Statistica Sinica **33** (2023), 1771-1788
doi:https://doi.org/10.5705/ss.202021.0419

UNCERTAINTY QUANTIFICATION IN DYNAMIC IMAGE RECONSTRUCTION WITH APPLICATIONS TO CRYO-EM

Tze Leung Lai[1], Shao-Hsuan Wang[2], Szu-Chi Chung[3], Wei-hau Chang[4]
and I-Ping Tu[4]

[1]*Stanford University,* [2]*National Central University,*
[3]*National Sun Yat-sen University and* [4]*Academia Sinica*

Abstract: Here, we propose combining empirical Bayes modeling with recent advances in Markov chain Monte Carlo filters for hidden Markov models. In doing so, we address long-standing problems in the reconstruction of 3D images, with uncertainty quantification, from noisy 2D pixels in cryogenic electron microscopy and other applications, such as brain network development in infants.

Key words and phrases: Change-points, cryogenic electron microscopy, empirical Bayes, hidden Markov models, Markov chain Monte Carlo, particle filters, stem cells, uncertainty quantification.

1. Introduction

Cryogenic electron microscopy (cryo-EM) is an imaging technique that uses transmitted electron waves to obtain projection images of a biological sample. In contrast to X-ray crystallography, cryo-EM imaging does not need crystals, and thus can determine the structure of proteins that are refractory to crystallization, including membrane proteins and yeast spliceosomes that exhibit dynamic patterns (Liao et al. (2013); Yan et al. (2015)). This ability to use cryo-EM to determine the high-resolution structure of isolated macromolecules in solution, earned Jacques Dubochet, Joachim Frank, and Richard Henderson the Nobel Prize in Chemistry in 2017.

These breakthroughs have established single-particle cryo-EM as the mainstream method for solving high-resolution 3D structure density maps of biomolecules. Nevertheless, performing single-particle cryo-EM is highly demanding, because cryo-EM images are extremely noisy, with a signal-to-noise ratio (SNR) often less than 0.1. As a result, a typical cryo-EM experiment tends to collect a large number of particle images (usually hundreds of thousands), and compensates for the noise contamination by using massive averaging. The size of

Corresponding author: I-Ping Tu, Institute of Statistical Science, Academia Sinica, Nan-Gang, Taipei 115, Taiwan. Email: iping@stat.sinica.edu.tw.

a cryo-EM particle image is often larger than 100 pixels, measured in each direction. Such images are characterized by strong noise contamination, a huge dimension, and a large sample size, making their processing and statistical analysis challenging.

In Section 4, we describe an empirical Bayes (EB) approach to address some of the challenges caused by a low SNR. This EB framework is now used in the open-source software RELION (REgularised LIkelihood OptimisatioN) (Scheres (2012b)). RELION is popular in the cryo-EM community for producing 3D density maps. However, the low SNR of cryo-EM images representing particle projections of unknown orientation makes image alignment and averaging difficult. When a reference structure is used to facilitate the extremely low SNR image alignment, the outcome is often dictated by the reference. This artifact, known as Einstein from noise, was identified and coined by Stewart and Grigorieff (2004). In a simulation experiment, they generated 1,000 white-noise images, aligning each of them to Einstein's facial image using rotation and translation. A blurred image of Einstein's face emerged after averaging the 1,000 aligned images. Henderson (2013) used this phenomenon to emphasize the importance of validating that a reconstructed 3D image is a reliable and faithful representation of the underlying molecule, which we discuss in Section 3.

From the perspective of statistical analysis, the "Einstein from noise" phenomenon can be formulated as a problem of model bias (also see the Supplementary Material S1). To capture the essence of model bias in image alignment, Wang et al. (2021) treat an image of p pixels as vector of dimension p, and a white-noise image as a random vector uniformly distributed on the $(p-1)$-dimensional unit sphere. Rather than rotating 1,000 images, Wang et al. (2021) generate millions of random vectors to compensate for the increased samples, and simplify the pixel-correlated format using the rotating process in image alignment. The cross-correlation (CC) of two images is defined as the inner product of the corresponding vectors, and is widely used as a similarity measure in image processing. Using this simplification, Wang et al. (2021) perform a simulation study with $n = 10^6$ white-noise images and pixel number $p = d_1 \times d_2 = 10^4$, where $d_1 = d_2 = 100$. Let $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ be independent and identically distributed (i.i.d.) $d_1 \times d_2$ random matrices, such that the $d_1 \times d_2$ components of each $\boldsymbol{Z}_i$ are iid standard normal. They call $\boldsymbol{X}_i = \mathrm{vec}(\boldsymbol{Z}_i)/\|\boldsymbol{Z}_i\|$ the ith white-noise image, where $\|\cdot\|$ denotes the Frobenius norm of a matrix or the Euclidean norm of a vector, because $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. uniformly distributed on the $(p-1)$-dimensional unit sphere $\mathcal{S}^{p-1}$. Let $\boldsymbol{R}$ be the $d_1 \times d_2$ reference matrix (the digital version of the reference image) that is normalized to have norm one so that $\boldsymbol{r} = \mathrm{vec}(\boldsymbol{R}) \in \mathcal{S}^{p-1}$.

The simulation study considers $\overline{\boldsymbol{X}}_m/\|\overline{\boldsymbol{X}}_m\| \in \mathcal{S}^{p-1}$, which is the normalized average of the m white-noise images that are most highly cross-correlated with the reference image, for $m = 100, 200, \ldots, 600$. Letting $\rho_{n,p,m} = \boldsymbol{r}^{\top}\overline{\boldsymbol{X}}_m/\|\overline{\boldsymbol{X}}_m\|$ be the cross-correlation of $\boldsymbol{r}$ and $\overline{\boldsymbol{X}}_m$, Wang et al. (2021) perform an elaborate asymptotic analysis of $\rho^2_{n,p,m}$ to explain the results in the simulation study, as summarized in the Supplementary Material S1.

To address the challenge of aligning very noisy images, instead of targeting a single, best orientation and class for each particle based on the cross-correlation coefficient between the references and the particle, Sigworth (1998) considered a maximum-likelihood approach. In this approach, one computes the posterior probabilities for all possible orientations and classes, where each particle contributes to all references and in all orientations according to a weighting scheme based on the posterior probabilities. Later, Scheres (2012b) extended this maximum-likelihood approach to cryo-EM analysis to Bayesian approach. Note that the Bayesian approach of RELION differs form other likelihood optimization approaches by introducing a regularization term to the likelihood function.

Compared with X-ray crystallography, a technique that measures an ensemble of particles, cryo-EM produces images of individual particles. These noisy images record individual particles at unknown conformation states, viewed from unknown directions. Thus, cryo-EM has the potential to compile 3D maps of the dynamical processes in which these macromolecules participate, which would, in turn, uncover the functionality of these macromolecules. Unfortunately, the presence of multiple, different 3D structures in the data presents challenges. When the number of functional states is small and the structural differences between them are sufficiently large, the 3D classification tool in the RELION software can sort particles into subsets that are structurally more homogeneous. However, this in silico structure purification, namely, struggles to identify homogeneous structure segmentation, offered by the EB on RELION subtle structural differences between function states associated with a continuum of states, rather than with distinct independent states. When a function domain of a macromolecular machine undergoes continuous conformation change, the 3D density map produced by RELION consists of two parts of different quality, a static core with remarkably high resolution, and a blurry region, usually at the periphery. In the worst case, the densities associated with a flexible domain are missing, making it impossible to even hypothesize the corresponding location. Recovering the densities of these nonrigid dynamic pieces is one of the great challenges in cryo-EM analysis.

Habeck (2017) shows that the Markov chain Monte Carlo (MCMC) framework enables us to characterize the uncertainty in numerical solutions, and provides a natural connection to atomic structures. Lederman, Anden and Singer (2019) propose using MCMC algorithms to model dynamic states to meet the aforementioned challenge. In Section 5, we propose using the MCMC with sequential substitutions (MCMC-SS) approach proposed by Lai et al. (2020) to strengthen the computation efficiency. This MCMC-SS approach provides adaptive filtering in hidden Markov models (HMMs) on general state spaces to allow us to approximate an intractable distribution of interest (or target distribution) using the empirical distribution of M representative atoms, chosen sequentially by an MCMC procedure. As a result, the empirical distribution converges to the target distribution after a sufficiently large number of iterations. The MCMC-SS approach has been applied to quantify the uncertainty in image reconstruction of the brain network development in infants using induced pluro-potential stem cells (Wu et al. (2021)).

2. The Development of Single Particle Analysis

In the early 1980s, the Swiss biophysicist Dubochet of the European Molecular Biology Laboratory in Heidelberg succeeded in cooling water so rapidly that it solidified around a biological molecule without forming distorting ice crystals.

According to Mossman (2007), Joachim Frank published two landmark papers in 1981 on electron microscopy, namely, van Heel and Frank (1981) and Frank, Verschoor and Boublik (1981), after receiving his PhD from the Technical University of Munich in 1970. His thesis, which proposed novel methods of aligning carbon films by using cross-correlations, led to his first publication in Optik and a Harkness Fellowship to spend two years at laboratories in Pasadena, Berkeley, and Cornell, working in the field of electron microscopy. He would later become a research scientist at Wadsworth Center. After his work with van Heel, Miloslav Boublik of the Roche Institute of Molecular Biology approached him to try his method on images of eukaryotic ribosomes. The resulting side-view images were greatly enhanced by averaging, leading to the work of Frank, Verschoor and Boublik (1981) and an NIH grant that he has held continuously since 1982. He moved to Columbia University in 2008 as Professor of Biochemistry and Molecular Biophysics and of Biological Sciences, and was awarded the Nobel Prize in Chemistry in 2017.

Marin van Heel finished his PhD thesis in Biophysics in 1981, after returning to the University of Groningen. His advisor was Emi van Bruggen of the

Department of Biochemistry, who had studied hemocyanin (cited in Reference [5] of van Heel and Frank (1981)) and provided the electron micrograph for the statistical analysis. He joined the Fritz Haber Institute in Berlin as head of the interdisciplinary structural biology group from 1982 to 1996, and moved to Imperial College London in 1996 as Professor of Structural Biology.

While still a student, Marin van Heel visited Frank's laboratory at Cornell University's Wadsworth Center in Albany, NY. He was an excellent programmer and contributed to the technique that uses correspondence analysis, as suggested by J. P. Bretaudiere of the NY State Department of Health, to automatically classify molecular images from EM experiments. This approach enabled them to order the particles into subsets, and thus the averages for each subset could be related to different positions of the molecules on the support film. Their experimental observations focused on hemocyanin half-molecules of the horseshoe crab Limulus polyphemus, with each digitized image of a half-molecule (consisting of four hexamers produced by dissociation of the whole molecule) represented by a 32×32 array, and the 46 independent sets of measurements arranged in a 1024×46 input matrix. A PCA is based on a Euclidean distance measure. In contrast, a correspondence analysis uses the $\mathcal{X}^2$ distance as a measure of the proximity between any two elements. The $\mathcal{X}^2$ distances between any two rows or columns of the input matrix are computed, and then the eigenvalues and eigenvectors of the symmetric matrix of these distances are determined and ordered by their relative importance.

Sjors Hendrik Willem Scheres received his undergraduate and doctoral degrees at Utrecht University in the Netherlands. His 2003 DPhil thesis was on conditional optimization for protein structure refinement. He spent 2003–2010 as a postdoctoral researcher at the Spanish National Center for Biotechnology, where he developed algorithms for cryo-EM images based on maximum likelihood estimation. In 2010, he was appointed group leader at the MRC laboratory of Molecular Biology at Cambridge Biomedical Campus of Cambridge University, where his lab specializes in visualizing proteins in health and disease. There, he extended his maximum likelihood methods to a general EB approach, and developed the computer program RELION. Moreover, at the MRC laboratory, he met frequently with Richard Henderson, who received his PhD there in 1970, and later returned as a group leader in 1975, after spending five postdoctoral years at Yale. Henderson would later be awarded the Nobel Prize in Chemistry.

3. Cryo-EM Validation Task Force

Henderson et al. (2012) describe the inaugural meeting of the EM Validation Task Force (`http://www.emdatabank.org/`, website currently under development and available at `http://www.emdataresource.org/`). It was a two-day workshop in September 2010, attended by 28 scientists from 19 academic institutions, worldwide, collecting EM data to determine 3DEM density maps and build molecular models. All 3DEM maps and model include some uncertainty, with five papers between 2002 and 2005 independently reporting different structures on the same receptor complex using the best available methods at the time. Furthermore, the absence of appropriate validation tools has made it impossible to prove whether the structures are correct or incorrect. Moreover, the current increase in the size, productivity, and effect of the 3DEM community require guidelines for validating, annotating, and depositing 3DEM maps and map-derived models. Here, recommendations has been proposed by the Map Group and by the Modeling Group. The Map Group lists validation criteria for experimentalists to use to assess their maps and report the methods used when depositing at EMDataBank. In particular, for a statistical assessment of a map, map variance and local resolution determination, such as bootstrap-based variance maps (Penczek, Frank and Spahn (2006)) and local Fourier shell correlation (FSC) measurements, can provide additional measures to help interpret structures.

Recommendations by the Modeling Group are described by (Henderson et al. (2012, pp. 210–212)). The granularity when representing the same model, the degrees of freedom in the search of a model, and the type of information in addition to the 3DEM map can vary between cryo-EM studies, which may use either a simple set of coordinates or an ensemble of coordinates. As a result, the Modeling Group argues for “criteria for assessing models” with respect to their “accuracy, for the development of methods for estimating model accuracy,” and for the creation of “community-wide benchmarks for modeling methods.”

Figure 1 is a visual presentation of the single particle cryo-EM image analysis workflow and AAV2 cryo-EM map analysis in Tan et al. (2018). The left-hand side column, marked by (A), represents the workflow. The right-hand side figures, marked by (B) and (C), are adapted from Figures 1 and 3 of Tan et al. (2018), with permission under the CC BY license. In Figure 1(B), we check the reliability of a map using the correlations between two maps from half of the data, and show the validation of the model as the FSC curve of Map-to-Model. In Figure 1(C), the maps of representative amino acids with embedded atomic models provide a visual check of the quality of the map, as claimed by the FSC.

4. EB Analysis of Cryo-EM Images in RELION

We begin with a brief review of the cryo-EM image reconstruction problem of estimating the 3D molecular structure $\phi : \mathbb{R}^3 \to \mathbb{R}$ associated with the 3D orientations of the particles embedded in the ice from the 2D images $I_1, \ldots, I_N$; each image I_i is formed by rotating ϕ by a 3D rotation R_{ω_i} and a 2D shift t_i. Although $\omega_1, \ldots, \omega_N$, $t_1, \ldots, t_N$ are unknown a priori, they are nuisance parameters, because their estimation is not the aim of reconstructing ϕ, which is possible up to three intrinsic ambiguities: a global 3D rotation, the (3D) location of the center of the molecule, and handedness. The first two are related to the nuisance parameters, and can be handled using stochastic modeling, as in the expectation-maximization algorithm of RELION. In proteins, the polypeptide chain forms a number of right- and left-handed helices and superhelices; this property is referred to as "handedness" or pseudo-chirality (Efimov (2018)). However, handedness cannot be determined from cryo-EM images alone, because the original 3D structure and its reflection yield identical sets of 2D projections.

Scheres (2012a) gives an overview of the open-source computer program RELION for cryo-EM structure determination, and Scheres (2012b) expresses the reconstruction problem as the optimization of a single target function. The latter notes that a fundamental difficulty with 3D structure reconstruction from cryo-EM data is "the lack of information about the relative orientations of all particles and, in the case of structural variability in the sample, also their assignment to a structurally unique class," because "these data are lost during the experiment, where molecules in distinct comformations coexist in solution and adopt random orientations in the ice." Hence, determining a cryo-EM structure needs to be "tackled by regularization, where the experimental data are complemented by prior information so that the two sources of information together fully determine a unique solution." In practice, experimental data often need to be supplemented with prior information because of a low SNR or an insufficiently large sample size. Thus, a Bayesian approach that assumes a Gaussian distribution on the Fourier components of the signal is used for a maximum a posteriori (MAP) estimation of the latent vector of the actual orientations of the images. MAP is a point estimate defined by the mode (i.e., $\text{argmax}_\theta f(\theta \mid \mathfrak{X}_n)$) of the posterior density $f(\cdot \mid \mathfrak{X}_n)$, given the observed sample $\mathfrak{X}_n$ of size n. Hence, it is the solution of a regularized likelihood maximization problem. For a loss function of the form $L(\phi, a) = \mathrm{I}_{\{\|\phi - a\| < c\}}$, the Bayes estimate $\hat{\phi}_c$ approaches MAP as $c \to 0$. Scheres (2012b, pp. 521–525) uses the Dempster–Laird–Rubin expectation-maximization

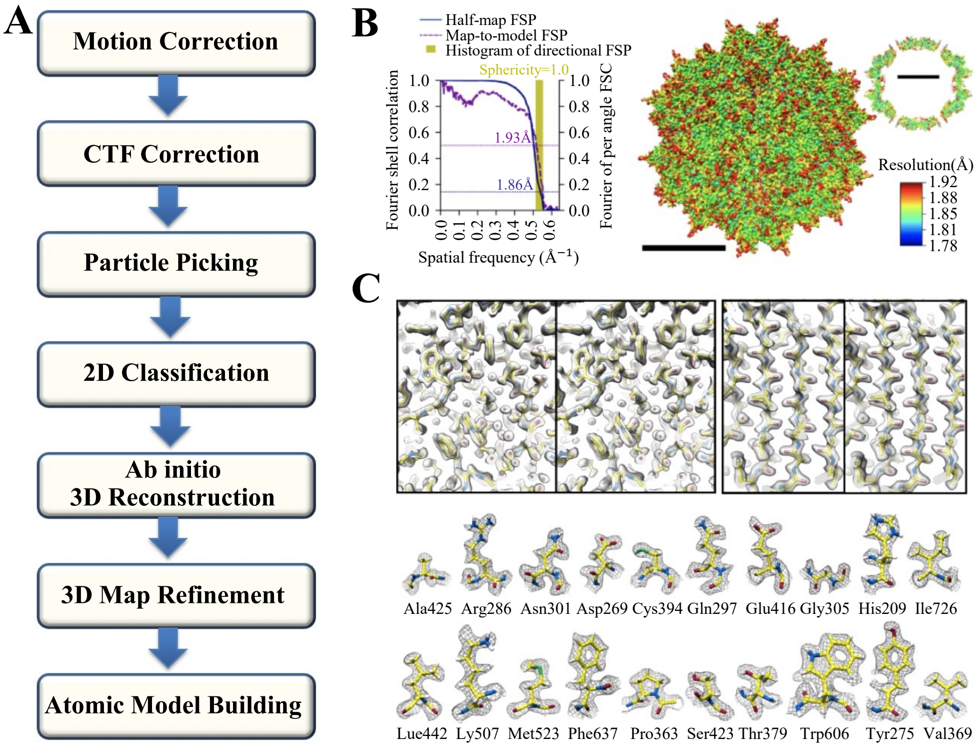

Figure 1. (**A**) An image processing workflow of single-particle cryo-EM reconstruction. (**B**) Sub-2 Åreconstruction of the AAV2 viral capsid using single-particle cryo-EM from Dmitry Lyumkis lab, adopted from Tan et al. (2018). In the upper panel, the right presents the reconstruction colored by local resolution (Hohn et al. (2007)). Both scale bars correspond to 100 Å. the left presents the Fourier shell correlation (FSC) curves describing the half-map (blue solid line) and map-to-model (purple solid line) resolutions, as well as a histogram of directional resolutions sampled evenly over the 3DFSC (Tan et al. (2017)) (yellow), and corresponding sphericity value. The resolution cut-offs of 0.143 (blue dotted line) and 0.5 (purple dotted line) are used. (**C**) Stereo view of a slice through the map and model containing both amino-acid residues and water molecules, and a stereo view through a beta sheet. Map densities for each of the 20 types of amino-acid residues, shown as a stick representation and colored according to atom type: C = yellow, O = red, N = blue, S = green, inside either a translucent solid density or black mesh density map. These figures are adopted from Tan et al. (2018), `https://doi.org/10.1038/s41467-018-06076-6`, with the permission under the terms of the Creative Commons CC BY license.

algorithm to evaluate MAP, using fast Fourier-space interpolation and adaptation to speed up the computations. Scheres (2016) reviews the processing of structurally heterogeneous cryo-EM data in RELION. He recognizes that the Bayesian approach in Scheres (2012a,b) is actually EB regularization, because the hyperparameters in the prior model are replaced by their estimates in the regularized likelihood: “Whereas in standard Bayesian methods the prior is fixed before any data are observed, inside RELION parameters of the prior are estimates from the data themselves. This type of algorithm is referred to as an empirical Bayes approach,” in which “both the likelihood and the prior are expressed in the Fourier domain” that “permits a convenient description of the effects of microscope optics and defocusing (by the so-called contrast transfer function, or CTF).” He notes that using the expectation-maximization algorithm to compute the MAP estimate in the Fourier domain “results in an update formula for the reconstruction that shows strong similarities with previously introduced Wiener filters,” which depend on estimates for the power and the noise as functions of spatial frequency. By using the EB approach to update these estimates from the data, “RELION effectively calculates the best possible filter, in the sense that it yields the least noisy reconstruction, at every iteration of the optimization process.” As pointed out by Scheres (2012b, p. 520), the MAP estimate is based on the following linear regression model in the Fourier space for the kth homogeneous structure group ($k = 1, \ldots, K$):

$$x_{ij} = c_{ij} \sum_{\ell=1}^{L} P_{j\ell}^{\phi_i} s_{k\ell} + \epsilon_{ij},$$

where ϕ_i contains the orientation parameters of the projection matrix P for the ith image of the particles in the kth homogeneous structure group, x_{ij} is the jth component ($j = 1, \ldots, J$) of the 2D Fourier transform of the ith image ($i = 1, \ldots, N$), c_{ij} is the jth component of the CTF for the ith image, $s_{k\ell}$ is the ℓth component of the 3D Fourier transform of the kth structure group, ϵ_{ij} is noise (usually assumed to be independent and normally distributed with mean zero and variance σ^2) in the complex plane, and $P^{\phi_i} = (P_{j\ell}^{\phi_i})_{1 \leq j \leq J, 1 \leq \ell \leq L}$ is a matrix that relates the 2D Fourier transform of the ith particle image to the 3D Fourier transform using the projection-slice theorem (also called the Fourier slice theorem). This theorem states that the 2D Fourier transform of the projection of ϕ_i belonging to a 3D manifold is the restriction of the 3D Fourier transform to a 2D plane. Hence, we can estimate s_k to a certain resolution if we have enough projections from known viewing directions. An example of mixed homogeneous structures from Sobti et al. (2020) is provided in Figure 2. For a review of the algorithms used to separate homogeneous structures, refer to Chang et al. (2021).

In the RELION framework, the user manually picks a few hundred particles (i.e., 2D projections from the noisy cryo-EM micrographs). The images are then grouped into 2D classes, which are used as templates for automatic template matching. However, Scheres (2015, pp. 120–121) notes that this does not pre-

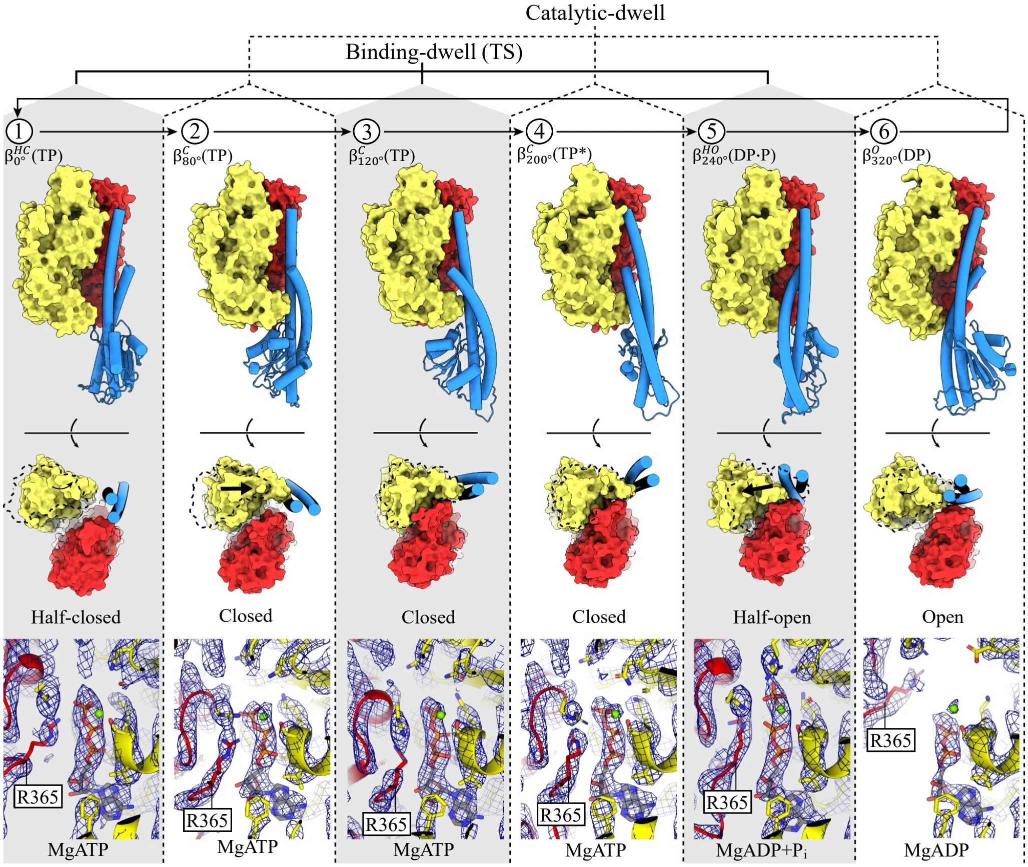

Figure 2. The six sequential conformations in the F1-ATPase rotary catalytic cycle represent six homogeneous cryo-EM structures (Sobti et al. (2020)). Top, $\alpha\beta$-pairs superimposed on the N termini ($\beta 2 - 82$) and viewed from the side (perpendicular to the membrane) and below (from the membrane). Subunit α in red, β in yellow, and γ in blue, with stencil outline of the β-subunit from the previous step in the scheme for comparison. Bottom, close-up of the catalytic nucleotide-binding sites, superimposed on residues around the nucleotides (β158-166, β336-342, and β412-421). Cryo-EM map shown as blue mesh. Nucleotides, Mg_2^+, and Pi are shown as sticks with CPK coloring and αR365 (the arginine finger) labeled in the movement from the binding dwell (TS) (states 1, 3, and 5) to the catalytic dwell (states 2, 4, and 6). This figure is adopted from Sobti et al. (2020), `https://doi.org/10.1038/s41467-020-16387-2`, without modification, with permission under the terms of the Creative Commons CC BY license.

clude the possible “Einstein from noise” problem raised by Stewart and Grigorieff (2004), and used by Henderson (2013) to emphasize the importance of verifying that a reconstructed 3D image is a reliable and faithful representation of the underlying molecule.

5. HMM-Based Approach to Cryo-EM Image Reconstruction with Uncertainty Quantification

Using the EB approach in RELION and the iterative expectation maximization algorithm of Dempster, Rubin, and Laird to simultaneously estimate the state vector and the hyperparameter vector seems appropriate when the observed data are measurements related to the latent states of an HMM with unknown parameters and we wish to jointly estimate the state and parameter (or adaptive filtering) in the HMM, which is a long-standing problem in STEM (Science, Technology, Engineering, Mathematics) fields. Lai et al. (2020) recently developed the MCMC-SS method for adaptive filtering in HMMs on general state spaces; see also Lai (2021). Their basic idea is to approximate a target distribution by the empirical distribution of M representative atoms, chosen sequentially using an MCMC scheme so that the distribution approximates the target distribution after a large number K of iterations. Using bounds on a weighted total variation norm of the difference between the target distribution and the empirical measure defined by the sample paths of the MCMC scheme, they also developed asymptotic theory for the MCMC-SS. This theory includes the asymptotic normality (as both K and M approach ∞) of the estimates of the functionals of the target distribution using MCMC-SS, together with a consistent estimation of their standard errors, and provides oracle properties that prove their asymptotic optimality. In particular, the convergence is guaranteed and automated for MCMC-SS, in contrast to standard MCMC schemes, which need manual checks of convergence. Moreover, the computation can be vectorized and accelerated using a GPU, and parallelized across multiple GPUs.

Wu et al. (2021) apply the MCMC-SS to uncertainty quantification in image reconstruction. They refer to Cotter et al. (2013), who propose using MCMC methods "whenever the target measure has density with respect to a Gaussian process or Gaussian random field reference measure." Numerous applications involving such a framework consider a Bayesian inference on a latent random field $\{u(x) : u \in D\} \subset \mathbb{R}^d$ generated by some stochastic partial differential equation (SPDE) in which D is a connected subset of $\mathbb{R}^d$, based on data generated by some nonlinear function of the random field. It is shown that after discretization and truncation to fit into this framework, the Radon–Nikodym derivative of the target measure P with respect to the reference measure Q has the form

$$\left(\frac{dQ}{dP}\right)(u) \propto \exp(-\ell(u)), \tag{5.1}$$

for some real-valued function ℓ, which Cotter et al. (2013) call "potential" in their substantive applications. The advantage of using a zero-mean Gaussian random field reference measure Q is that it is specified by the covariance operator C, the eigenvalues λ_i and orthonormal eigenfunctions ϕ_i of which yield the Karhunen–Loeve expansion $u(x) = \sum_{i=1}^{\infty} \xi_i \phi_i(x)$, with i.i.d. ξ_i that are $N(0, \lambda_i^2)$ and $\sum_{i=1}^{\infty} \lambda_i^2 < \infty$. Cotter et al. (2013) use a random truncation τ with a sieve prior to convert the infinite-dimensional expansion to a finite sum $u(x) = \sum_{i=1}^{\infty} \xi_i \phi_i(x)$. In addition, they use a discrete approximation of the random field $u(x)$, with x taken over a mesh of width δ in each coordinate.

MCMC-SS uses a parametric family of Gaussian proposal measures $Q(\gamma)$, rather than the a single Q in Cotter et al. (2013). Putting $1/L(\theta) = \exp(-\ell(u(x)))$, Wu et al. (2021) incorporate the random truncation τ and possibly also other random effects ρ into the state $\theta = (\tau, \zeta_1, \ldots, \zeta_\tau, \rho)$, where $\zeta_j = \mathcal{G}(u(x_j))$, for $j = 1, \ldots, \tau$, and $\mathcal{G}$ is an operator associated with the SPDE and the discretization scheme for which x_j belongs to a discrete subset of D. With this definition of θ, MCMC-SS uses the updating procedure described in the Supplementary Material S2. Cotter et al. (2013, Sec. 4.2) argue that simply applying an MCMC to a discretized random field leads to a singular reference measure with respect to the target measure. However, the MCMC procedure they consider is the random walk Metropolis algorithm, which involves the acceptance probability $a(u, v) = \min\{1, (d\eta^*/d\eta)(u, v)\}$, where η is a measure defined by the transition kernel $q(u, v)$ of the MCMC algorithm (i.e., $v \mid u \sim q(u, \cdot)$), and η^* is the measure obtained by reversing the roles of u and v in the definition of η. Their Theorem 6.3 shows that after discretization, η^* is singular with respect to η and, therefore, "all proposal moves are rejected with probability 1" for the random walk Metropolis algorithm, which proposes $v^{(k)} = u^{(k)} + \beta \xi^{(k)}$, with $\xi^{(k)} \sim N(0, C)$, and chooses $u^{(k+1)} = v^{(k)}$ with probability $a(u^{(k)}, v^{(k)})$, setting $u^{(k+1)} = u^{(k)}$ if $v^{(k)}$ is rejected. To get around this difficulty, they introduce a preconditioned Crank–Nicolson (pCN) adjustment, which proposes $v^{(k)} = \sqrt{1-\beta^2} u^{(k)} + \beta \xi^{(k)}$. Here, $\beta = 8\delta/(2+\delta)^2$ and C is the covariance matrix (after truncation and discretization) of the covariance operator C for the Gaussian proposal measure. Because MCMC-SS does not involve η and η^*, it does not require the pCN adjustments. Cryo-EM image analysis actually involves a multilevel Bayesian model, with the first level consisting of homogeneous segments. Moreover, the homogeneous segments have been determined by Chung et al. (2020) using a two-stage dimension reduction analysis procedure based on a principal component analysis of the order-two tensors introduced by Hung et al. (2012).

For the homogeneous segments in the first level of the EB model, we rephrase the hybrid resampling approach to the credible intervals (i.e., posterior confidence intervals) of Chen and Lai (2007) for change-point ARX models and those of Dai and Tsang (2021) for change-point ARX-GARCH models in the context of 3D cryo-EM image reconstruction. We begin with a brief review of the exact, bootstrap, and hybrid resampling methods for constructing confidence intervals for univariate functionals of a parameter belonging to some (possibly multidimensional) Euclidean space. If $\{F_\theta : \theta \in \Theta\}$ is a family of distribution functions indexed by a real-valued parameter θ, an exact equal-tailed confidence set can always be found via the duality between the hypothesis tests and the confidence sets: Let $R(\boldsymbol{Y}, \theta_0)$ be some real-valued test statistic of the null hypothesis $\theta = \theta_0$ based on the observed data vector $\boldsymbol{Y}$. Let $u_\alpha(\theta_0)$ be the α-quantile of $R(\boldsymbol{Y}, \theta_0)$ under the null hypothesis, which is accepted at level 2α if $u_\alpha(\theta_0) < R(\boldsymbol{Y}, \theta_0) < u_{1-\alpha}(\theta)$. An exact equal-tailed confidence set with coverage probability $1-2\alpha$ consists of all θ_0 not rejected by the test, and is therefore given by $\{\theta : u_\alpha(\theta) < R(\boldsymbol{Y}, \theta) < u_{1-\alpha}(\theta)\}$. The exact method assumes no nuisance parameters. The bootstrap method replaces the quantiles $u_\alpha(\theta)$ and $u_{1-\alpha}(\theta)$ with the approximate quantiles u^*_α and $u^*_{1-\alpha}$, respectively, obtained in the following manner. Based on $\boldsymbol{Y}$, the quantile u^*_α is defined as the α-quantile of the distribution of $R(\boldsymbol{Y}^*, \hat{\theta})$, with $\boldsymbol{Y}^*$ generated from the empirical distribution $\hat{F}$ of $\boldsymbol{Y}$ and $\hat{\theta} = \theta(\hat{F})$, yielding the confidence set $\{\theta : u^*_\alpha < R(\boldsymbol{Y}, \theta) < u^*_{1-\alpha}\}$, analogously to the exact method. The hybrid resampling method, introduced by Chuang and Lai (2000), is a hybrid of the exact and bootstrap methods for constructing confidence sets when the data-generating process is too complex for us to apply the exact approach and for the bootstrap method to be valid. It uses a family of distributions $\{\hat{F}_\theta, \theta \in \Theta\}$ as the resampling family, in which θ is the unknown parameter of interest. Let $\hat{u}_\alpha(\theta)$ be the α-quantile of the sampling distribution of $R(\boldsymbol{Y}, \theta)$ under the assumption that $\boldsymbol{Y}$ has distribution $\hat{F}_\theta$. The hybrid confidence set, with approximate coverage probability $1 - 2\alpha$, is given by $\{\theta : \hat{u}_\alpha(\theta) < R(\boldsymbol{Y}, \theta) < \hat{u}_{1-\alpha}(\theta)\}$. Sections 7.2–7.5 of Bartroff, Lai and Shih (2013) describe the development, up to 2012, of hybrid resampling, including valid confidence intervals for population means following group sequential tests, secondary endpoints, and time-sequential survival outcomes; for more recent work, see Dai and Tsang (2021), and the references therein. The change-point model for cryo-EM image segmentation adopts the EB model of multiple change-points of Lai and Xing (2011), cited by Dai and Tsang (2021).

The model assumes an exponential family of density functions $f_{\boldsymbol{\theta}}(\boldsymbol{y}) = \exp(\boldsymbol{\theta}^\top \boldsymbol{y} - \psi(\boldsymbol{\theta}))$ with respect to some measure on $\mathbb{R}^d$ for observation $\boldsymbol{y}$, a prior density function

$$\pi_{a_0,\boldsymbol{\mu}_0} = c(a_0, \boldsymbol{\mu}_0) \exp\{a_0 \boldsymbol{\mu}_0^\top \boldsymbol{\theta} - a_0 \psi(\boldsymbol{\theta})\}$$

on the parameter space $\Theta \in \mathbb{R}^d$. The posterior density of $\boldsymbol{\theta}$ given observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ is also of this form and, therefore, $\pi_{a,\boldsymbol{\mu}}$ is a conjugate family of prior density functions so that a_0 in $\pi_{a_0,\boldsymbol{\mu}_0}$ can be interpreted as an additional sample size, and $\boldsymbol{\mu}_0$ as the mean of that sample. In addition, the parameter vector $\boldsymbol{\theta}_t$ may undergo occasional changes such that for $t > 1$, the indicator variables $I_{\{\boldsymbol{\theta}_t \neq \boldsymbol{\theta}_{t'}\}}$ are independent Bernoulli with $\mathrm{P}(I_t = 1) = p$. Note that $\boldsymbol{\mu}_t = \nabla\psi(\boldsymbol{\theta}_t)$. Thus, the hyperparameters of the change-point model are $a_0, \boldsymbol{\mu}_0$, and p.

Given the hyperparameters, there are explicit formulae for the posterior distribution of $\boldsymbol{\theta}_t$ given $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, for $1 \le t \le n$; see Section 2.2 of Lai and Xing (2011). By combining the recursive formulae for forward filters in their Section 2.1 with corresponding recursions for backward filters, the posterior density of $\boldsymbol{\theta}_t$ given $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ is the mixture of $\pi_{a_0+j-i+1,\bar{Y}_{i,j}}$ over $i \le t \le j$:

$$\sum_{1 \le i \le t \le j \le n} \beta_{ijt}\, \pi_{a_0+j-i+1.\bar{Y}_{i,j}}(\boldsymbol{\theta}_t),$$

where $\bar{Y}_{i,j} = (a_0\boldsymbol{\mu}_0 + \sum_{k=i}^{j} \boldsymbol{y}_k)/(a_0 + j - i + 1)$, for $j \ge i$, and $\beta_{ijt} = \beta^*_{ijt}/(p + \sum_{1 \le i' \le t \le j' \le n} \beta^*_{i'j't})$, with

$$\beta^*_{ijt} = \begin{cases} p\dfrac{c(a_0, \boldsymbol{\mu}_0)}{c(a_0 + 1, \bar{Y}_{t,t})} & \text{if } i \le t = j \\ (1-p) q^*_{j,t+1}\, p^*_{i,t} & \text{if } i \le t < j, \end{cases}$$

in which $p^*_{i,t}$ is given by the recursion

$$p^*_{i,t} = (1-p)\left(\frac{p^*_{i,t-1}}{\sum_{k=1}^{t-1} p^*_{k,t-1}}\right)\frac{c(a_0 + i - t + 1, \bar{Y}_{i,t-1})}{c(a_0 + i - t, \bar{Y}_{i,t})},$$

and $q^*_{j,t+1}$ is given by the recursion

$$(1-p)\frac{q^*_{j,t+1}}{\sum_{k=t+1}^{n} q^*_{k,t+1}} = q^*_{j,t}\frac{c(a_0 + j - t + 1, \bar{Y}_{t,j})}{c(a_0 + j - t, \bar{Y}_{t+1,j})}.$$

Therefore, we can apply MCMC-SS to jointly estimate the hyperparameters $a_0, \boldsymbol{\mu}_0$, and p and the states using the mixture of $\pi_{a_0+j-i,\bar{Y}_{i,j}}$ with the weight β_{ijt}, as in Chen and Lai (2007, Sec. II B and III A) and Dai and Tsang (2021, Sec. 2.2, 2.3, and 2.4). The hyperparameters a_0, $\boldsymbol{\mu}_0$, and p can be estimated consistently by applying the method of moments, described by Lai et al. (2020) and Lai (2021, Sec. 3) for this simulation-based method to implement recursive MCMC-SS particle filters.

The EB approach to cryo-EM image analysis is associated with a multilevel Bayesian model, where the first level consists of homogeneous segments, and the second level specifies the prior distributions of the hyperparameters of each homogeneous segment. This is akin to Yao's (1984) EB estimation of a step function (with multiple change-points) when we observe a step function plus Gaussian noise, and its subsequent refinements and extensions; see Chen and Lai (2007), Lai, Xing and Zhang (2008), Lai and Xing (2011, 2013), and Dai and Tsang (2021).

6. Supplementary Materials

The online Supplementary Material contains an introduction to model bias in image alignment and HMM-based approach to Cryo-EM image Reconstruction with Uncertainty Quantification.

7. Acknowledgments

Lai acknowledges the support of the National Science Foundation under grant DMS-1811818, and helpful comments from Wah Chiu, the Wallenberg-Bienstock Professor of Bioengineering, Microbiology, and Immunology, and the SLAC Cryo-EM Center. Tu acknowledges the support proposed by Academia Sinica under grant AS-IA-110-M05 and the Ministry of Science and Technology under grant 109-2118-M-001-006, as well as critical discussions with Sunney I. Chan, the Distinguished Visiting Fellow of Institute of Chemistry, Academia Sinica, and the Distinguished Chair Professor of Department of Chemistry, National Taiwan University. Chung acknowledges the support of the Ministry of Science and Technology under grant 110-2118-M-110 -003. Wang acknowledges the support of the Ministry of Science and Technology under grant 109-2118-M-008-005.

References

Bartroff, J., Lai, T. L. and Shih, M. C. (2013). *Sequential Experimentation in Clinical Trials: Design and Analysis.* Springer, New York.

Chang, W.-H., Huang, S.-H., Lin, H.-H., Chung, S.-C. and Tu, I.-P. (2021). Cryo-EM analyses permit visualization of structural polymorphism of biological macromolecules. *Frontiers in Bioinformatics* **1**, 788308.

Chen, Y. and Lai, T. L. (2007). Identification and adaptive control of change-point ARX models via Rao-Blackwellized particle filters. *IEEE Transactions on Automatic Control* **52**, 67–72.

Chuang, C. S. and Lai, T. L. (2000). Hybrid resampling methods for confidence intervals, with discussion and rejoinder. *Statistica Sinica* **10**, 1–50.

Chung, S. C., Wang, S. H., Niu, P. Y., Huang, S. Y., Chang, W. H. and Tu, I. P. (2020). Two-stage dimension reduction for noisy high-dimensional images and application to cryogenic electron microscopy. *Annals of Mathematical Sciences and Applications* **5**, 283–316.

Cotter, S. L., Roberts, G. O., Stuart, A. M. and White, D. (2013). MCMC methods for functions: Modifying old algorithms to make them faster. *Statistical Science* **28**, 424–446.

Dai, W. and Tsang, K. W. (2021). Hybrid resampling confidence intervals for change-point or stationary high-dimensional stochastic regression models. *Statistica Sinica* **31**, 1–17.

Efimov, A. V. (2018). Chirality and handedness of protein structures. *Biochemistry (Moscow)* **83**, S103–S110.

Frank, J., Verschoor, A. and Boublik, M. (1981). Computer averaging of electron micrographs of 40s ribosomal subunits. *Science* **214**, 1353–1355.

Habeck, M. (2017). Bayesian modeling of biomolecular assemblies with cryo-EM maps. *Frontiers in Molecular Biosciences* **4**, 15.

Henderson, R. (2013). Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 18037–18041.

Henderson, R., Sali, A., Baker, M. L., Carragher, B., Devkota, B., Downing, K. H. et al. (2012). Outcome of the first electron microscopy validation task force meeting. *Structure* **20**, 205–214.

Hohn, M., Tang, G., Goodyear, G., Baldwin, P. R., Huang, Z., Penczek, P. A. et al. (2007). SPARX, a new environment for Cryo-EM image processing. *Journal of Structural Biology* **157**, 47–55.

Hung, H., Wu, P.-S., Tu, I P. and Huang, S.-Y. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika* **99**, 569–583.

Lai, T. L. (2021). Recursive particle filters for joint state and parameter estimation in hidden Markov model with multifaceted applications. In *Proceedings of International Congress of Chinese Mathematicians, Beijing 2019.* International Press of Boston, Somerville.

Lai, T. L. and Xing, H. (2011). A simple Bayesian approach to multiple change-points. *Statistica Sinica* **21**, 539–569.

Lai, T. L. and Xing, H. (2013). Stochastic change-point ARX-GARCH models and their applications to econometric time series. *Statistica Sinica* **23**, 1573–1594.

Lai, T. L., Xing, H. and Zhang, N. (2008). Stochastic segmentation models for array-based comparative genomic hybridization data analysis. *Biostatistics* **9**, 290–307.

Lai, T. L., Xu, H., Zhu, M. H. and Chan, H. P. (2020). *MCMC with Sequential Substitutions for Joint State and Parameter Estimation in Hidden Markov Models.* Technical Report. Department of Statistics, Stanford University, Stanford.

Lederman, R. R., Anden, J. and Singer, A. (2019). Hyper-molecules: On the representation and recovery of dynamical structures, with application to flexible macro-molecular structures in cryo-EM. *arXiv:1907.01589v1.*

Liao, M., Cao, E., Julius, D. and Cheng, Y. (2013). Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**, 107–112.

Mossman, K. (2007). Profile of joachim frank. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 19668–19670.

Penczek, P. A., Frank, J. and Spahn, C. M. (2006). A method of focused classification, based on the bootstrap 3D variance analysis, and its application to EF-G-dependent translocation. *Journal of Structural Biology*, 184–194.

Scheres, S. H. W. (2012a). A Bayesian view on cryo-EM structure determination. *Journal of Molecular Biology* **415**, 406–418.

Scheres, S. H. W. (2012b). RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology* **180**, 519–530.

Scheres, S. H. W. (2015). Semi-automated selection of cryo-EM particles in RELION-1.3. *Journal of Structural Biology* **189**, 114–122.

Scheres, S. H. W. (2016). Processing of structurally heterogeneous cryo-EM data in RELION. *Methods in Enzymology* **579**, 125–157.

Sigworth, F. (1998). A maximum-likelihood approach to single-particle image refinement. *Journal of Structural Biology* **122**, 328–339.

Sobti, M., Walshe, J., Wu, D., Ishmukhametov, R., Zeng, Y. C., Robinson, C. V. et al. (2020). Cryo-EM structures provide insight into how E. coli F1Fo ATP synthase accommodates symmetry mismatch. *Nature Communications* **11**, 2615.

Stewart, A. and Grigorieff, N. (2004). Noise bias in the refinement of structures derived from single particles. *Ultramicroscopy* **102**, 67–84.

Tan, Y. Z., Aiyer, S., Mietzsch, M., Hull, J. A., McKenna, R., Grieger, J. et al. (2018). Sub-2 Å Ewald curvature corrected structure of an AAV2 capsid variant. *Nature Communications* **9**, 3628.

Tan, Y. Z., Baldwin, P. R., Davis, J. H., Williamson, J. R., Potter, C. S., Carragher, B. et al. (2017). Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nature Methods* **14**, 793–796.

van Heel, M. and Frank, J. (1981). Use of multivariate statistics in analysing the images of biological macromolecules. *Ultramicroscopy* **6**, 187–194.

Wang, S. H., Yao, Y. C., Chang, W. H. and Tu, I P. (2021). Quantification of model bias underlying the phenomenon of "Einstein from noise". *Statistica Sinica* **31**, Online Special Issue, 2355–2379.

Wu, H. T., Lai, T. L., Haddad, G. G. and Muotri, A. (2021). Oscillatory biomedical signals: Frontiers in mathematical models and statistical analysis. *Frontiers in Applied Mathematics and Statistics* **7**. doi: `10.3389/fams.2021.689991`.

Yan, C., Hang, J., Wan, R., Huang, M., Wong, C. C. and Shi, Y. (2015). Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* **349**, 1182–1191.

Yao, Y. C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annals of Statistics* **12**, 1434–1447.

Tze Leung Lai

Department of Statistics, Stanford University, Stanford, CA 94305, USA.

E-mail: lait@stanford.edu

Shao-Hsuan Wang

Institute of Statistics, National Central University, Zhongli District, Taoyuan City 320317, Taiwan.

E-mail: pico@stat.sinica.edu.tw

Szu-Chi Chung

Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung 804201, Taiwan.

E-mail: steve2003121@gmail.com

Wei-hau Chang

Institute of Chemistry, Academia Sinica, Nankang, Taipei 11529, Taiwan.

E-mail: weihau@chem.sinica.edu.tw

I-Ping Tu

Institute of Statistical Science, Academia Sinica, Nankang, Taipei 11529, Taiwan.

E-mail: iping@stat.sinica.edu.tw

(Received November 2021; accepted August 2022)