

# Image Restoration and 3D Reconstruction from Cryo-EM Images

Tejal Bhamre

Final Public Oral Exam, Princeton University  
Advisers: Amit Singer and Joshua Shaevitz

May 9, 2017

# Outline

- 1 Introduction
- 2 Motivation
- 3 Contributions
- 4 Part 1: Covariance Estimation from Noisy Measurements
- 5 Part 2: 3D Homology Modeling
- 6 Appendix

# Introduction

# Cryo-Electron Microscopy (Cryo-EM)

The screenshot shows a news feature from the journal 'nature' (International weekly journal of science). The URL is <http://www.nature.com/nature/journal/v525/n7568/full/7568news.html>. The page title is 'The revolution will not be crystallized: a new method sweeps through structural biology'. Below the title is a sub-headline: 'Move over X-ray crystallography. Cryo-electron microscopy is kicking up a storm by revealing the hidden machinery of the cell.' The author is Ewen Callaway, and the date is 09 September 2015. There are links for PDF and Rights & Permissions.

NATURE | NEWS FEATURE



## The revolution will not be crystallized: a new method sweeps through structural biology

Move over X-ray crystallography. Cryo-electron microscopy is kicking up a storm by revealing the hidden machinery of the cell.

Ewen Callaway

09 September 2015

PDF Rights & Permissions



Illustration by Viktor Koen

Tejal Bhamre

Image Restoration and 3D Reconstruction

# Cryo-EM Revolution

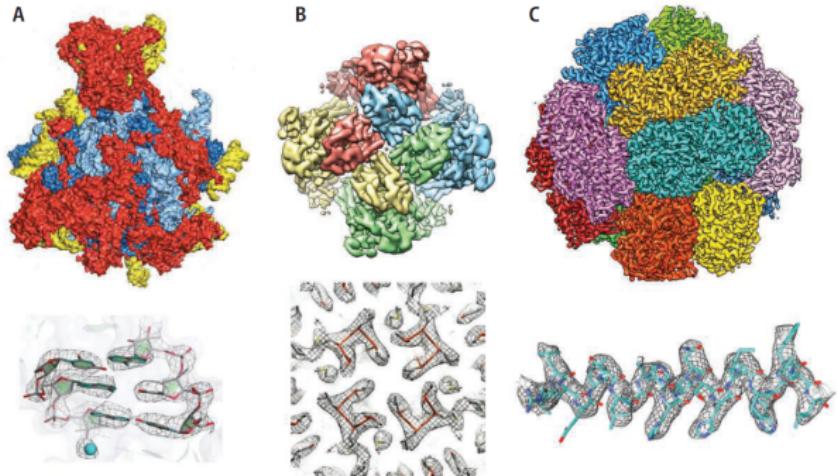
## The Resolution Revolution

Werner Kühlbrandt

Precise knowledge of the structure of macromolecules in the cell is essential for understanding how they function. Structures of large macromolecules can now be obtained at near-atomic resolution by averaging thousands of electron microscope images recorded before radiation damage accumulates. This is what Amunts *et al.* have done in their research article on page 1485 of this issue (1), reporting the structure of the large subunit of the mitochondrial ribosome at 3.2 Å resolution by electron cryo-microscopy (cryo-EM). Together with other recent high-resolution cryo-EM structures (2–4) (see the figure), this achievement heralds the beginning of a new era in molecular biology, where structures at near-atomic resolution are no longer the prerogative of x-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy.

Ribosomes are ancient, massive protein-

Advances in detector technology and image processing are yielding high-resolution electron cryo-microscopy structures of biomolecules.



**Near-atomic resolution with cryo-EM.** (A) The large subunit of the yeast mitochondrial ribosome at 3.2 Å reported by Amunts *et al.* In the detailed view below, the base pairs of an RNA double helix and a magnesium ion (blue) are clearly resolved. (B) TRPV1 ion channel at 3.4 Å (2), with a detailed view of residues lining the

# Cryo-EM Database

**EMBL-EBI**

**EMPIAR** Electron Microscopy Public Image Archive

[EMPIAR home](#) | [Deposition](#) | [REST API](#) | [FAQ](#) | [About EMPIAR](#)

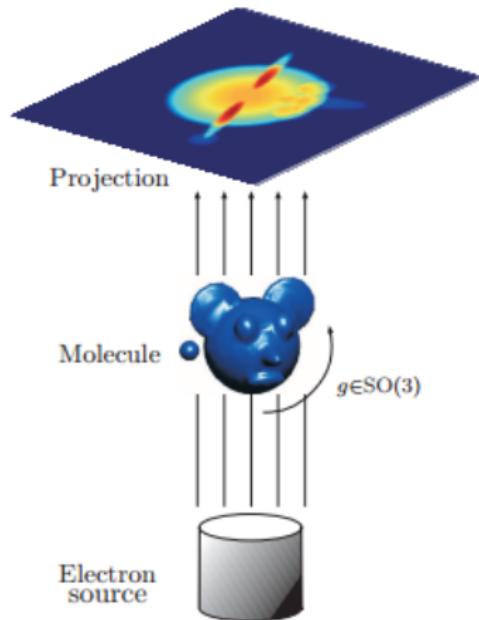
EMPIAR, the Electron Microscopy Public Image Archive, is a public resource for raw, 2D electron microscopy images. Here, you can browse, upload, and download and reprocess the thousands of raw, 2D images used to build a 3D structure. [More ...](#)

[Deposit your data](#) in EMPIAR to share it with the structural biology community.

Browse and [download](#) EMPIAR datasets using the table below.

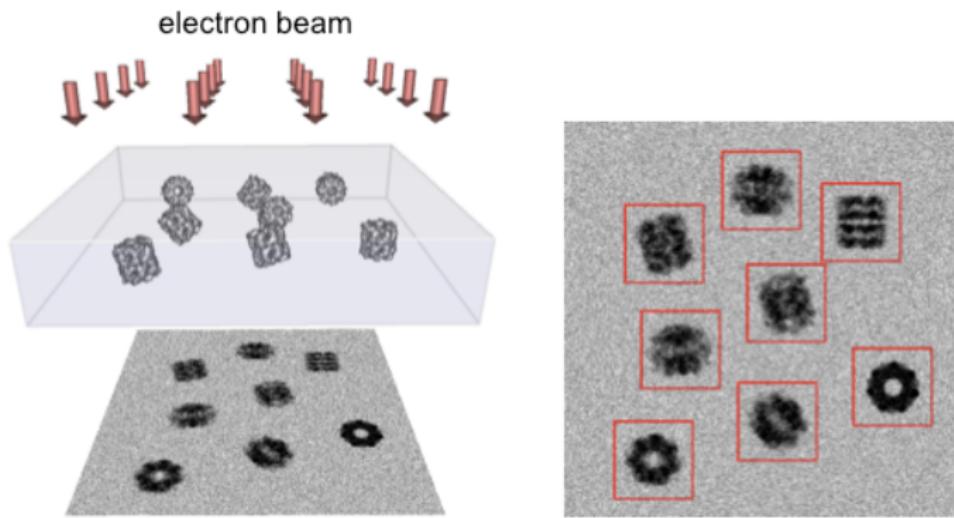
Show <a href="#">50</a> entries	<a href="#">Search:</a>			
Dataset	Title	Authors	Related EMDB/PDB entries	Size
<a href="#">EMPIAR-10087</a> 	Soft X-ray tomography of Plasmodium falciparum infected human erythrocytes stalled in egress by the inhibitors Compound 2 and E64 [ in MRC format]	Hale VL, Saibil HR, Duke E, Fleck RA, Blackman MJ [PubMed: <a href="#">28292906</a> ] [DOI: <a href="#">10.1073/pnas.1619441114</a> ]	<a href="#">EMD-3586</a> , <a href="#">EMD-3587</a> , <a href="#">EMD-3606</a> , <a href="#">EMD-3610</a>	280.6 MB
<a href="#">EMPIAR-10084</a> 	Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate [2261 multi-frame micrographs composed of 40 frames each in TIFF format]	Khoshouei M, Radjainia M, Baumeister W, Danev R [DOI: <a href="#">10.1101/087841</a> ]	<a href="#">EMD-3488</a> , <a href="#">5me2</a>	237.1 GB
<a href="#">EMPIAR-10083</a> 	Bacteriophage P22 mature virion capsid protein [stack of 45150 particles in IMAGIC format]	Hryc CF, Chen D-H, Afonine PV, Jakana J, Wang Z, Haase-Pettingill C, Jiang W, Adams PD, King JA, Schmid MF, Chiu W [PubMed: <a href="#">28270620</a> ] [DOI: <a href="#">10.1073/pnas.1621152114</a> ]	<a href="#">EMD-8606</a> , <a href="#">Suu5</a>	159.4 GB

# Single Particle Reconstruction (SPR)

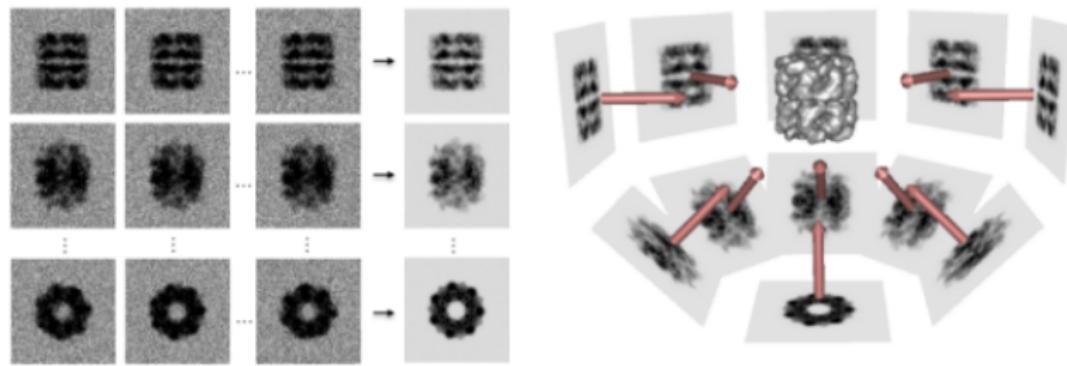


- GOAL: 3D electron density map
- $\sim 10^5 - 10^6$  particles frozen in a thin, vitreous ice layer
- Top view images with EM
- Each particle assumes a random, unknown orientation

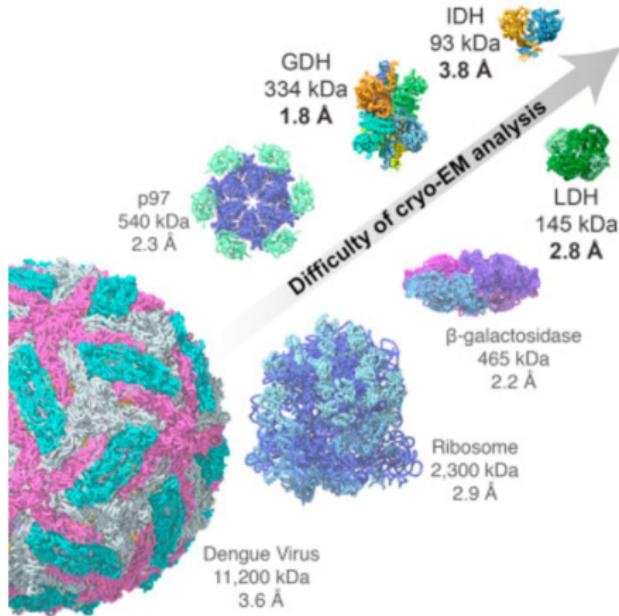
## Single Particle Reconstruction (SPR)



# Single Particle Reconstruction (SPR)

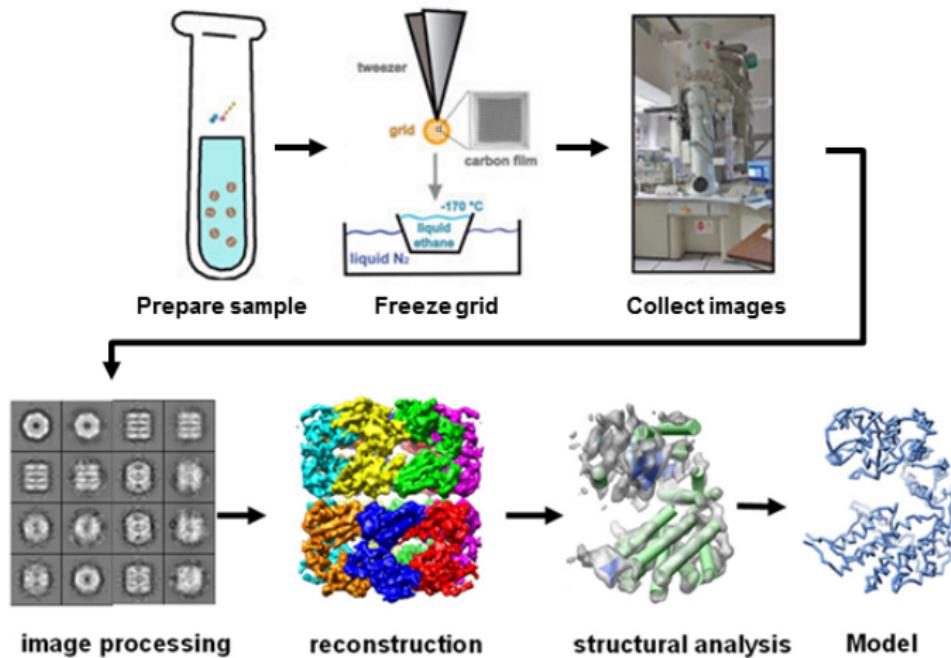


# Cryo-EM Database

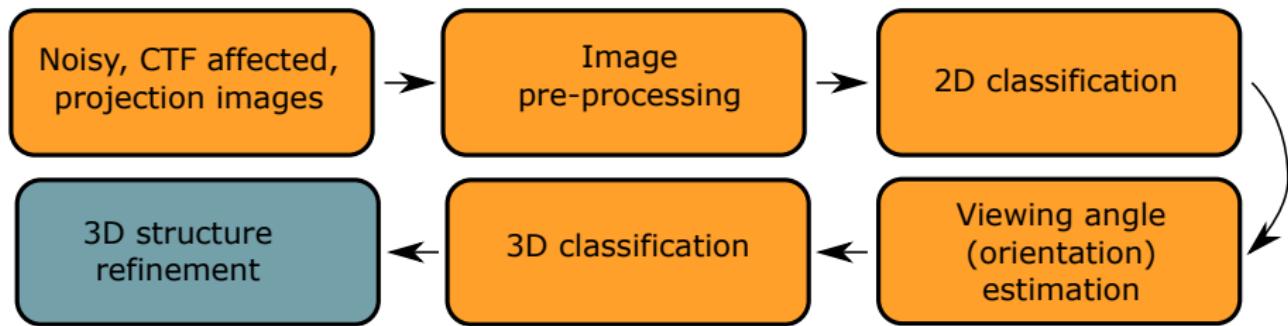


# Motivation

# SPR Pipeline

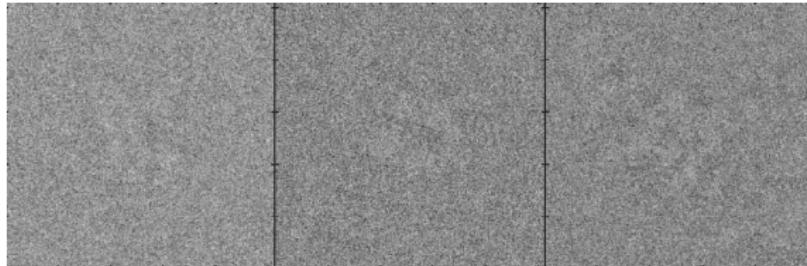


# Existing Computational Pipeline



# Contributions

# Challenges



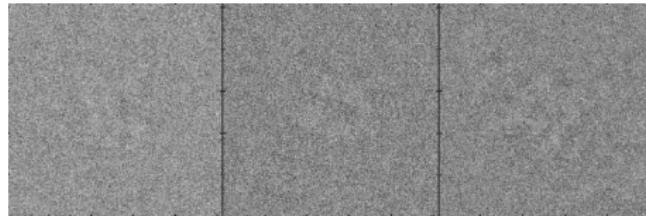
Raw images from an experimental dataset of TRPV1 \*

- Radiation damage: very low signal to noise ratio (SNR)
- Information loss: contrast transfer function (CTF) of microscope
- Estimating unknown orientations of 2D images: challenging non-convex optimization

---

\* M. Liao et al.(2013)

# Challenge: Covariance Estimation



- Existing CTF correction suboptimal
- Need better denoising for preliminary inspection, outlier detection, without class averaging
- Covariance matrix estimate needed for our 3D homology approach

# Challenge: Initial Model for 3D Refinement



- ‘Guess’ low resolution initial model
- 3D refinement sensitive to initial model
- Data driven ab-initio model <sup>a</sup> using common-lines for orientation estimation
- Reliable estimation (common-lines etc.) difficult at low SNR without averaging

---

<sup>a</sup> A. Singer et al. (2011)

S. Dutta et al.(2014)

# Contributions

## ① Covariance estimation from noisy images

- Denoising and ‘optimal’ CTF correction
- Improve 2D classification
- Outlier detection

## ② Homology modeling for 3D reconstruction

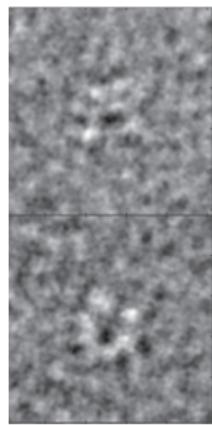
- Reliable orientation estimation difficult at very low SNR
- Need data-driven low resolution initial model
- Use existing structures, **skip 2D classification (averaging) and orientation estimation**

Algorithms validated on both synthetic and real experimental datasets

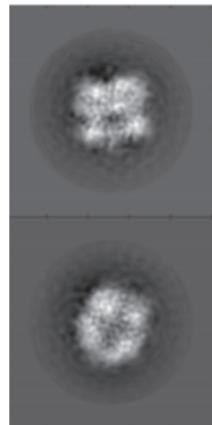
# Denoising: Real TRPV1 dataset



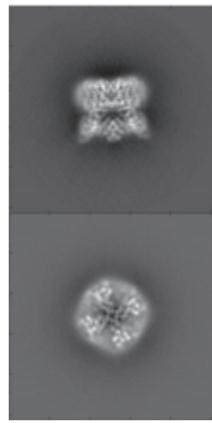
Raw



Existing

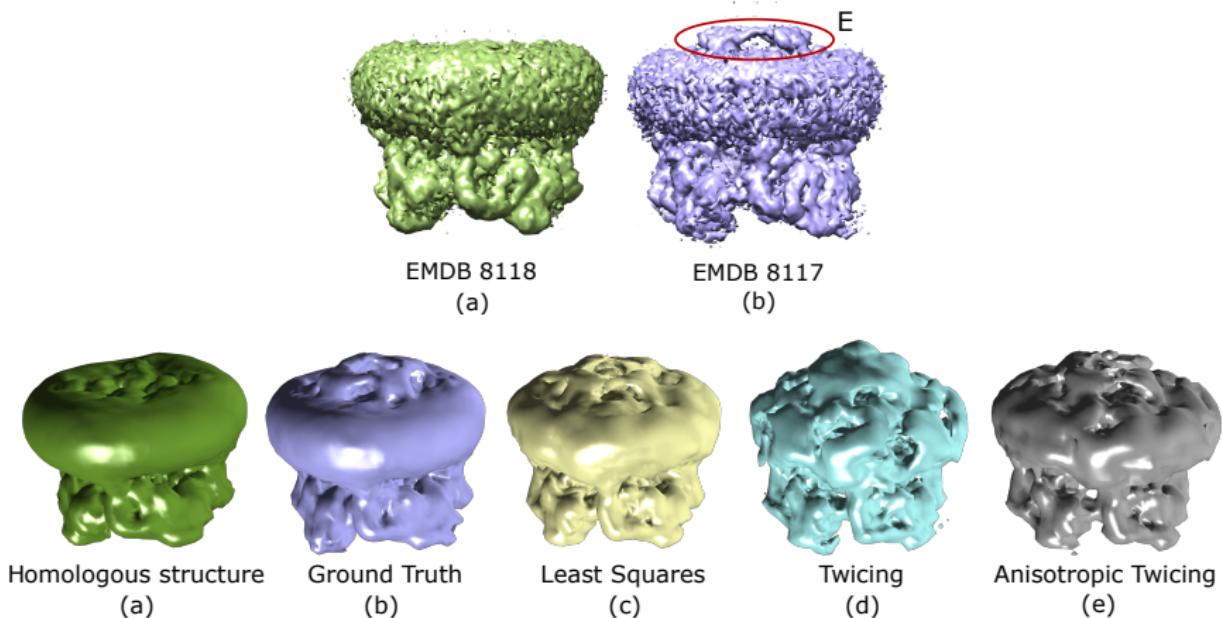


This work



Closest

# Homology modeling: synthetic TRPV1 dataset



# Publications

## Work with Amit Singer, Jane Zhao, Teng Zhang

- *Orthogonal matrix retrieval in cryo-electron microscopy*, T.B., T. Zhang, and A. Singer, 12th IEEE International Symposium on Biomedical Imaging (2015)
- *Denoising and Covariance Estimation of Single Particle Cryo-EM Images*, T.B., T. Zhang, and A. Singer, Journal of Structural Biology (2016)
- *Mahalanobis Distance for Class Averaging of Cryo-EM Image*, T.B., Z. Zhao, and A. Singer, 14th IEEE International Symposium on Biomedical Imaging (2017)
- *Anisotropic Twicing for Single Particle Reconstruction using Autocorrelation Analysis*, T.B., T. Zhang, and A. Singer, submitted (2017).

# Code

Open source software toolbox for cryo-EM: [spr.math.princeton.edu](http://spr.math.princeton.edu)



# ASPIRE

Algorithms for Single Particle Reconstruction

[Home](#)[Methods](#)[Publications](#)[Download](#)[Getting Started](#)[External Links](#)

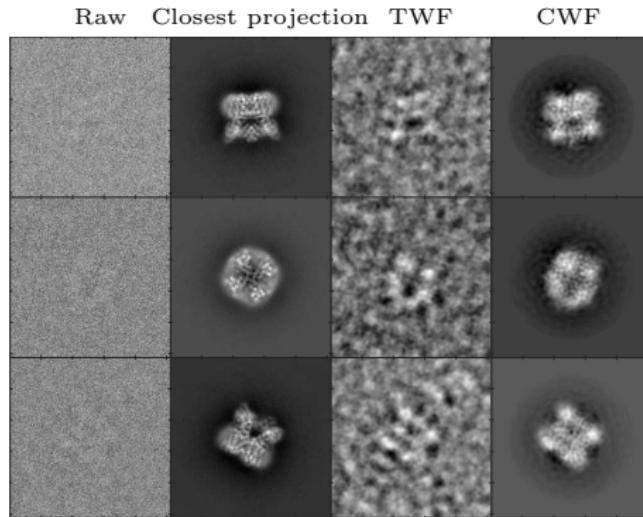
## Algorithms for Single Particle Reconstruction

Download our program here.

[Download](#)

## Part 1: Covariance Estimation from Noisy Measurements

# Experimental data - TRPV1

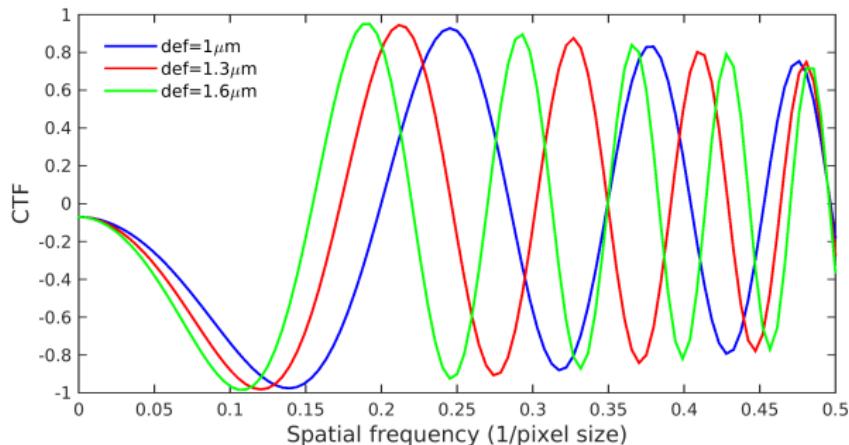


- K2 direct electron detector
- 35645 motion corrected, picked particle images of  $256 \times 256$  pixels

# Motivation

- **Denoising:** visualize underlying particles without class averaging
- **Image restoration:** CTF correction and denoising in a single step
- **Automated outlier detection**
- Bonus: Improved class averaging

# CTF Correction



- Suppresses/loses information and inverts contrast
- Not invertible (zero crossings)
- Information lost from one defocus group could be recovered from another.

# Current Image Restoration Techniques

- **Phase flipping + steerable PCA (sPCA):**
  - Flip sign of the Fourier coefficients at frequencies for which the CTF is negative
  - Preserves noise statistics
  - Data adaptive basis: eigenvectors of the sample covariance matrix
  - Phase flipping corrects only phases
- **Traditional Wiener Filtering (TWF):**
  - Corrects both phases and amplitudes
  - Requires prior estimation of the spectral signal to noise ratio (SSNR)
  - Cannot restore information at zero crossings of the CTF
  - Not in a data adaptive basis (restricted to Fourier basis)

# Covariance Wiener Filtering (CWF)

- Estimate the CTF-corrected covariance matrix of the underlying clean 2D projection images
- Wiener filtering to solve the image restoration deconvolution problem
- No averaging, act on each image separately
- CTF correction and denoising in a single step

# Covariance Wiener Filtering (CWF)

Table : Comparison of CTF Correction/Denoising Methods

Property	Phaseflip + sPCA	TWF	CWF
Applicable at preliminary stage	✓	✓	✓
Data dependent basis	✓	✗	✓
Correct both phases and amplitudes	✗	✓	✓
CTF corrected covariance estimate	✗	✗	✓

# The Model: Real space

Linear, weak phase approximation

$$y_i = a_i * x_i + \epsilon_i, \quad i = 1, 2, \dots, n \quad (1)$$

$n$ : number of images

$*$ : convolution operation

$y_i$ : noisy, CTF filtered  $i$ 'th image in real space

$x_i$ : underlying clean projection image in real space

$a_i$ : the point spread function of the microscope

$\epsilon_i$ : additive Gaussian noise that corrupts the image

# The Model: Fourier space

$$Y_i = A_i X_i + \xi_i, \quad i = 1, 2, \dots, n \quad (2)$$

- $A_i$ : diagonal operator (CTF)
- $X_1, \dots, X_n$ : vectors in  $\mathbb{C}^p$ , ( $p$  is the number of pixels)
- i.i.d. samples from a distribution with mean  $\mathbb{E}[\mathbf{X}] = \mu$  and covariance  $\Sigma = \mathbb{E}[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T]$

“All models are wrong but some are useful” - George Box

# The Model

$$\mathbb{E}[\mathbf{Y}_i] = A_i \mathbb{E}[\mathbf{X}_i], \quad i = 1, 2, \dots, n. \quad (3)$$

$$\begin{aligned}\mathbb{E}[(\mathbf{Y}_i - \mathbb{E}[\mathbf{Y}_i])(\mathbf{Y}_i - \mathbb{E}[\mathbf{Y}_i])^T] &= \mathbb{E}[A_i(\mathbf{X}_i - \mu)(\mathbf{X}_i - \mu)^T A_i^T] + \sigma^2 I \\ &= A_i \Sigma A_i^T + \sigma^2 I.\end{aligned} \quad (4)$$

Relates the second order statistics of the noisy images with the population covariance  $\Sigma$  of the clean images

# Mean Estimation

$$\hat{\mu} = \arg \min_{\mu} \sum_{i=1}^n \|(Y_i - A_i \mu)\|_2^2 + \lambda \|\mu\|_2^2 \quad (5)$$

$$\hat{\mu} = (\sum_{i=1}^n A_i^T A_i + \lambda I)^{-1} (\sum_{i=1}^n A_i^T Y_i). \quad (6)$$

# Covariance Estimation

$$\begin{aligned}\hat{\Sigma} &= \arg \min_{\Sigma} \sum_{i=1}^n \|(Y_i - \mathbb{E}[\mathbf{Y}_i])(Y_i - \mathbb{E}[\mathbf{Y}_i])^T - (A_i \Sigma A_i^T + \sigma^2 I)\|_F^2 \\ &= \arg \min_{\Sigma} \sum_{i=1}^n \|A_i \Sigma A_i^T + \sigma^2 I - C_i\|_F^2\end{aligned}\tag{7}$$

where  $C_i = (Y_i - A_i \mu)(Y_i - A_i \mu)^T$  and  $\|\cdot\|_F$  is the Frobenius matrix norm.

# Solving using Conjugate Gradient

System of linear equations for the elements of the matrix  $\hat{\Sigma}$

$$\sum_{i=1}^n A_i^T A_i \hat{\Sigma} A_i^T A_i = \sum_{i=1}^n A_i^T C_i A_i - \sum_{i=1}^n \sigma^2 A_i^T A_i \quad (8)$$

$$L(\hat{\Sigma}) = B \quad (9)$$

where  $L : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^{p \times p}$  is the linear operator acting on  $\hat{\Sigma}$  defined by the left hand side of eqn. 8, and  $B$  is the right hand side.

- Direct inversion of this linear system is slow for large image sizes
- Applying  $L$  only involves matrix multiplications: fast!
- Conjugate gradient

# Eigenvalue Thresholding

- $L(\hat{\Sigma})$  is a PSD matrix whenever  $\hat{\Sigma}$  is PSD (as a sum of PSD matrices)
- $B$  may not necessarily be PSD due to finite sample fluctuations
- Project  $B$  onto the cone of PSD matrices
- Compute the spectral decomposition of  $B$  and set all negative eigenvalues to 0 (eigenvalue thresholding)

# Eigenvalue Shrinkage: Spiked Covariance Model

- Eigenvalues corresponding to the signal can only be detected if they reside outside of the support of the Marčenko Pastur (MP) distribution
- Kritchman Nadler (KN) \* rank estimation to determine the number of eigenvalues corresponding to the signal
- Apply operator norm eigenvalue shrinkage procedure \*\* to those eigenvalues, while setting all other eigenvalues to 0

---

\* Kritchman and Nadler (2008)

\*\* Donoho et al.(2013)

# Deconvolution by Wiener Filtering

- White noise: estimate  $X_i$  as

$$\hat{X}_i = (I - H_i A_i) \hat{\mu} + H_i Y_i \quad (10)$$

where  $H_i = \hat{\Sigma} A_i^T (A_i \hat{\Sigma} A_i^T + \sigma^2 I)^{-1}$  is the linear Wiener filter

- Colored noise: estimate  $X_i$  as

$$\hat{X}_i = (I - H_i W A_i) \hat{\mu} + H_i Y_i \quad (11)$$

with  $H_i = \hat{\Sigma} A_i^T W^T (W A_i \hat{\Sigma} A_i^T W^T + \sigma^2 I)^{-1}$

# Fourier-Bessel Steerable Basis

- The population covariance matrix  $\Sigma$  must be invariant under in-plane rotation of the projection images
- Block diagonal in any steerable basis in which the basis elements are outer products of radial functions and angular Fourier modes
- Suffices to estimate each diagonal block of  $\Sigma$ , corresponding to the angular frequency  $k$ , separately
- **Nearly unitary transformation**

# Computational Complexity

$O(TDL^4 + nL^3)$ , where  $T$  is the number of conjugate gradient iterations

- $D$  defocus groups with  $d_i$  images in group  $i$
- Images of size  $L \times L$
- $n$  images

# Computational Complexity: Timings

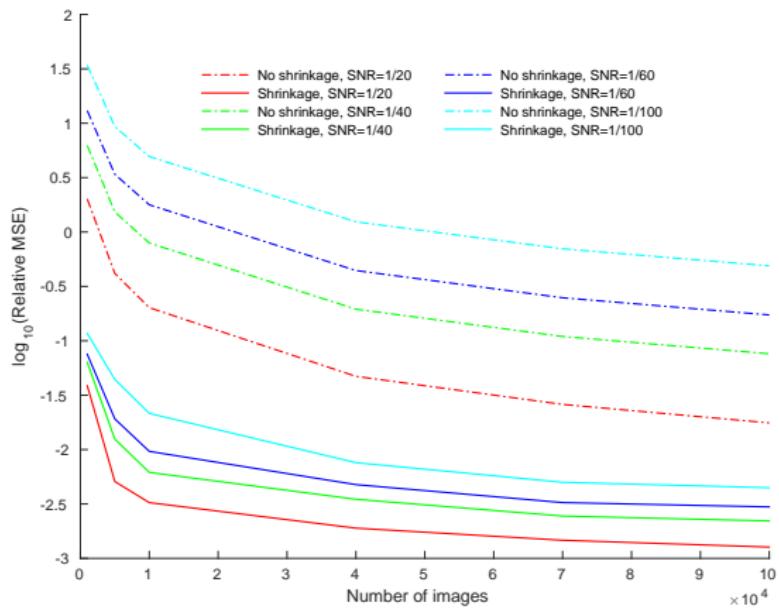
$n$  images of size  $L \times L$

UNIX environment with 60 cores, running at 2.3 GHz, with total RAM of 1.5TB

Table : Timing in seconds

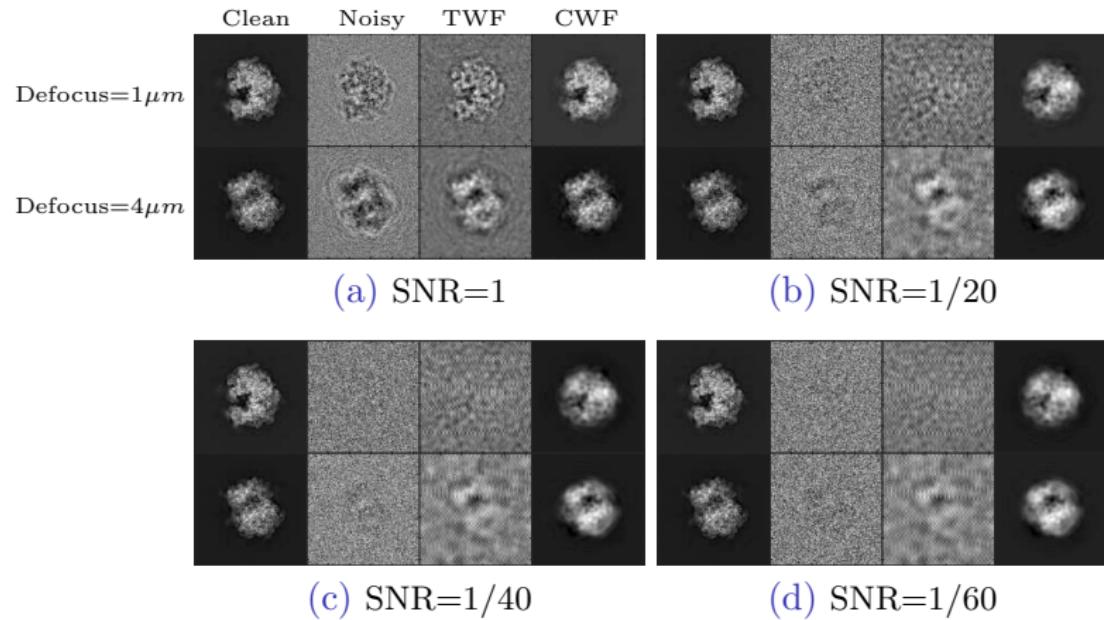
Dataset	$L$	$n$	Basis coeffs	CWF
TRPV1	256	35645	312s	574s
80s	360	30000	731s	385s
IP3R1	256	37382	429s	589s
70s	250	99979	1174s	113s

# Relative error of estimated covariance



The estimator  $\hat{\Sigma}$  can be shown to be consistent in the large sample limit  
 $n \rightarrow \infty$

## Simulations with white noise: 80S ribosome (EMDB-6454)



# Outlier Detection

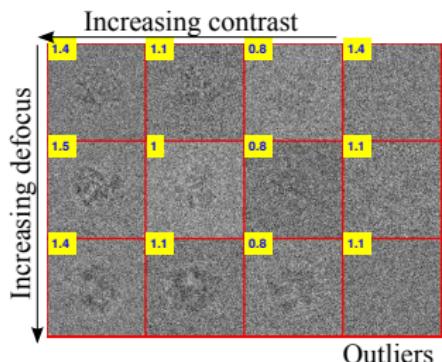
- Current method: manual visual inspection after particle picking
- CWF: automatic way to classify picked particles
- Specimen particles at various depths in the ice layer: acquired projection images can have different contrasts
- 

$$Y_i = \alpha_i A_i X_i + \xi_i, \quad i = 1, 2, \dots, n \quad (12)$$

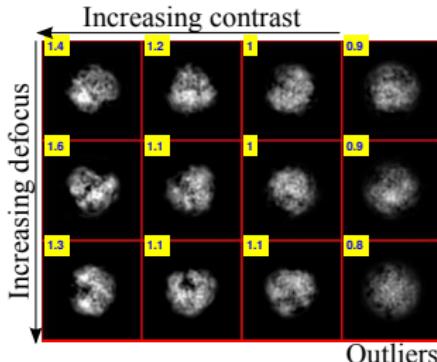
- Absorb  $\alpha$  into  $\mathbf{X}$  and estimate  $\alpha_i X_i$
- Outlier images typically have low contrast after denoising: linear classifier after CWF

# Outlier Detection: 80S ribosome (EMDB-6454)

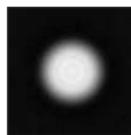
SNR=1/20  $\alpha \in [0.75, 1.5]$  10% images are pure noise



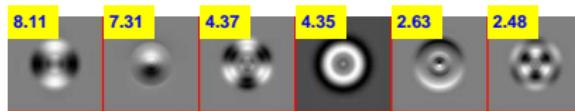
(e)



(f)

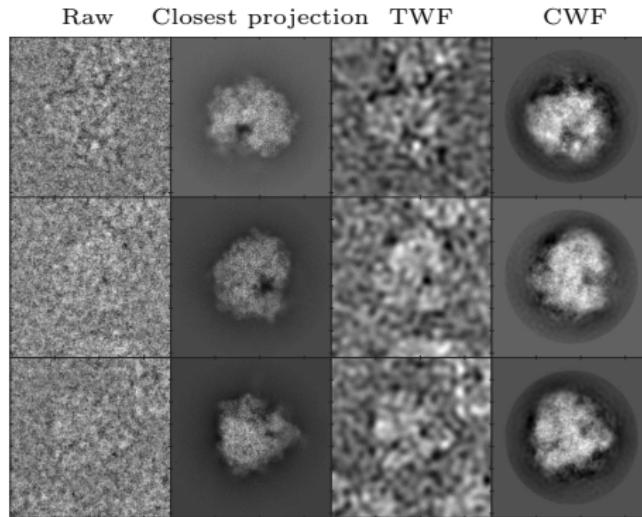


(g)



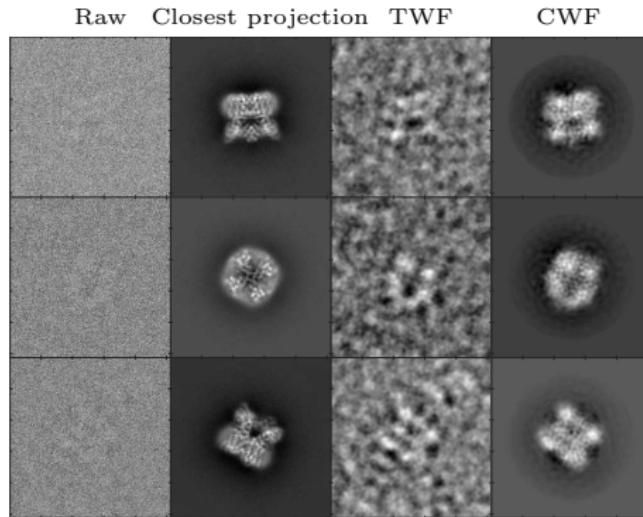
(h)

# Experimental data - 80S ribosome



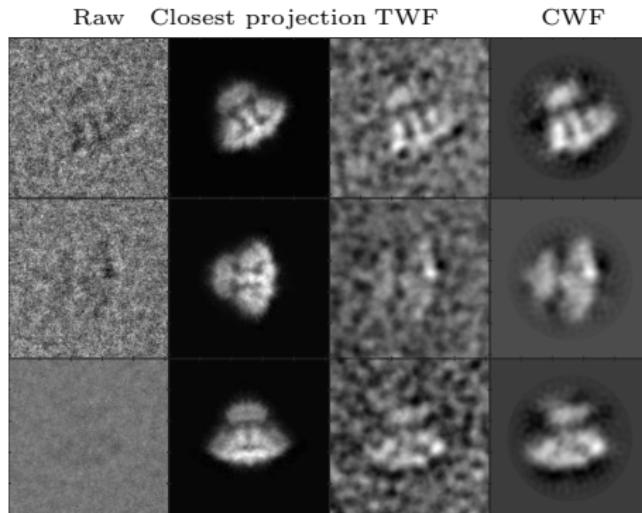
- FALCON II  $4k \times 4k$  direct electron detector
- 105247 motion corrected, picked particle images of  $360 \times 360$  pixels

# Experimental data - TRPV1



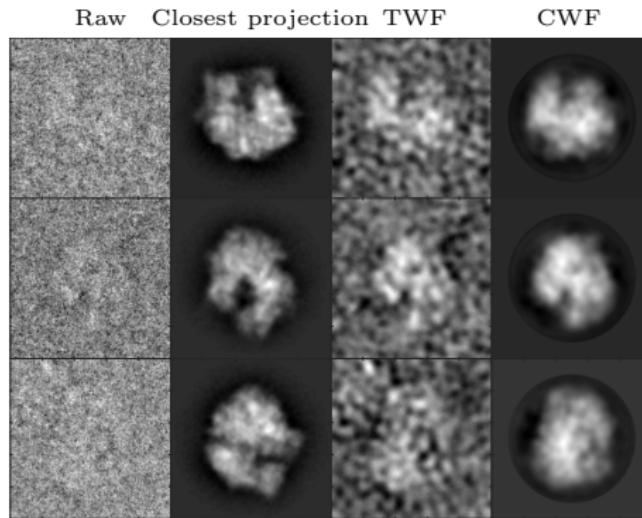
- K2 direct electron detector
- 35645 motion corrected, picked particle images of  $256 \times 256$  pixels

# Experimental data -IP<sub>3</sub>R1



- Gatan 4k×4k CCD
- 37382 picked particle images of 256×256 pixels

# Experimental data - 70S ribosome



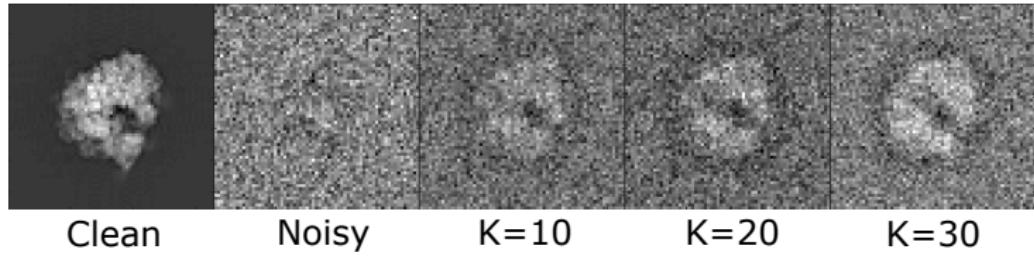
- TVIPS TEMCAM-F415 (4k x 4k) CCD
- 216517 picked particle images of  $250 \times 250$  pixels

# Application: Mahalanobis Affinity

Mahalanobis affinity or likelihood for the underlying clean images  $x_i$  and  $x_j$  to originate from the same viewing direction

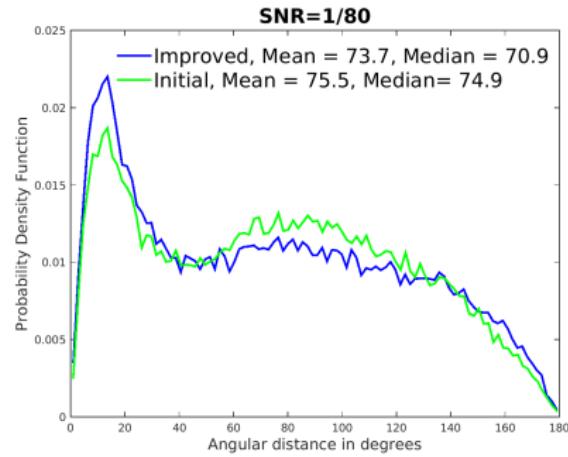
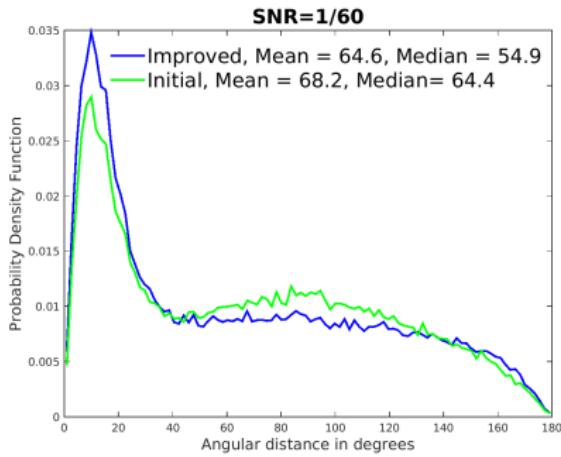
$$\Pr(||X_{ij}||_p < \epsilon | Y_i = y_i, Y_j = y_j) \\ = \frac{\epsilon^d \text{Vol}(B_p(0, 1))}{(2\pi)^{\frac{d}{2}} |L_i + L_j|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2} \alpha_{ij}^T (L_i + L_j)^{-1} \alpha_{ij}\right\} + \mathcal{O}(\epsilon^{d+1})$$

# Application: Mahalanobis Affinity



Class averages with the improved algorithm using the anisotropic affinity, using  $K = 10, 20, 30$ .

# Application: Mahalanobis Affinity



Improved class averaging using Mahalanobis affinity to detect nearest neighbors

## Part 2: 3D Homology Modeling

# Existing Methods: Ab Initio Modeling



- **Random conical tilt** (Radermacher et al.): tilting to acquire two sets of micrographs
- **Method of moments** (Goncharov et al., Salzman, et al.): sensitive to errors in data
- **Common lines approach** (Singer et al.): needs class averaging

# Motivation: Molecular Replacement (MR)

- Missing phase problem in X-ray crystallography
- Use previously solved homologous structure
- Fourier magnitudes from unknown structure, phases from homologous structure

# A Tail of Two Cats

- ‘Twicing’ for magnitude correction
- $2\hat{\mathbf{A}}_{LS} - \mathbf{B}$  for  $\mathbf{A} \in \mathbb{C}^{1 \times 1}$  is an unbiased estimator



---

J. Tukey (1977)

P. Main

K. Cowtan (2014)

# New approach for homology modeling

- Two new algorithms to predict structure **directly from raw images (no averaging)**
- Use previously solved homologous structure
- Use Kam's autocorrelation analysis

# Kam's Autocorrelation Analysis

$$\mathbf{C}_l = \mathbf{A}_l \mathbf{A}_l^* \quad l = 0, 1, \dots \quad (13)$$

- $\mathbf{C}_l$ : autocorrelation matrix over  $SO(3)$  (“magnitude”)
- $bA_l$ : Expansion coefficients of 3D Fourier volume
- $\mathbf{C}_l$  can be computed from the covariance matrix  $\Sigma$  of the 2D projection images
- Requirement: Uniformly distributed viewing angles over the sphere

# Outline of Our Approach

$$\mathbf{C}_l = \mathbf{A}_l \mathbf{A}_l^* \quad l = 0, 1, \dots \quad (14)$$

- $\mathbf{C}_l$  can be computed from the covariance matrix  $\Sigma$
- Use estimated  $\hat{\Sigma}$  from Covariance Wiener Filtering to compute  $\mathbf{C}_l$  (“magnitude”)
- Determines  $\mathbf{A}_l$  upto an orthogonal matrix (“missing phase”)

# Orthogonal Matrix Retrieval Problem

- $\Phi_A : \mathbb{R}^3 \rightarrow \mathbb{R}$ : electron scattering density of the unknown structure.
- $\mathcal{F}(\Phi_A) : \mathbb{R}^3 \rightarrow \mathbb{C}$ : 3D Fourier transform

$$\mathcal{F}(\Phi_A)(k, \theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l A_{lm}(k) Y_l^m(\theta, \varphi)$$

- $k$ : radial frequency
- $Y_l^m$ : real spherical harmonics

# Kam's Autocorrelation Analysis

$$\mathcal{F}(\Phi_A)(k, \theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l A_{lm}(k) Y_l^m(\theta, \varphi)$$

$$C_l(k_1, k_2) = \sum_{m=-l}^l A_{lm}(k_1) \overline{A_{lm}(k_2)}, \quad l = 0, 1, \dots$$

- $C_l$ : autocorrelation matrix over  $SO(3)$
- $C_l$  can be estimated from the covariance matrix  $\Sigma$  of the 2D projection images
- Requirement: Uniformly distributed viewing angles over the sphere

# Resolution Limit

- Finite pixel grid: Nyquist criterion

$$\mathcal{F}(\Phi_A)(k, \theta, \varphi) = \sum_{l=0}^L \sum_{m=-l}^l A_{lm}(k) Y_l^m(\theta, \varphi), \quad l = 0, 1, \dots, L$$
$$A_{lm}(k) = \sum_{s=1}^{S_l} a_{lms} j_{ls}(k). \quad (15)$$

# Resolution Limit

Normalized spherical Bessel functions

$$j_{ls}(k) = \frac{1}{c\sqrt{\pi}|j_{l+1}(R_{l,s})|} j_l(R_{l,s} \frac{k}{c}), \quad 0 < k < c, \quad s = 1, 2, \dots, S_l$$

- $c$ : bandlimit of the images
- $R_{l,s}$ :  $s$ 'th positive root of the equation  $j_l(x) = 0$
- $S_l$ : largest integer  $s$  that satisfies the Nyquist criterion

$$R_{l,(s+1)} \leq 2\pi c R$$

- $L$ : largest integer  $l$  for which  $S_l$  is at least 1.

# Orthogonal Matrix Retrieval Problem

$$\mathbf{C}_l = \mathbf{A}_l \mathbf{A}_l^*$$

- $\mathbf{A}_l$  is a matrix of size  $S_l \times (2l + 1)$
- $\mathbf{A}_l$  known up to an orthogonal matrix  $\mathbf{O}_l \in O(2l + 1)$  (Cholesky decomposition of  $\mathbf{C}_l$ )
- Recover missing orthogonal matrices → expansion coefficients → 3D structure

# Orthogonal Extension (OE)

- Determine the coefficient matrices  $\mathbf{A}_l$
- Known, homologous structure  $\Phi_B$

$$\mathcal{F}(\Phi_B)(k, \theta, \varphi) = \sum_{l=0}^{L_B} \sum_{m=-l}^l B_{lm}(k) Y_l^m(\theta, \varphi) \quad (16)$$

- $\mathbf{F}_l$ : any matrix of size  $S_l \times 2l + 1$  satisfying  $\mathbf{C}_l = \mathbf{F}_l \mathbf{F}_l^*$

$$\mathbf{A}_l = \mathbf{F}_l \mathbf{O}_l \quad (17)$$

# Orthogonal Extension (OE)

- Determine the coefficient matrices  $\mathbf{A}_l$
- Known, homologous structure  $\Phi_B$

$$\mathcal{F}(\Phi_B)(k, \theta, \varphi) = \sum_{l=0}^{L_B} \sum_{m=-l}^l B_{lm}(k) Y_l^m(\theta, \varphi) \quad (18)$$

- $\mathbf{F}_l$ : any matrix of size  $S_l \times 2l + 1$  satisfying  $\mathbf{C}_l = \mathbf{F}_l \mathbf{F}_l^*$

$$\mathbf{A}_l = \mathbf{F}_l \mathbf{O}_l \quad (19)$$

# Orthogonal Extension (OE)

- $\mathbf{A}_l \approx \mathbf{B}_l$  (homologous assumption)

$$\mathbf{O}_l = \underset{\mathbf{O} \in \mathrm{O}(2l+1)}{\arg \min} \|\mathbf{F}_l \mathbf{O} - \mathbf{B}_l\|_F^2$$

- Closed form solution via singular value decomposition (SVD)

$$\mathbf{B}_l^* \mathbf{F}_l = \mathbf{U}_l \boldsymbol{\Sigma}_l \mathbf{V}_l^T$$

$$\mathbf{O}_l = \mathbf{V}_l \mathbf{U}_l^T$$

---

J. Keller (1975)

TB, T. Zhang, A. Singer (2015)

# OE with Least Squares (OE-LS)

- Least squares estimator

$$\hat{\mathbf{A}}_{l, \text{LS}} = \mathbf{F}_l \mathbf{V}_l \mathbf{U}_l^T \quad (20)$$

- Twicing in practice:  $2\hat{\mathbf{A}}_{l, \text{LS}} - \mathbf{B}_l$

# Orthogonal Replacement (OR)

- Known, homologous structure might not exist
- Difference between two structures  $\Phi_A$  and  $\Phi_B$  might be known e.g. antibody fragment binding to a protein
- Known structure  $\Delta\Phi = \Phi_B - \Phi_A$ , cryo-EM images of  $\Phi_A$ ,  $\Phi_A$

$$A_l^{(B)} - A_l^{(A)} = F_l^{(B)} O_l^{(B)} - F_l^{(B)} O_l^{(A)}$$

- $F_l^{(B)}$  and  $F_l^{(A)}$  computed from the autocorrelation matrix (using 2D covariance matrix from cryo-EM images)

# Orthogonal Replacement (OR)

- Known, homologous structure might not exist
- Difference between two structures  $\Phi_A$  and  $\Phi_B$  might be known e.g. antibody fragment binding to a protein
- Known structure  $\Delta\Phi = \Phi_B - \Phi_A$ , cryo-EM images of  $\Phi_A$ ,  $\Phi_A$

$$A_l^{(B)} - A_l^{(A)} = F_l^{(B)} O_l^{(B)} - F_l^{(B)} O_l^{(A)}$$

- $F_l^{(B)}$  and  $F_l^{(A)}$  computed from the autocorrelation matrix (using 2D covariance matrix from cryo-EM images)

# Relaxation to a Semidefinite Program (SDP)

$$\min_{O_l^{(1)}, O_l^{(2)} \in \mathcal{O}(2l+1)} \|A_l^{(2)} - A_l^{(1)} - F_l^{(2)} O_l^{(2)} + F_l^{(1)} O_l^{(1)}\|_F^2$$

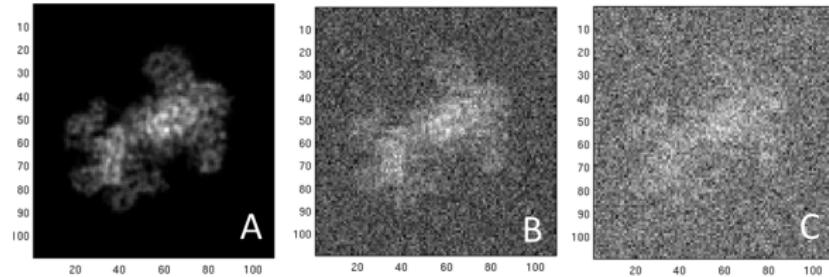
can be written as

$$\min_Q \text{Tr}(WQ)$$

- $Q \in \mathbb{R}^{3(2l+1) \times 3(2l+1)}$
- subject to  $Q_{ii} = I$ ,  $\text{rank}(Q) = 2l + 1$  and  $Q \succeq 0$ ,
- $W$  can be written in terms of  $A_l^{(B)} - A_l^{(A)}$ ,  $F_l^{(A)}$  and  $F_l^{(B)}$
- Relax rank constraint: SDP (polynomial time in  $l$ )

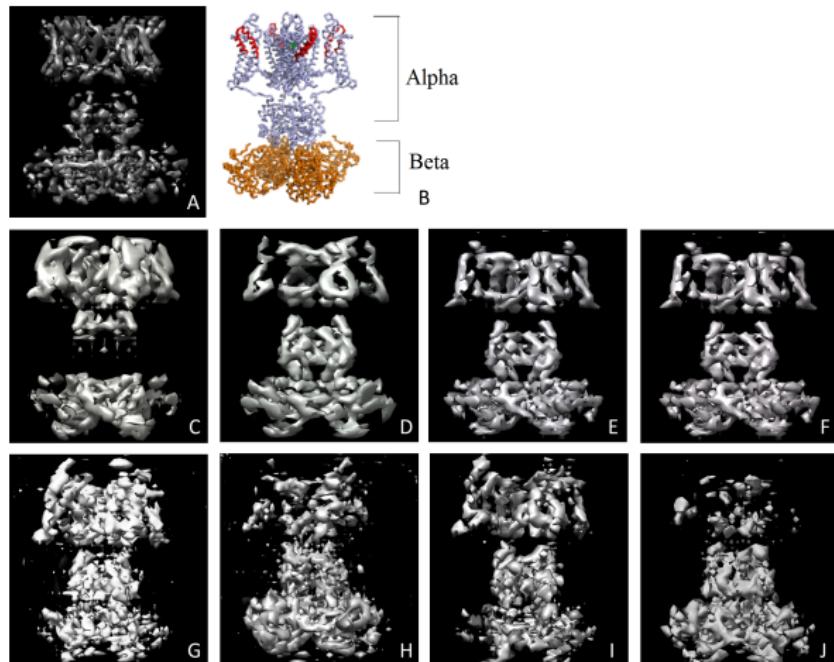
# Preliminary Result: Kv1.2 Potassium Channel

Synthetic images with SNR= 0.7, 0.35 (no CTF)



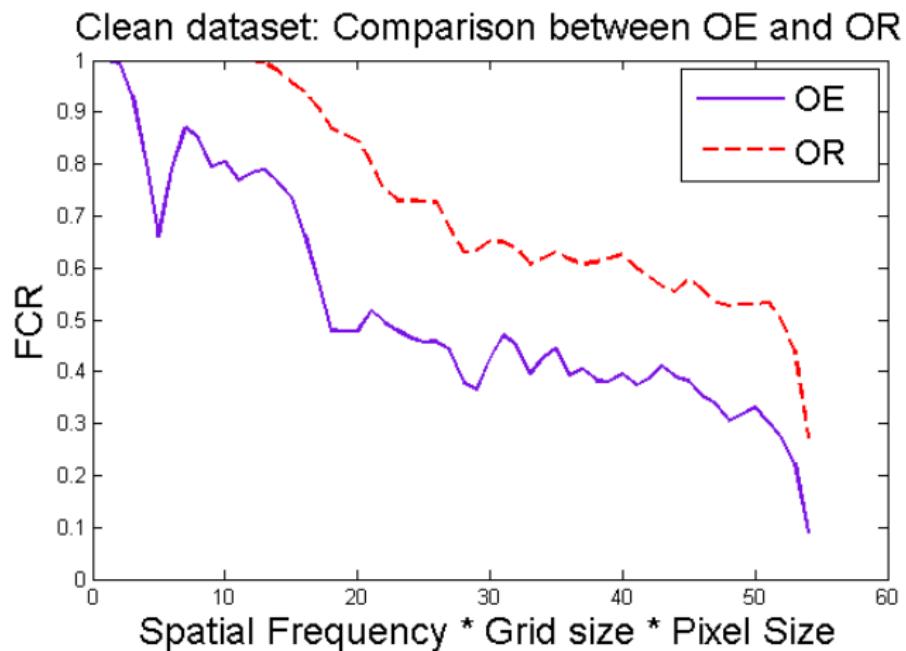
# Preliminary Result: Kv1.2 Potassium Channel

Synthetic images with noise (no CTF)

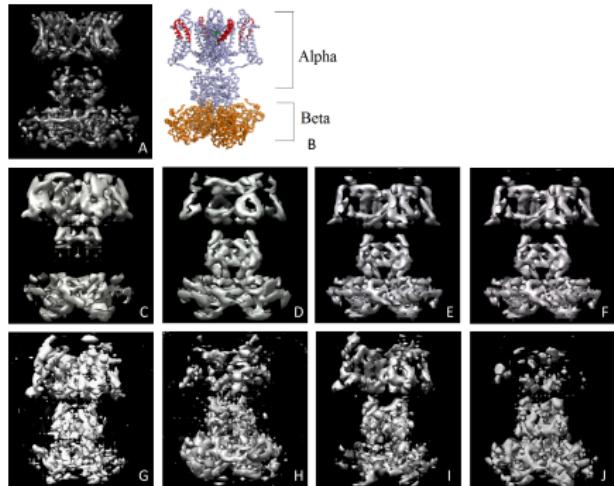


# Validation: Fourier Cross Resolution

Synthetic images with noise (no CTF)



# Revisiting OE



- Preliminary result: no CTF
- Expect improvement using estimated covariance  $\Sigma$
- Test on real data with CTF and low SNR

# Revisiting OE: Unbiased Estimator



- Twicing  $2\hat{\mathbf{A}}_{\text{LS}} - \mathbf{B}$ : unbiased estimator for scalar case  $\mathbf{A} \in \mathbb{C}^{1 \times 1}$
- Recovers unknown subunit better
- How to estimate  $\mathbf{A} \in \mathbb{R}^{N \times D}$  (or  $\mathbb{C}^{N \times D}$ ) from  $\mathbf{C}$  and  $\mathbf{B}$ , where  $\mathbf{C} = \mathbf{A}\mathbf{A}^*$  and  $\mathbf{A} = \mathbf{B} + \mathbf{E}$  for a matrix  $\mathbf{E}$  of small magnitude?

# Unbiased Estimator: Anisotropic Twicing (AT)

- Spectral decomposition  $\mathbf{C} = \mathbf{U} \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_D) \mathbf{U}^*$
- Asymptotically unbiased estimator of  $\mathbf{A}$  when  $N = D$  is given by

$$\hat{\mathbf{A}}_{\text{AT}} = \mathbf{B} + \mathbf{U} \mathbf{W} \mathbf{U}^* (\hat{\mathbf{A}}_{\text{LS}} - \mathbf{B})$$

where  $\mathbf{T}$  is a diagonal matrix with

$$\mathbf{T}_{ii} = \begin{cases} \frac{1}{D} \left[ -\frac{1}{2} + \sum_{1 \leq j \leq D} \frac{\lambda_i^2}{\lambda_i^2 + \lambda_j^2} \right] & \text{when } \mathbf{A}, \mathbf{C} \in \mathbb{R}^{D \times D}, \\ \frac{1}{D} \sum_{1 \leq j \leq D} \frac{\lambda_i^2}{\lambda_i^2 + \lambda_j^2} & \text{when } \mathbf{A}, \mathbf{C} \in \mathbb{C}^{D \times D}, \end{cases}$$

and  $\mathbf{W} = (\mathbf{I} - \mathbf{T})^{-1}$

# Unbiased Estimator: Anisotropic Twicing (AT)

Estimate autocorrelation matrices using CTF-corrected covariance from CWF

---

## Algorithm 2 Orthogonal Extension

---

1: **procedure** ORTHOGONAL EXTENSION BY ANISOTROPIC TWICING (OE-AT):  
ESTIMATE  $\mathbf{A}$  GIVEN  $\mathbf{B} \approx \mathbf{A}$ , SUBJECT TO  $\mathbf{C} = \mathbf{AA}^*$

2:

**Input:**  $\mathbf{B} \in \mathbb{C}^{N \times D}$ ,  $\mathbf{C} \in \mathbb{C}^{N \times N}$

3: Find any  $\mathbf{F} \in \mathbb{C}^{N \times D}$  such that  $\mathbf{C} = \mathbf{FF}^*$

4: Calculate  $\mathbf{B}^*\mathbf{F}$  and calculate its singular value decomposition  $\mathbf{B}^*\mathbf{F} = \mathbf{U}_0 \Sigma_0 \mathbf{V}_0^*$ .

5: Calculate the OE-LS estimator is  $\hat{\mathbf{A}}_{\text{LS}} = \mathbf{F}\mathbf{V}_0\mathbf{U}_0^*$ , (see Algorithm 1).

6: For  $N = D$ , the OE-AT estimator is given by  $\hat{\mathbf{A}}_{\text{AT}} = \mathbf{B} + \mathbf{UWU}^*(\hat{\mathbf{A}}_{\text{LS}} - \mathbf{B})$ .

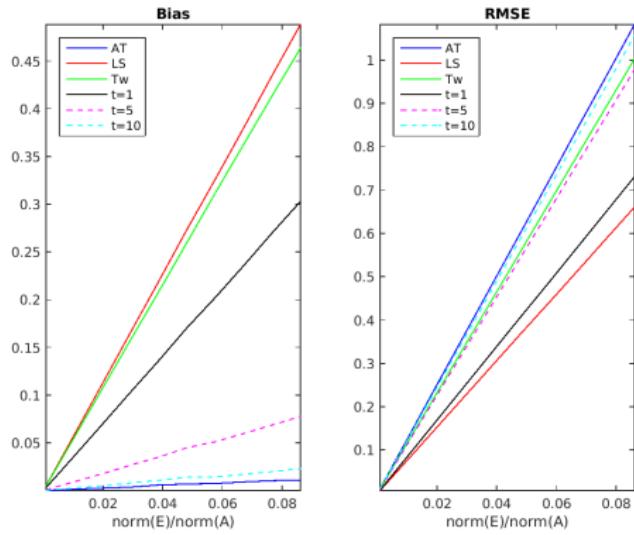
7: For  $N > D$ , assuming that  $\mathbf{P}$  is the projector of size  $N \times D$  to the  $D$ -dimensional subspace spanned by the columns of  $\mathbf{C}$ ,  $\hat{\mathbf{A}}_{\text{AT}} = \mathbf{P}\hat{\mathbf{A}}_{\text{AT}}^{(0)}$ .

8: For  $N < D$ , assuming that  $\mathbf{P}$  is the projector of size  $D \times N$  to the  $N$ -dimensional subspace in  $\mathbb{R}^D$  spanned by the rows of  $\mathbf{B}$ ,  $\hat{\mathbf{A}}_{\text{AT}} = \hat{\mathbf{A}}_{\text{AT}}^{(0)}\mathbf{P}^*$ .

---

## Unbiased Estimator: Anisotropic Twicing (AT)

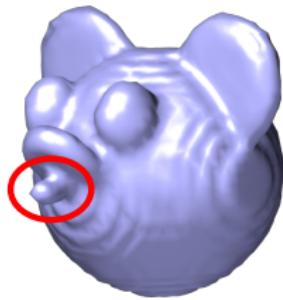
$$\text{MSE} = \mathbb{E}[||\theta - \hat{\theta}||^2] = \text{Bias}^2 + \text{Var}$$



# Synthetic Dataset: Toy Molecule

Clean, 1000 images

Relative error in unknown subunit: AT 19%, Twicing 31%, LS 59%



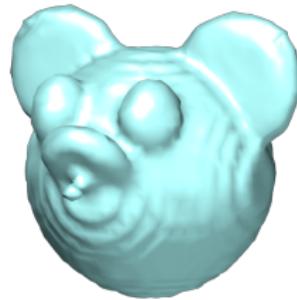
Original molecule with  
additional subunit  
marked in red

(a)



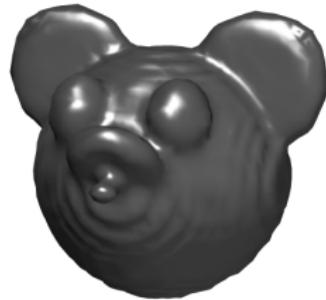
Least Squares

(b)



Twicing

(c)



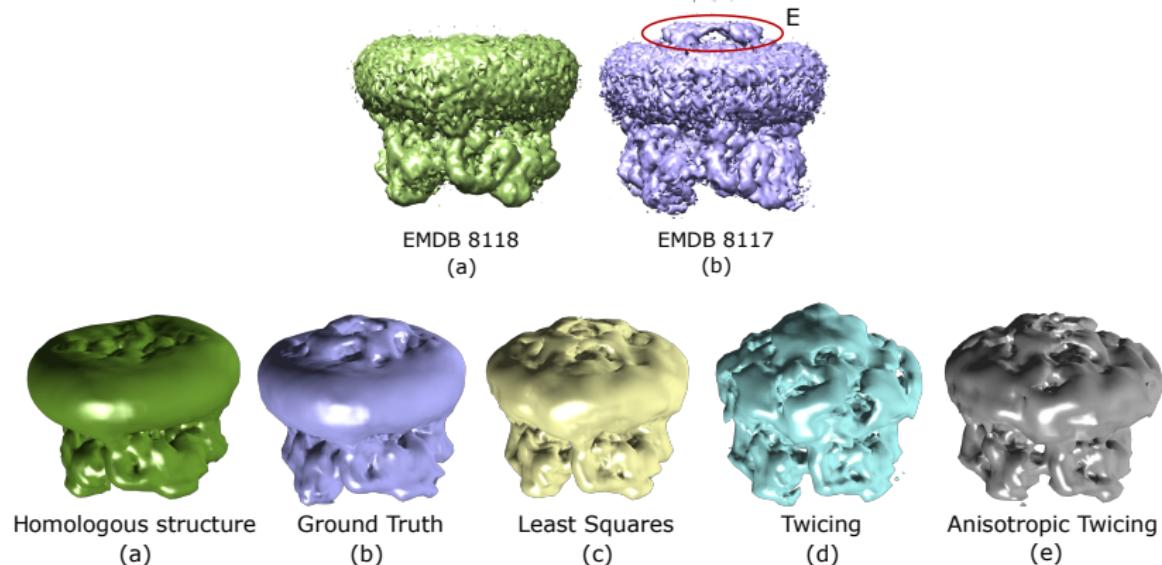
Anisotropic Twicing

(d)

# Synthetic Dataset: TRPV1

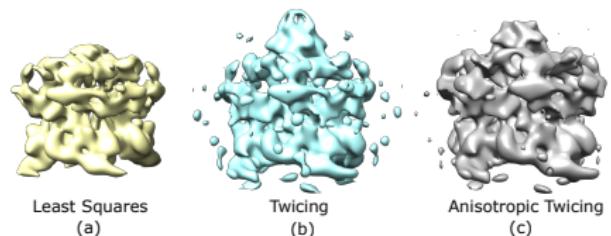
SNR= 1/40, 26000 images, 10 defocus groups.

Relative error in unknown subunit: AT 30%, Twicing 56%, LS 43%

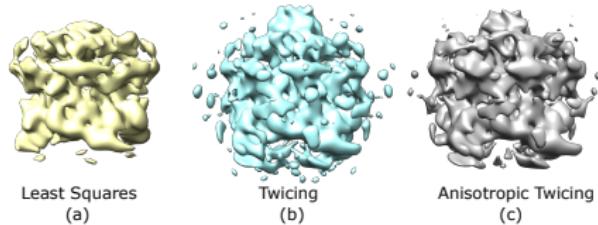


# Experimental Dataset: TRPV1 with DkTx and RTx

EMPIAR 10059: 73000 motion corrected images



Non-uniform viewing angles



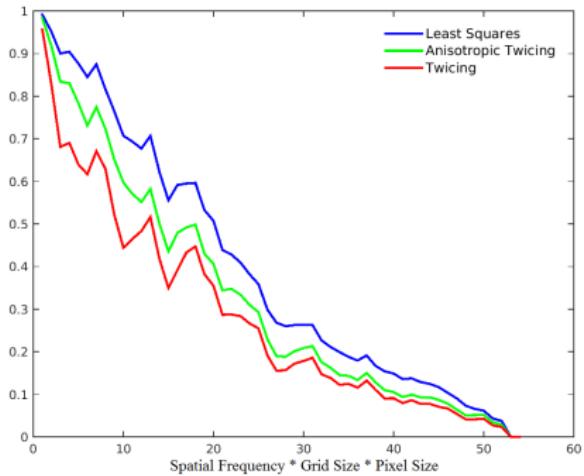
Selecting uniform viewing angles

Robust to viewing angle distribution

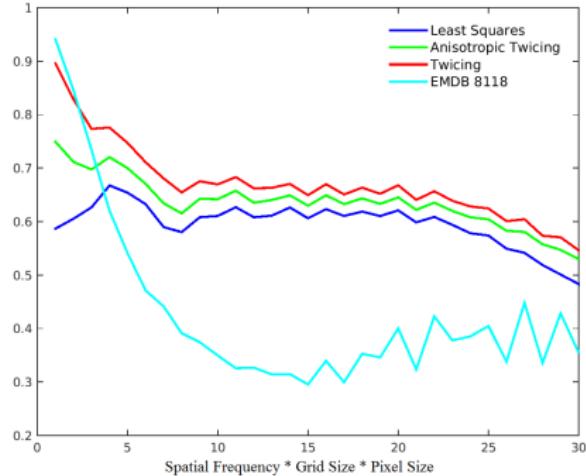
TB, T. Zhang, A. Singer (2017)

## Validation: FCR

## Dataset: EMPIAR 10059



## Full volume



## Unknown subunit

# Summary

- **Covariance estimation:** denoising, outlier detection
- **Mahalanobis affinity:** nearest neighbor detection for class averaging
- **Homology modeling:** 3D model directly from raw data

# Other applications

- **Covariance estimation:** other kinds of data with or without blurring kernels
- **Mahalanobis affinity:** extension to other imaging modalities with different blurring kernels
- **Homology modeling:** 3D model directly from raw data
- Model validation
- Extension to SPR with X-ray free electron lasers (XFEL)

# Acknowledgement

Amit Singer (Princeton)

David Huse (Princeton)

Fred Sigworth (Yale)

Joakim Anden (Princeton)

Joshua Shaevitz (Princeton)

Paul Steinhardt (Princeton)

Teng Zhang (UCF)

Xiuyuan Cheng (Yale)

Yoel Shkolnisky (Tel-Aviv)

Zhizhen Zhao (UIUC)

Adam Frost (UCSF)

Daniel Asarnow (UCSF)

David Julius (UCSF)

Eugene Palovcak (UCSF)

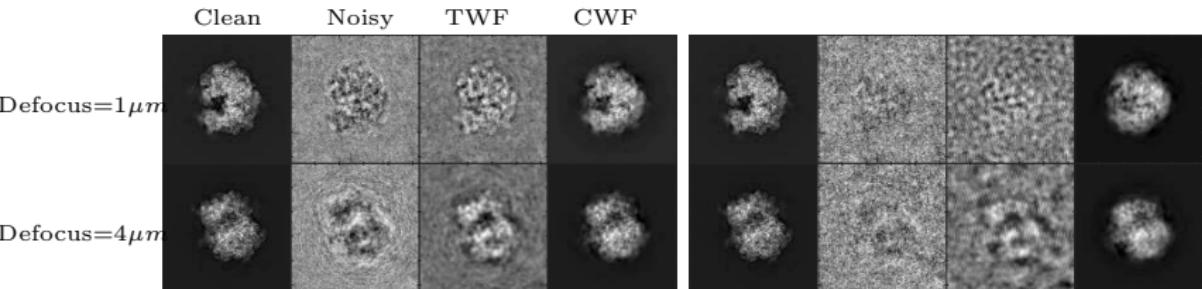
Garib Murshudov (MRC Cambridge)

Xiochen Bai (UTS)

Yuan Gao (UCSF)

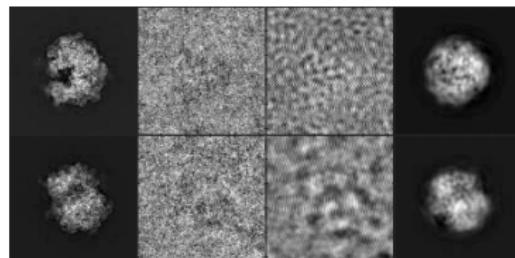
# Appendix

# Simulations with colored noise: 80S ribosome (EMDB-6454)



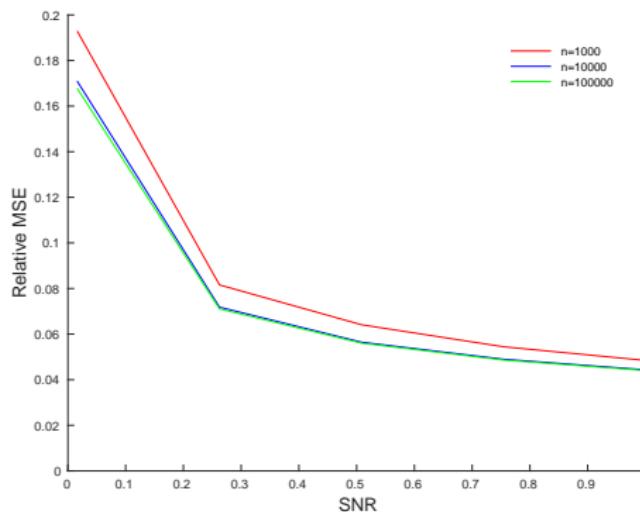
(a) SNR=1

(b) SNR=1/10



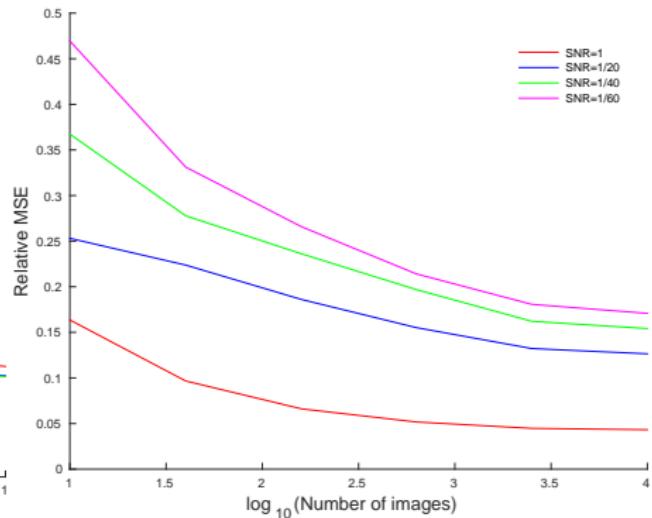
(c) SNR=1/20

# Relative error of estimated clean images



(d)

(a) Fixed number of images



(e)

(b) Fixed SNR