Post-Processing and Validation of 3D EM Reconstructions

- Model-based Refinement & the Need for Validation
- Validation at Low Resolution
- Tilt-Pair Validation
- Labels/Biochemistry
- Validation at High Resolution
- What should my map look like?
- Validating models built into EM maps
- Model stereochemistry

Peter Rosenthal

EMBO Image Processing 2019

Why is validation important?

- Map validation
- Model validation
- In course of structure determination, may indicate whether the experiment or data processing needs to be optimized.
- Data deposition
- Peer review
- Still a subject of research!

Map Validation

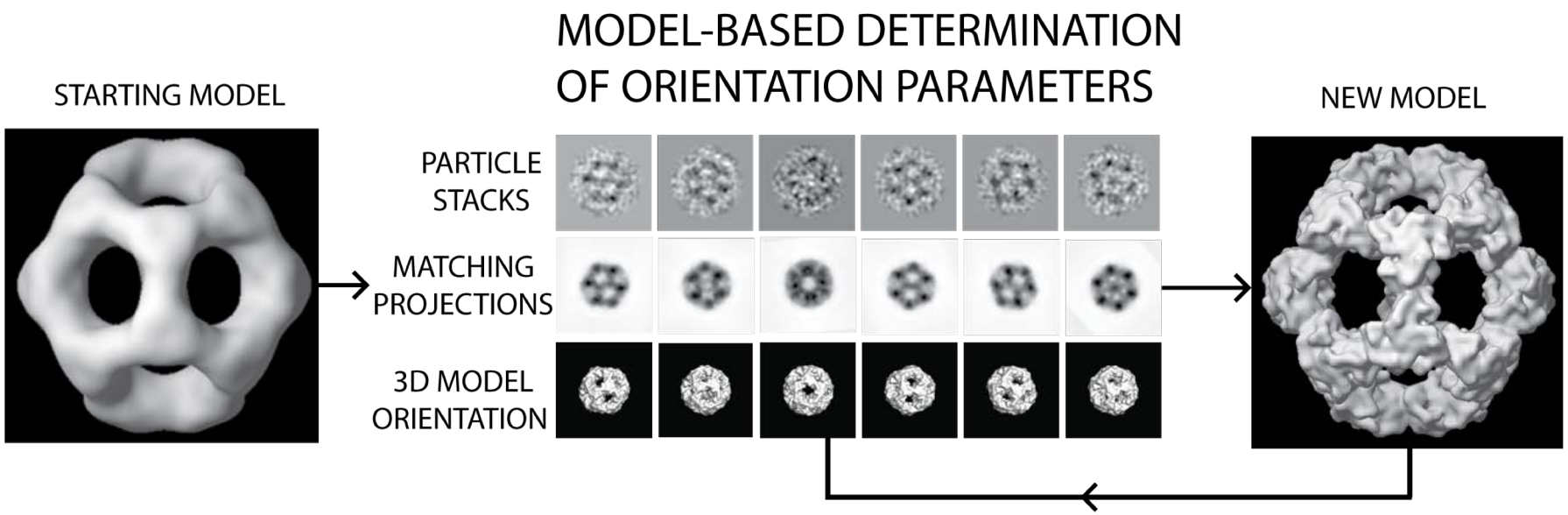
- Is map correct (incorrect) at low resolution?
- What should the map look like at high resolution? (Contrast Restoration)
- Is resolution assessment exaggerated?
- Simple tests to demonstrate validity of map
- User still decides how to process the data

Structure
Meeting Review

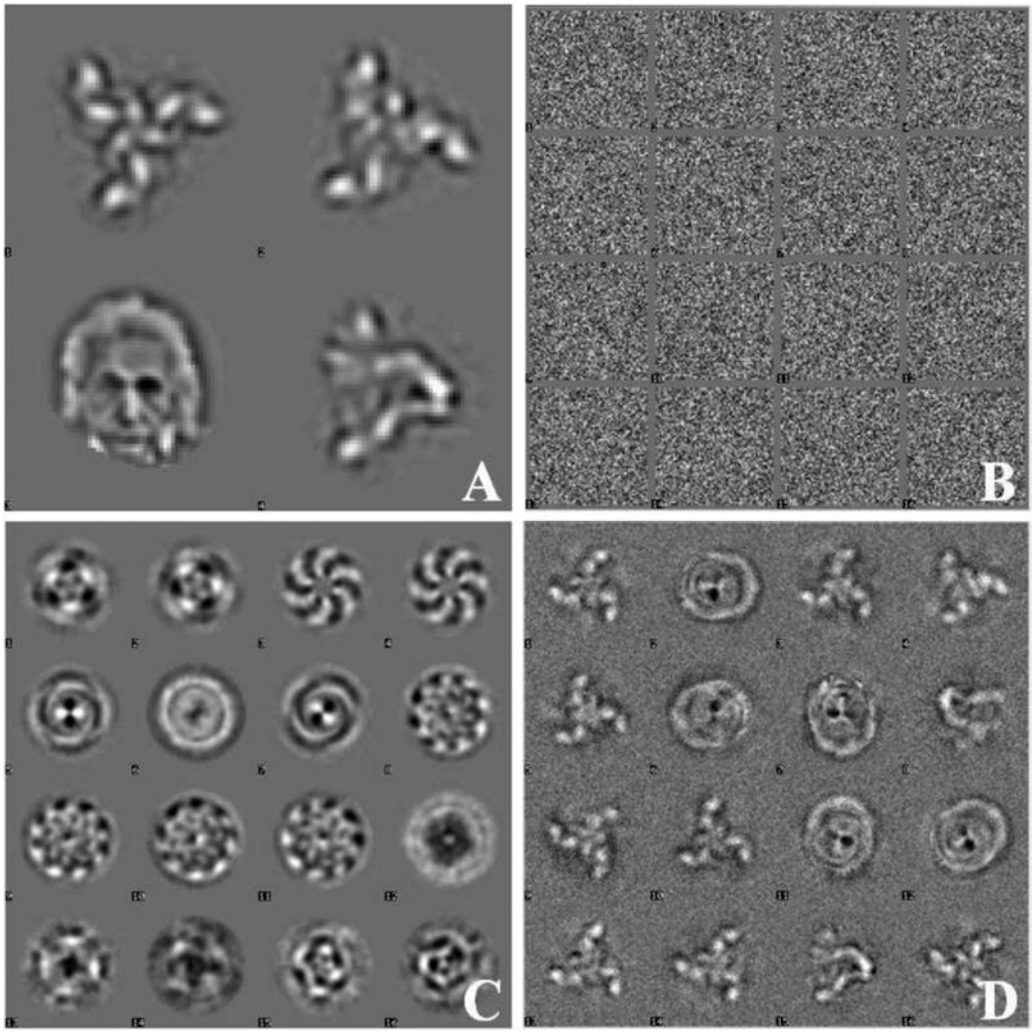
Outcome of the First Electron Microscopy Validation Task Force Meeting

Richard Henderson,[1] Andrej Sali,[2] Matthew L. Baker,[3] Bridget Carragher,[4] Batsal Devkota,[5] Kenneth H. Downing,[6] Edward H. Egelman,[7] Zukang Feng,[5] Joachim Frank,[8,9] Nikolaus Grigorieff,[10] Wen Jiang,[11] Steven J. Ludtke,[3] Ohad Medalia,[12,21] Pawel A. Penczek,[13] Peter B. Rosenthal,[14] Michael G. Rossmann,[15] Michael F. Schmid,[3] Gunnar F. Schröder,[16] Alasdair C. Steven,[17] David L. Stokes,[18] John D. Westbrook,[5] Willy Wriggers,[19] Huanwang Yang,[5] Jasmine Young,[5] Helen M. Berman,[5] Wah Chiu,[3] Gerard J. Kleywegt,[20] and Catherine L. Lawson[5,*]

Structure. 2012 Feb 8;20(2):205-14.
doi: 10.1016/j.str.2011.12.014.


MODEL-BASED DETERMINATION
OF ORIENTATION PARAMETERS
STARTING MODEL
NEW MODEL
PARTICLE
STACKS
MATCHING
PROJECTIONS
3D MODEL
ORIENTATION

(A) Four reference images (each 64 × 64 pixels) used for picking from 1,024 random noise images (of 1,024 × 1,024 pixels).



van Heel M PNAS 2013;110:E4175-E4177

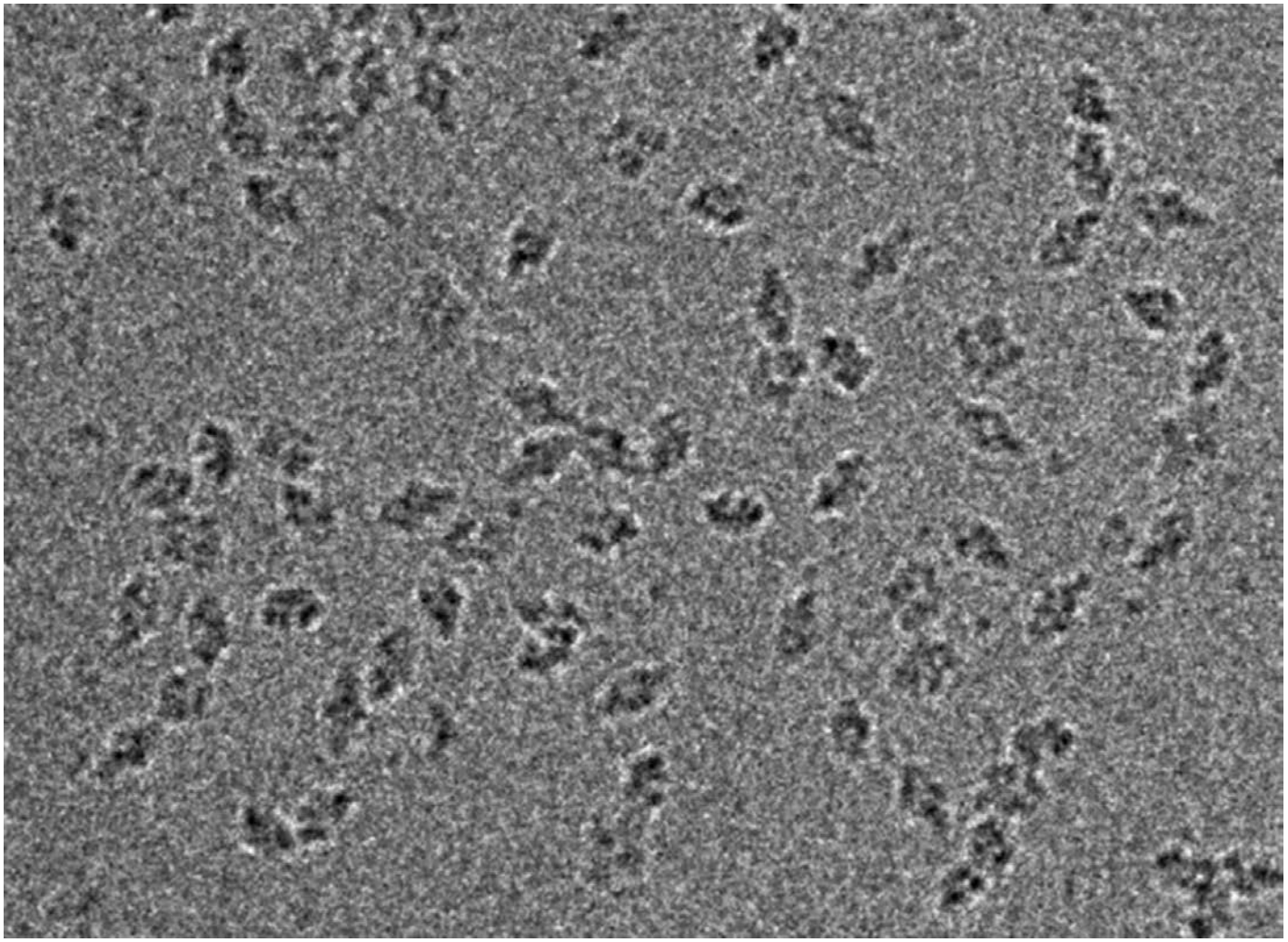
model bias: the persistence of an incorrect map or map features during refinement

over-refinement or over-fitting: causes build-up of spurious noise features in the map during iterative refinement.

e.g. the map matches signal in the images at low resolution, but at high resolution predominantly matches noise.

over-refined maps can possess features that resemble side-chain densities but in incorrect locations.

Cryo-EM image of a field of view of β-galactosidase single particles (molecular weight, 450 kDa).



Henderson R PNAS 2013;110:18037-18041

De novo determination of "Starting Map"

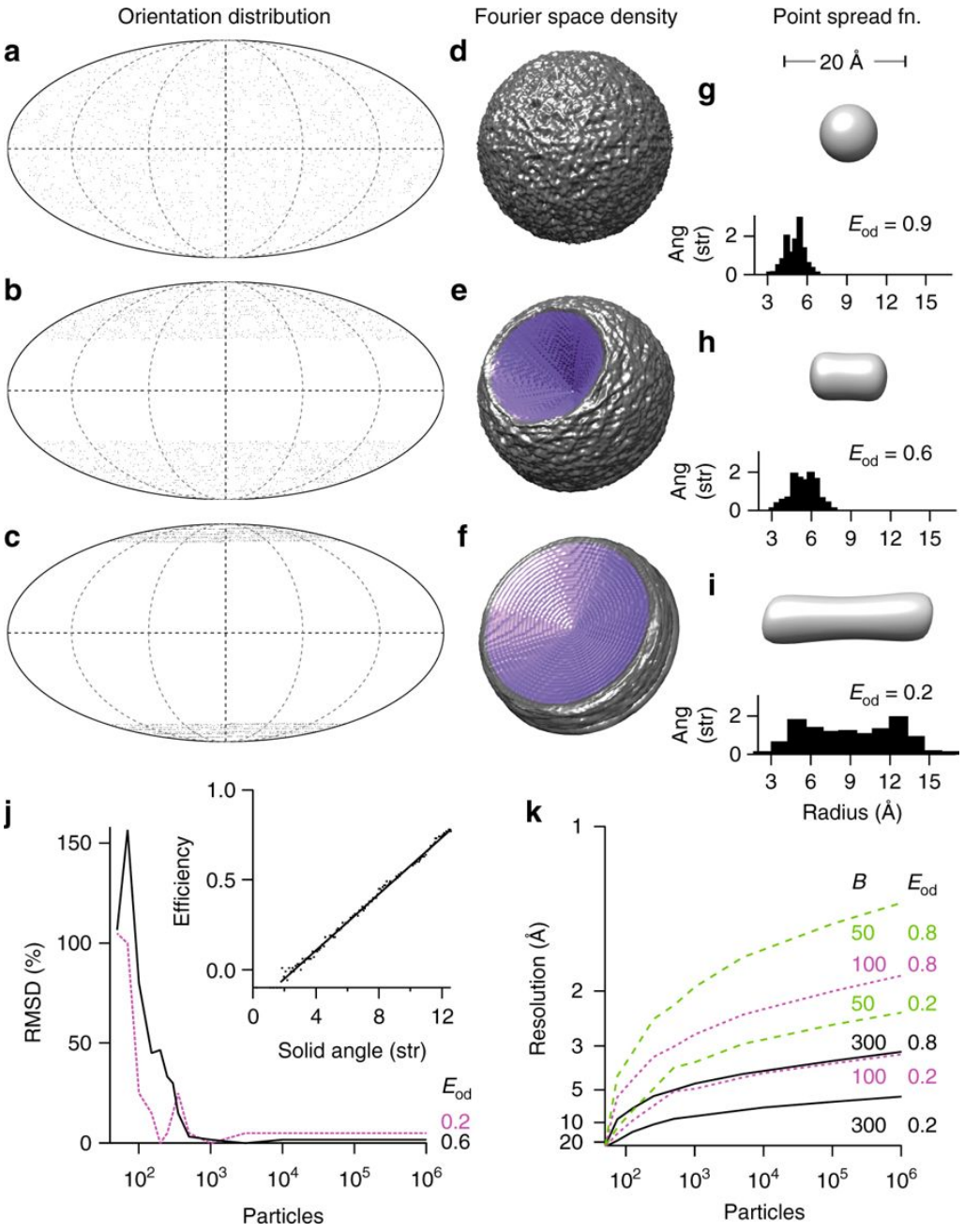
Map projections agree with individual raw images and class averages ("reference free")
Distribution of particle orientations
Symmetry
Absolute Hand (Experimental Determination-Tilting)

Other sources for starting map:

Derived from another highly similar structure?

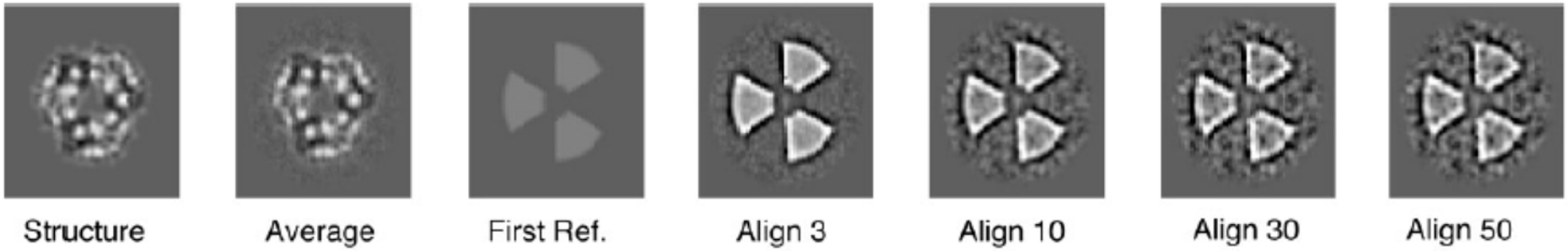
Density from X-ray model? Low-pass filtered

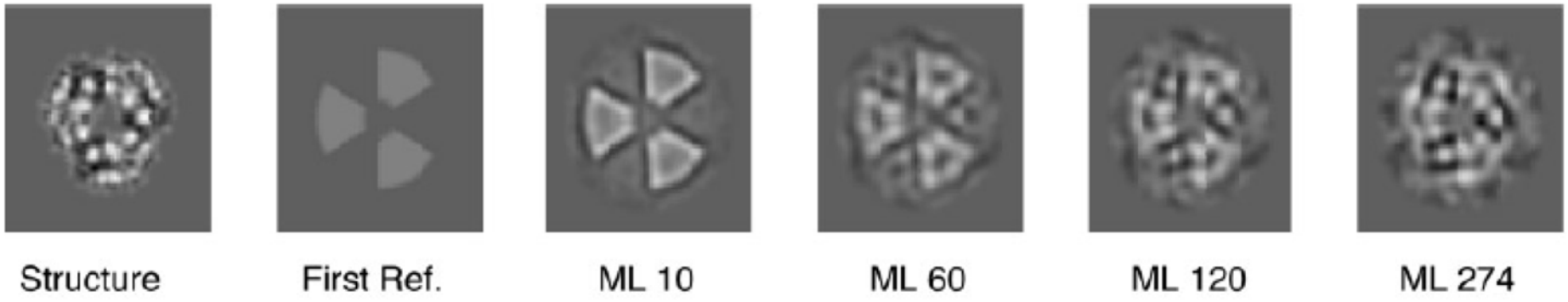
A spherical or cylindrical blob? Icosahedral or helical symmetry



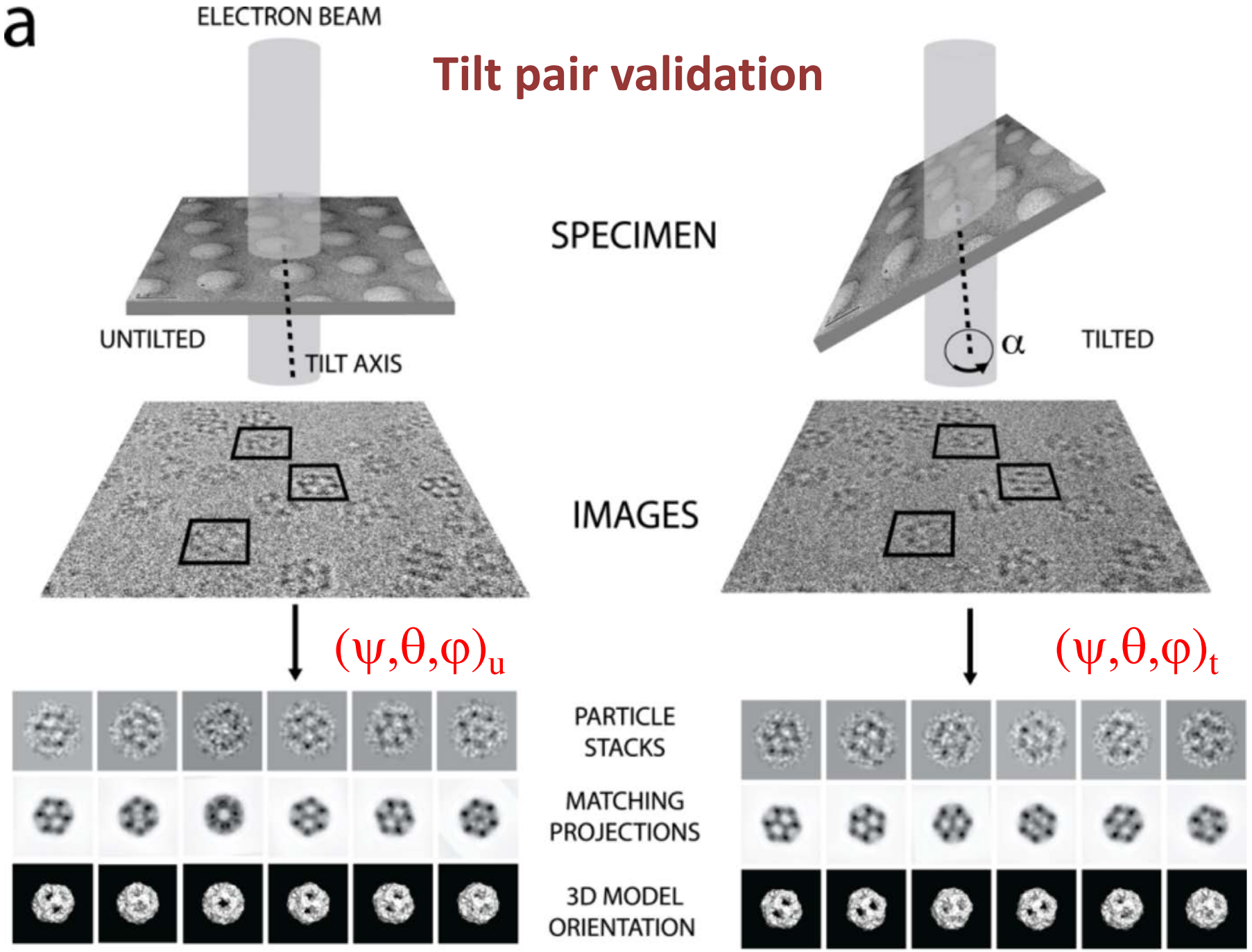
Naydenova & Russo, Nature Communications, 2017

Maximum Likelihood (ML) vs. CC (Align)

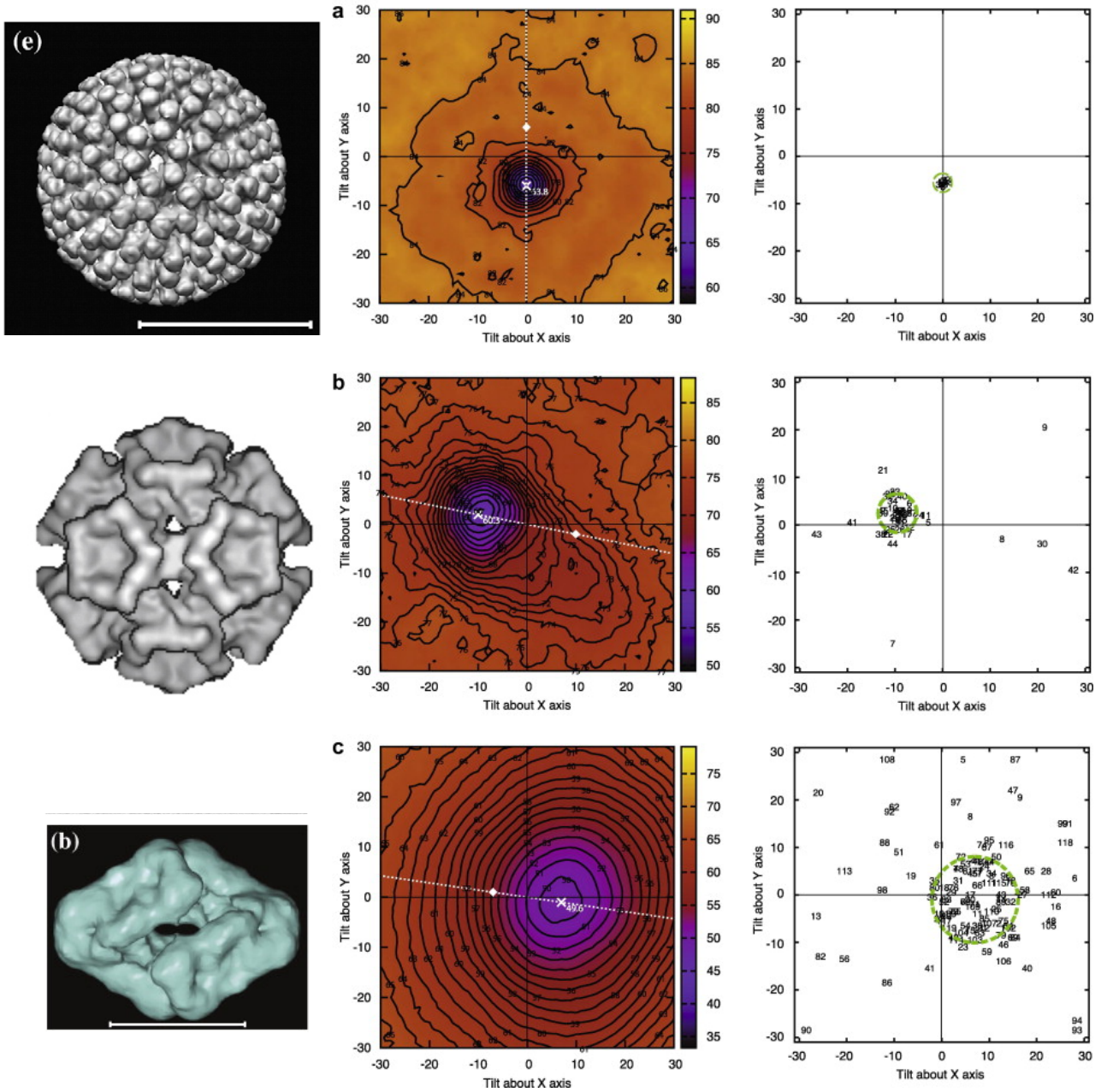




F.J. Sigworth (1998) JSB, 328-339

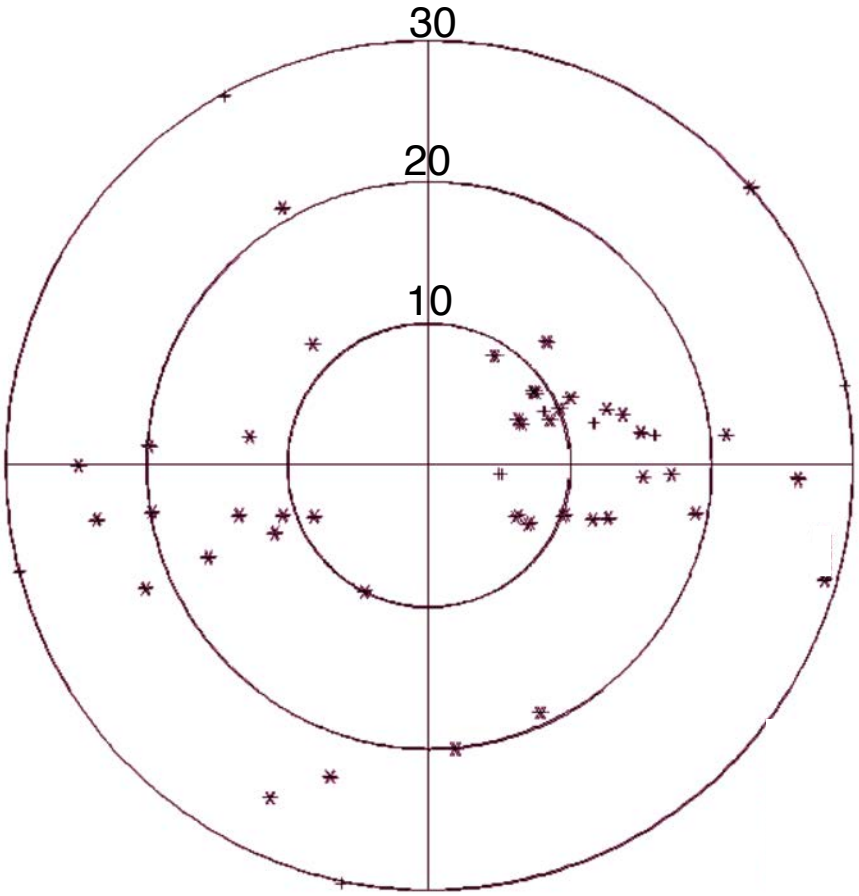

a
ELECTRON BEAM
Tilt pair validation
SPECIMEN
UNTILTED
TILT AXIS
α
TILTED
IMAGES
$(\psi,\theta,\varphi)_u$
$(\psi,\theta,\varphi)_t$
PARTICLE STACKS
MATCHING PROJECTIONS
3D MODEL ORIENTATION

Henderson et al. (2011)
Wasilewski and Rosenthal (2014)

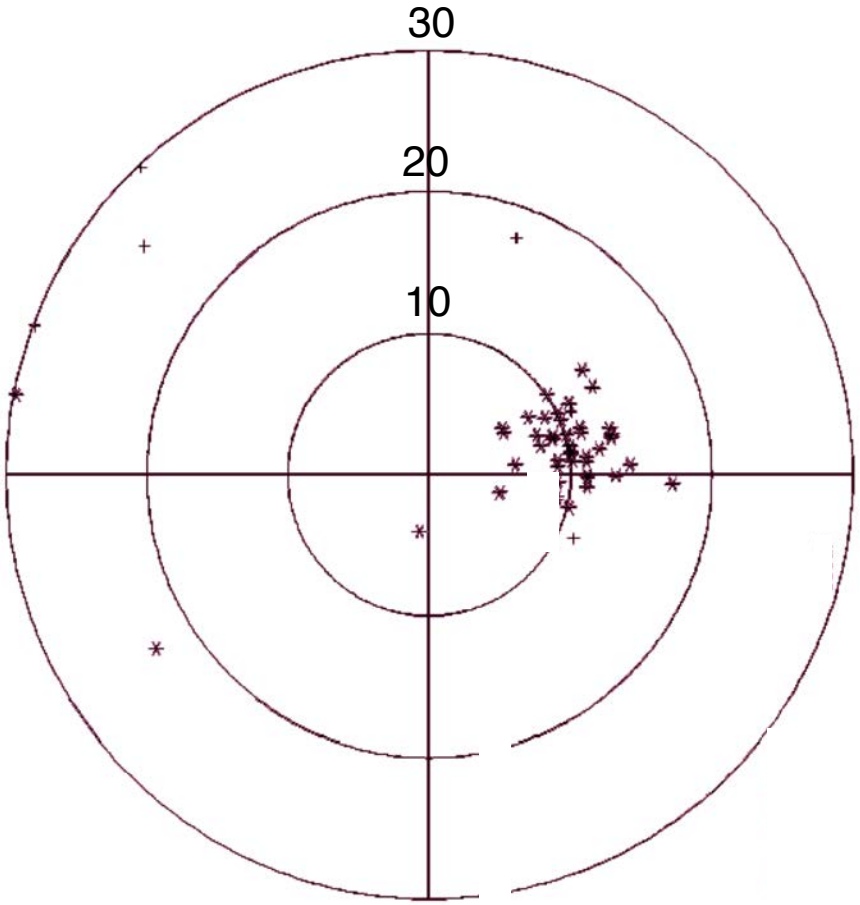


TILT AXIS FOR EACH PARTICLE PAIR

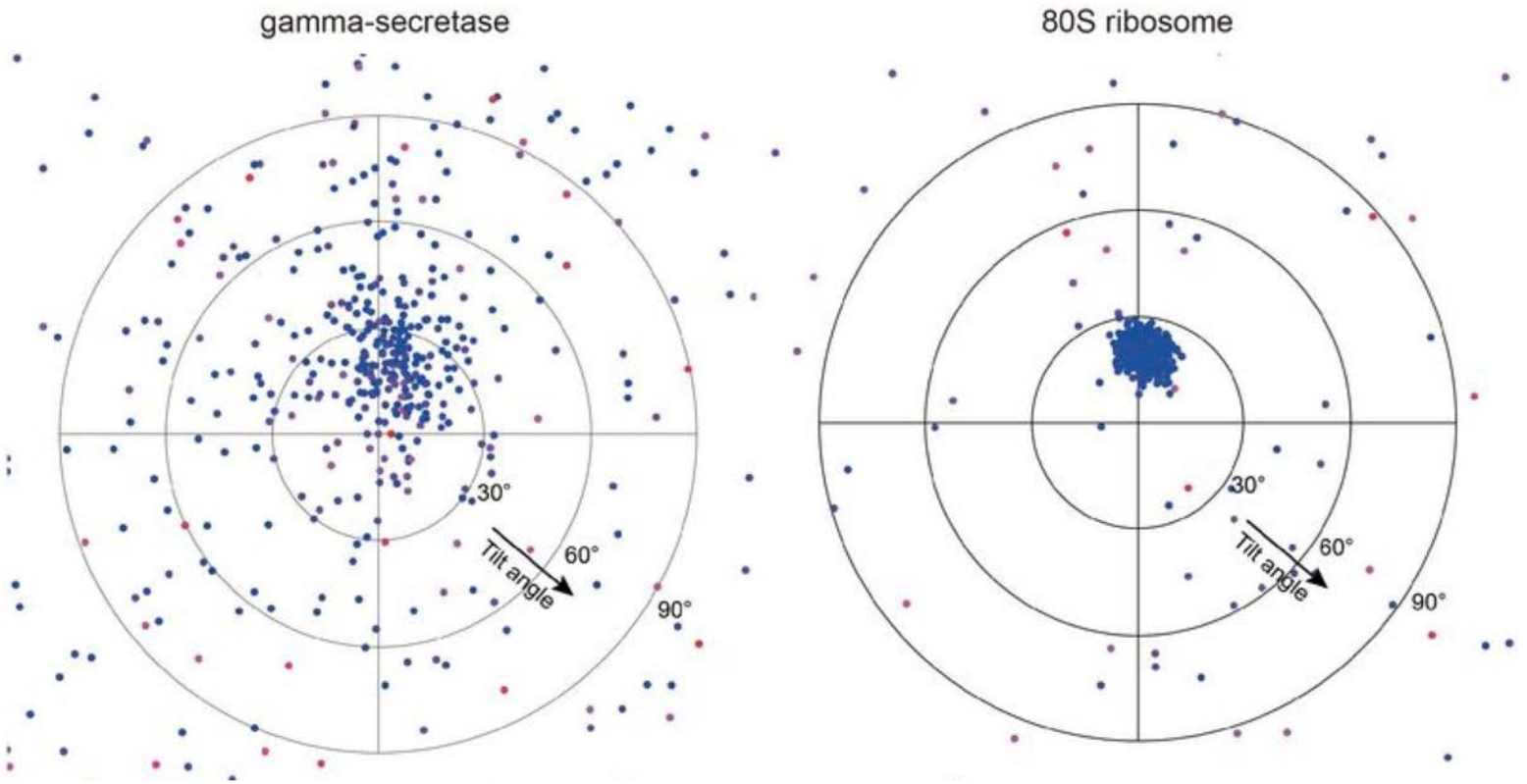
BEFORE OPTIMIZATION

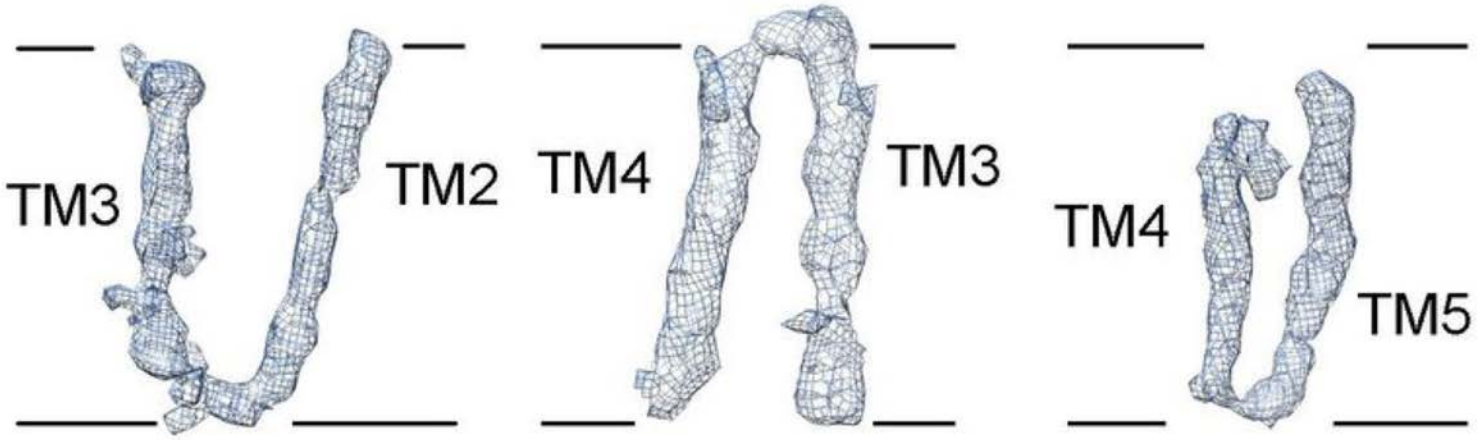


AFTER OPTIMIZATION



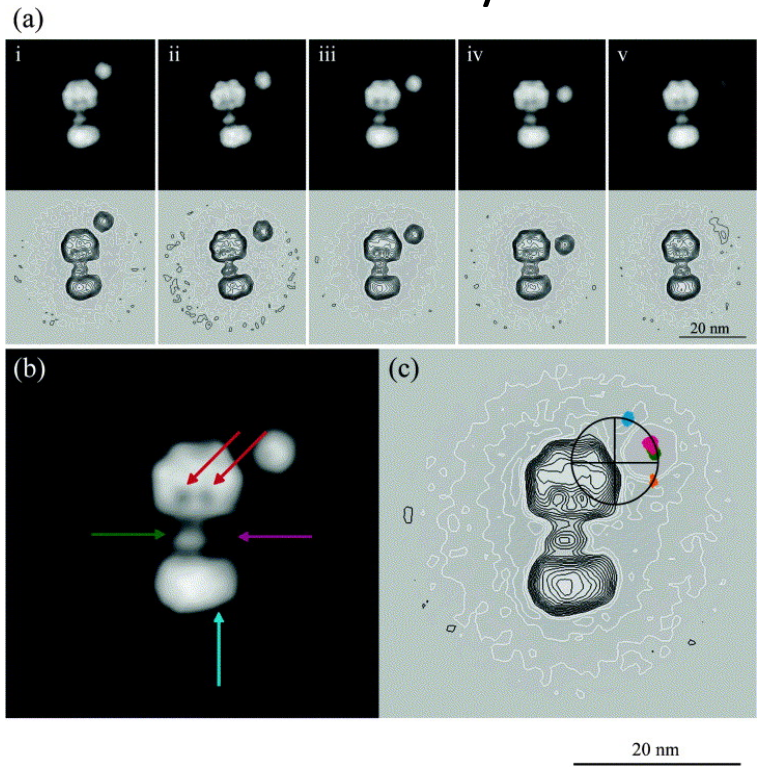
Lu et al. Nature 2014 Tilt-pair for Hand Determination





Identifying Biochemical Labels in Maps

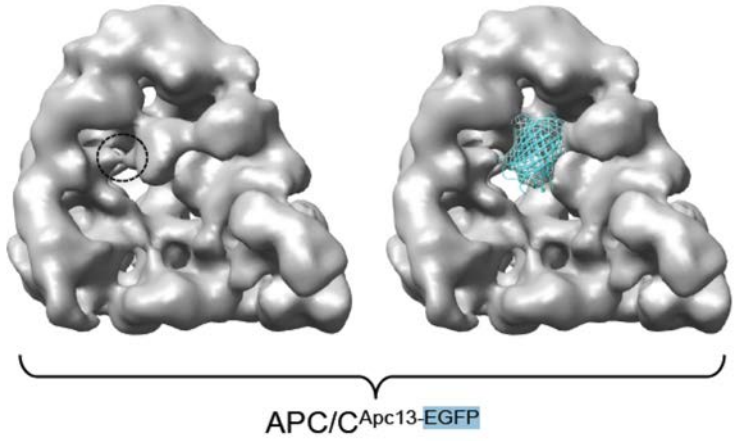
Avidin/biotin



ATP synthase from Saccharomyces cerevisiae: location of subunit h in the peripheral stalk region.

Rubinstein et al. (2005) JMB

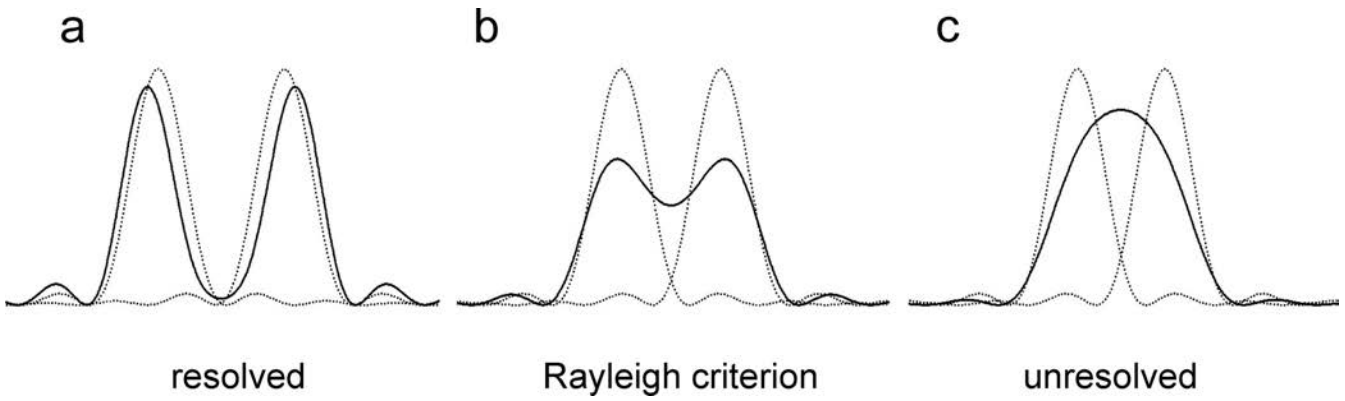
GFP



Structural basis for subunit assembly in the anaphase-promoting complex.

Schreiber et al. (2011) Nature

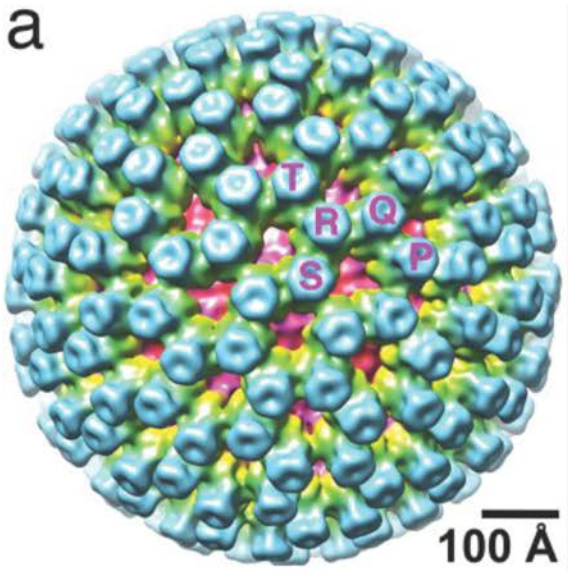
MAP RESOLUTION

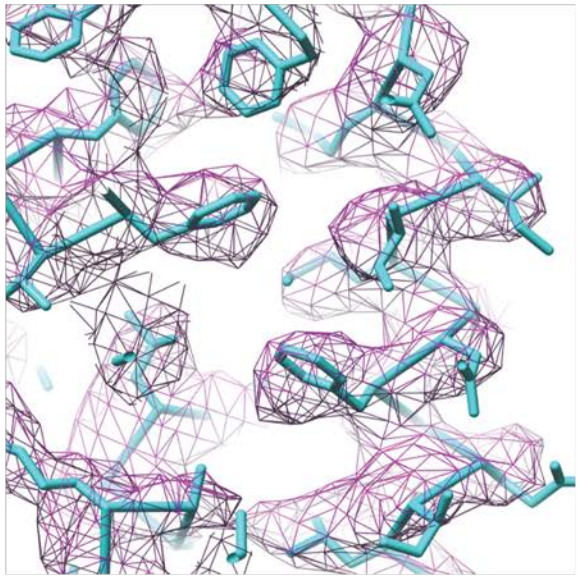


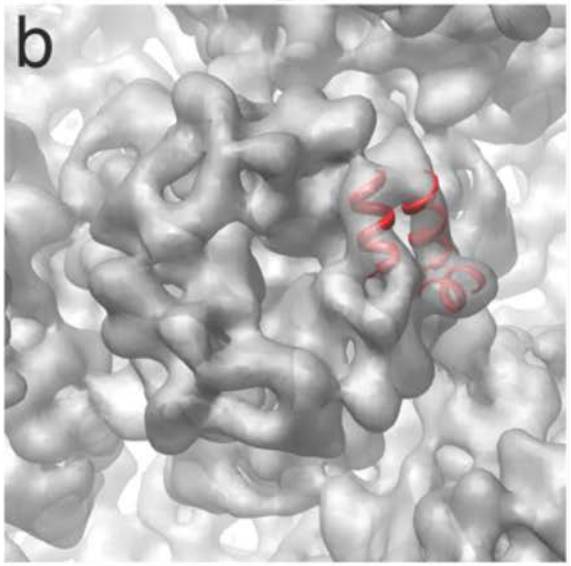
Resolution definition by separation of features. (a) When two points are far apart, there is a deep trough of density between them. (b) Two points are regarded as just resolved when the peak of one point spread function overlaps the first minimum of the other (Rayleigh criterion), see ref 60. (c) The point spread functions of two dots close together overlap to form one maximum, so that the points are not resolved.

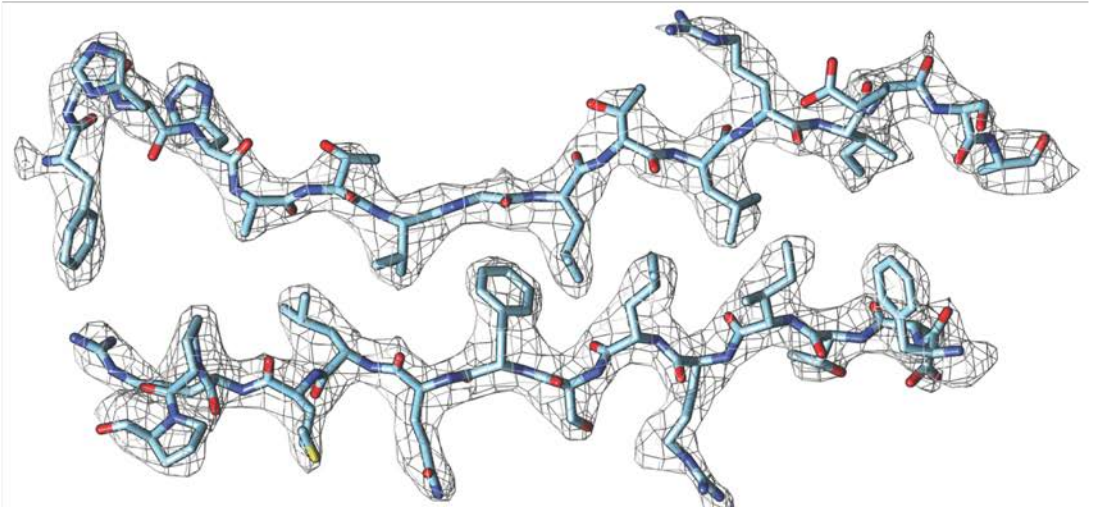
Published in: E. V. Orlova; H. R. Saibil; *Chem. Rev.* Article ASAP

The observed features of a map should be consistent with the resolution assessment





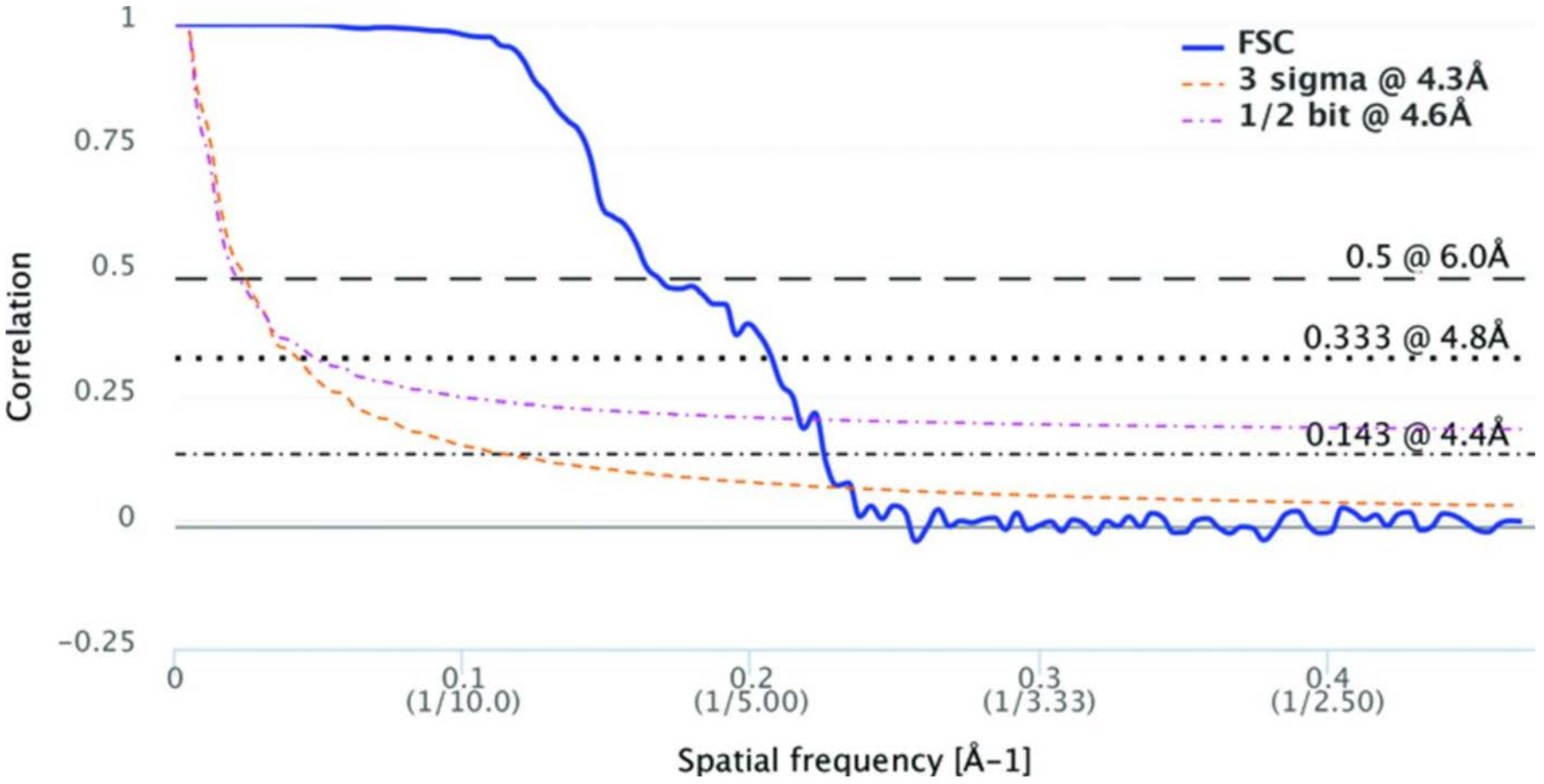




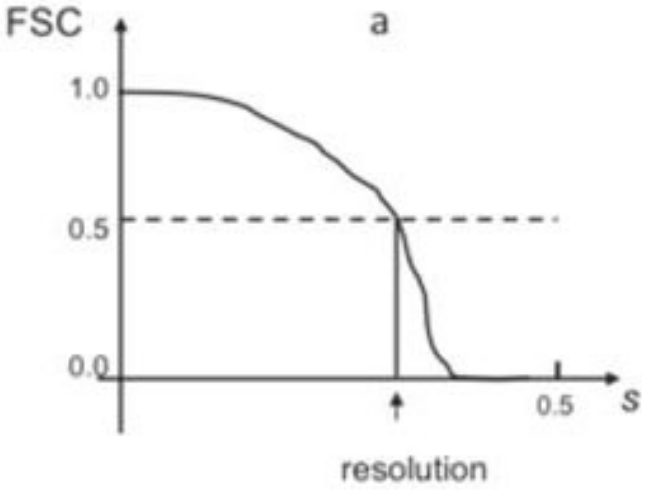
Map Resolution Should Be Reported, and Visible Structural Features Should Be in Accordance with the Claimed Resolution

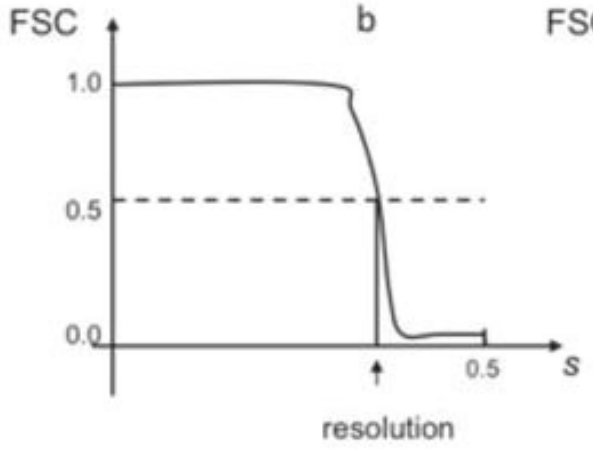
Visibility of expected features – α-helices visible at 9 Å resolution? β-strands at 4.8 Å resolution? Side-chains beyond 4 Å?

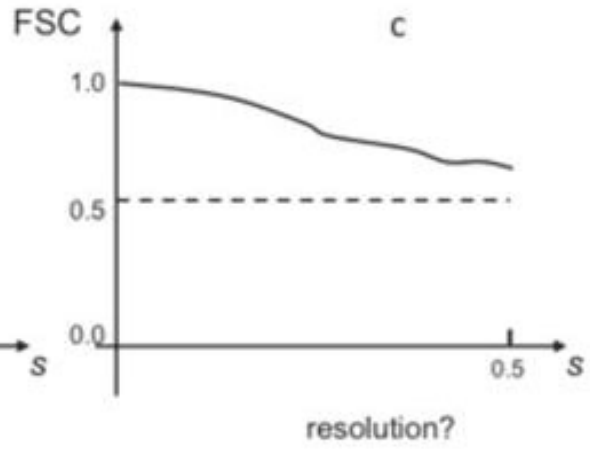
Fourier Shell Correlation: Show the whole curve.

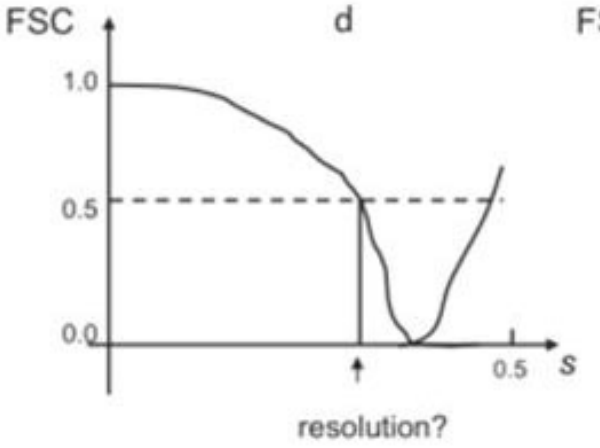


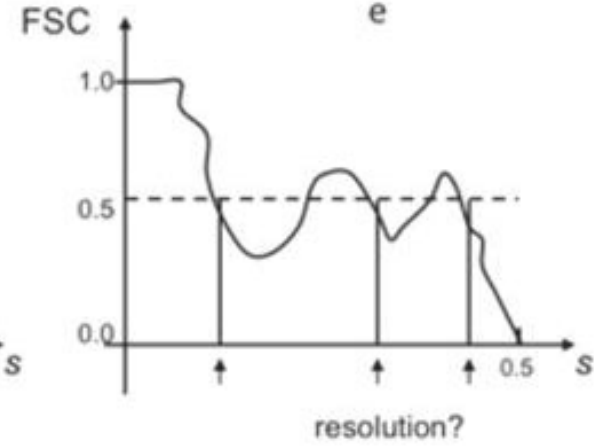
From: EMDataBank unified data resource for 3DEM
Nucleic Acids Res. 2015;44(D1):D396-D403. doi:10.1093/nar/gkv1126









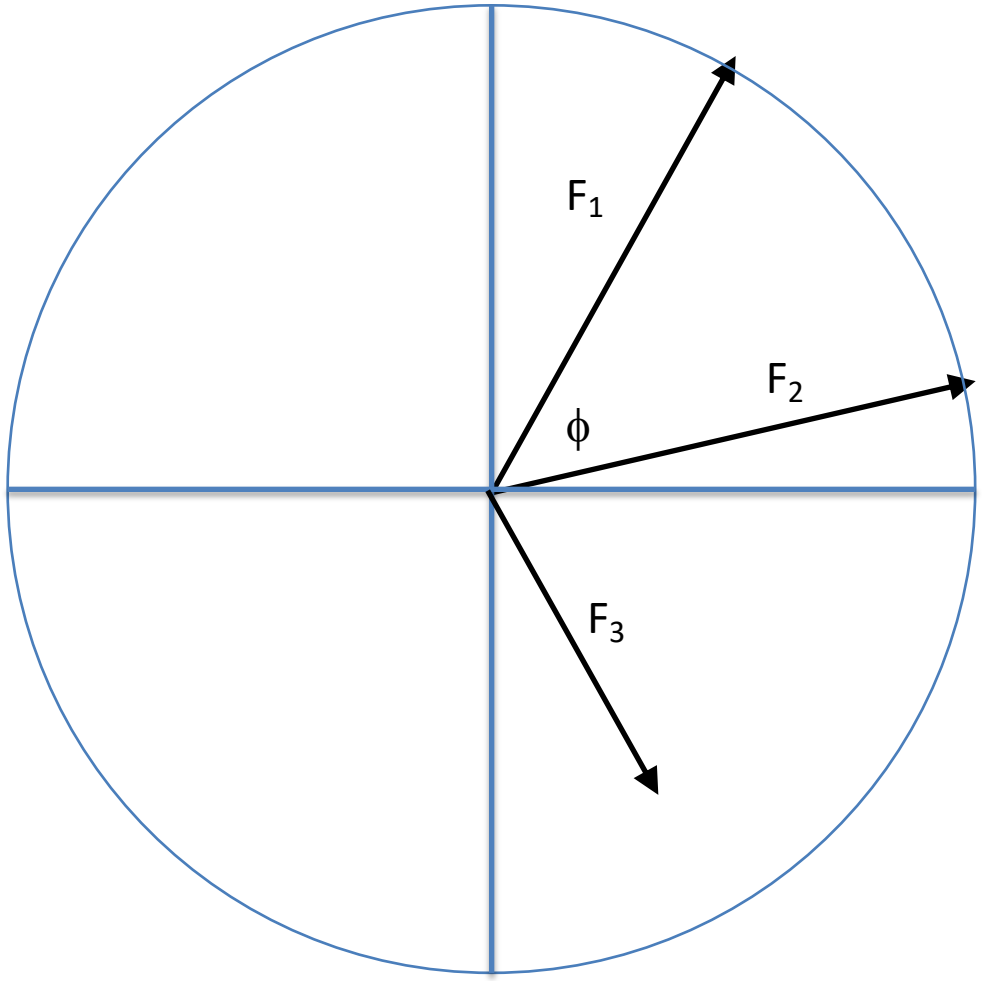


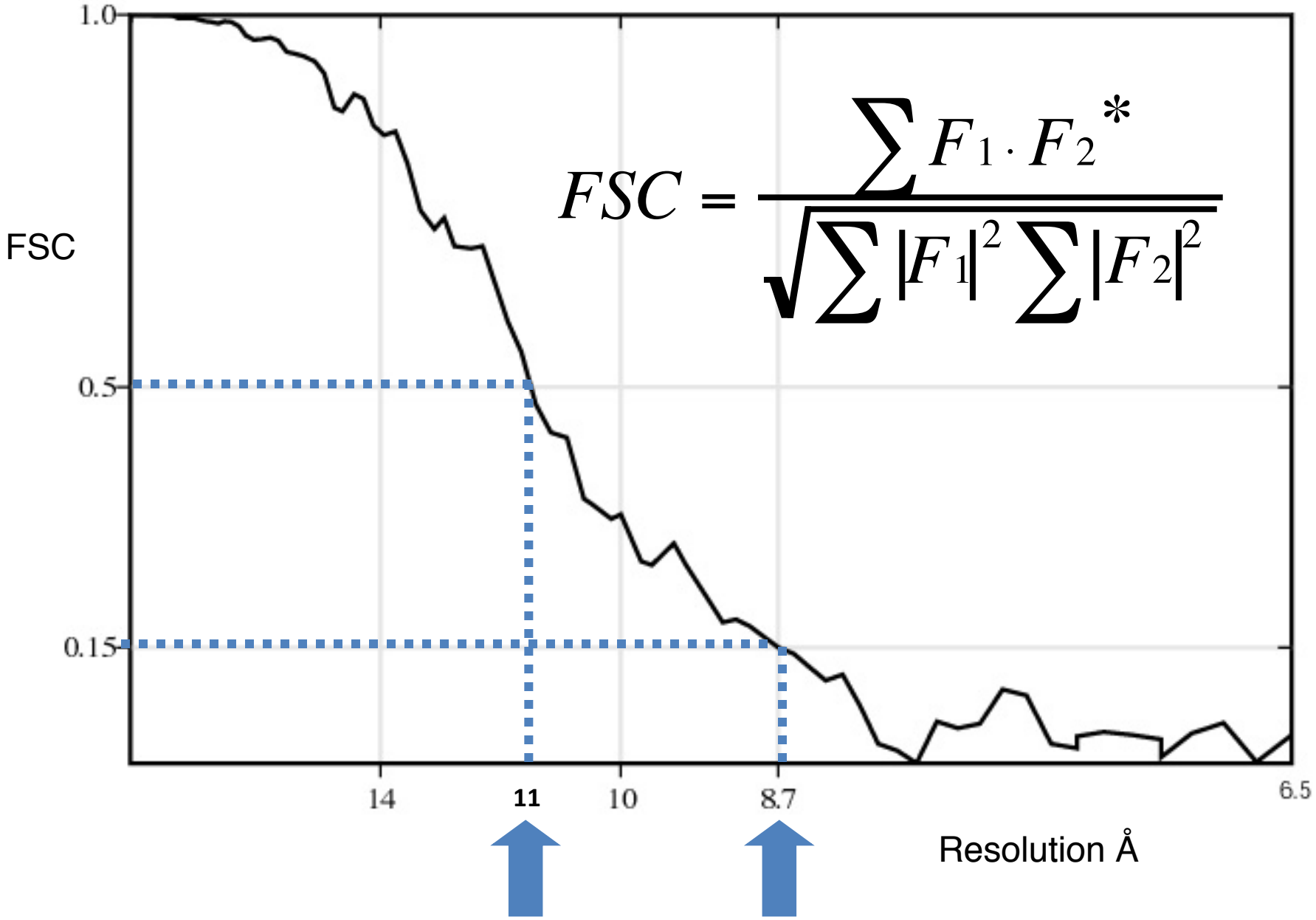
Resolution measures in molecular electron microscopy

Pawel A. Penczek

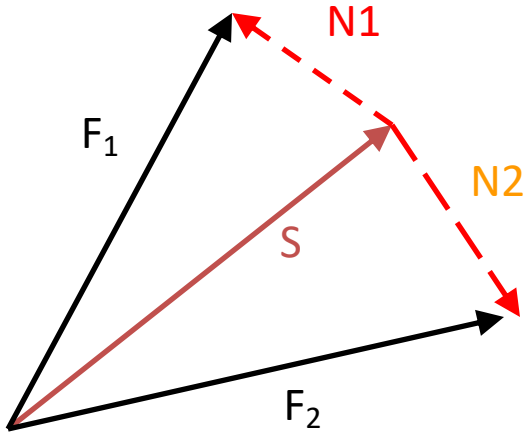
Methods Enzymol. 2010; 482: 73–100.

What is a structure factor?




1.0
0.5
0.15
FSC
$FSC = \frac{\sum F_1 \cdot F_2^*}{\sqrt{\sum |F_1|^2 \sum |F_2|^2}}$
14
11
10
8.7
6.5
Resolution Å

$$FSC = \frac{\sum F_1 \cdot F_2^*}{\sqrt{\sum |F_1|^2 \sum |F_2|^2}}$$

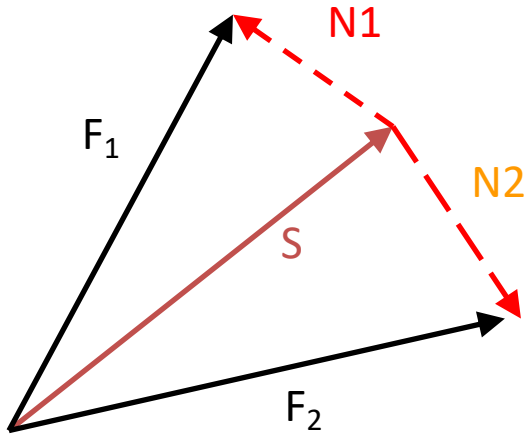


$$= \frac{\sum (S + N1) \cdot (S + N2)^*}{\sqrt{\sum |S + N1|^2 \sum |S + N2|^2}}$$

$$FSC = \frac{\sum |S|^2}{\sum |S^2 + N^2|}$$

S~N FSC=0.5

$$FSC = \frac{\sum F_1 \cdot F_2^*}{\sqrt{\sum |F_1|^2 \sum |F_2|^2}}$$



Signal~Noise FSC=0.5

$$FSC_{full} = \frac{2FSC}{1+FSC}$$

Correlation with a perfect reference

$$C_{ref} = \frac{\sum F_1 \cdot F_{ref}}{\sqrt{\sum |F_1|^2 \sum |F_{ref}|^2}} = \frac{\sum |F_1||F_{ref}|\cos(\Delta\varphi)}{\sqrt{\sum |F_1|^2 \sum |F_{ref}|^2}}$$

Looks like figure-of-merit (Blow and Crick,1959)

Estimating C_{ref}

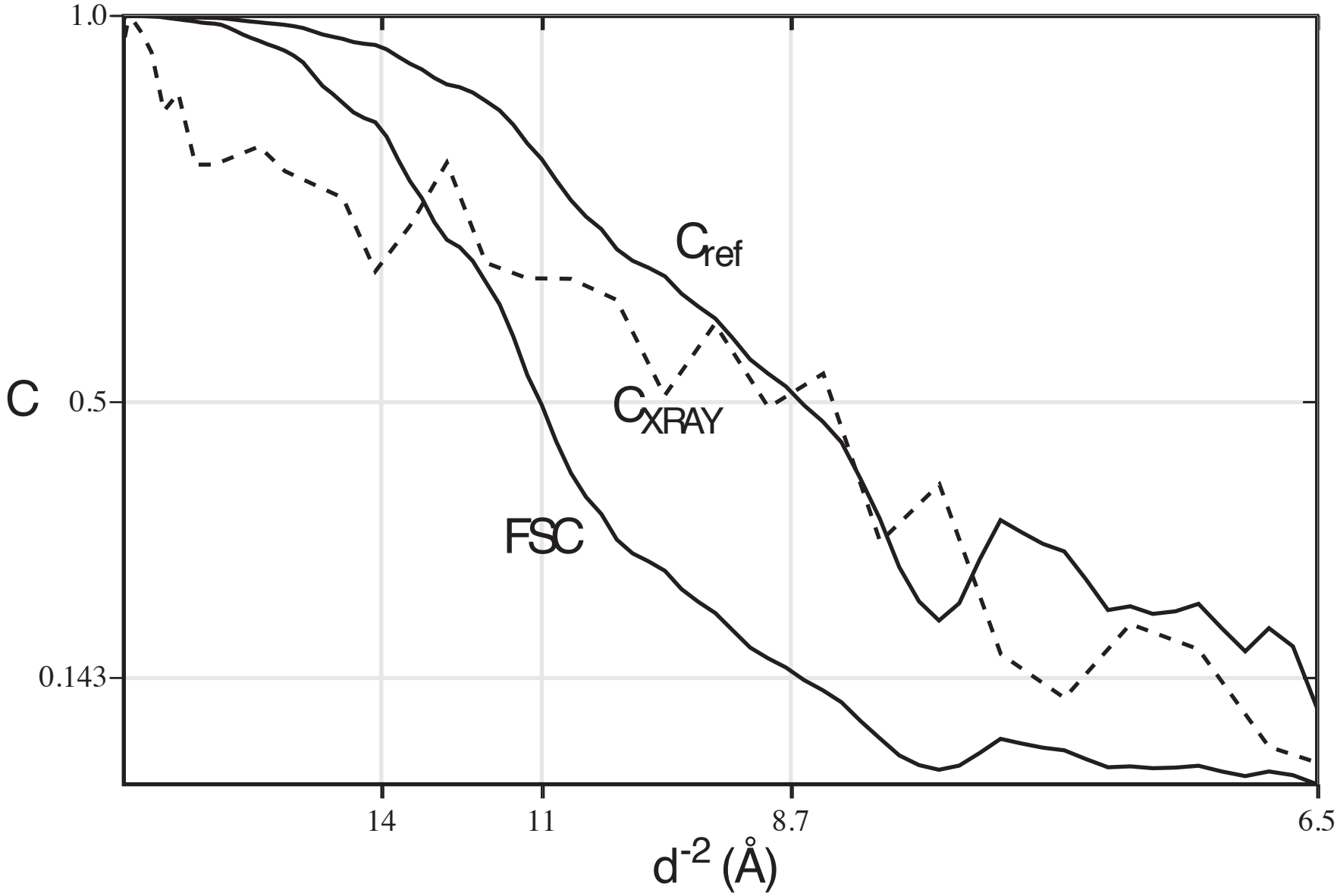
$$C_{ref} = \sqrt{\frac{2FSC}{1+FSC}}$$

C_{ref}=0.5 FSC=0.14

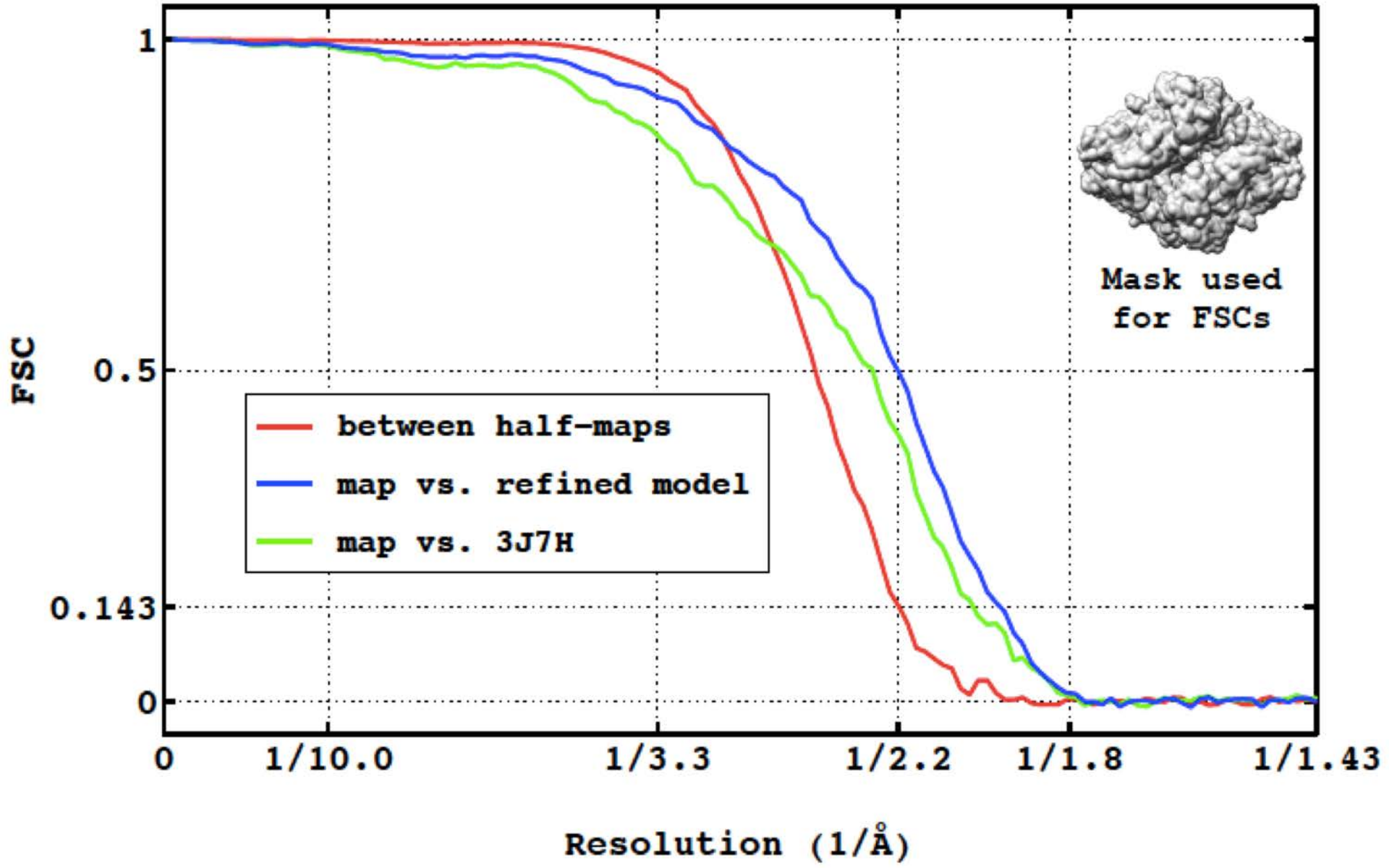
C_{ref} corresponds to crystallographic FOM "Figure of Merit"

C_{ref}=0.5 mean phase error 60 ° (last shell) interpretable by an atomic model

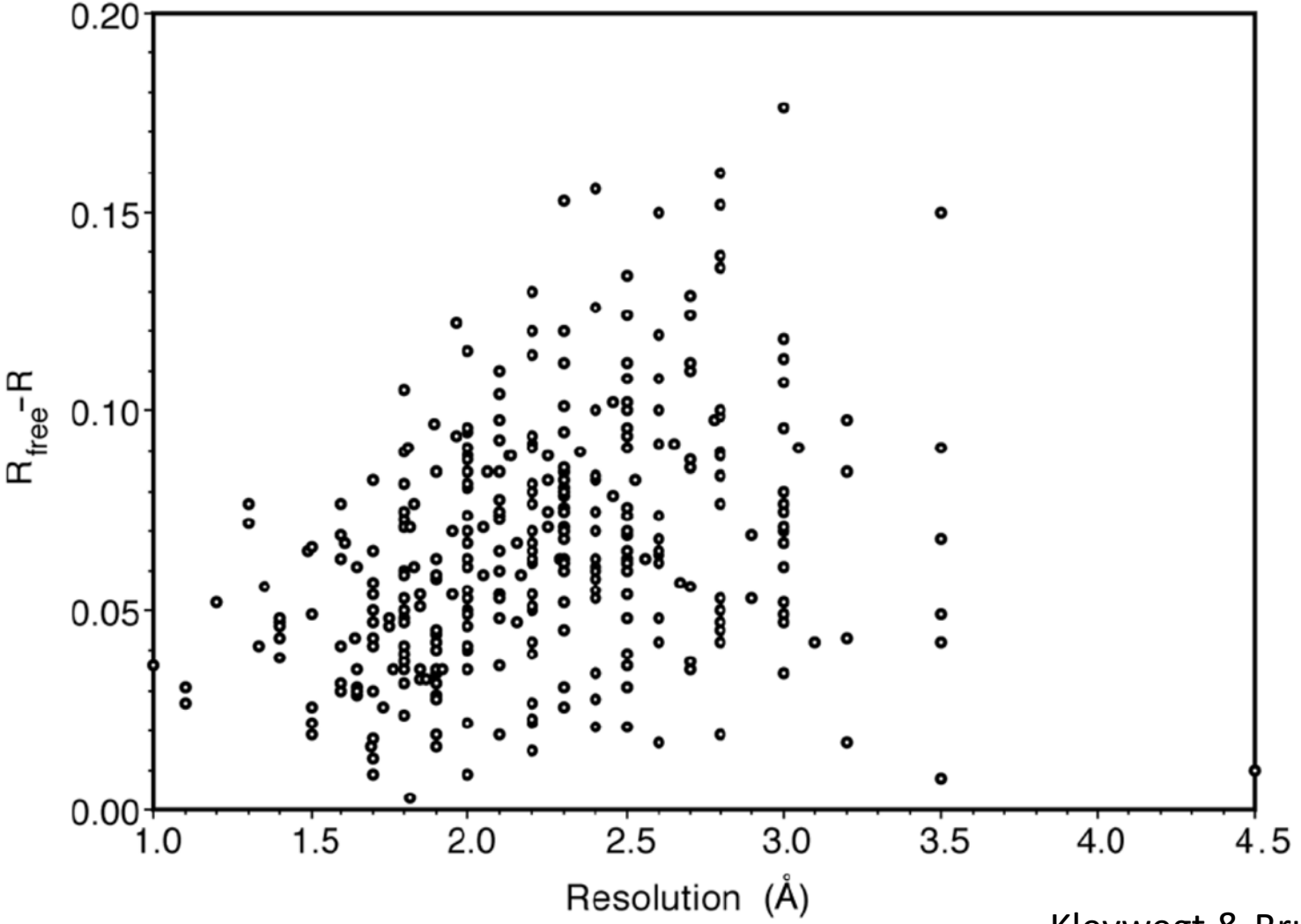
FSC	FSC_{FULL}	C_{REF}	PHASE ERROR	$S/N_{1/2}$
0.50	0.67	0.82	35°	1.00
0.33	0.50	0.71	45°	0.71
0.14	0.25	0.50	60°	0.41


1.0
0.5
0.143
C
C_{ref}
C_{XRAY}
FSC
14
11
8.7
6.5
d^{-2} (Å)

Beta-galactosidease cryo-EM map Bartesaghi et al. 2015

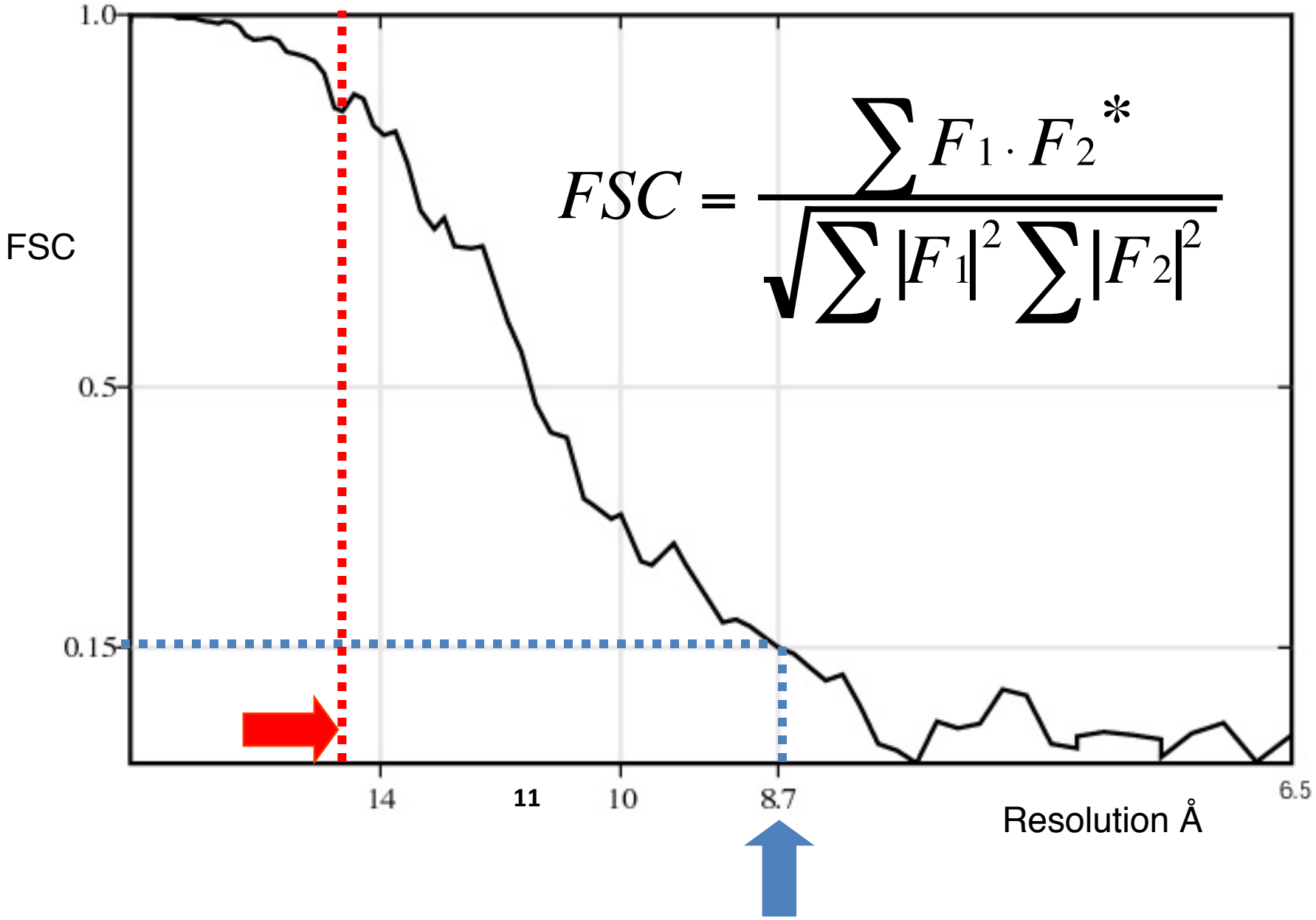


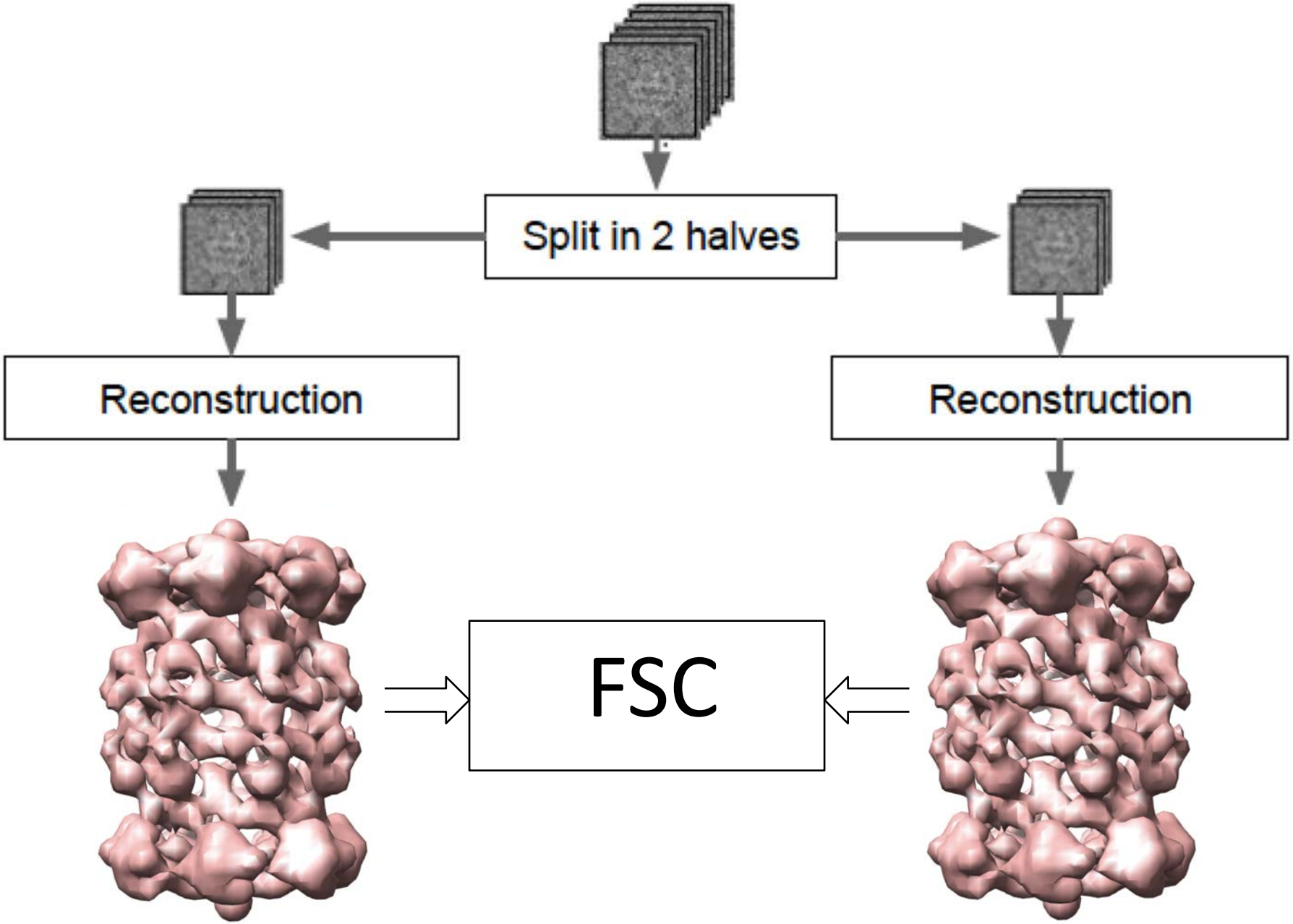
R_{work} , R_{free} in X-ray cross-validation to detect over-fitting



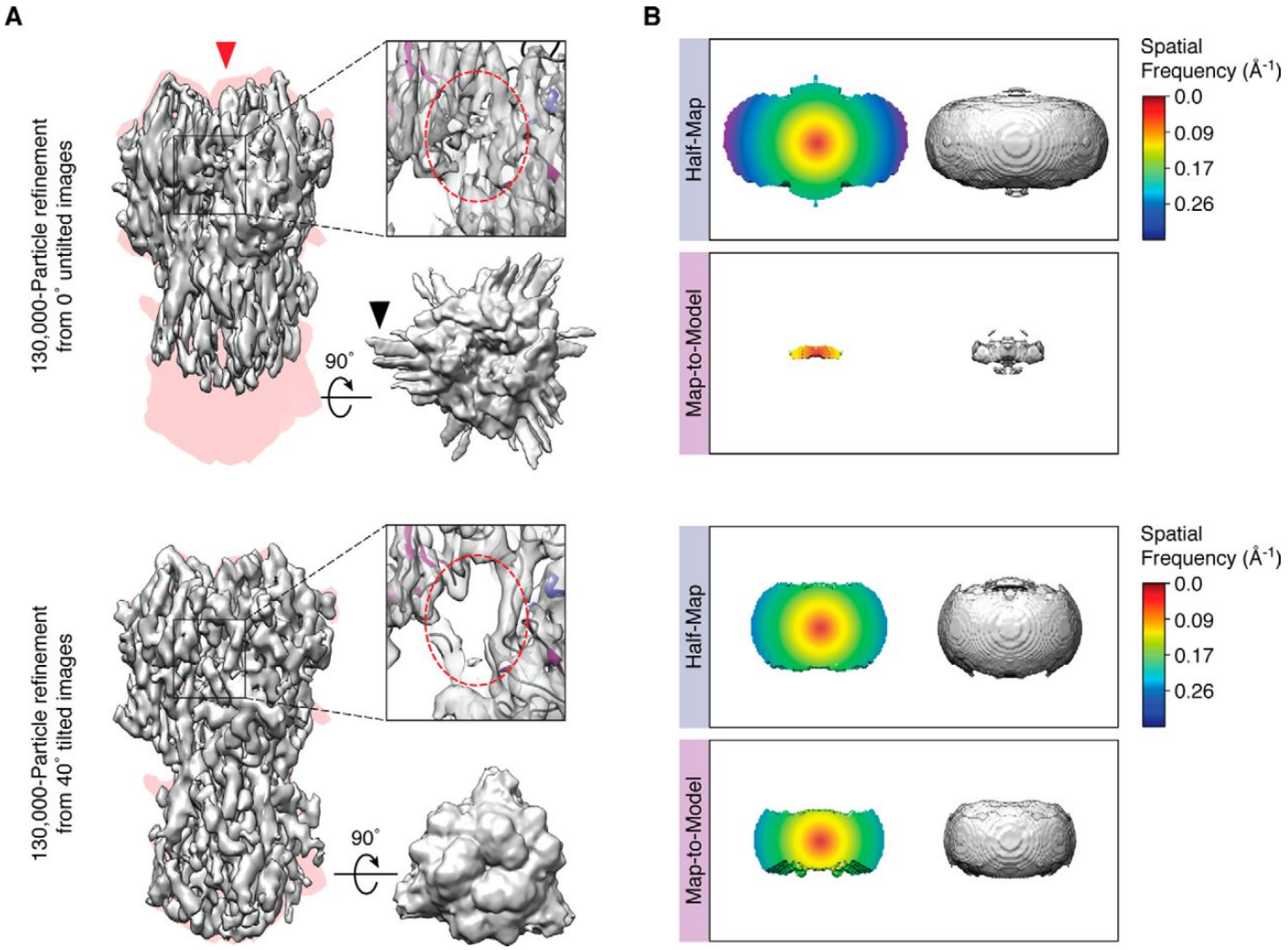
Kleywegt & Brunger 1996

Free shells at high-resolution not included in refinement




Split in 2 halves
Reconstruction
Reconstruction
FSC

Experimental results showing the utility of a tilted data collection strategy.



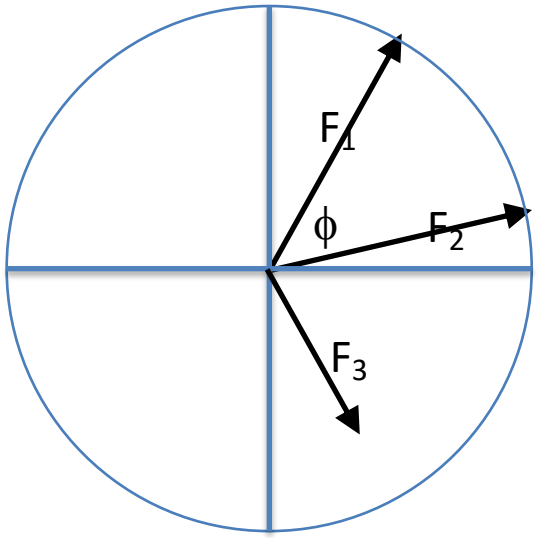
Dmitry Lyumkis J. Biol. Chem. 2019;294:5181-5197

JBC JOURNAL OF BIOLOGICAL CHEMISTRY

HIGH RESOLUTION NOISE SUBSTITUTION

Chen et al. *Ultramicroscopy* **135** (2013) 24-35

Perform Single Particle EM analysis.



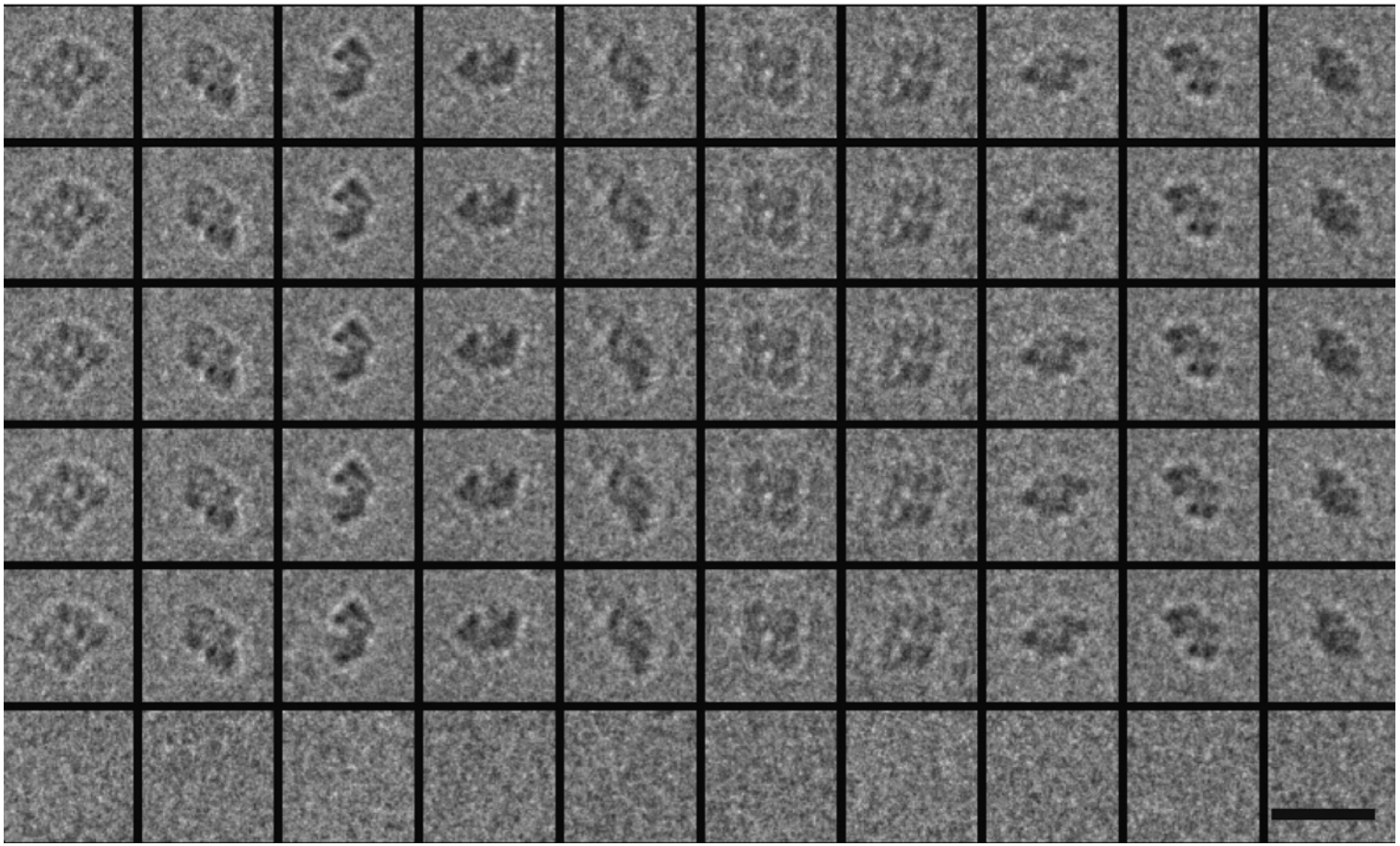
Repeat but substitute random phases beyond a selected resolution (HR-noise)

Any overfitted noise will show up as non-zero FSC.

Genuine information will show up as the area between the two curves

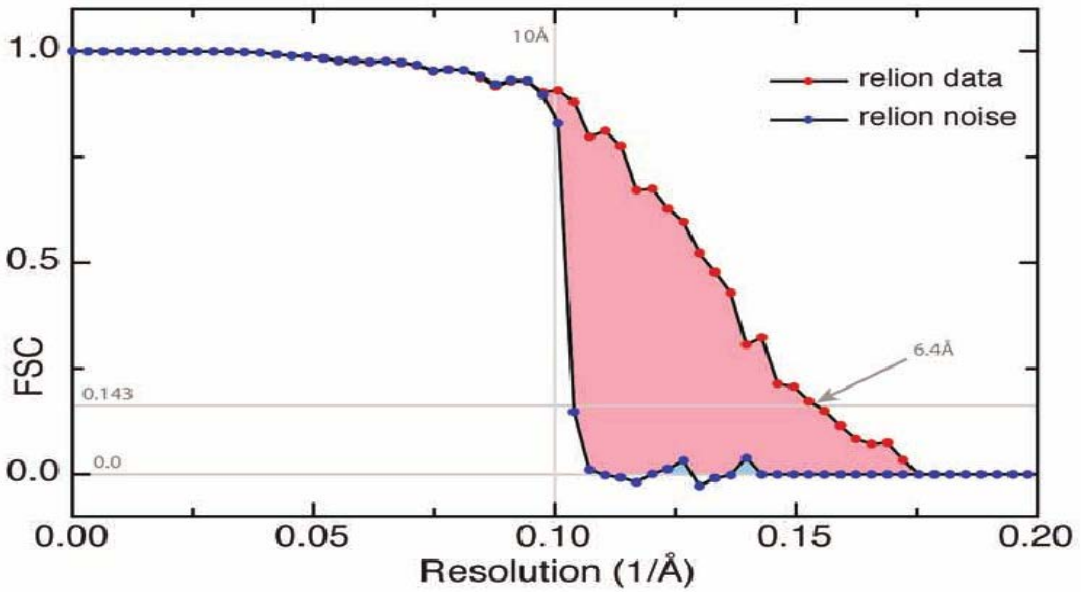
HIGH RESOLUTION NOISE SUBSTITUTION

Chen et al. *Ultramicroscopy* **135** (2013) 24-35

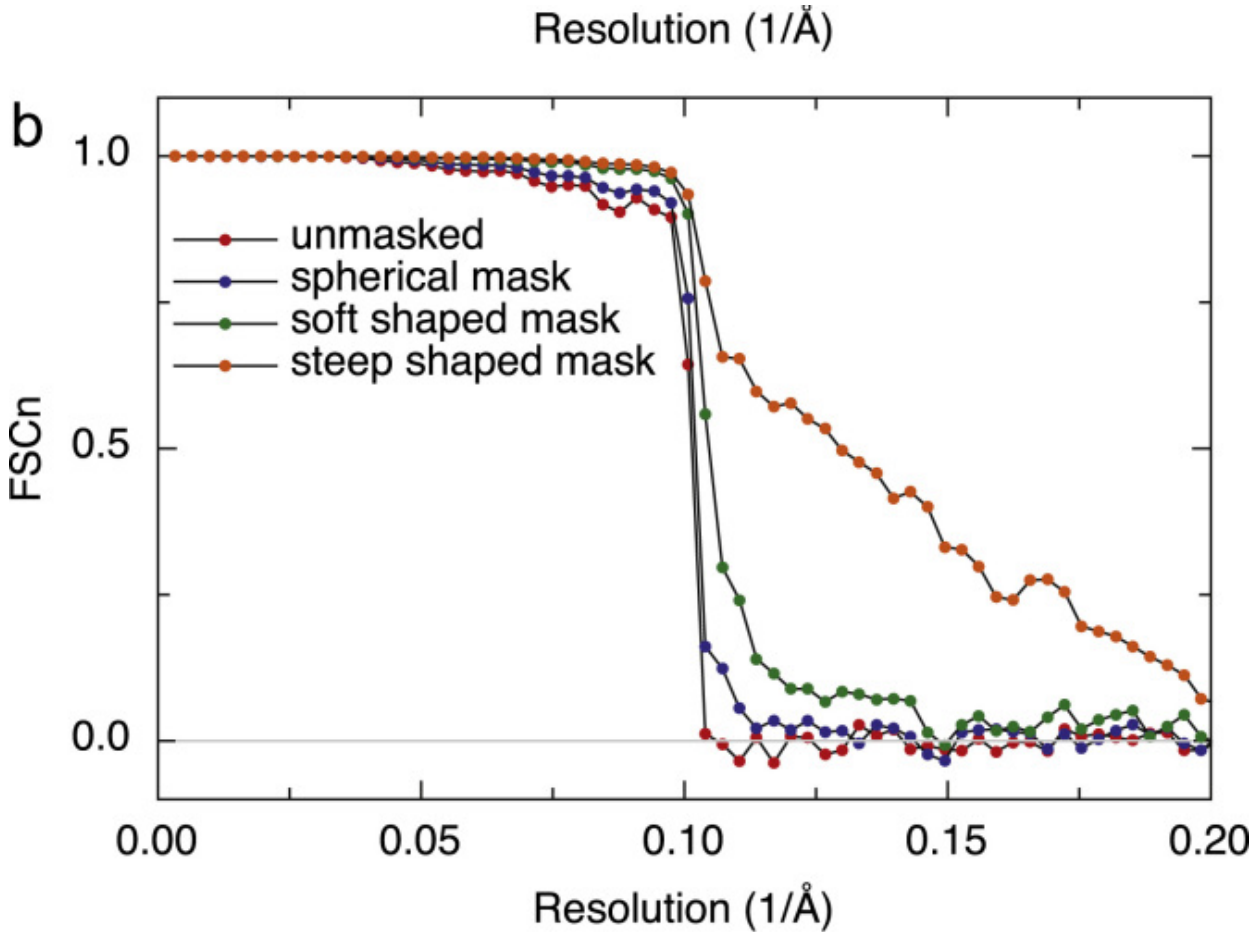


Assess overfitting

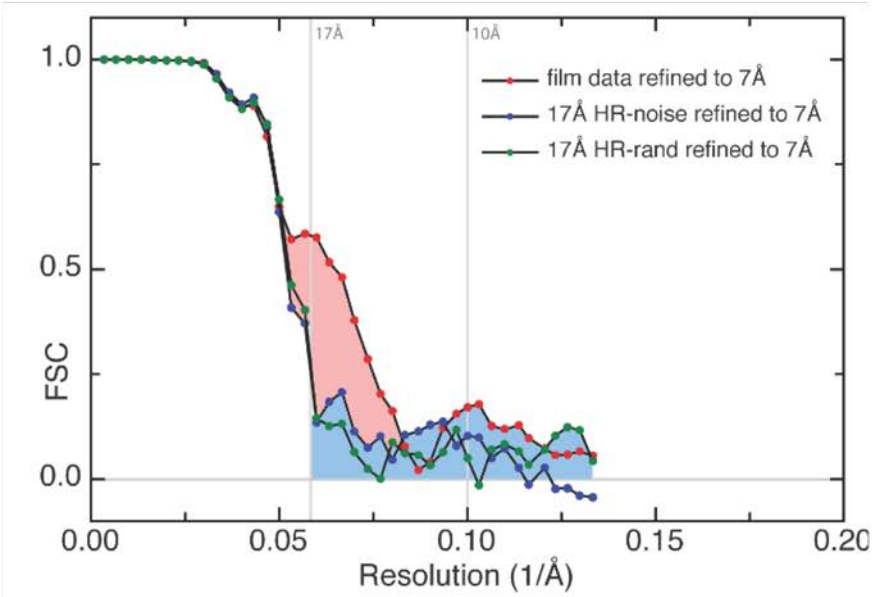
- Especially useful for new programs/methods
  - high-resolution noise substitution test

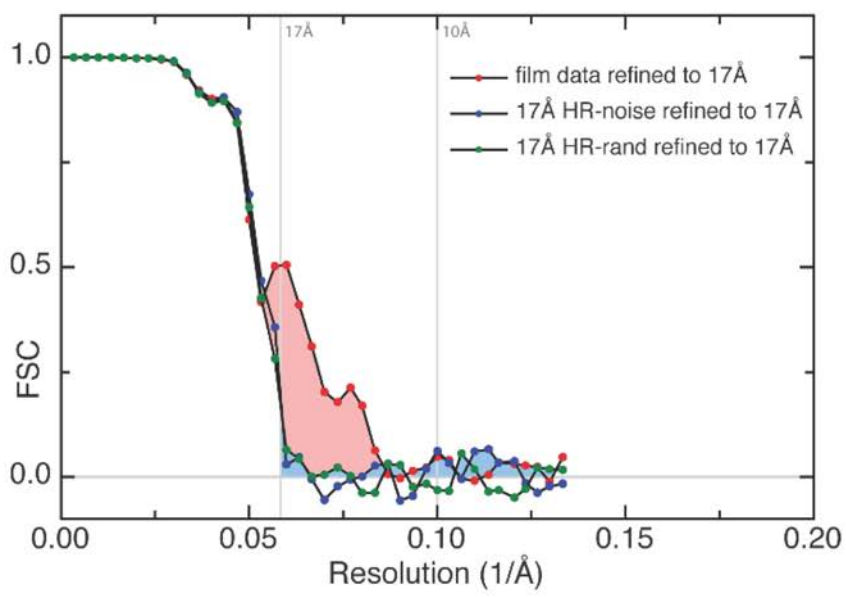


Chen et al. *Ultramicroscopy* **135** (2013) 24-35



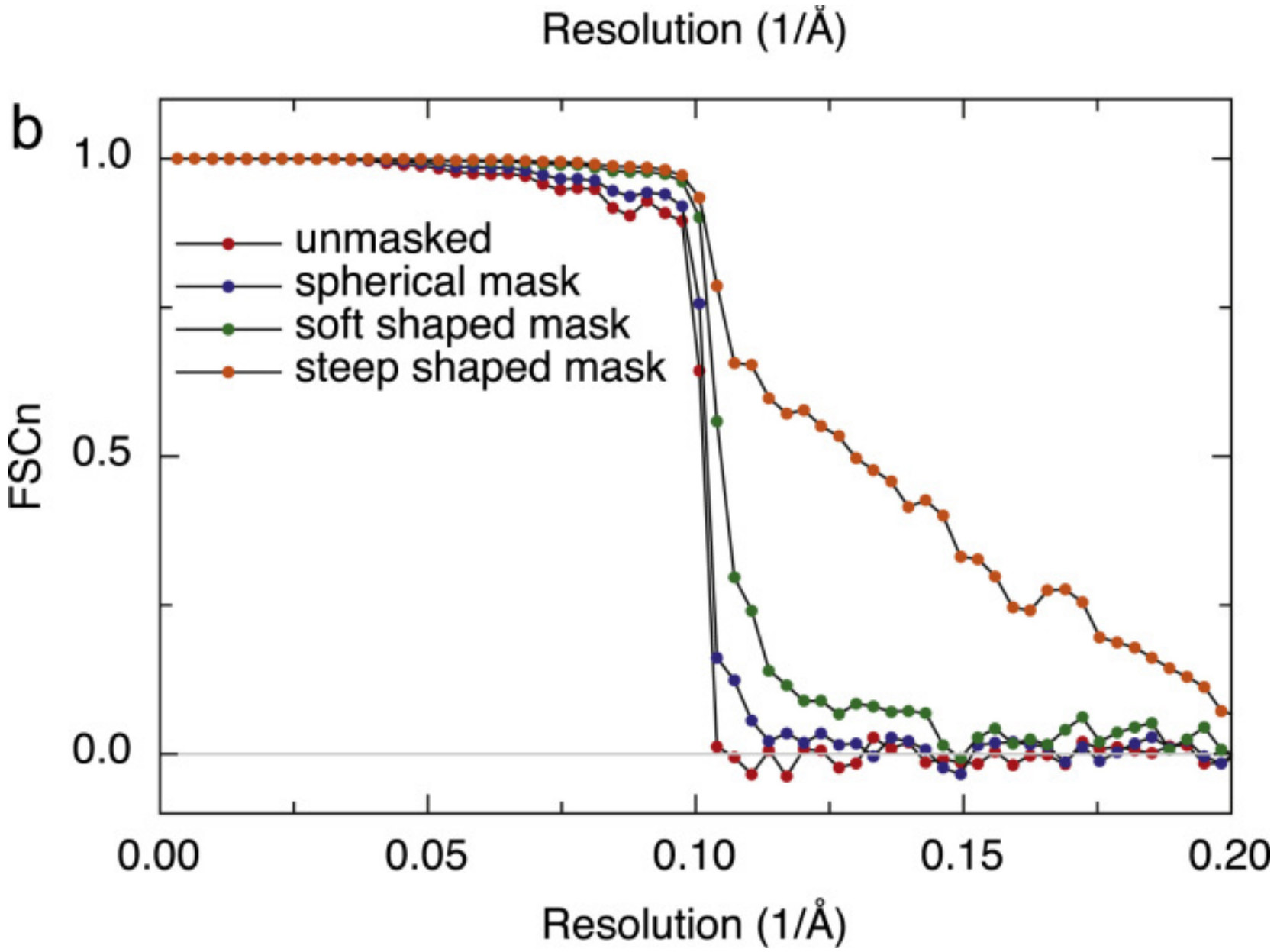
Chen et al. *Ultramicroscopy* **135** (2013) 24-35





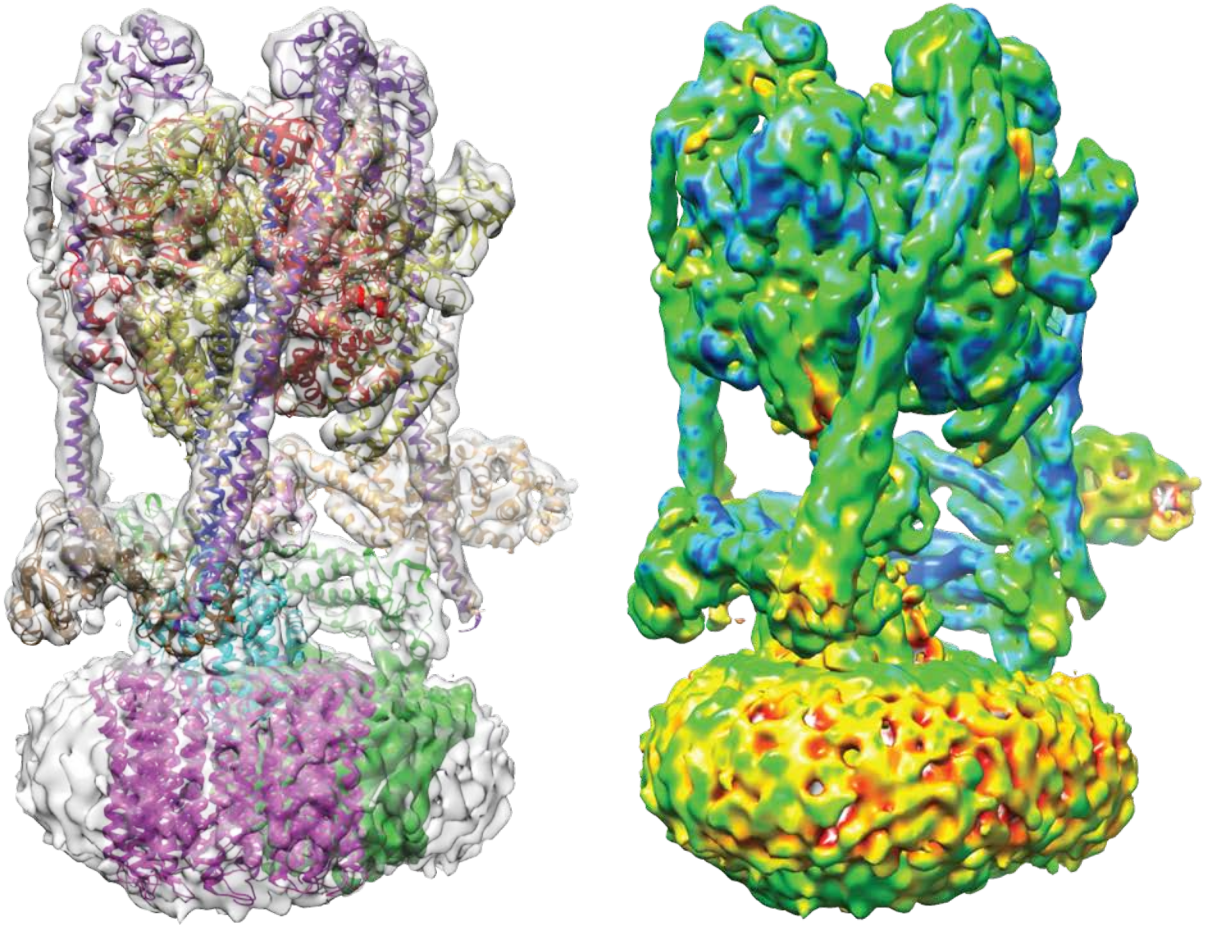
The difference between the FSC from the original data (FSCt) and the FSC with high-resolution noise substitution or phase randomization (FSCn) represents the true high-resolution signal in the map (FSCtrue)

Chen et al. *Ultramicroscopy* **135** (2013) 24-35

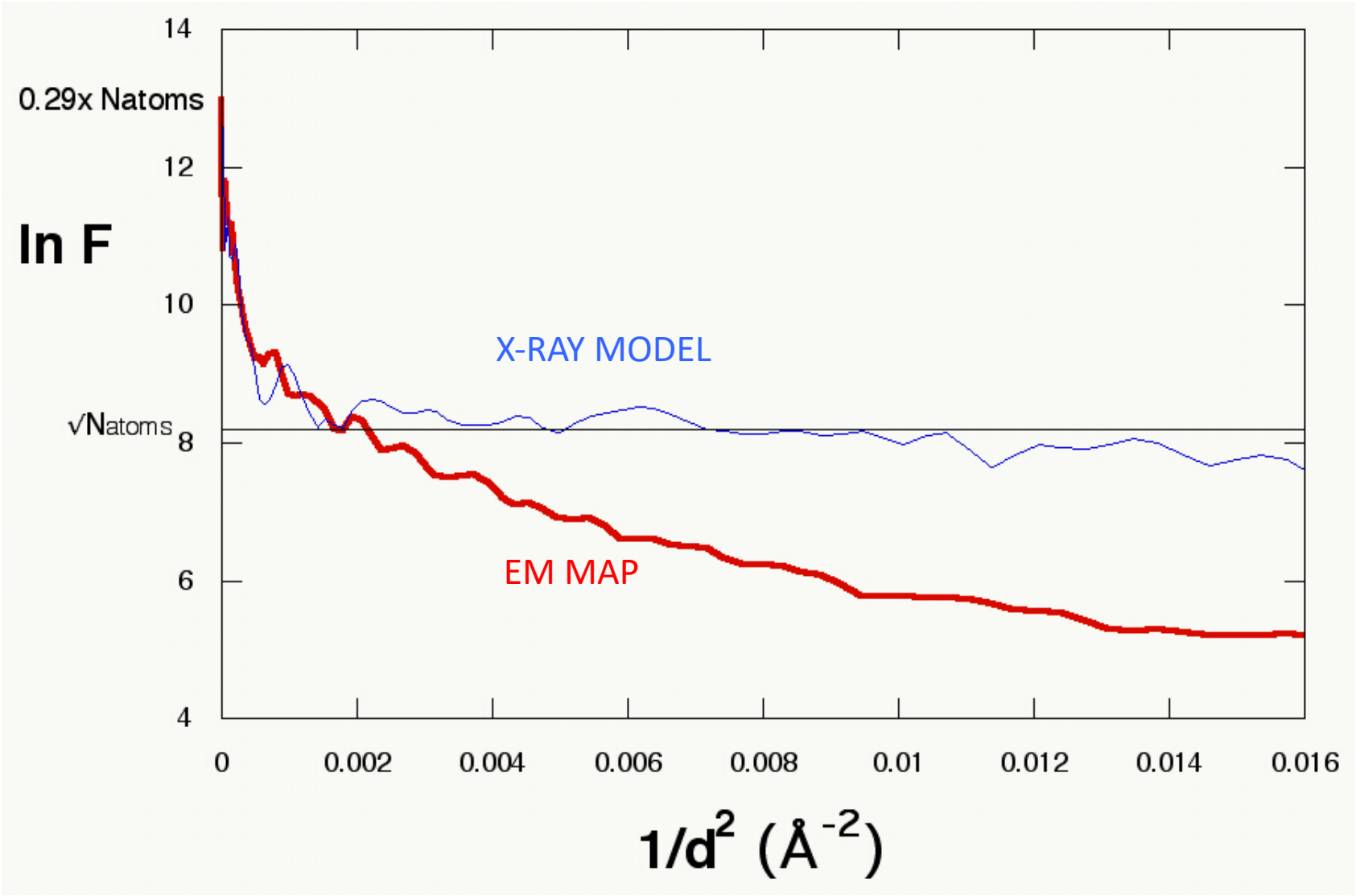


Chen et al. *Ultramicroscopy* **135** (2013) 24-35

LOCAL RESOLUTION: RESMAP



RADIALLY-AVERAGED EM MAP AMPLITUDES COMPARED TO X-RAY



Contrast Restoration

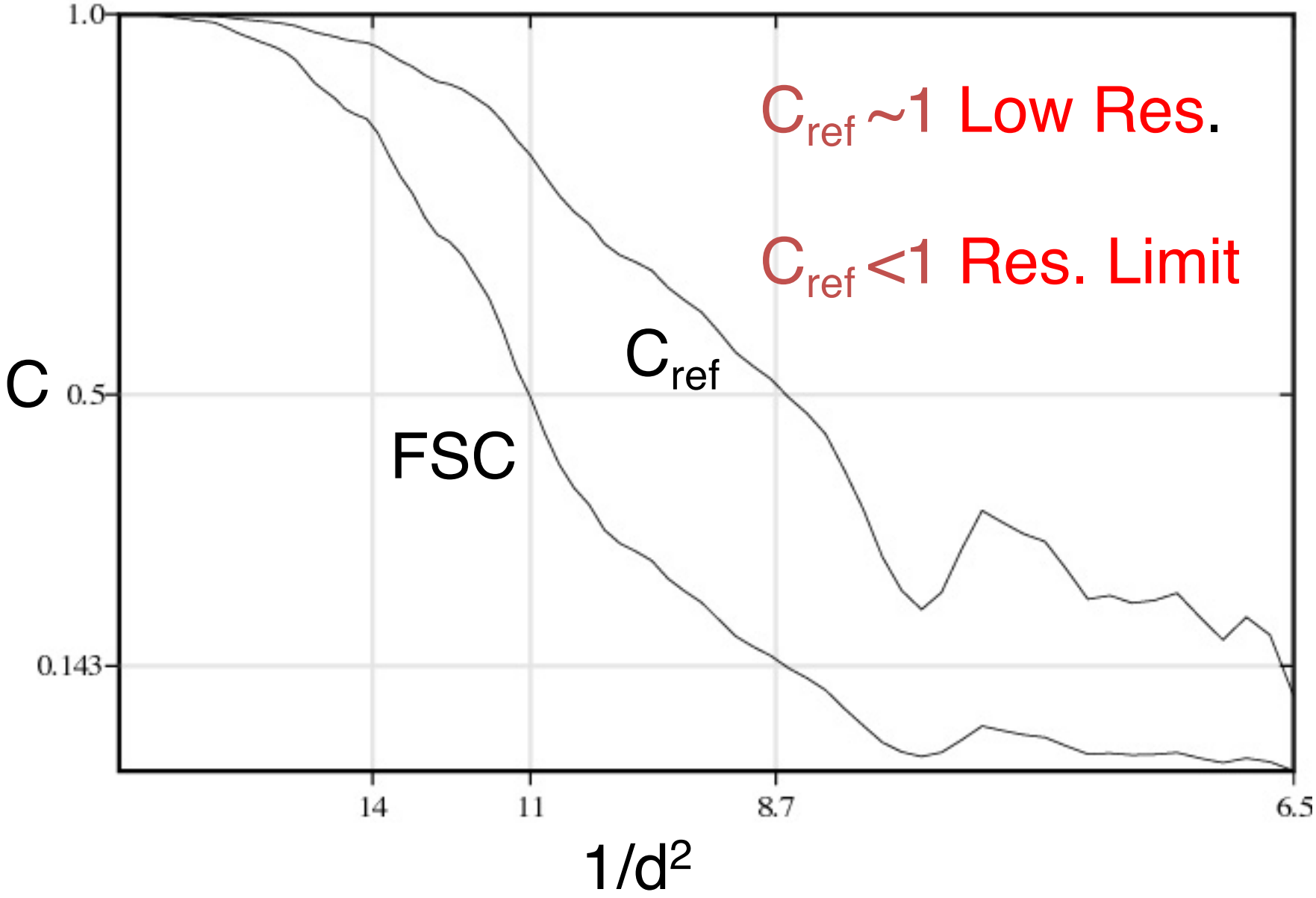
(Rosenthal and Henderson 2003)

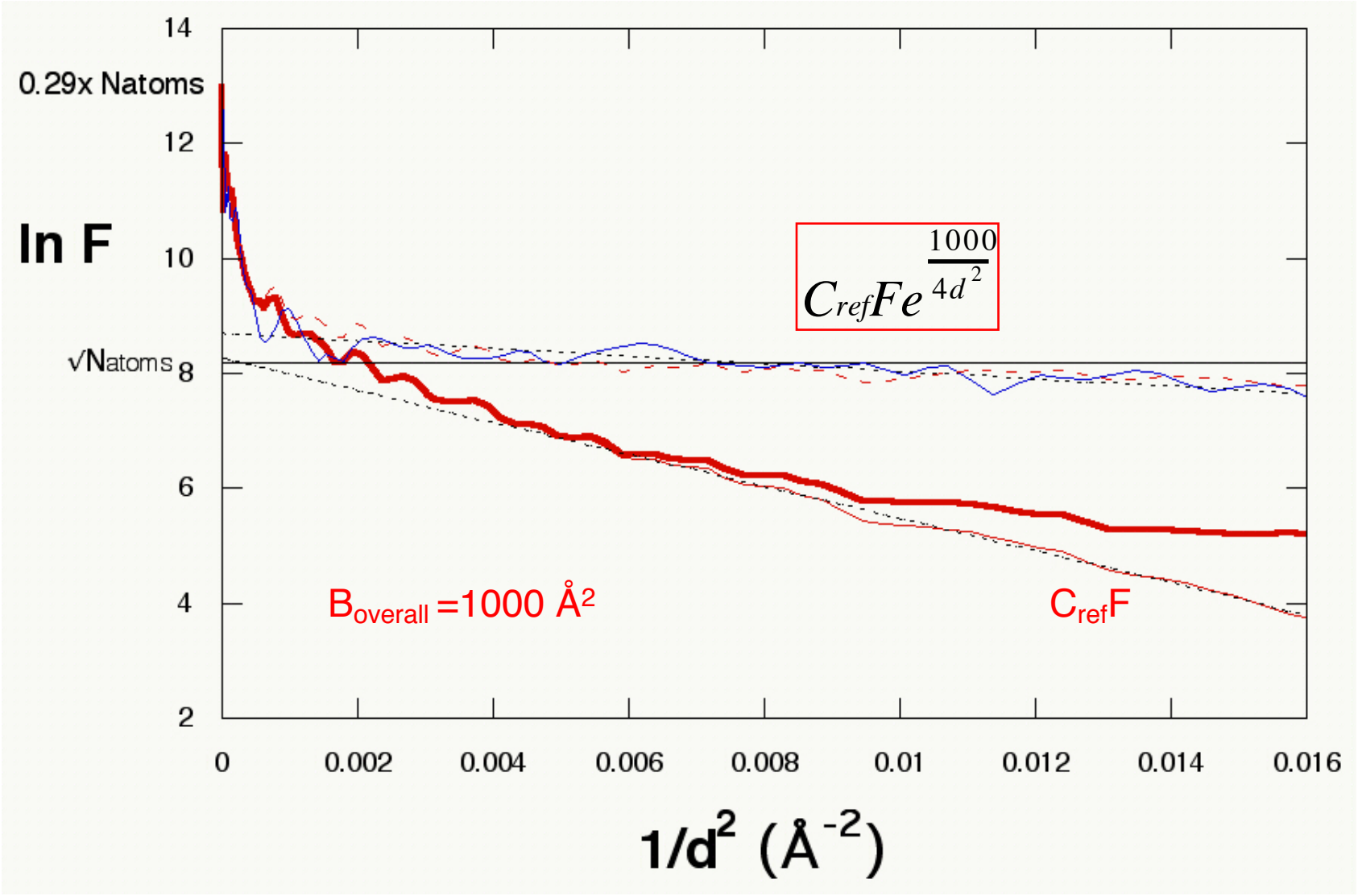
Incorrect Scaling of High and Low Resolution Amplitudes makes map looks featureless

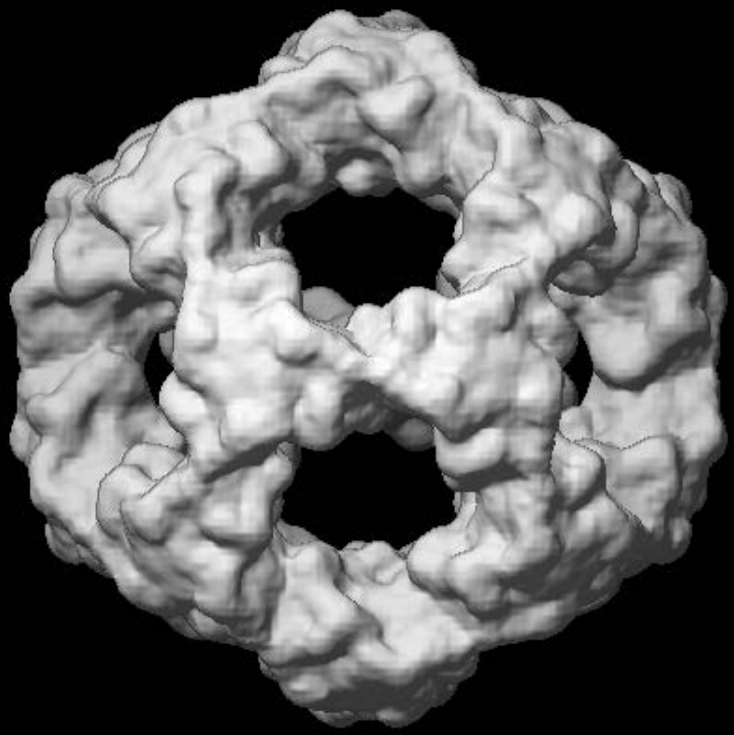
Problem: Application of a negative temperature factor amplifies both signal and noise

Suppress Noise by using
Noise-Weighted Structure Factors $C_{ref}F$
Similar to Figure-of-Merit weighting

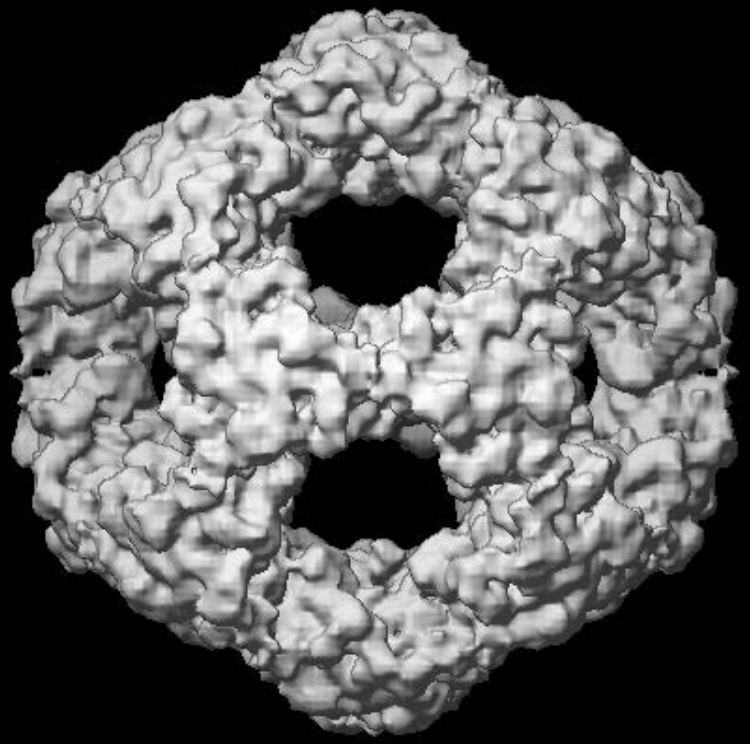
$$C_{ref} = \sqrt{\frac{2FSC}{1+FSC}}$$


1.0
0.5
0.143
C
14
11
8.7
6.5
1/d²
FSC
C_ref
C_ref ~1 Low Res.
C_ref <1 Res. Limit

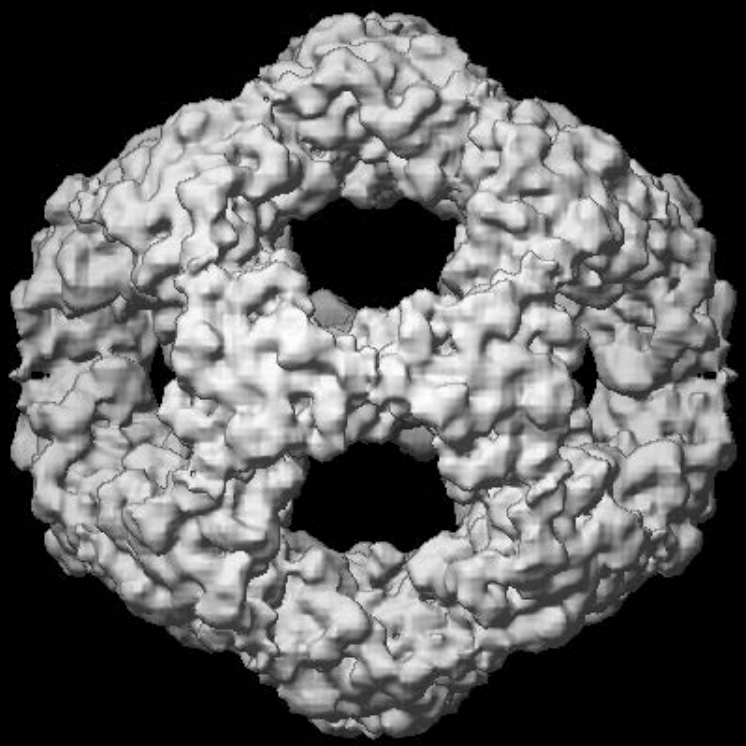

14
0.29x Natoms
12
ln F
10
√Natoms
8
6
4
2
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014
0.016
1/d² (Å⁻²)
$C_{ref}Fe^{\frac{1000}{4d^2}}$
B_overall =1000 Å²
C_ref F

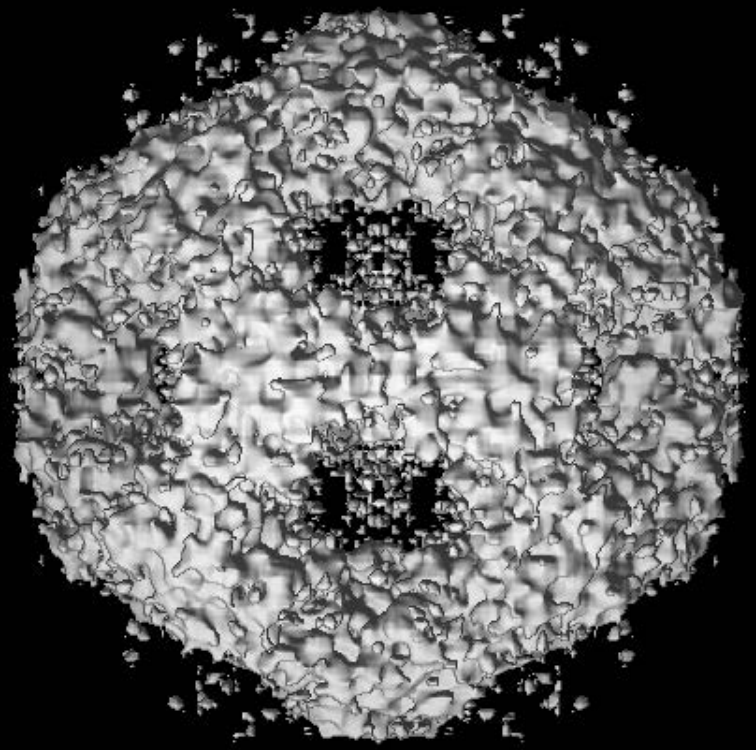


F



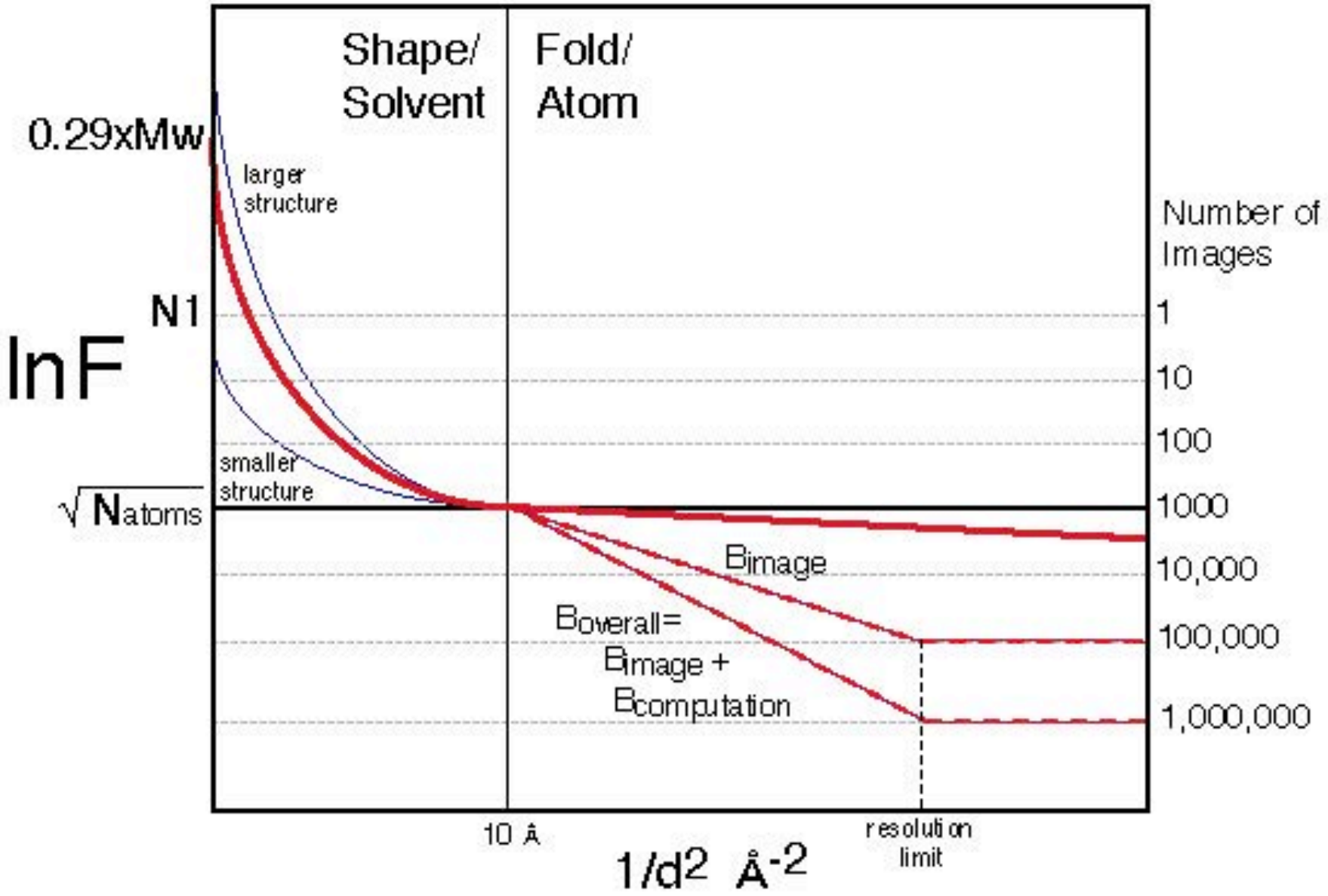
$C_{ref}Fe^{(1000/4d2)}$

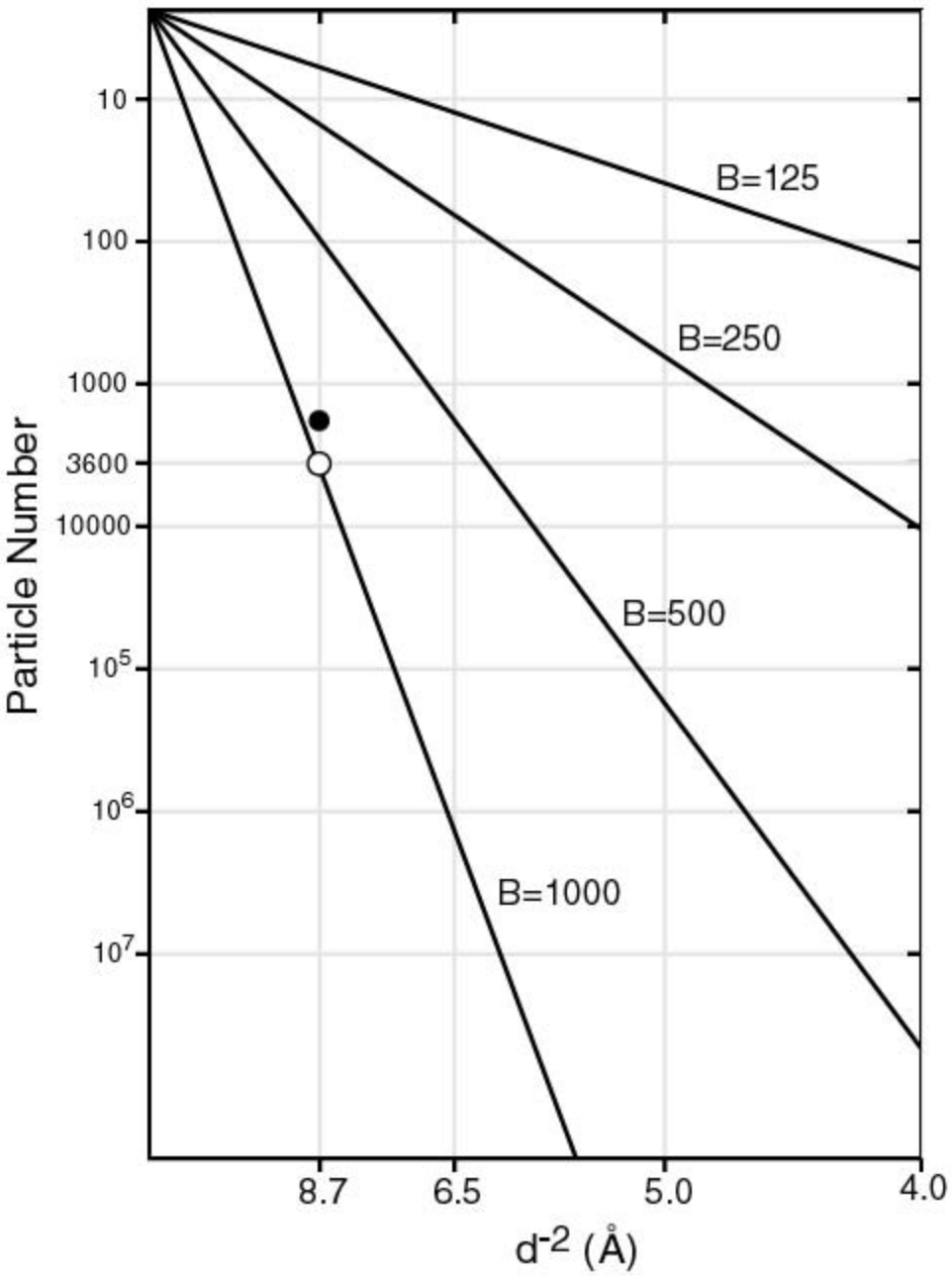




$C_{ref}Fe^{(1000/4d2)}$

$Fe^{(1000/4d2)}$


Shape/
Solvent
Fold/
Atom
0.29xMw
larger
structure
N1
lnF
smaller
structure
√ Natoms
Bimage
Boverall=
Bimage +
Bcomputation
Number of
Images
1
10
100
1000
10,000
100,000
1,000,000
10 Å
resolution
limit
1/d2 Å-2


Particle Number
10
100
1000
3600
10000
10^5
10^6
10^7
B=125
B=250
B=500
B=1000
8.7
6.5
5.0
4.0
d^{-2} (Å)

Summary – Map Validation

- Recommend performing as many tests as possible to communicate the validity of a map. Furthermore, validation tools may be applied to optimizing map calculation or for testing new computational procedures.

- Heterogeneous particles raise several new problems in cryo-EM for which new validation tools may be required. High-resolution maps may be calculated from a small subset of the data (e.g. ~5%). Challenge!