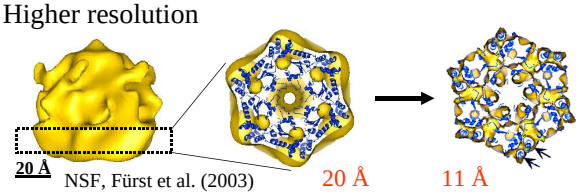
Refinement Strategies for Single Particle Structure Determination

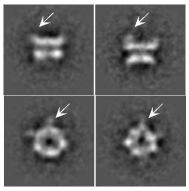
N. Grigorieff

Goals

- Higher resolution



- Sorting of structural heterogeneity



The Prophecy

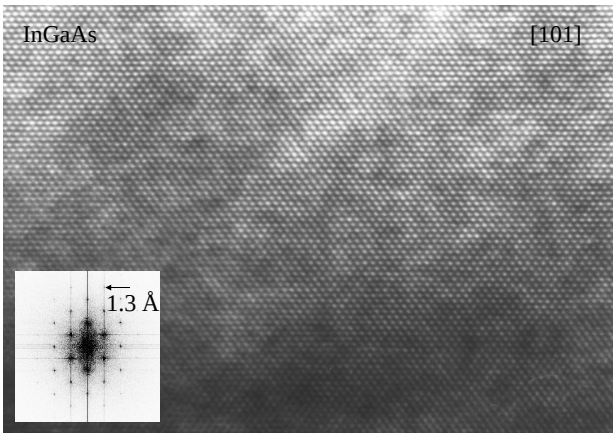
King Richard hath decreed... (QRB, 1995)

- Use 5 e^- per $Å^2$
- Demand a signal-to-noise ratio of 9 or better
- Aim for 3 Å resolution

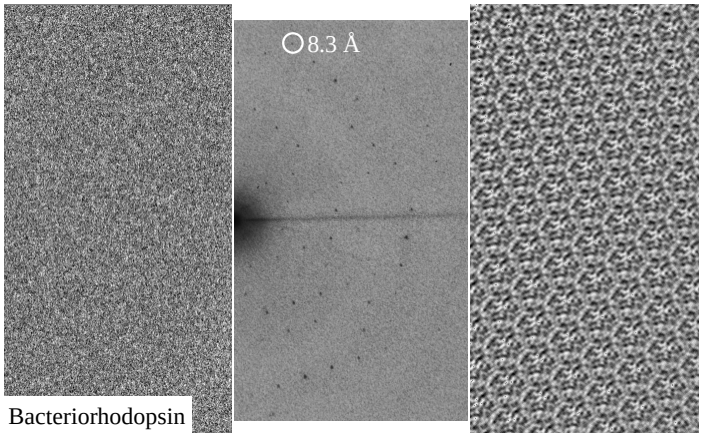
②Thou shall need to image 13,000 molecules

②For 6 Å, thou shall need only 7,000 images

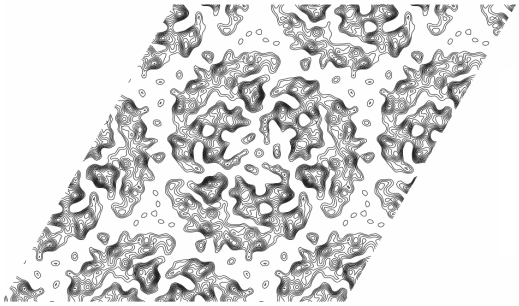
Resolving Power



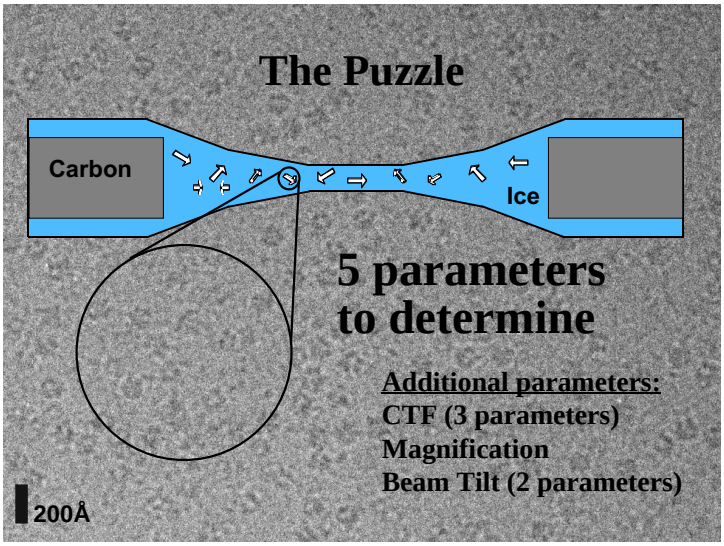
Protein Crystals



Purple Membrane



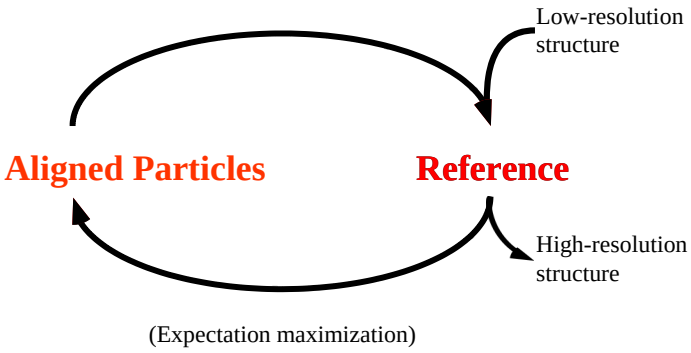
2.6 Å resolution



A Crazy Idea

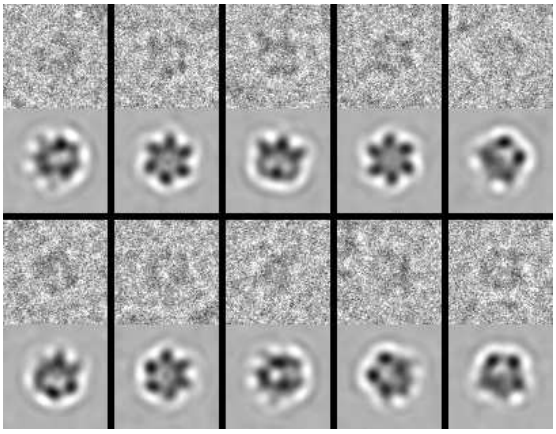
- Assume reliable resolution measure
- Search entire parameter space for highest resolution
- Given enough images, atomic resolution is reached
- Example:
  3 angles, 1 deg step; two coordinates, 1 pixel step:
  $360 \times 360 \times 360 \times 100 \times 100 = 5 \times 10^{11}$
  13000 particles: $(5 \times 10^{11})^{13000}$ structures to search
- This is a big number!

Refinement



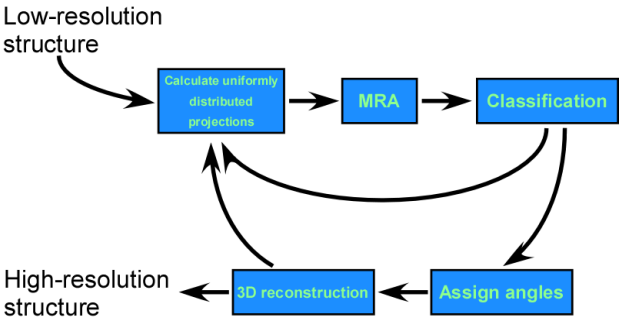
(Expectation maximization)

Strategy 1: Projection Matching



Strategy 2: Alignment in Reciprocal Space

Strategy 3: MRA and Classification



Strategy 4: Maximum Likelihood

Structure for *n*+1 iteration

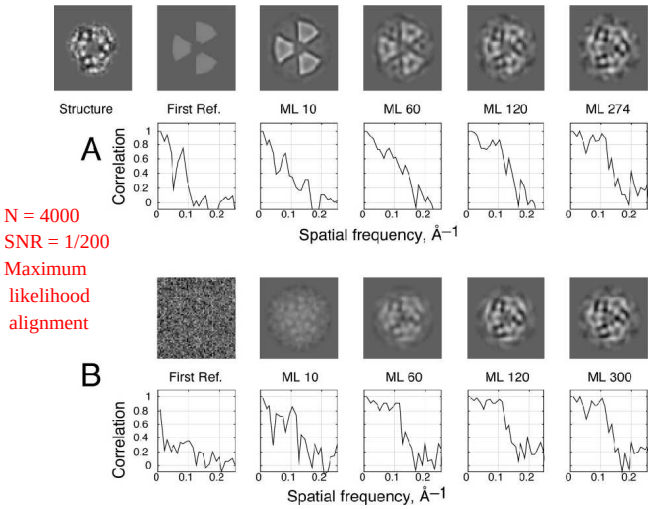
$$A^{(n+1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{\int X_i(\phi)p_i(\phi,\Theta^{(n)})\,d\phi}{\int p_i(\phi,\Theta^{(n)})\,d\phi}$$

Probability function

$$p_i(\phi,\Theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^M \exp\left[-\frac{|X_i(\phi)-A|^2}{2\sigma^2}\right] f(\phi\,|\,\Theta)$$

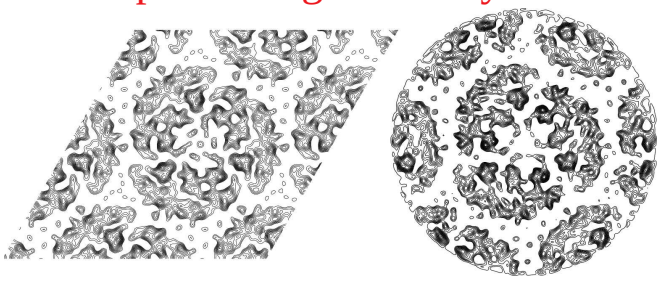
X_i: *i*th image **N: # of images** **ϕ: alignment parameters**
Θ: model parameters **σ: noise in images** **f: positional probab.**

Sigworth (1998), J. Struct. Biol. 122, 328-339



Sigworth (1998), J. Struct. Biol. 122, 328-339

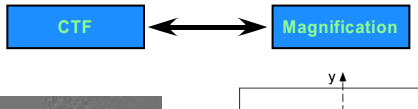
ML processing of 2D crystals

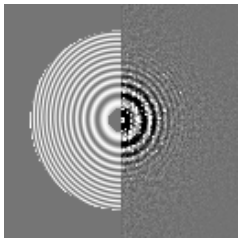


Crystallography

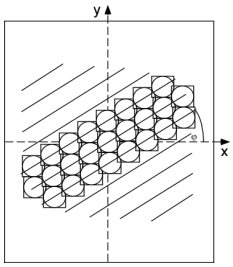
Alignment of individual unit cells using ML approach

Defocus/Astigmatism and Magnification



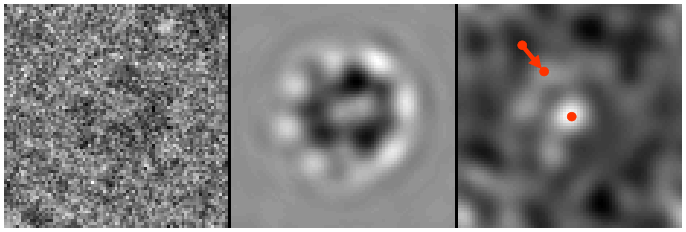


CTFFIND3



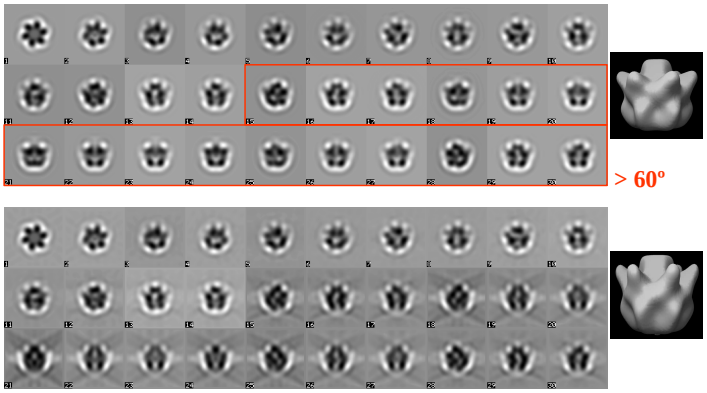
CTFTILT

Problem 1: Local Optima

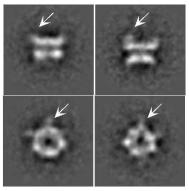


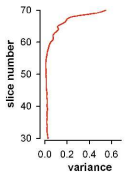
Particle Reference

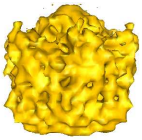
Problem 2: Missing Views



Problem 3: Heterogeneity







- Misalignment of particles
- Lower resolution in disordered regions
- Loss of features

Classification Using ML

Structure for *n*+1 iteration

$$A_k^{(n+1)} = \frac{1}{\sum_i q_i^k(\Theta)} \sum_{i=1}^{N} \frac{\int X_i(\phi) p_i^k(\phi, \Theta^{(n)})\, d\phi}{\sum_k \int p_i^k(\phi, \Theta^{(n)})\, d\phi}$$

Probability function

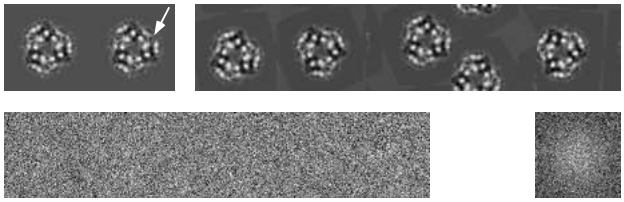
$$p_i^k(\phi, \Theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^M \exp\left[-\frac{|X_i(\phi) - A_k|^2}{2\sigma^2}\right] f(\phi \mid \Theta)$$

Probability for class *k*

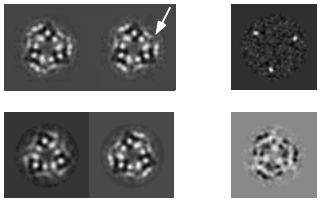
$$q_i^k(\Theta) = \int p_i^k(\phi, \Theta)\, d\phi$$

X_i: *i*th image **N: # of images** **ϕ: alignment parameters**
Θ: model parameters **σ: noise in images** **f: positional probab.**

Classification Using ML



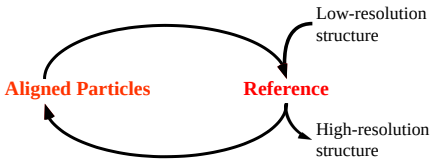
SNR = 1/50
N = 2000



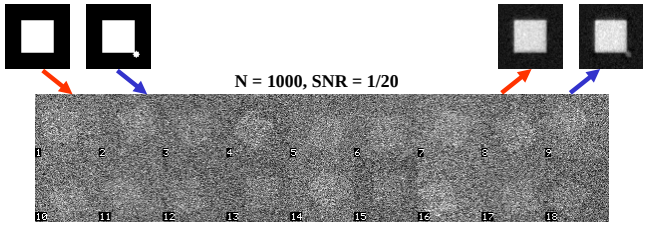
Difference map

Correlation alignment

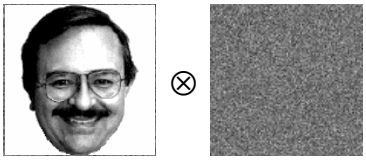
Problem 4: Processing Artifacts



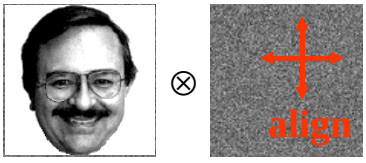
- Interpolation errors
- Masking
- Negative B-factor
- …



Problem 5: Noise Bias



= **0 on average**



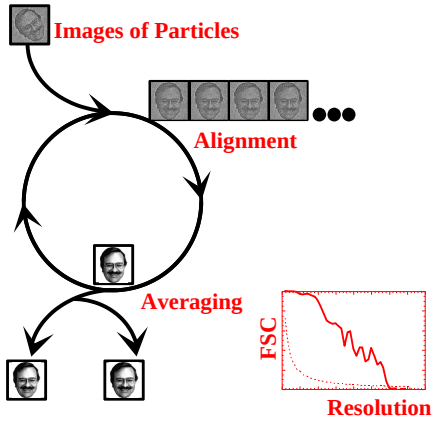
> **0**

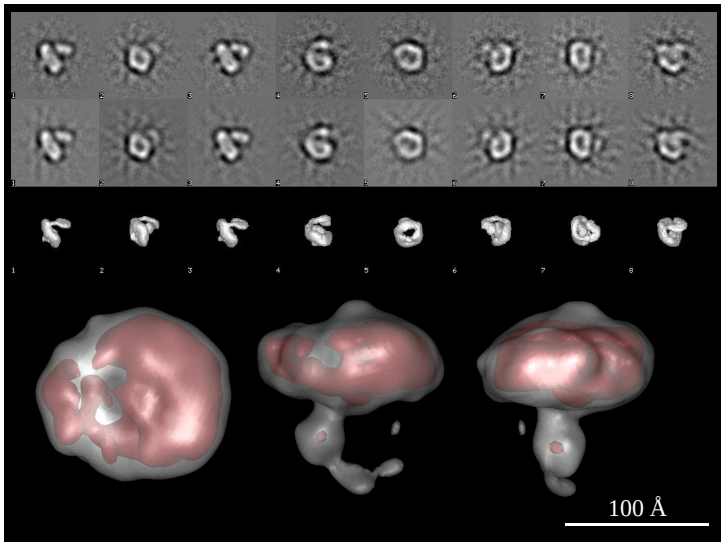
for 64x64 image: average correlation = 0.064

Seeing is NOT Always Believing

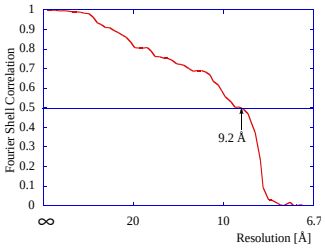
100 Images 1000 Images Reference

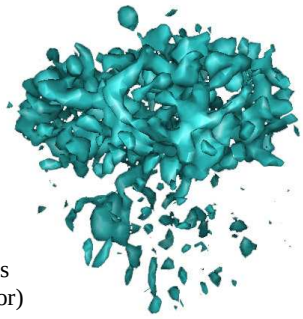
Resolution Measurement





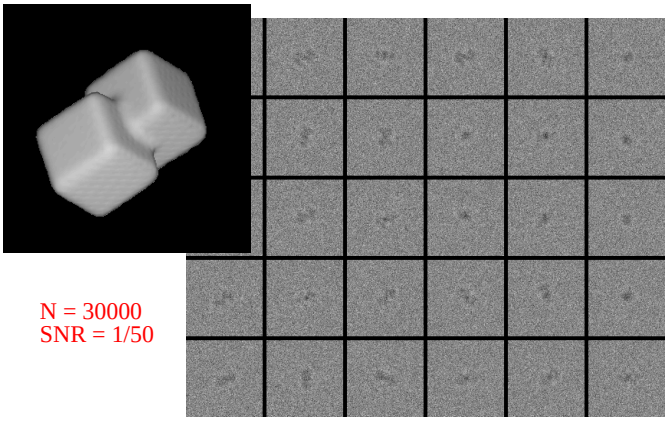
Swiss Cheese





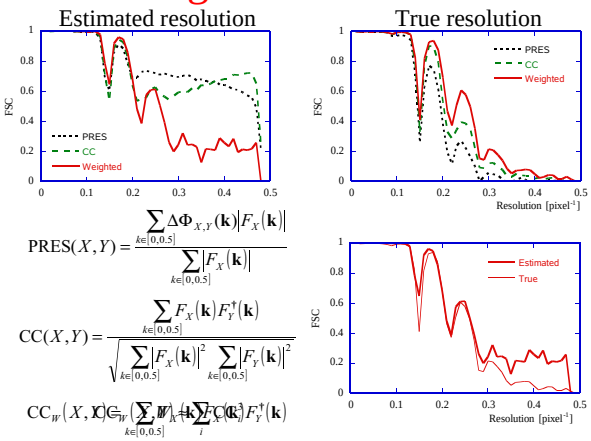
Dangerous:
Boosting of high-resolution terms
(application of a negative B-factor)

Gedanken Experiments



N = 30000
SNR = 1/50

Weighted Correlation

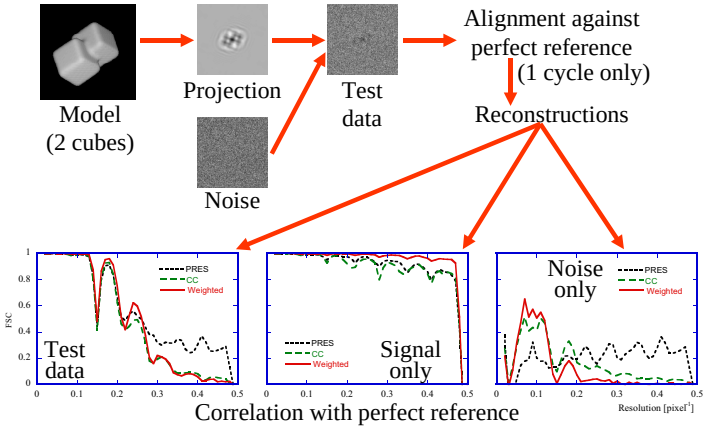


$$\mathrm{PRES}(X,Y)=\frac{\sum_{k\in[0,0.5]}\Delta\Phi_{X,Y}(\mathbf{k})\left|F_X(\mathbf{k})\right|}{\sum_{k\in[0,0.5]}\left|F_X(\mathbf{k})\right|}$$

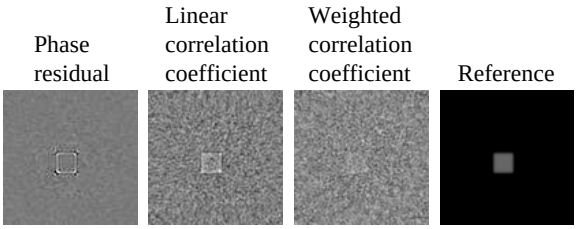
$$\mathrm{CC}(X,Y)=\frac{\sum_{k\in[0,0.5]}F_X(\mathbf{k})F_Y^{*}(\mathbf{k})}{\sqrt{\sum_{k\in[0,0.5]}\left|F_X(\mathbf{k})\right|^2\sum_{k\in[0,0.5]}\left|F_Y(\mathbf{k})\right|^2}}$$

$$\mathrm{CC}_W(X,Y)=\frac{\sum_{k\in[0,0.5]}W(\mathbf{k})F_X(\mathbf{k})F_Y^{*}(\mathbf{k})}{\sqrt{\sum_{k\in[0,0.5]}\left|F_X(\mathbf{k})\right|^2\sum_{k\in[0,0.5]}\left|F_Y(\mathbf{k})\right|^2}}\quad W(\mathbf{k})=\frac{\mathrm{FSC}(\mathbf{k})}{1+\sum_i\mathrm{FSC}(\mathbf{k})}$$

Noise Bias

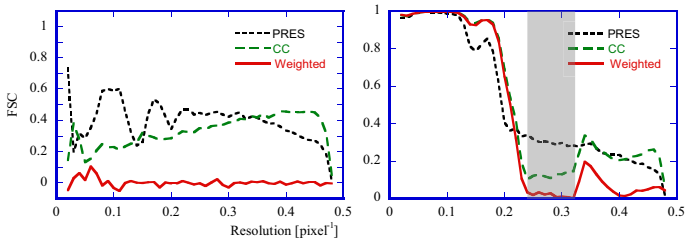


Noise Reconstruction



Coherence Constraint

$$CC_W(X,Y) = \sum_i |CC|_i^3$$



Acknowledgements

- Ca Channel — Matthias Wolf (Glossmann/Striessnig)
- NSF/20S — Johannes Fürst (Axel Brünger)
- Noisy Face — David DeRosier
- **Financial Support:** **HHMI, NIH, NSF**