($Å^2$)	-119^+	-107	-170	-85
Original movie processing:				
Running average frames	7	7	7	5
CPU time (hr)	3,720	690	16,060	8,030
Resolution (Å)	5.4	4.4	5.7	3.23
B-factor ($Å^2$)	-199	-166	-228	-76
New movie processing:				
Running average frames	7	7	7	5
σ_{NB}	300	300	200	100
CPU time (hr)	940	470	5,960	1,300
Resolution (Å)	4.5	4.0	4.8	3.3
B-factor ($Å^2$)	-85	-95	-143	-54

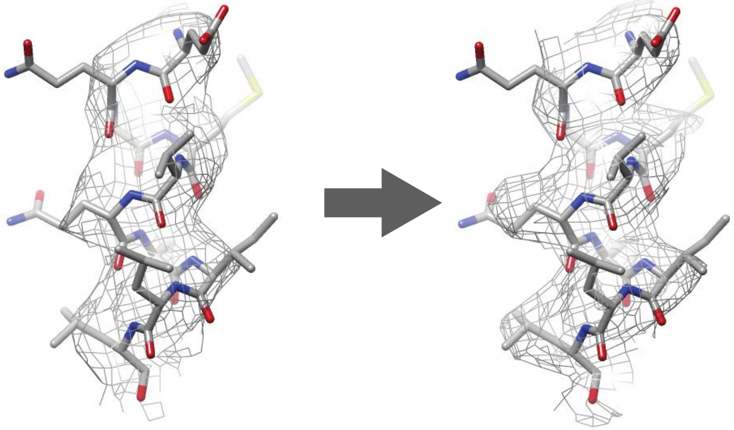
Before -> after "particle polishing"

g-secretase

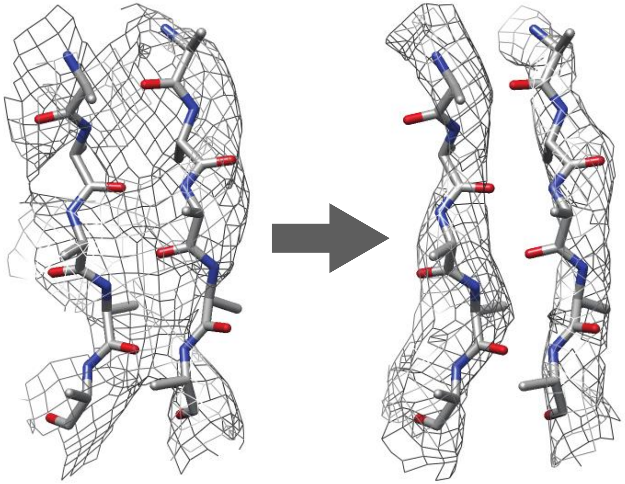


(after UCSF scripts)

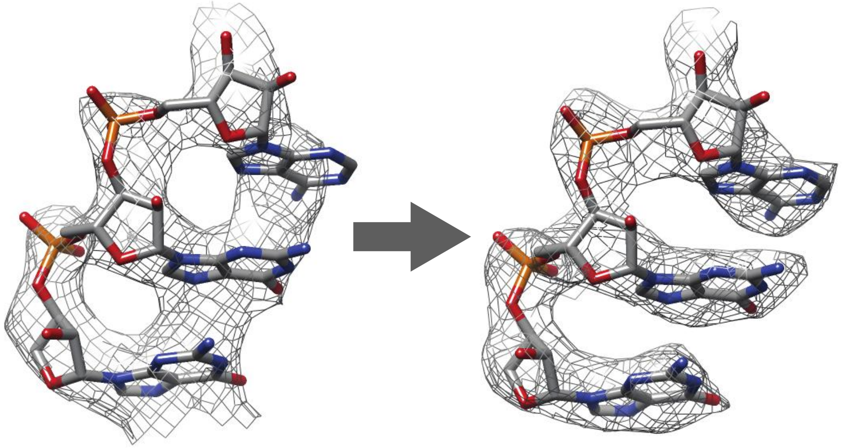
b-galactosidase



complex-I



mitoribosome



Scheres, eLife, 2014

Introduction and new approaches

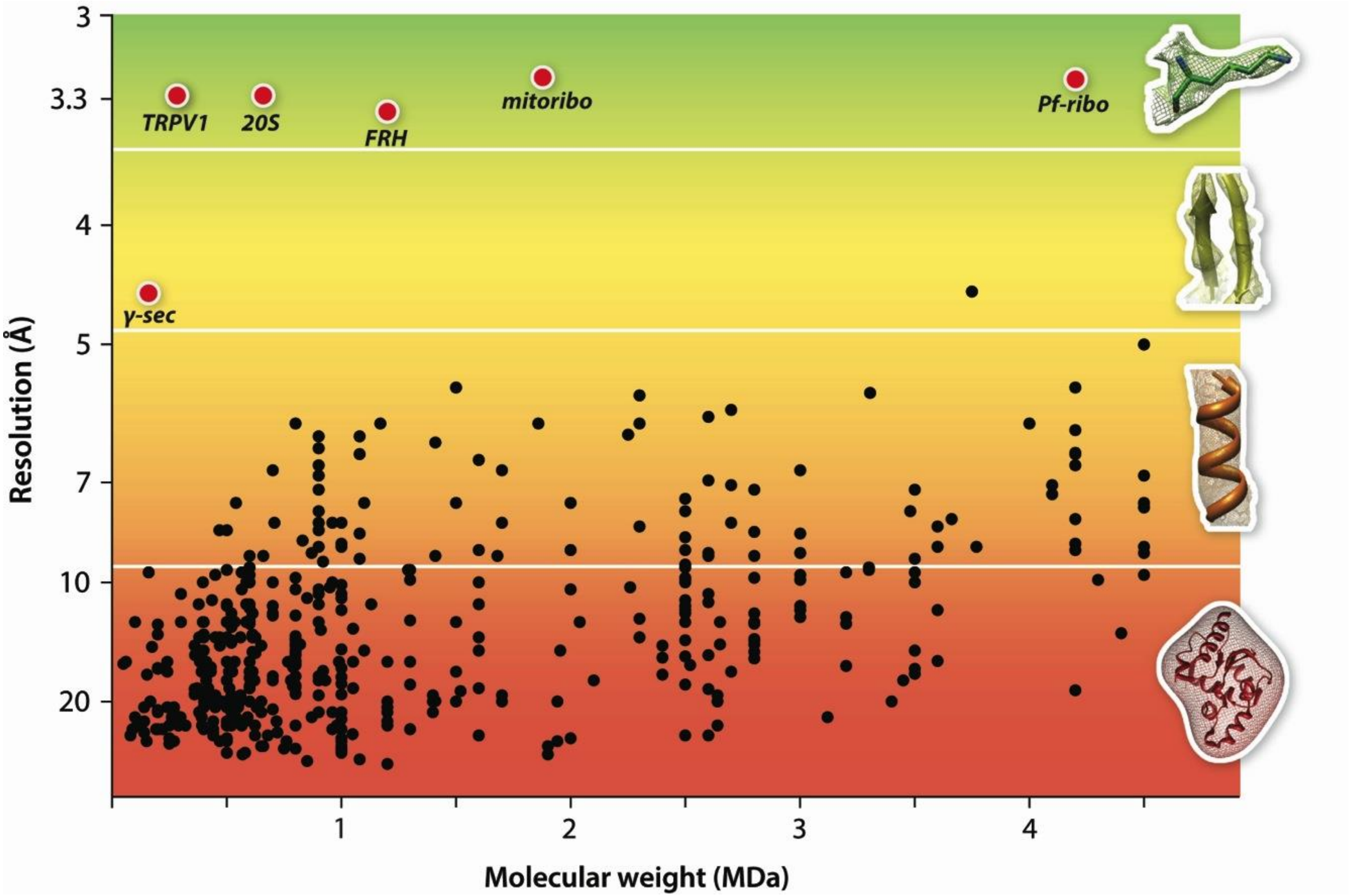
A comprehensive overview of the major advances that have taken place in the last few years that have enabled maps to achieve "atomic" resolution.

Topics to be covered include:

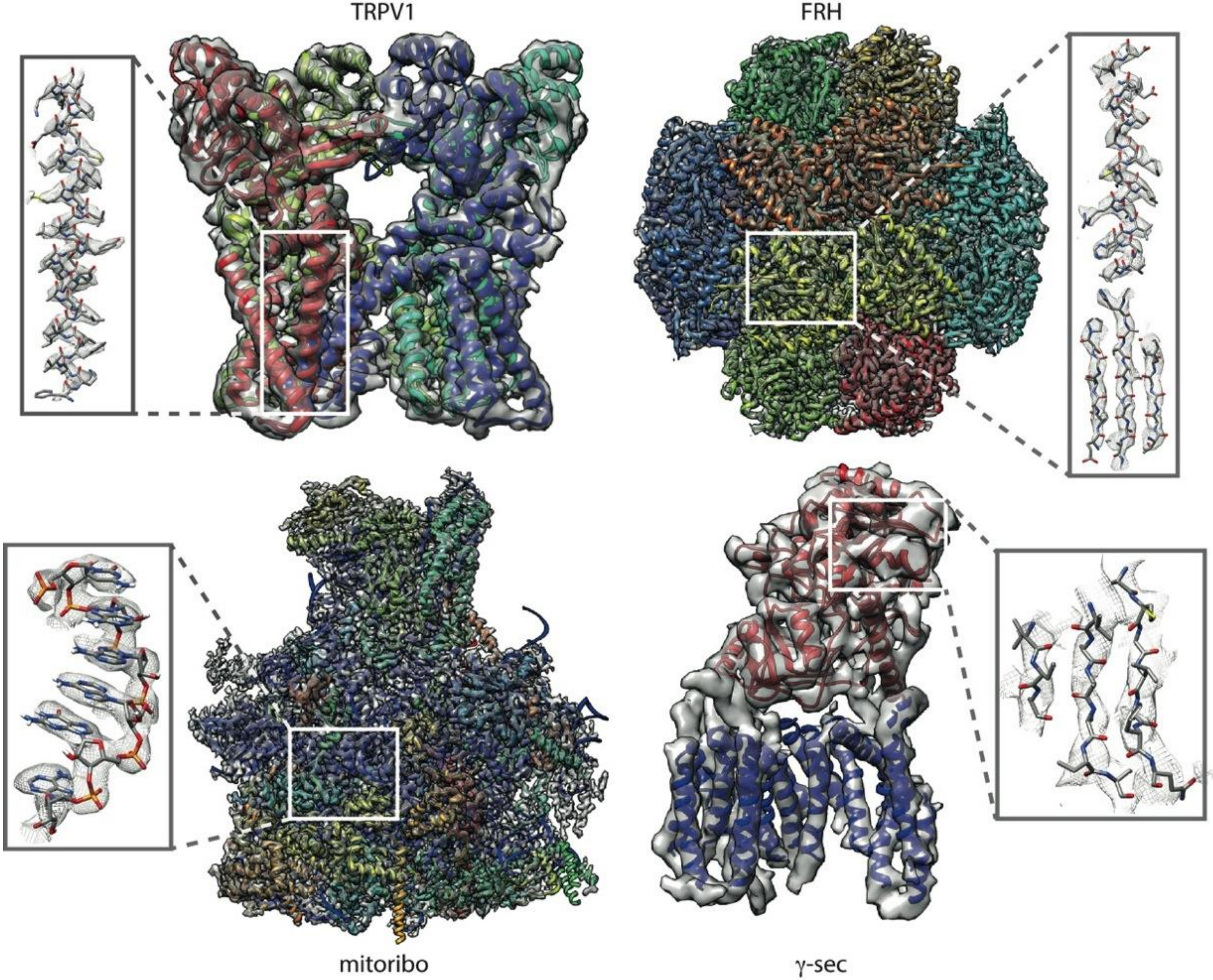
- 3D reconstruction
- image restoration techniques
- how to deal with heterogeneous populations.

- What are the hot topics in processing?
- What are the major mathematical approaches and available software?
- What are the success stories and the failures?
- Where are the greatest challenges right now and how are we approaching these?
- Do we need completely new algorithms or just incremental improvements on the current ones?
- Mistakes to avoid!

Success Stories



Success Stories



Introduction and ne

A comprehe
last few ye

Topics

- 3D r
- ima
- how to de populations

- What are the hot topics in p essing?
- What are the major mathema al app aches and available software?
- What are the success stories and the failures?
- Where are the greatest challenges right now ho these?
- Do we need completely new algorithms or jus on the current ones?
- Mistakes to avoid!

You never hear about these……..

We have them very often!
Mostly related to sample or grid preparation….

We don’t like:
negative stain &
cross-linking

Introduction and new approaches

A comprehensive overview of the major advances that have taken place in the last few years that have enabled maps to achieve "atomic" resolution.

Topics to be covered include:

- 3D reconstruction
- image restoration techniques
- how to deal with heterogeneous populations.

- What are the hot topics in processing?
- What are the major mathematical approaches and available software?
- What are the success stories and the failures?
- Where are the greatest challenges right now and how are we approaching these?
- Do we need completely new algorithms or just incremental improvements on the current ones?
- Mistakes to avoid!

Challenges

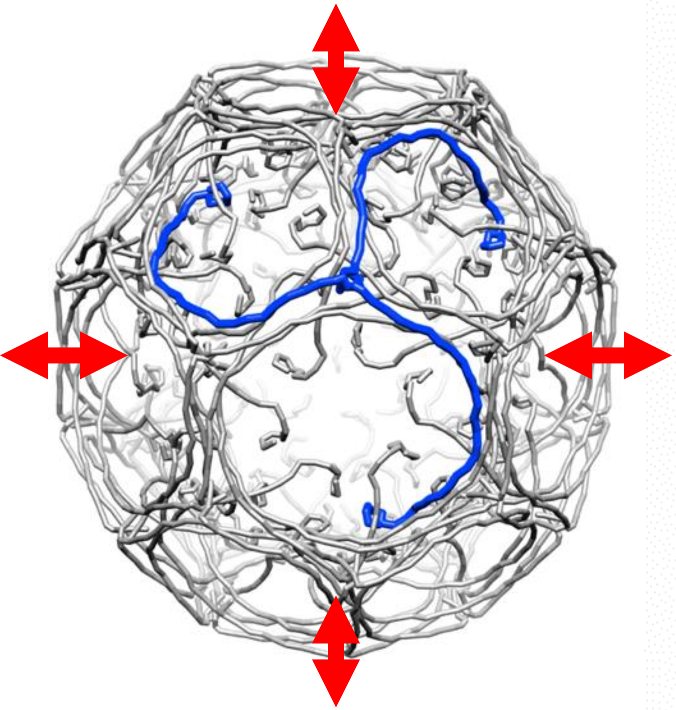
- Go significantly beyond 3 Å for many samples
  - Already getting there for some samples (poster: Tim Grant & Niko)
  - Cs-corrector (Holger Stark)

- STRUCTURAL HETEROGENEITY

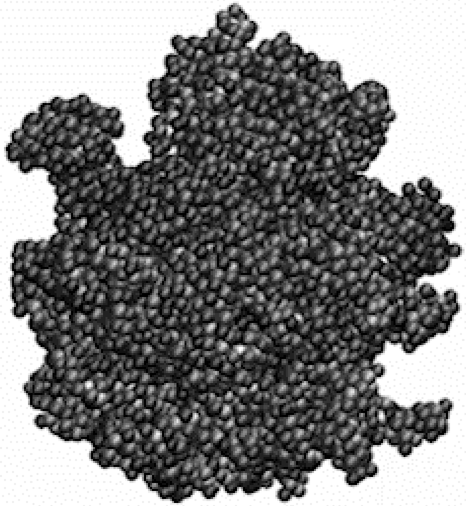
Dealing With Heterogeneity

Talk by Niko

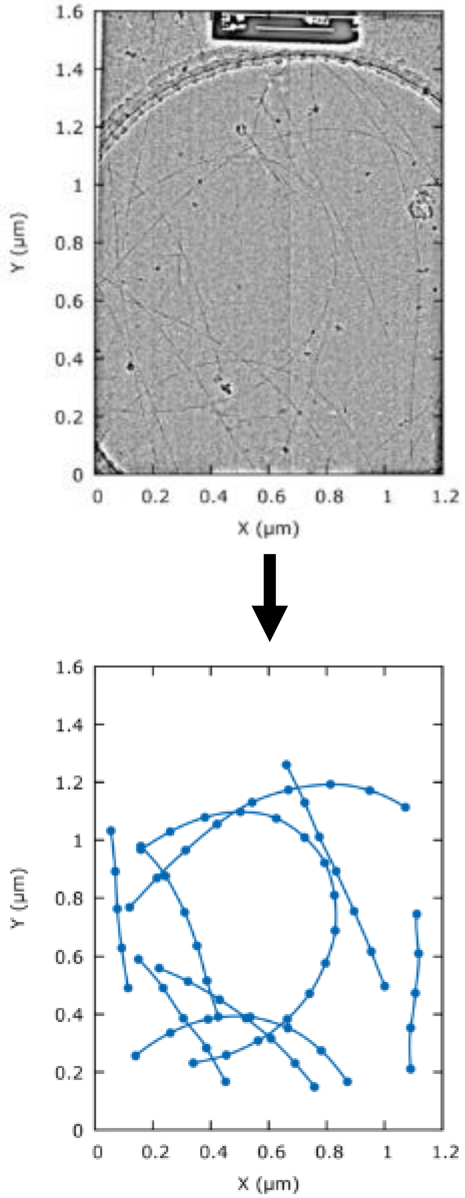
Deformable particles



Molecular machines



Flexible filaments



Introduction and new approaches

A comprehensive overview of the major advances that have taken place in the last few years that have enabled maps to achieve "atomic" resolution.

Topics to be covered include:

- 3D reconstruction
- image restoration techniques
- how to deal with heterogeneous populations.

- What are the hot topics in processing?
- What are the major mathematical approaches and available software?
- What are the success stories and the failures?
- Where are the greatest challenges right now and how are we approaching these?
- Do we need completely new algorithms or just incremental improvements on the current ones?
- Mistakes to avoid!

Making existing algorithms better

- Raw data quality assessment
  - Only make reconstructions with the best particles

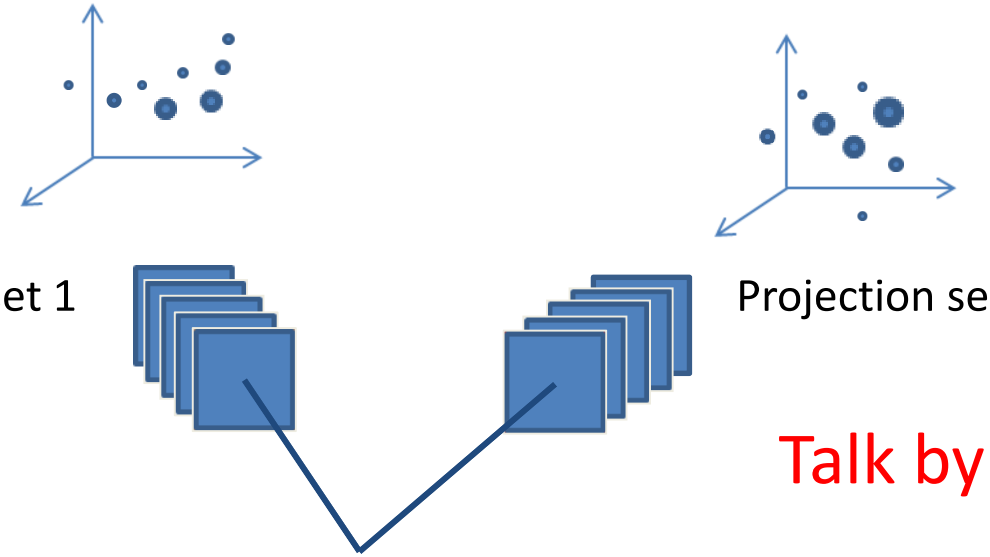
- New similarity metrics?

Talk by Steve

Or something completely new?

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

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


Manifold 1

Manifold 2

Projection set 1

Projection set 2

Talk by Joachim

Premise: variation of particle image due to conf. changes is small compared to its variation due to changes in projection direction. Step 1: sort particles by orientation.

Set of projections in direction 1 forms an N-dim. manifold where N is the number of degrees of freedom.
Set of projections in direction 2 forms another N-dim manifold that is quite different since conf. variations manifest themselves differently in different projection directions.
How are the two manifolds related to one another? More generally, is there a mapping operation (a "synchronization") that allows us to "collect" ***all*** *particle snapshots, from* ***all*** *directions, that originate from particles in the same conformational state? And then do the same thing for all conformations encountered?*

Introduction and new approaches

A comprehensive overview of the major advances that have taken place in the last few years that have enabled maps to achieve "atomic" resolution.

Topics to be covered include:

- 3D reconstruction
- image restoration techniques
- how to deal with heterogeneous populations.

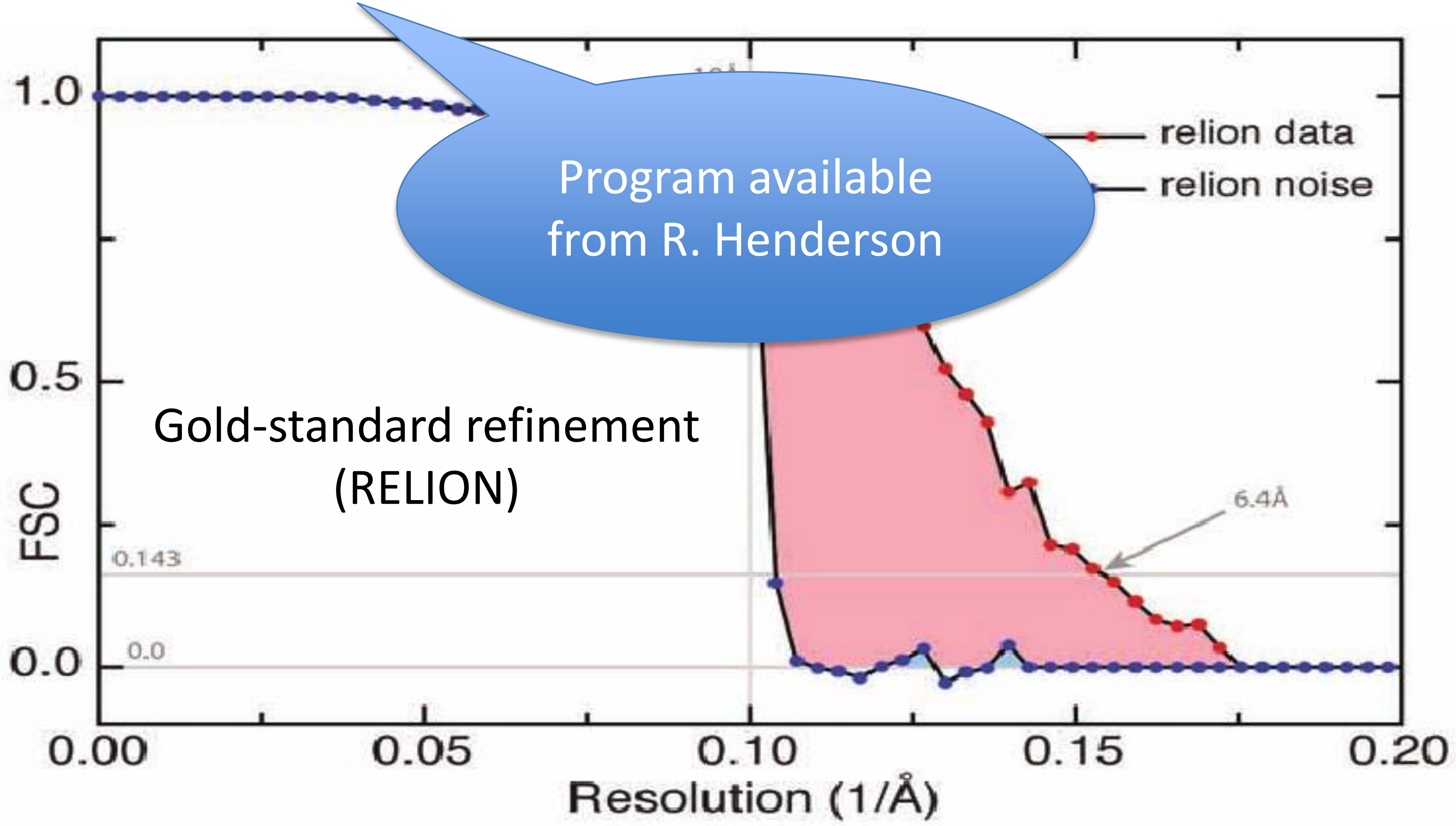
- What are the hot topics in processing?
- What are the major mathematical approaches and available software?
- What are the success stories and the failures?
- Where are the greatest challenges right now and how are we approaching these?
- Do we need completely new algorithms or just incremental improvements on the current ones?
- Mistakes to avoid!

Mistakes to avoid (I)

- 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



Program available from R. Henderson

Gold-standard refinement (RELION)

Mistakes to avoid (II)

- Get stuck with a wrong initial model

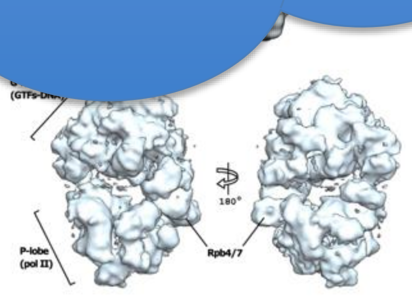
Human RNA polymerase II PIC

He et al & Nogales, Nature (2013)

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

3)

Should we stop publishing blobs?

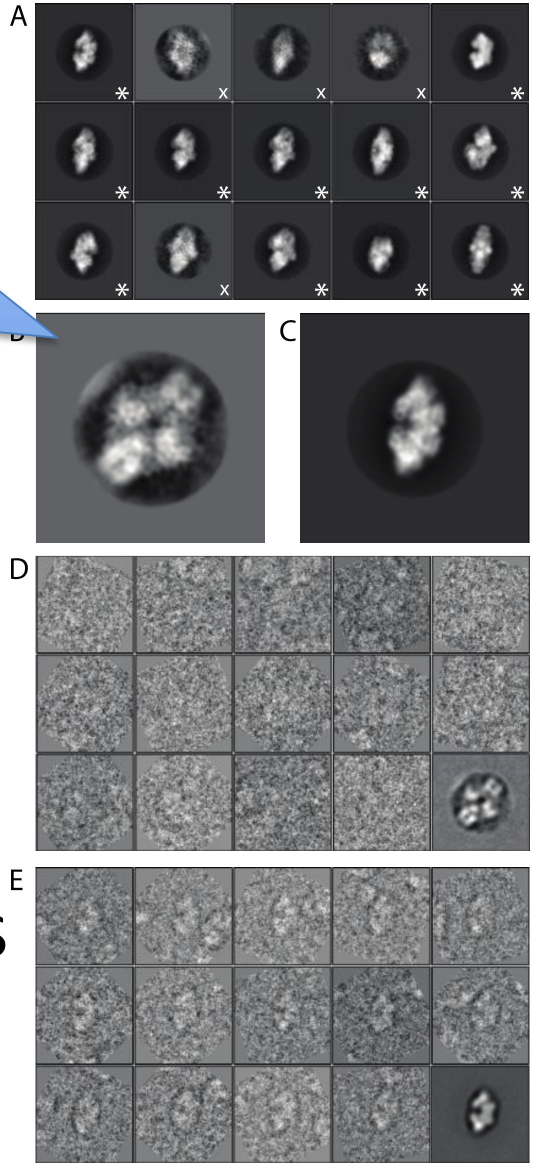


Validation session tomorrow!

(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

Mistakes to avoid (IV)

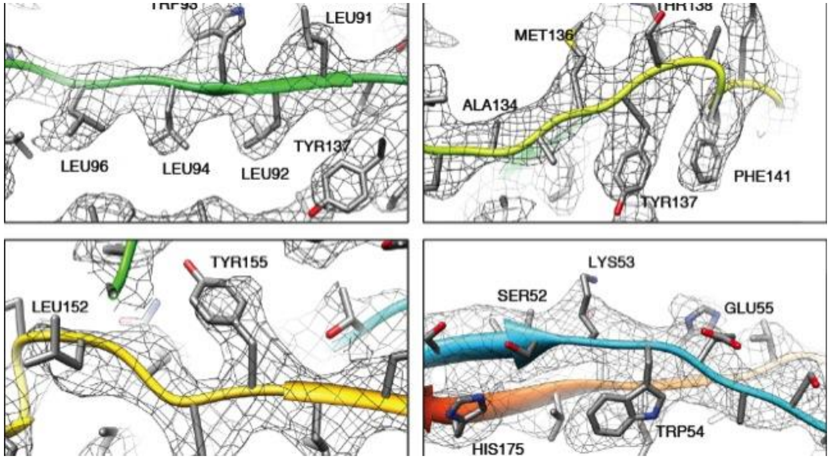
Monoculture

Microscopes: FEI, Jeol, Zeiss, …

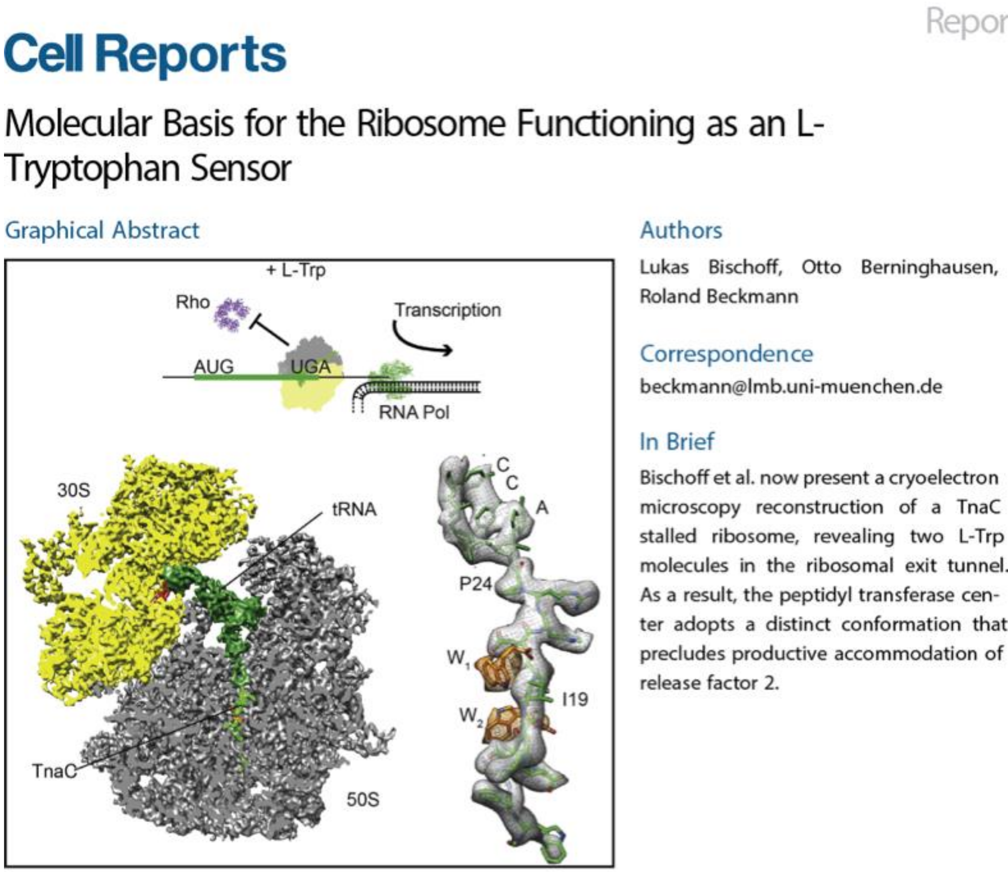
Detectors: K2, Falcon, DE, TVIPS, …

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

Wang et al (2014) Nat Comm.



JEOL3200, DE-12, EMAN (3.8 Å)



Repor

Cell Reports

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

Graphical Abstract

Authors

Lukas Bischoff, Otto Berninghausen, Roland Beckmann

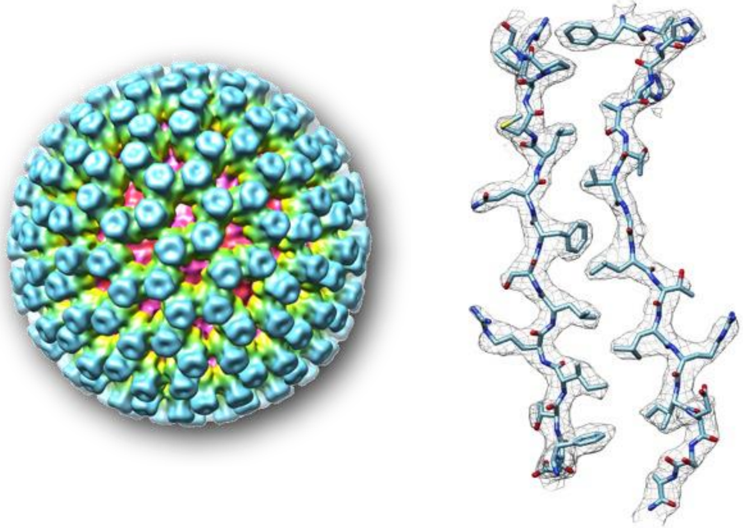
Correspondence

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) unpublished, see poster!

Titan Krios, K2, FREALIGN (2.6 Å)

Conclusions

- Image pr
  field

- As has ju
  continue to ha

Chris Tate, Pfizer and myself
are looking for post-docs with experience in
cell culture expression and/or
membrane protein biochemistry…

- Making good samples already was crucial, but will be ever more important!

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

I am leaving this afternoon. If you have any more RELION-questions, ask me this morning