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

Introduction

Cryo-Electron Microscopy (Cryo-EM)


nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio &

Archive > Volume 525 > Issue 7568 > News Feature > Article

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

- Understanding function, mechanisms, drug discovery
- X-ray crystallography has limitations
- Structure of biological macromolecules *in-vivo*, without crystallization

Cryo-EM Revolution

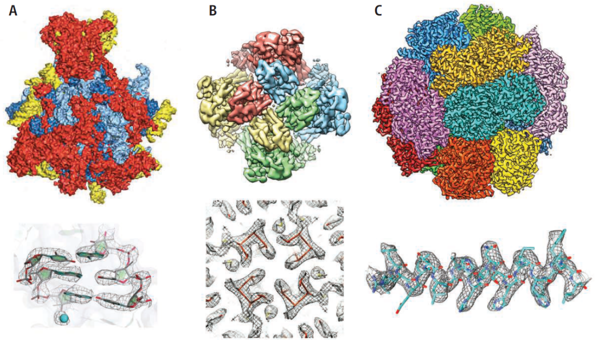
The Resolution Revolution

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

Werner Kühlbrandt

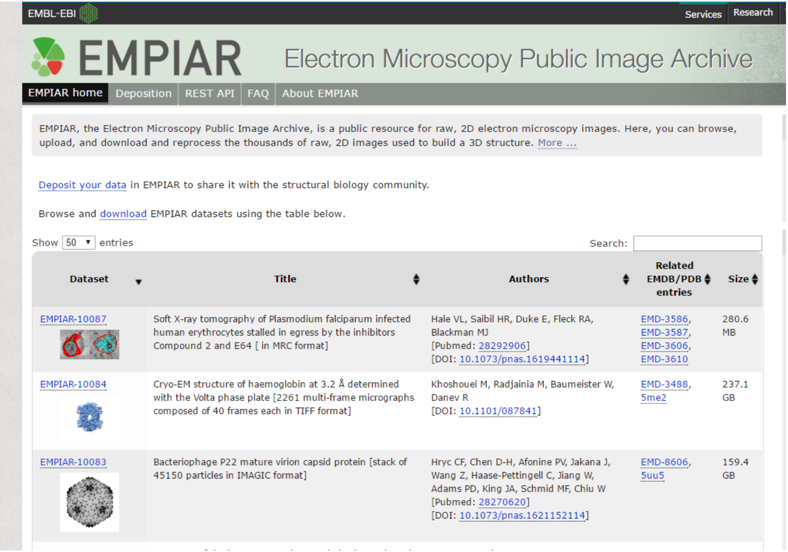
Precise knowledge of the structure of macromolecules in the cell is essential for understanding how they function. Structures of large macromolecules can now be obtained at near-atomic resolution by averaging thousands of electron microscope images recorded before radiation damage accumulates. This is 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-



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 Services Research

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 ...

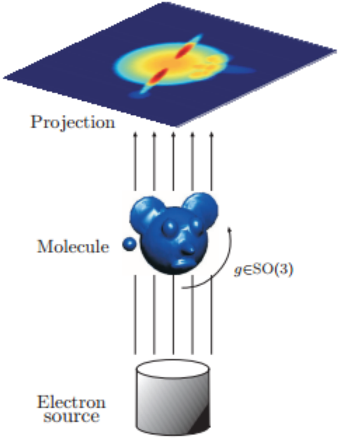
Deposit your data in EMPIAR to share it with the structural biology community.

Browse and download EMPIAR datasets using the table below.

Show 50 entries Search:

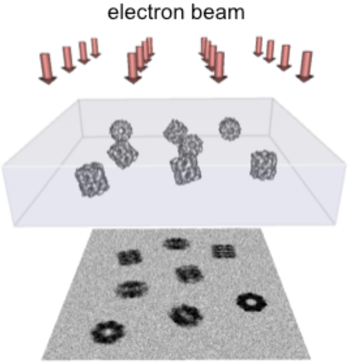
Dataset	Title	Authors	Related EMDB/PDB entries	Size
EMPIAR-10087	Soft X-ray tomography of Plasmodium falciparum infected human erythrocytes stalled in egress by the inhibitors Compound 2 and E64 [in MRC format]	Hale VL, Saibil HR, Duke E, Fleck RA, Blackman MJ [Pubmed: 28292906] [DOI: 10.1073/pnas.1619441114]	EMD-3586, EMD-3587, EMD-3606, EMD-3610	280.6 MB
EMPIAR-10084	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: 10.1101/087841]	EMD-3488, 5me2	237.1 GB
EMPIAR-10083	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-Pettingell C, Jiang W, Adams PD, King JA, Schmid MF, Chiu W [Pubmed: 28270620] [DOI: 10.1073/pnas.1621152114]	EMD-8606, 5uu5	159.4 GB

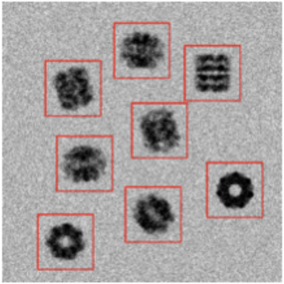
Single Particle Reconstruction (SPR)



- 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

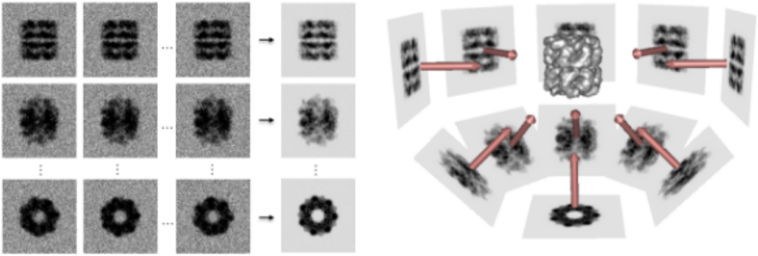
Single Particle Reconstruction (SPR)





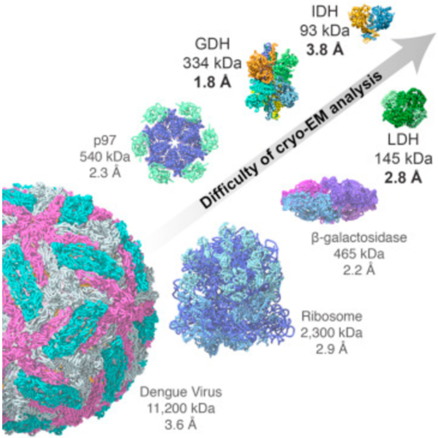
https://people.csail.mit.edu/gdp/cryoem.html

Single Particle Reconstruction (SPR)



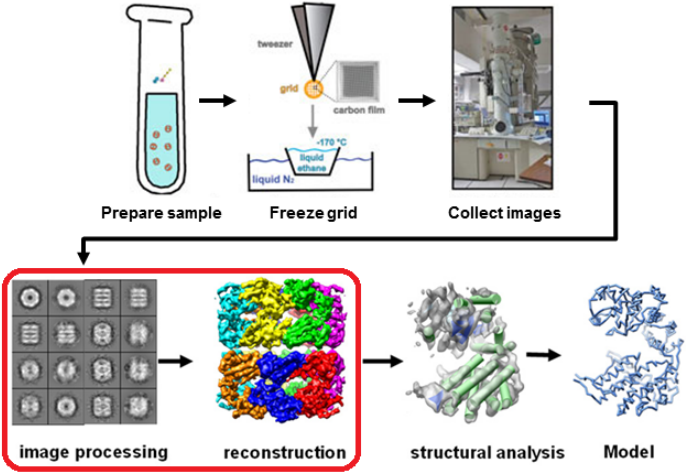
https://people.csail.mit.edu/gdp/cryoem.html

Cryo-EM Length Scales

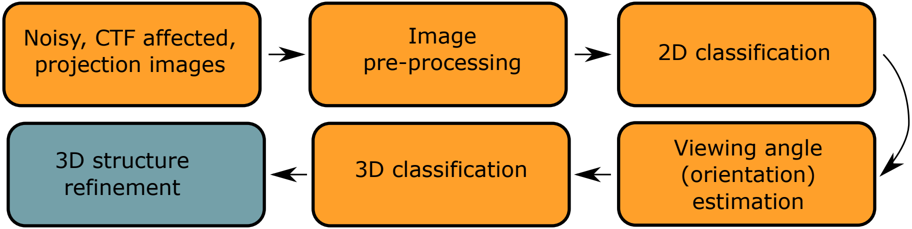


Merk et al. (2016)

SPR Pipeline

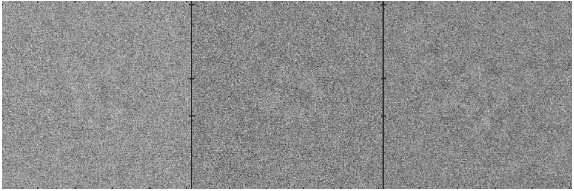


Existing Computational Pipeline



Challenges and Contributions

Challenges

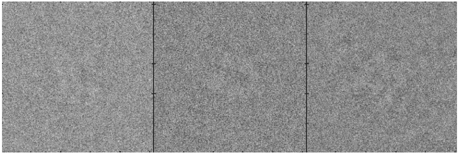


Raw images from an experimental dataset of TRPV1 *

- 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

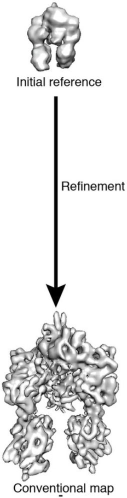
* M. Liao et al.(2013)

Challenge: Covariance Estimation



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

Challenge: Initial Model for 3D Refinement



S. Dutta et al.(2014)

- 'Guess' low resolution initial model
- 3D refinement sensitive to initial model
- Data driven ab-initio model [a] using common-lines for orientation estimation
- Reliable estimation (common-lines etc.) difficult at low SNR without averaging

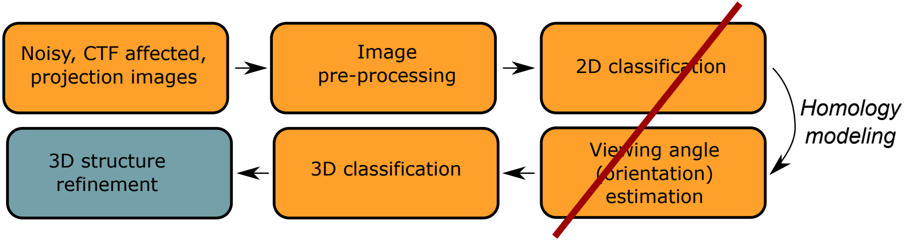
[a] A. Singer et al. (2011)

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

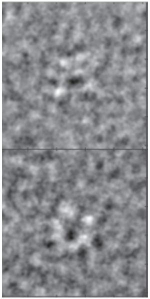
Algorithms validated on both synthetic and real experimental datasets

Proposed Computational Pipeline

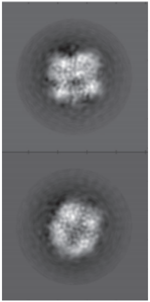


Denoising: Real TRPV1 dataset

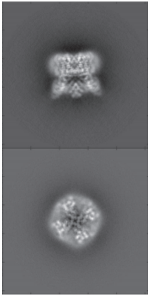
Raw



Existing



This work

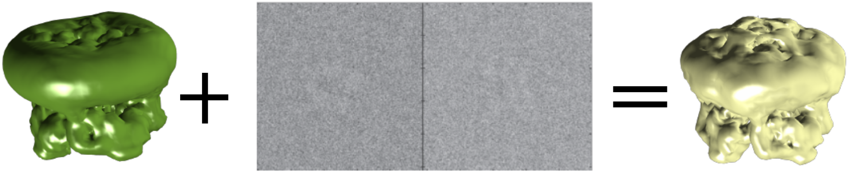


Closest

T.B., T. Zhang, A. Singer (2016)

Homology modeling: synthetic TRPV1 dataset

Molecule with unknown structure



Homologous molecule and projection images of unknown molecule

Recovered molecule

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

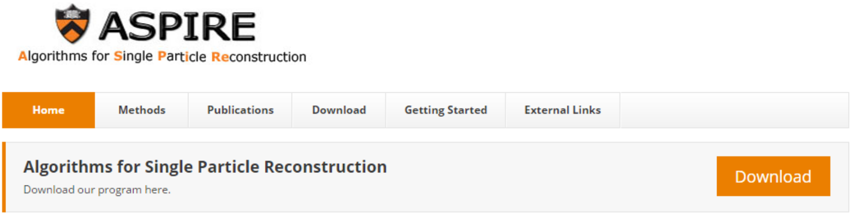
Publications

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).

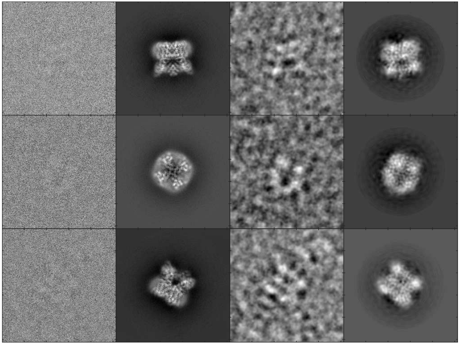
Code

Open source software toolbox for cryo-EM: `spr.math.princeton.edu`



Part 1: Covariance Estimation from Noisy Measurements

Experimental data - TRPV1

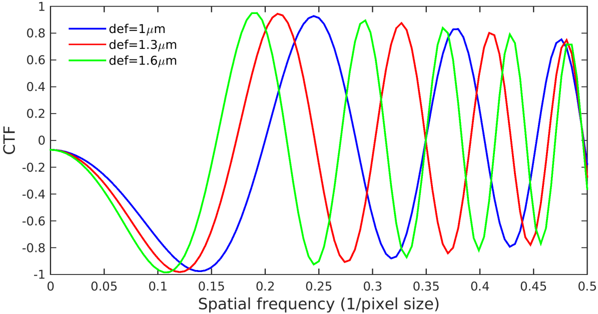


- K2 direct electron detector
- 35645 motion corrected, picked particle images of 256×256 pixels

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

$$y_i = a_i * x_i + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{1}$$

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
ϵ_i: additive Gaussian noise that corrupts the image

The Model: Fourier space

$$Y_i = A_i X_i + \xi_i, \quad i = 1, 2, \ldots, n \tag{2}$$

- A_i: diagonal operator (CTF)
- $X_1, \ldots, X_n$: 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

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

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

Relates the second order statistics of the noisy images with the population covariance Σ of the clean images

Mean Estimation

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

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

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 $||.||_F$ is the Frobenius matrix norm.

Solving using Conjugate Gradient

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

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

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

where $L : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ 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

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

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

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

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 Σ 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 Σ, corresponding to the angular frequency k, separately
- **Nearly unitary transformation**

Z. Zhao, Y. Shkolnisky and A. Singer (2016)

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

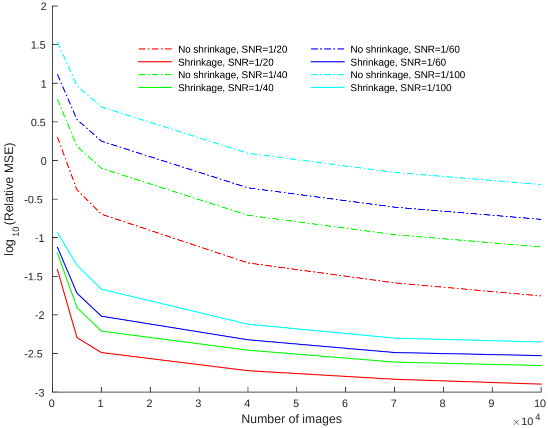
Computational Complexity: Timings

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

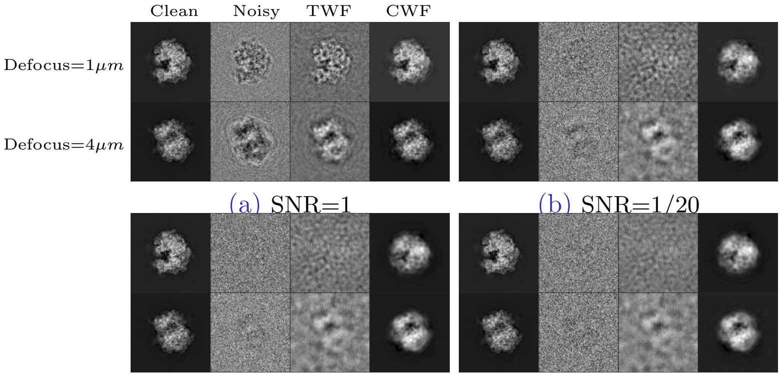
Dataset	L	n	Basis coeffs	CWF
TRPV1	256	35645	312s	574s
80s	360	30000	731s	385s
IP3R1	256	37382	429s	589s
70s	250	99979	1174s	113s

Relative error of estimated covariance



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

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



(a) SNR=1 (b) SNR=1/20

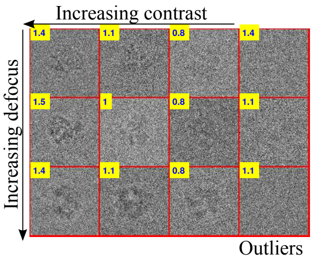
(c) SNR=1/40 (d) SNR=1/60

Outlier Detection

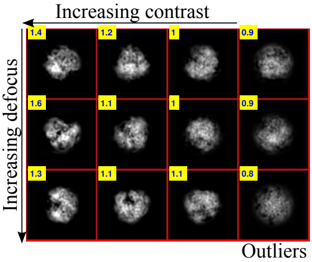
- Current method: manual visual inspection after particle picking
- CWF: automatic way to classify picked particles
- Specimen particles at various depths in the ice layer: acquired projection images can have different contrasts
- $$Y_i = \alpha_i A_i X_i + \xi_i, \quad i = 1, 2, \ldots, n \tag{12}$$
- Absorb α 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

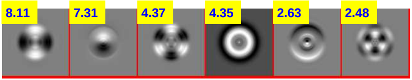


(e)



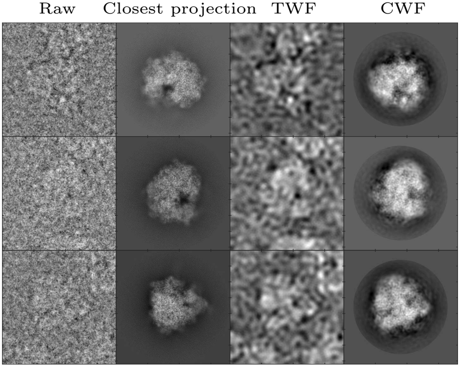
(f)

(g)



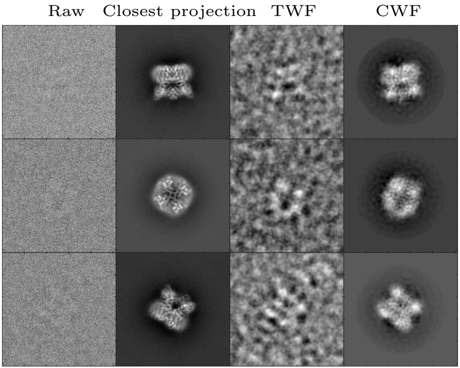
(h)

Experimental data - 80S ribosome



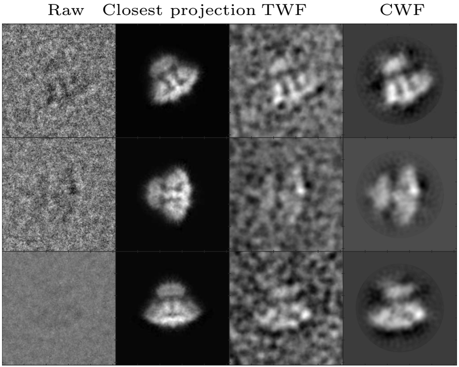
- FALCON II 4k×4k direct electron detector
- 105247 motion corrected, picked particle images of 360×360 pixels

Experimental data - TRPV1



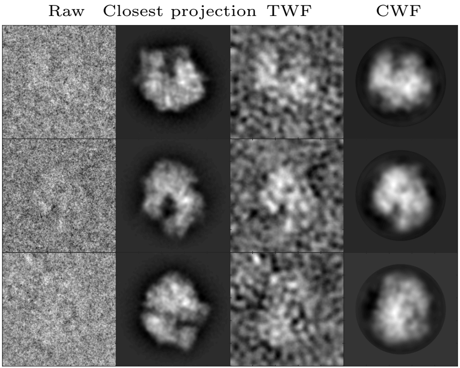
- K2 direct electron detector
- 35645 motion corrected, picked particle images of 256×256 pixels

Experimental data -IP_3R1



- 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 250×250 pixels

Application: Mahalanobis Affinity

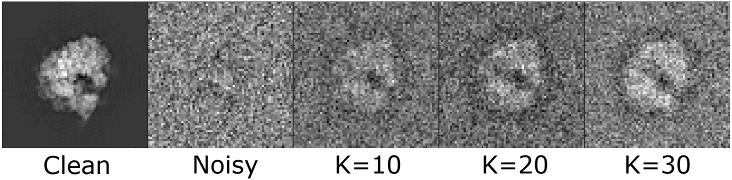
Mahalanobis affinity or likelihood for the underlying clean images x_i and x_j to originate from the same viewing direction

$$\Pr(||X_{ij}||_p < \epsilon | Y_i = y_i, Y_j = y_j)$$

$$= \frac{\epsilon^d \mathrm{Vol}(B_p(0,1))}{(2\pi)^{\frac{d}{2}} |L_i + L_j|^{\frac{1}{2}}} \exp\{-\frac{1}{2}\alpha_{ij}^T (L_i + L_j)^{-1} \alpha_{ij}\} + \mathcal{O}(\epsilon^{d+1})$$

TB, Z. Zhao and A. Singer (2017)

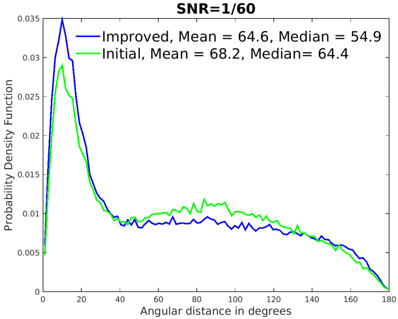
Application: Mahalanobis Affinity

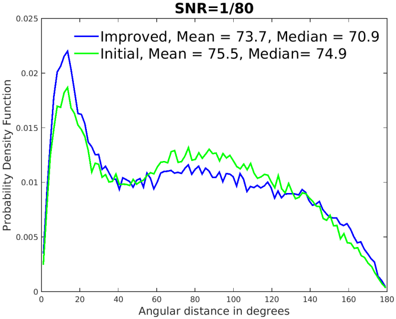


Class averages with the improved algorithm using the anisotropic affinity, using $K = 10,\ 20,\ 30$.

TB, Z. Zhao and A. Singer (2017)

Application: Mahalanobis Affinity





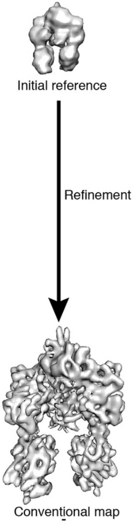
Improved class averaging using Mahalanobis affinity to detect nearest neighbors

---

TB, Z. Zhao and A. Singer (2017)

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
- $2\hat{\boldsymbol{A}}_{\mathrm{LS}} - \boldsymbol{B}$ for $\boldsymbol{A} \in \mathbb{C}^{1\times 1}$ 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

$$\boldsymbol{C_l} = \boldsymbol{A_l}\boldsymbol{A_l^*} \quad l = 0, 1, \ldots \tag{13}$$

- $\boldsymbol{C_l}$: autocorrelation matrix over $SO(3)$ ("magnitude")
- bA_l: Expansion coefficients of 3D Fourier volume
- $\boldsymbol{C_l}$ can be computed from the covariance matrix Σ of the 2D projection images
- Requirement: Uniformly distributed viewing angles over the sphere

Z. Kam (1980, 1985)

Outline of Our Approach

$$\boldsymbol{C_l} = \boldsymbol{A_l}\boldsymbol{A_l^*} \quad l = 0, 1, \ldots \tag{14}$$

- $\boldsymbol{C_l}$ can be computed from the covariance matrix Σ
- Use estimated $\hat{\Sigma}$ from Covariance Wiener Filtering to compute $\boldsymbol{C_l}$ ("magnitude")
- Determines $\boldsymbol{A_l}$ upto an orthogonal matrix ("missing phase")

Orthogonal Matrix Retrieval Problem

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

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

Kam's Autocorrelation Analysis

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

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

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

Z. Kam (1980)

Resolution Limit

- Finite pixel grid: Nyquist criterion

$$\mathcal{F}(\Phi_A)(k,\theta,\varphi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} A_{lm}(k) Y_l^m(\theta,\varphi), \quad l = 0, 1, \ldots, L$$

$$A_{lm}(k) = \sum_{s=1}^{S_l} a_{lms} j_{ls}(k). \tag{15}$$

Resolution Limit

Normalized spherical Bessel functions

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

- 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.

X. Cheng (2013)

Orthogonal Matrix Retrieval Problem

$$\boldsymbol{C_l} = \boldsymbol{A_l}\boldsymbol{A_l^*}$$

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

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

Orthogonal Extension (OE)

- Determine the coefficient matrices $\boldsymbol{A_l}$
- Known, homologous structure Φ_B

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

- $\boldsymbol{F_l}$: any matrix of size $S_l \times 2l+1$ satisfying $\boldsymbol{C_l} = \boldsymbol{F_l}\boldsymbol{F_l^*}$

$$\boldsymbol{A_l} = \boldsymbol{F_l}\boldsymbol{O_l} \tag{17}$$

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

Orthogonal Extension (OE)

- Determine the coefficient matrices $\boldsymbol{A_l}$
- Known, homologous structure Φ_B

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

- $\boldsymbol{F_l}$: any matrix of size $S_l \times 2l+1$ satisfying $\boldsymbol{C_l} = \boldsymbol{F_l}\boldsymbol{F_l^*}$

$$\boldsymbol{A_l} = \boldsymbol{F_l}\boldsymbol{O_l} \tag{19}$$

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

Orthogonal Extension (OE)

- $\boldsymbol{A_l} \approx \boldsymbol{B_l}$ (homologous assumption)

$$\boldsymbol{O_l} = \underset{\boldsymbol{O} \in \mathrm{O}(2l+1)}{\arg\min} \|\boldsymbol{F_l O} - \boldsymbol{B_l}\|_F^2$$

- Closed form solution via singular value decomposition (SVD)

$$\boldsymbol{B}_{\boldsymbol{l}}^{*}\boldsymbol{F_l} = \boldsymbol{U_l \Sigma_l V_l^T}$$

$$\boldsymbol{O_l} = \boldsymbol{V_l U_l^T}$$

J. Keller (1975)

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

OE with Least Squares (OE-LS)

- Least squares estimator

$$\hat{\boldsymbol{A}}_{\boldsymbol{l},\mathrm{LS}} = \boldsymbol{F}_{\boldsymbol{l}}\boldsymbol{V}_{\boldsymbol{l}}\boldsymbol{U}_{\boldsymbol{l}}^T \tag{20}$$

- Twicing in practice: $2\hat{\boldsymbol{A}}_{\boldsymbol{l},\mathrm{LS}} - \boldsymbol{B}_{\boldsymbol{l}}$

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

Orthogonal Replacement (OR)

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

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

Orthogonal Replacement (OR)

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

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

Relaxation to a Semidefinite Program (SDP)

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

can be written as

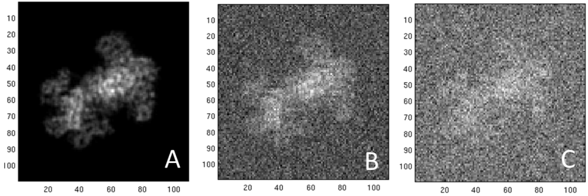
$$\min_Q \mathrm{Tr}(WQ)$$

- $Q \in \mathbb{R}^{3(2l+1) \times 3(2l+1)}$
- subject to $Q_{ii} = I$, $\mathrm{rank}(Q) = 2l+1$ 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)

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

Preliminary Result: Kv1.2 Potassium Channel

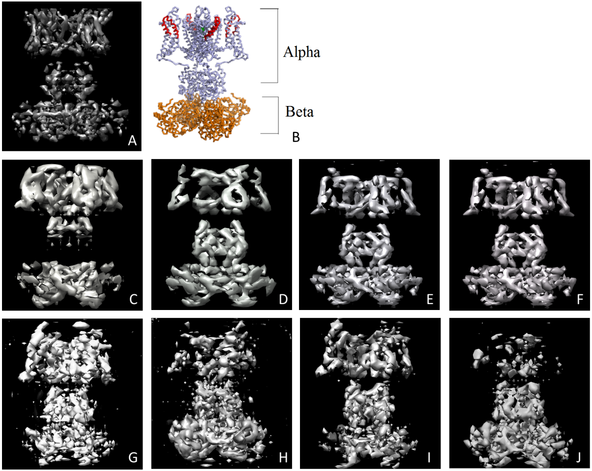
Synthetic images with SNR$= 0.7, 0.35$ (no CTF)



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

Preliminary Result: Kv1.2 Potassium Channel

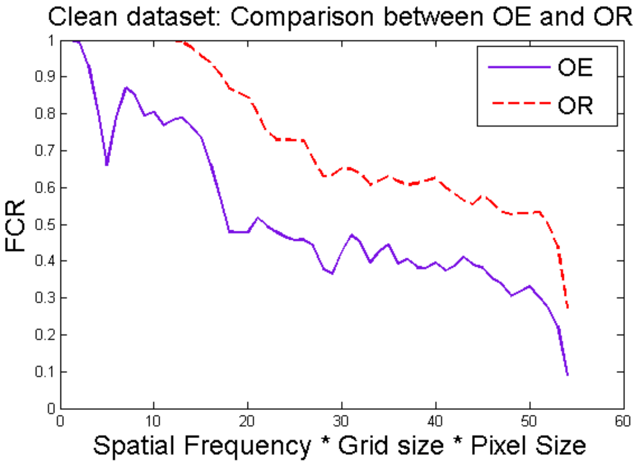
Synthetic images with noise (no CTF)



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

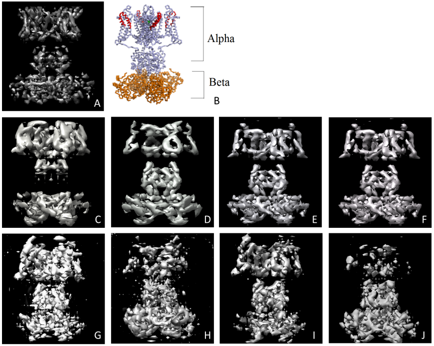
Validation: Fourier Cross Resolution

Synthetic images with noise (no CTF)



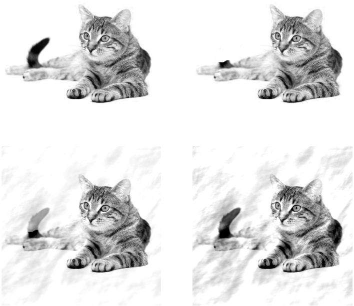
FCR curve for reconstruction from β_4 (clean images)

Revisiting OE



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

Revisiting OE: Unbiased Estimator



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

Unbiased Estimator: Anisotropic Twicing (AT)

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

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

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

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

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

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

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 $\boldsymbol{A}$ GIVEN $\boldsymbol{B} \approx \boldsymbol{A}$, SUBJECT TO $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{A}^*$

2:

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

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

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

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

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

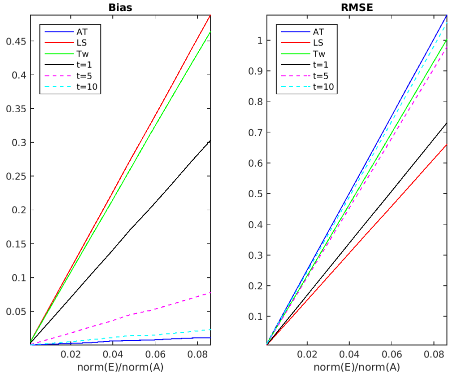
7: For $N > D$, assuming that $\boldsymbol{P}$ is the projector of size $N \times D$ to the D-dimensional subspace spanned by the columns of $\boldsymbol{C}$, $\hat{\boldsymbol{A}}_{\mathrm{AT}} = \boldsymbol{P}\hat{\boldsymbol{A}}_{\mathrm{AT}}^{(0)}$.

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

TB, T. Zhang, A. Singer (2016)
TB, T. Zhang, A. Singer (2017)

Unbiased Estimator: Anisotropic Twicing (AT)

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

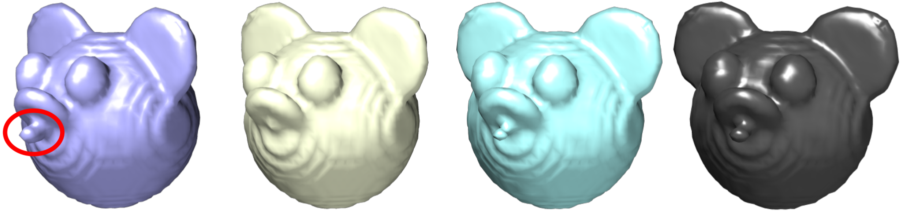


TB, T. Zhang, A. Singer (2017)

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)

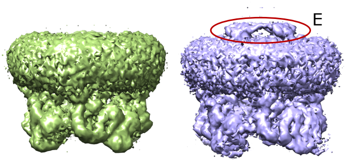
Twicing
(c)

Anisotropic Twicing
(d)

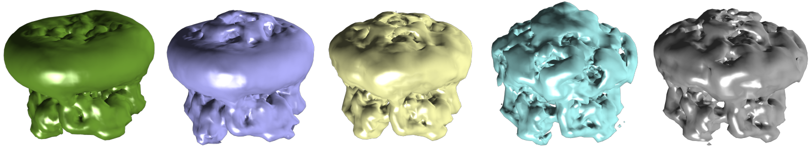
TB, T. Zhang, A. Singer (2017)

Synthetic Dataset: TRPV1

SNR= 1/40, 26000 images, 10 defocus groups.
Relative error in unknown subunit: AT 30%, Twicing 56%, LS 43%



EMDB 8118 (a) EMDB 8117 (b)

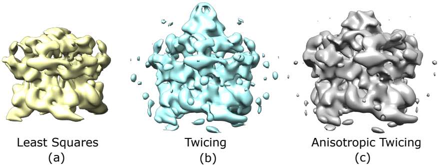


Homologous structure (a) Ground Truth (b) Least Squares (c) Twicing (d) Anisotropic Twicing (e)

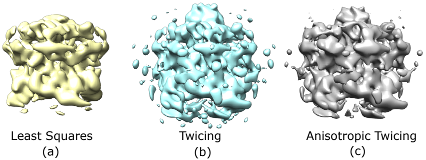
TB, T. Zhang, A. Singer (2017)

Experimental Dataset: TRPV1 with DkTx and RTx

EMPIAR 10059: 73000 motion corrected images



Non-uniform viewing angles



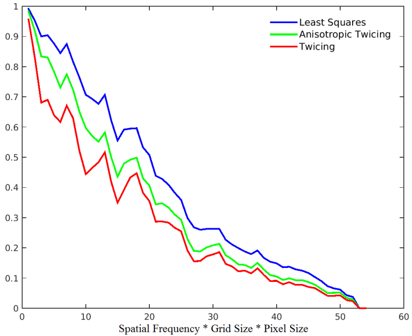
Selecting uniform viewing angles

Robust to viewing angle distribution

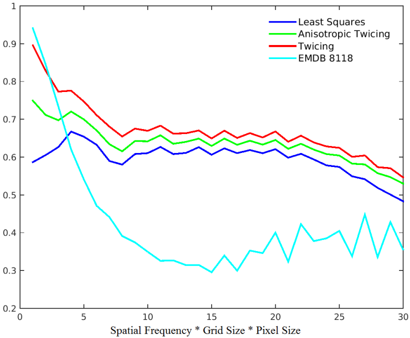
TB, T. Zhang, A. Singer (2017)

Validation: FCR

Dataset: EMPIAR 10059



Full volume



Unknown subunit

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

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)

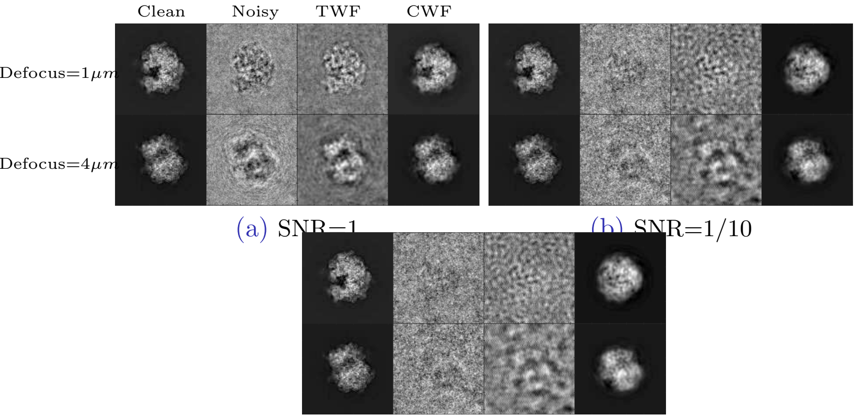
Acknowledgement

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)

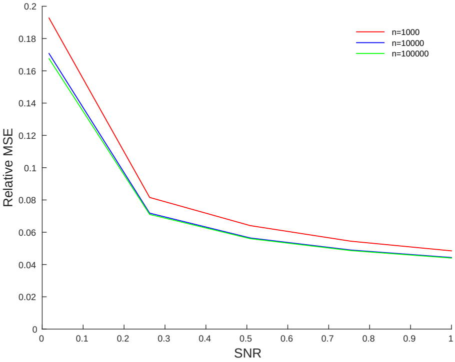


(a) SNR=1

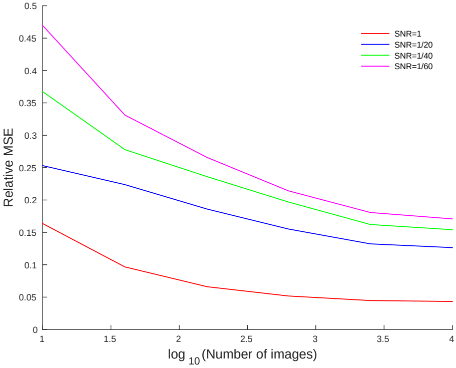
(b) SNR=1/10

(c) SNR=1/20

Relative error of estimated clean images



(d)



(e)

(a) Fixed number of images (b) Fixed SNR