RESEARCH ARTICLE

Massively parallel unsupervised single-particle cryo-EM data clustering via statistical manifold learning

Jiayi Wu[1,2], **Yong-Bei Ma**[2], **Charles Congdon**[3], **Bevin Brett**[3], **Shuobing Chen**[1,2], **Yaofang Xu**[2,4], **Qi Ouyang**[1,5], **Youdong Mao**[1,2,6]*

1 State Key Laboratory for Artificial Microstructure and Mesoscopic Physics, Institute of Condensed Matter Physics, School of Physics, Center for Quantitative Biology, Peking University, Beijing, China, **2** Intel Parallel Computing Center for Structural Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, **3** Software and Services Group, Intel Corporation, Santa Clara, California, United States of America, **4** Department of Biophysics, Peking University Health Science Center, Beijing, China, **5** Peking-Tsinghua Joint Center for Life Sciences, Peking University, Beijing, China, **6** Department of Microbiology and Immunobiology, Harvard Medical School, Boston, Massachusetts, United States of America

* youdong_mao@dfci.harvard.edu

Abstract

Structural heterogeneity in single-particle cryo-electron microscopy (cryo-EM) data represents a major challenge for high-resolution structure determination. Unsupervised classification may serve as the first step in the assessment of structural heterogeneity. However, traditional algorithms for unsupervised classification, such as K-means clustering and maximum likelihood optimization, may classify images into wrong classes with decreasing signal-to-noise-ratio (SNR) in the image data, yet demand increased computational costs. Overcoming these limitations requires further development of clustering algorithms for high-performance cryo-EM data processing. Here we introduce an unsupervised single-particle clustering algorithm derived from a statistical manifold learning framework called generative topographic mapping (GTM). We show that unsupervised GTM clustering improves classification accuracy by about 40% in the absence of input references for data with lower SNRs. Applications to several experimental datasets suggest that our algorithm can detect subtle structural differences among classes via a hierarchical clustering strategy. After code optimization over a high-performance computing (HPC) environment, our software implementation was able to generate thousands of reference-free class averages within hours in a massively parallel fashion, which allows a significant improvement on *ab initio* 3D reconstruction and assists in the computational purification of homogeneous datasets for high-resolution visualization.

Introduction

Single-particle cryo-EM is evolving into a mainstream approach in visualizing the three-dimensional (3D) structures of biomolecules in their native functional states at near-atomic

resolutions [1,2]. Since individual biomolecules may sample multiple conformations, a prerequisite for high-resolution structure determination is to obtain a 'pure' dataset that only includes particle images in the same conformational state [3]. To distinguish different conformations in raw particle images, unsupervised two-dimensional (2D) clustering is commonly used as the first step in the evaluation of structural heterogeneity [1,4–6], or used as an intermediate step during the alteration of 2D and 3D classifications for *in silico* purification [7,8]. Besides, high-quality unsupervised 2D class averages are essential for reliable initial reconstruction and its verification [1,9,10].

Although several classification schemes were proposed for single-particle analysis [11–15], there are mainly two approaches for unsupervised 2D clustering of single-particles implemented in multiple end-user software packages [16–21] and widely used in cryo-EM structure determination: (1) K-means clustering following reference-free alignment through cross-correlation (CC) and multivariate statistical analysis (MSA) [1,22–24], and (2) unsupervised maximum-likelihood (ML) or *maximum a posteriori* (MAP) classification [25–27]. In the former approach, the classification accuracy is affected by the noise-induced misalignment resulting from false peaks in cross-correlation computation. Noise may also introduce errors in the distance calculation in K-means clustering. Hence, its performance is dramatically reduced as the signal-to-noise ratio (SNR) decreases. By contrast, the ML-based approach explores optimal probability in measuring image similarity, and exhibits robust resistance to noise-induced misalignment [27]. However, a prominent drawback lies in that the likelihood matching insufficiently differentiates structural heterogeneity among similar but critically different views. In each ML-classified group of single-particles one could find a mixture of heterogeneous 2D projection structures with a large variation of likelihood. This effect causes a decrease in the effective number of classes with increasing the number of iterations during ML optimization. To date, there has been a lack of unsupervised single-particle clustering methods that can efficiently sort out highly heterogeneous single-particles into thousands of homogenous 2D classes while still keeping computation efficient [5].

Manifold learning encompasses the disciplines of geometry, computation, and statistics, and has become an important research frontier in data mining and machine learning. It is a class of algorithms devised for recovering a low-dimensional manifold embedded in a high-dimensional data space. Perhaps the most popular algorithm for linear dimensionality reduction is principal component analysis (PCA). Given a data set, PCA finds the directions along which the data have maximum variance in addition to the relative importance of these directions. Most real-world high-dimensional data are intrinsically governed by hidden variables through nonlinear relationships. When linear approximation in dimensionality reduction fails in finding a good low-dimensional representation of high-dimensional data, nonlinear dimensionality reduction can be exploited as an alternative solution. Major breakthroughs in methods for recovering low-dimensional nonlinear embeddings of high-dimensional data [28,29] have led to the construction of a number of algorithms carrying out nonlinear manifold learning. Unlike PCA, nonlinear manifold learning attempts to use nonlinear kernel functions or mapping to find the directions along which the data have significant variance. To date, several nonlinear manifold learning frameworks have been proposed, such as isometric feature mapping (Isomap) [28,30–32], generative topographic mapping (GTM) [33–35], locally linear embedding (LLE) [29,36,37], and semidefinite embedding (SDE) [38]. Furthermore, several algorithms for manifold learning have been successfully applied to high-dimensional data for specific tasks, such as discriminant image clustering [39], image feature extraction [40] and person-independent precise 3D pose estimation [41].

Recently, manifold embedding via diffusion map was applied to study the continuous conformational changes of the 80S ribosome by single-particle cryo-EM [42]. In this approach, a

one-dimensional manifold was used to describe a continuous conformational spectrum that can be discerned through a narrow angular aperture. The approach begins with a classification of data into different orientations with respect to a common 3D reference, with the assumption that the changes in the structure are relatively small so that a projection-matching based refinement algorithm with a single 3D reference can be used. However, this assumption does not necessarily hold for molecules exhibiting dramatic conformational changes, in which a small angular aperture is not sufficient to distinguish the conformational spectrum. Hence, such an approach is yet to be adapted for more general conditions.

Here we introduce an unsupervised clustering method for single-particle cryo-EM data based on GTM [33–35], which is a framework for statistical manifold learning, without any assumptions regarding how the molecules assume their conformations in the cryo-EM data [33–35]. We implemented the GTM-based unsupervised single-particle clustering algorithm in a software package named ROME (Refinement and Optimization based on Machine lEarning), which has been optimized for both Intel$^{®}$ Xeon$^{®}$ multi-core processors and Intel$^{®}$ Many-Integrated Core (MIC) architecture [43] to enable efficient computation of thousands of reference-free class averages in a highly affordable fashion. Algorithmic implementations of unsupervised clustering methods in the existing software packages [16–21] normally allow the efficient computation of tens to a few hundred of classes. In this study, through adapting GTM to cryo-EM image data and optimizing its computational efficiency against Intel$^{®}$ MIC architecture, we can achieve unsupervised data clustering with the number of classes on the order of magnitude of 10^3 or higher within several hours in a highly parallel fashion. Our unsupervised clustering approach can markedly improve the quality and resolution of *ab initio* 3D models with angular reconstitution[44]. We further tested our approach using several cryo-EM datasets to demonstrate the advantage of GTM-based unsupervised clustering in discerning subtle structural differences directly from 2D class averages corresponding to distinct conformations.

Methods

Mathematical model for statistical manifold learning

The goal of GTM is to find a representation for the distribution $p(t)$ of a dataset in a J-dimensional data space $t = (t_1, \ldots, t_J)$ in terms of L-dimensional latent variables $s = (s_1, \ldots, s_L)$. One may consider a non-linear, parametric function $t = \mathbf{A}(s; \mathbf{W})$, which maps points s in the latent space into corresponding points $\mathbf{A}(s; \mathbf{W})$ in the data space [33] (Fig 1A). The mapping is controlled by a set of parameters $\mathbf{W}$, which represent the weights and biases in the case of a feed-forward neural network as the mapping. In the situation in which the dimensionality L of the latent space is less than the dimensionality J of the data space, the nonlinear transformation $\mathbf{A}(s; \mathbf{W})$ maps the latent space onto an L-dimensional non-Euclidean manifold $\mathcal{S}$ embedded within the data space. Previous studies have established GTM as an alternative to self-organizing maps (SOM) [45] and that the GTM framework overcomes most limitations of SOM while introducing no significant disadvantages [33,34].

To adapt the general framework of GTM to the single-particle cryo-EM data clustering problem, we built the contrast transfer function (CTF) into the non-linear mapping function in Fourier or reciprocal space, which allows for an efficient correction of the aberration effect of the objective lenses in electron microscopy [1]. A vector in data space x_i represents the Fourier transform of a particle image. A latent variables s reflects the inherent degrees of freedom controlling the 2D structural differences observed in single-particle images, which arise either from distinct conformational states of the imaged biomolecules or from different viewing angles. For the sake of simplicity, we used t_i to represent the translated and rotated version of
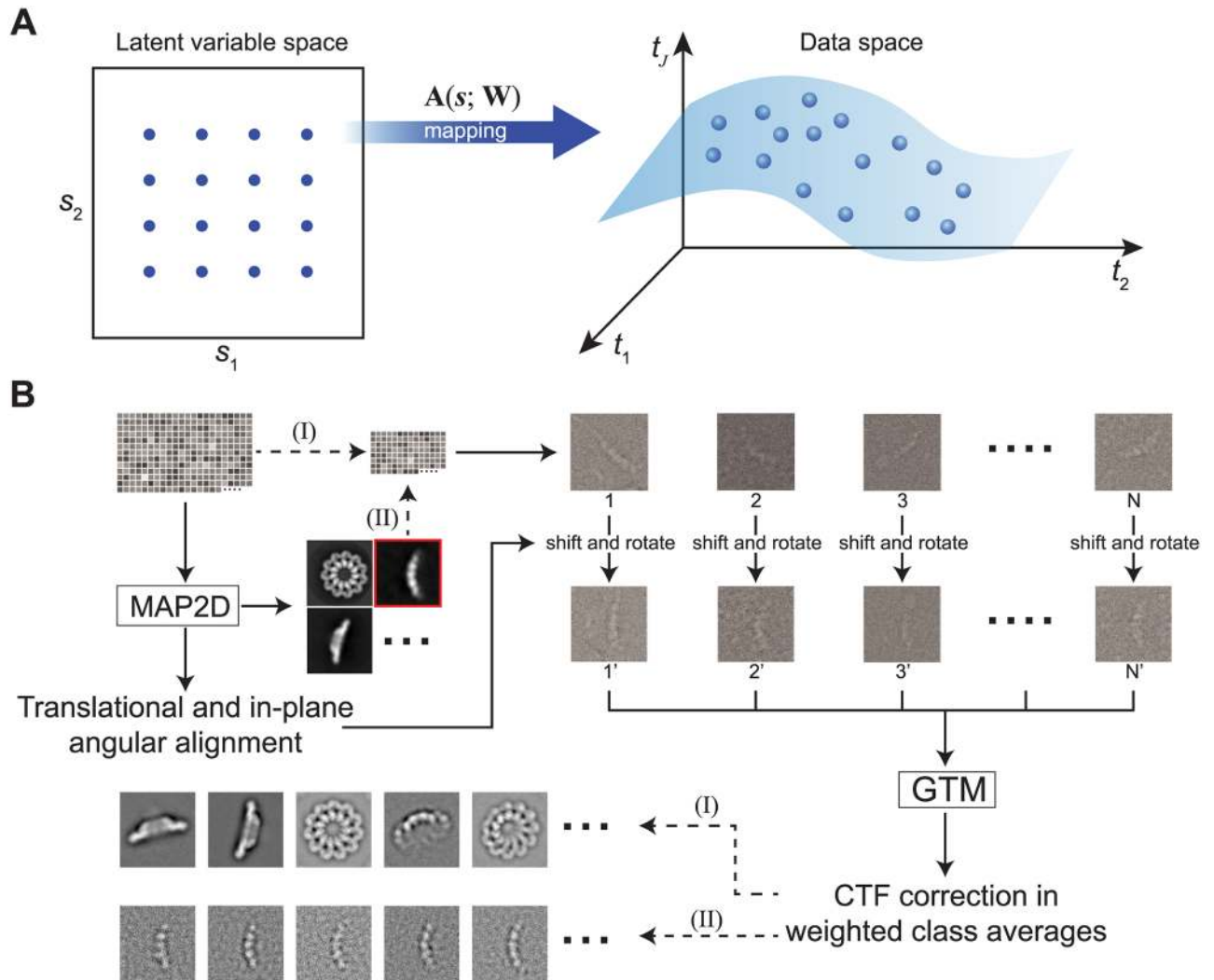
Fig 1. Strategy for unsupervised single-particle clustering via statistical manifold learning. (A) The fundamental principle of GTM is to establish a numerical relationship between variables in the latent space and a non-Euclidean manifold composed of the Fourier transformed image data in the data space. The manifold embedding can be determined by a set of nonlinear basis functions and a weighted parametric matrix. The likelihood function for the nonlinear mapping is solved by the expectation-maximization algorithm. **(B)** The workflow of implementing the unsupervised clustering strategies in ROME is as follows: (I) All images are aligned using MAP2D in a reference-free manner, and are subsequently classified into many groups by unsupervised GTM. (II) The unsupervised classes obtained in step (I) are further classified into many sub-classes by unsupervised GTM in a hierarchical fashion.

x_i, so that $t_i = T_{-\tau}(x_i)$, where $T_{-\tau}$ denotes the operator of in-plane image rotation and translation; $\tau = (\theta, r_x, r_y)$ is the in-plane rotation angle and translation, which can be determined through a 2D image alignment procedure, such as the regularized ML method [25–27]. Thus, an image t_i in the frequency domain can be modelled as:

$$t_{ij} = \mathrm{CTF}_{ij} A_j(s_k; \mathbf{W}) + n_{ij} \qquad (1)$$

where t_{ij} is the j-th component of the 2D Fourier transform of the i-th experimental image; CTF_{ij} is the j-th component of the contrast transfer function for the i-th image; n_{ij} is Gaussian noise; and $A_j(s_k; \mathbf{W})$ is the j-th component of the 2D Fourier transform of the underlying k-th molecular projection.

The nonlinear function $\mathbf{A}(s; \mathbf{W})$ is expanded by a set of basis functions $\{\phi_1, \phi_2, \ldots, \phi_M\}$ through the weight matrix $\mathbf{W}$:

$$\mathbf{A}(s; \mathbf{W}) = \mathbf{\Phi}(s)\mathbf{W} \tag{2}$$

where $\mathbf{A}(s; \mathbf{W})$ is a $K \times J$ matrix with elements $A_j(s_k; \mathbf{W})$; $\mathbf{\Phi}(s)$ is a $K \times M$ matrix with elements $\phi_{km} = \phi_m(s_k)$; and $\mathbf{W}$ is an $M \times J$ matrix containing the weight and bias parameters. In our algorithmic design, we used a combination of one fixed basis function and many Gaussians basis functions in the form:

$$\phi_m(s_k) = \begin{cases} \exp\left\{ -\dfrac{\|s_k - \mu_m\|^2}{2\sigma^2} \right\}, m \leq M_{NL} \\ 1, m = M_{NL} + 1 \end{cases} \tag{3}$$

where M_{NL} is the number of Gaussian basis functions; μ_m is the mean of the Gaussian distribution; and σ is the variance of the Gaussian distribution.

Although the noise in cryo-EM data may take multiple forms, potentially including both white and non-white noises, previous studies have established that cryo-EM noise can be largely approximated with the normal distribution without obvious detrimental effects on data analysis [1,25–27]. Thus, in our algorithm we chose the Gaussian distribution for the probability density function in the data space:

$$p(t_i|s, \mathbf{W}, \mathbf{\beta}) = \prod_{j=1}^{J} \left(\frac{\beta_{ij}}{2} \right)^{\frac{1}{2}} \exp\left\{ -\frac{\beta_{ij}}{2} \left(t_{ij} - \mathrm{CTF}_{ij} A_j(s; \mathbf{W}) \right)^2 \right\} \tag{4}$$

where $1/\beta_{ij}$ is the variance of noise for the j-th pixel of the i-th image.

Since we manage to classify a dataset consisting of N images $\{t_1, t_2, \ldots, t_N\}$ into K classes, this can be translated into the problem of finding K points $\{s_1, s_2, \ldots, s_K\}$ in the latent space that are mapped onto $\{t_1, t_2, \ldots, t_N\}$ in the data space through Eq (1). To make the problem tractable, we considered a specific form for the probability distribution $p(s)$ given by a sum of delta functions centered on K nodes of a regular grid in the latent space:

$$p(s) = \frac{1}{K} \sum_{k=1}^{K} \delta(s - s_k) \tag{5}$$

By integrating over the s-distribution on the manifold, the distribution $p(t_i|\mathbf{W}, \mathbf{\beta})$ in the data space for a given value of $\mathbf{W}$ is:

$$p(t_i|\mathbf{W}, \mathbf{\beta}) = \int p(t_i|s, \mathbf{W}, \mathbf{\beta})p(s)ds = \frac{1}{K} \sum_{k=1}^{K} p(t_i|s_k, \mathbf{W}, \mathbf{\beta}) \tag{6}$$

The joint probability density of observing the collection of N images $\mathbf{t} = \{t_1, t_2, \ldots, t_N\}$ is $p(t|\mathbf{W}, \mathbf{\beta}) = \prod_{i=1}^{N} p(t_i|\mathbf{W}, \mathbf{\beta})$. The ML estimate of the model parameters $\mathbf{\Theta} = \{\mathbf{W}, \mathbf{\beta}\}$ can be found by maximizing the logarithmic form of the joint probability, namely, $\widehat{\mathbf{\Theta}} = \arg\max_{\mathbf{\Theta}} \sum_{i=1}^{N} \ln p(t_i|\mathbf{\Theta})$. As we would like to consider the prior probability of the weight matrix, a regularized ML estimator, also called the *maximum a posteriori* (MAP) estimate, with respect to $\mathbf{\Theta}$ can be used to describe the nonlinear mapping problem. Thus, the model

parameters are sought to maximize the MAP:

$$\widehat{\boldsymbol{\Theta}} = \arg \max_{\boldsymbol{\Theta}} \left[\sum_{i=1}^{N} \ln p(\boldsymbol{t}_i|\boldsymbol{\Theta}) + \ln p(\boldsymbol{\Theta}) \right] \tag{7}$$

Here we assume a Gaussian prior distribution over **W**:

$$p(\mathbf{W}) = \left(\frac{\alpha}{2\pi}\right)^{W/2} \exp\left(-\frac{\alpha}{2}\|\mathbf{W}\|^2\right) \tag{8}$$

Where $1/\alpha$ is the variance of the Gaussian prior distribution over the weight matrix **W**, and W is the total number of elements in the matrix **W**.

Expectation-maximization algorithm

The unsupervised clustering problem is ill-posed due to a high level of noise, potentially missing orientations and discontinuity between different conformational states in single-particle cryo-EM data. To enable effective computation, the expectation-maximization (E-M) algorithm can be used to estimate the model parameters $\boldsymbol{\Theta}$ corresponding to the optimal clustering. The E-M algorithm alternates between two steps: the expectation step (E-step) and the maximization step (M-step). In the E-step, the old model parameters $\mathbf{W}^{(n)}$ and are used to evaluate the posterior probabilities, or responsibilities, $R_{ki}(\boldsymbol{\Theta}^{(n)})$ of each latent variable $\boldsymbol{s}_k$ for every image $\boldsymbol{t}_i$ using Bayes' theorem:

$$R_{ki}(\boldsymbol{\Theta}^{(n)}) = p(\boldsymbol{s}_k|\boldsymbol{t}_i, \mathbf{W}^{(n)}, \boldsymbol{\beta}^{(n)}) = \frac{p(\boldsymbol{t}_i|\boldsymbol{s}_k, \mathbf{W}^{(n)}, \boldsymbol{\beta}^{(n)})p(\boldsymbol{s}_k)}{\sum_{k'=1}^{K} p(\boldsymbol{t}_i|\boldsymbol{s}_{k'}, \mathbf{W}^{(n)}, \boldsymbol{\beta}^{(n)})p(\boldsymbol{s}_{k'})} \tag{9}$$

In the M-step, $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(n)})$ is maximized with respect to the model parameters $\boldsymbol{\Theta}$:

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(n)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} R_{ki}(\boldsymbol{\Theta}^{(n)}) \ln p(\boldsymbol{t}_i|\boldsymbol{\Theta}) + \ln p(\boldsymbol{\Theta}) \tag{10}$$

To this end, the partial derivatives of Eq (10) with respect to each model parameter should equal zero when $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(n)})$ is maximized. This allows us to estimate the model parameters iteratively. Maximizing $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(n)})$ with respect to **W** thus gives rise to the following equation, which uses $R_{ki}(\boldsymbol{\Theta}^{(n)})$, $\beta_{ij}^{(n)}$ to update the weight matrix $\mathbf{W}^{(n+1)}$ in the $(n+1)$-th iteration:

$$\sum_{i=1}^{N} \sum_{k=1}^{K} R_{ki}(\boldsymbol{\Theta}^{(n)}) \beta_{ij}^{(n)} \phi_m(\boldsymbol{s}_k) \mathrm{CTF}_{ij} \left(t_{ij} - \mathrm{CTF}_{ij} \sum_{m'=1}^{M} \phi_{m'}(\boldsymbol{s}_k) W_{m'j}^{(n+1)} \right) - \alpha W_{mj}^{(n+1)} = 0 \tag{11}$$

Similarly, by maximizing Eq (10) with respect to $\boldsymbol{\beta}$, we obtained the following re-estimation formula that uses $R_{ki}(\boldsymbol{\Theta}^{(n)})$ and $\mathbf{W}^{(n+1)}$ to update $\beta_{ij}^{(n+1)}$:

$$\frac{1}{\beta_{ij}^{(n+1)}} = \sum_{k=1}^{K} R_{ki}(\boldsymbol{\Theta}^{(n)})(t_{ij} - \mathrm{CTF}_{ij} A_j(\boldsymbol{s}_k; \mathbf{W}^{(n+1)}))^2 \tag{12}$$

During the iteration of the E-M algorithm, no class averages need to be computed. Nonetheless, upon convergence of the E-M algorithm, one can calculate the class averages using the

posterior probability R_{ki}:

$$A_{kj} = \frac{\sum_{i=1}^{N} R_{ki}(\Theta)\beta_{ij}\mathrm{CTF}_{ij}t_{ij}}{\sum_{i=1}^{N} R_{ki}(\Theta)\beta_{ij}\mathrm{CTF}_{ij}^2 + \alpha} \tag{13}$$

Note that this equation is derived from Eqs (11) and (12) and resembles the Wiener filter.

Note that Eq (11) is a system of linear equations. Solving this system of equations is equivalent to separately solving J systems of linear equations $\mathbf{C}_j \boldsymbol{w}_j = \boldsymbol{b}_j$, $j = 1, \ldots, J$, where

$$\mathbf{C}_j = \begin{pmatrix} \sum_{i=1}^{N}\sum_{k=1}^{K} R_{ki}(\Theta^{(n)})\beta_{ij}^{(n)}\phi_1^2(\boldsymbol{s}_k)\mathrm{CTF}_{ij}^2 + \alpha & \cdots & \sum_{i=1}^{N}\sum_{k=1}^{K} R_{ki}(\Theta^{(n)})\beta_{ij}^{(n)}\phi_1(\boldsymbol{s}_k)\mathrm{CTF}_{ij}^2\phi_m(\boldsymbol{s}_k) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{N}\sum_{k=1}^{K} R_{ki}(\Theta^{(n)})\beta_{ij}^{(n)}\phi_m(\boldsymbol{s}_k)\mathrm{CTF}_{ij}^2\phi_1(\boldsymbol{s}_k) & \cdots & \sum_{i=1}^{N}\sum_{k=1}^{K} R_{ki}(\Theta^{(n)})\beta_{ij}^{(n)}\phi_m^2(\boldsymbol{s}_k)\mathrm{CTF}_{ij}^2 + \alpha \end{pmatrix}$$

$$\boldsymbol{w}_j = \begin{pmatrix} W_{1j} \\ \vdots \\ W_{Mj} \end{pmatrix}, \boldsymbol{b}_j = \begin{pmatrix} B_{1j} \\ \vdots \\ B_{Mj} \end{pmatrix}, \text{ and } B_{mj} = \sum_{i=1}^{N}\sum_{K=1}^{K} R_{ki}(\Theta^{(n)})\beta_{ij}^{(n)}\phi_m(\boldsymbol{s}_k)\mathrm{CTF}_{ij}t_{ij}. \text{ This}$$

decomposition allows us to distribute the numerical computation of Eq (11) to many computer nodes in parallel. To further reduce the computational cost, we used a mean to replace $\beta_{ij}^{(n+1)}$ in the numerical solution of Eq (11), as follows:

$$\overline{\beta_{ij}^{(n+1)}} = \frac{N \times J}{\sum_{i=1}^{N}\sum_{j=1}^{J} \frac{1}{\beta_{ij}^{(n+1)}}} \tag{14}$$

where N is the total number of the images and J is the total number of pixels in each image.

The convergence of the E-M algorithm can be monitored by the weighted loss function

$$\mathcal{L}^{(n)} = \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K} R_{ki}(\Theta^{(n)})(t_{ij} - \mathrm{CTF}_{ij}A_j(\boldsymbol{s}_k; \mathbf{W}^{(n)}))^2 \tag{15}$$

Thus, a convergence criterion can be devised based on the weighted loss function $\mathcal{L}^{(n)}$, i.e., the program is terminated when $\mathcal{L}^{(n)}$ does not change more than 0.1% for six consecutive iterations.

Implementation of the GTM algorithm

Based on the above mathematical framework, we have implemented the following algorithm in the ROME software package.

Input: Particle images $\boldsymbol{t}_i$, and their corresponding $\mathbf{CTF}_i$, $i = 1, 2, 3, \ldots, N$.

Output: Class assignment and class averages A_k

1. Set the fixed number of latent points $\{\boldsymbol{s}_k\}$, $k = 1, 2, 3 \ldots K$. The space between two adjacent points is 1.

2. Set the values of basis function centers $\{\mu_m\}$, m = 1, 2, 3 \ldots M. $\mu_m = \frac{M}{M-1} + \frac{(K-1)mM}{(M-1)^2}$

3. Select the value of basis function width $\sigma = \frac{(K-1)M}{(M-1)^2}$.

4. Calculate the basis function ϕ from Eq (3).

5. Initialize the weighted matrix $\mathbf{W}$ from a Gaussian circle and $\boldsymbol{\beta}$; and set α to a fixed value of 0.01.

6. Iterate until the convergence criterion, $\left| \frac{\mathcal{L}^{(n-i)} - \mathcal{L}^{(n-i-1)}}{\mathcal{L}^{(n-i-1)}} \right| < 0.1\%$ ($i = 0, 1, \ldots, 5$), is satisfied:

   a. Set iteration number $n = n + 1$;

   b. E-step:

      i. Compute the δ_{ij} distance between the latent space and data space

$$\delta_{ij} = \left(t_{ij} - \mathrm{CTF}_{ij} \sum_{m=1}^{M} \phi_m(\boldsymbol{s}_k) W_{mj} \right)^2;$$

      ii. Compute the probability matrix $\mathbf{P}$ from Eq (4) using ϕ, δ_{ij} and $\mathbf{W}$;

      iii. Compute the responsibility matrix $\mathbf{R}$ from Eq (9) using $\mathbf{P}$, δ_{ij} and $\mathbf{W}$.

   c. M-step:

      i. Use Eq (11) and $\mathbf{R}$ to update $\mathbf{W}$;

      ii. Use $\mathbf{R}$ and δ_{ij} to update β_{ij} and calculate its mean from Eq (14).

7. Calculate class averages from Eq (13).

Hierarchical strategy for unsupervised clustering

Although our GTM-based algorithm may be advantageous for unsupervised classification, it is not computationally efficient for image alignment, a procedure that determines three geometrical parameters for each image: x-y translation and in-plan rotation. Thus, we employed the adaptive MAP-based 2D alignment method (hereafter referred to as MAP2D) [19,46] to align single-particle images prior to the GTM-based clustering procedure (Fig 1B). Averages from random subsets of unaligned images were used to initialize the MAP2D-based image alignment. Furthermore, a Gaussian model was used to initialize the parameters $\mathbf{W}$ of the E-M algorithm in GTM solution. Thus, in both steps of the MAP2D-based alignment and GTM-based clustering, no external initial model or reference was needed, ensuring the unsupervised nature of our approach.

The unsupervised single-particle clustering may be conducted in a hierarchical fashion. First, all particles are aligned based on the translations and rotations determined by MAP2D. Then GTM is applied to partition these shifted and rotated particles into different classes. However, some classes might be mixtures of non-identical conformational states. For these classes, the aligned particles in each class can be further classified into tens or hundreds of reference-free sub-classes by GTM clustering. This hierarchical clustering strategy can be further iterated and used in different scenarios. For example, if large numbers of classes are needed to assess sample heterogeneity, the strategy can be used to obtain relatively more unsupervised classes without any human intervention. If one needs to verify the structural homogeneity within individual classes, the hierarchical clustering strategy can help examine structural variability within specific classes at a deeper level.

Comparative benchmarking with simulated data

Each simulated dataset, including 50,000 projections, was generated by projecting the density map of the *Escherichia coli* 70S ribosome along random orientations [20]. We uniformly chose 500 orientations covering half a sphere. Each orientation was regarded as a Gaussian center, around which 100 projections were generated with a Gaussian distribution. To simulate realistic situations, we added random shifts in the range of -2.0 to 2.0 pixels relative to the image center. We further generated defocus values in the range of -1.0 to -3.5 μm and modified the projection images with the CTF in the Fourier domain. The projections were then additively contaminated with Gaussian noise at SNRs of 1/50, 1/100, 1/150, and 1/200, which allowed us to benchmark the proposed algorithm at different noise levels.

Using the simulated data, we compared the results from our approach with those obtained using RELION [19], EMAN2 [47] and SPIDER [20]. All procedures were set to classify these 50,000 particles into 500 classes. In all cases described below, the alignment or classification was run for 30 rounds. Both in our approach and in the unsupervised 2D classification in RELION, we set up an in-plane rotational search with an angular increment of 2.5˚, a translational search in the range of -3 to 3 pixels, and 100 classes for MAP2D alignment. We also conducted a classification based on an alternative method in RELION, which was implemented using the 'skip_align' option to classify pre-aligned particles without doing further alignment during classification. To enable a fair comparison, in the pre-alignment step, we set up the same control parameters for MAP2D alignment as the ones without using the 'skip_align' option in RELION.

EMAN2 uses multivariate statistical analysis (MSA) [1,22–24] with multi-reference alignment (MRA) method [24,47]. PCA was used in the MSA to reduce the dimensionality of images [47]. The particles were first CTF-corrected through phase flipping. Then, translational and rotational invariants were calculated and used for initial clustering. Next, PCA in the MSA step was used to capture the principal features, which were then used as multiple references for image alignment and K-means clustering in the MRA step. We set up 20 MSA basis vectors, 12 alignment references and 8 iterations for the EMAN2 procedure.

SPIDER combines reference-free alignment (RFA) with PCA and K-means clustering [48]. It involves three steps. First, all CTF-corrected noisy images were filtered through a low-pass Butterworth filter, with the pass band and stop band frequencies at 0.08 and 0.12, respectively. These particle images were then aligned globally in a reference-free manner, in order to find rotations and translations for all images that minimize the sum of squared deviations from their mean using the AP SR command in SPIDER. Second, PCA was used to solve the eigenvalue and eigenvector of all rotated and shifted images in the MSA step. Third, all particles were classified by K-means clustering [20].

Experimental cryo-EM datasets

We used three experimental cryo-EM datasets to examine if our clustering approach could yield useful classes and detect subtle structural difference among classes. The experimental cryo-EM datasets were collected using an FEI Tecnai Arctica microscope, operated under an acceleration voltage of 200 kV and equipped with Gatan K2 Summit direct electron detector. The inflammasome [8] and 26S proteasome [7] were imaged at a nominal magnification of 21,000 times. The proteasomal 19S regulatory particle (RP) [49] was imaged at a nominal magnification of 54,000 times. Cryo-EM data were collected semi-automatically using Leginon [50]. Movies comprising 36 frames per exposure were collected for the inflammasome and the 19S RP with an accumulated dose of 50 electrons/Å^2. For the 26S proteasome, movies of 30 frames per exposure and an accumulated dose of 30 electrons/Å^2 were obtained. The physical

pixel size and the super-resolution pixel size of the inflammasome and 26S proteasome data-sets were 1.72 Å and 0.86 Å, respectively. The calibrated physical pixel size of 19S RP was 0.98 Å. The defocus in data collection were set in the range of -1.0 to -3.0 μm. See S1 Fig for typical micrographs and class averages.

Ab initio 3D reconstruction

To examine the effect of GTM-based clustering on *ab initio* 3D reconstruction, 117,471 RP complex particles were classified, by unsupervised GTM clustering implemented in ROME, into 1,000 classes (S2 Fig). We set the in-plane rotational search to a 2.5˚ increment, the transla-tional search in the range of -4 to 4 pixels and 100 classes for MAP2D alignment. We used 234 class averages of the highest quality from the 1,000 GTM-produced classes to generate an initial model of the RP complex using e2initialmodel.py in EMAN2. In this step, 8 refinement itera-tions were performed with outputs of 10 initial models in search of a good one. As a control, we used the traditional method to obtain an initial model of the RP complex in EMAN2 from 10,000 raw RP complex images. These particles were first classified into 100 classes using e2refine2d.py. 12 MSA basis vectors, 5 alignment references and 8 iterations were set up for this EMAN2 procedure. 72 good classes were used to generate an initial reconstruction of RP com-plex by e2initialmodel.py in EMAN2. In this step, 8 refinement iterations were performed with outputs of 10 initial reconstructions. The best reconstruction was chosen for further analysis.

Parallel implementation and code optimization

Both the MAP2D and GTM implementations in ROME are optimized specifically for Intel® Xeon® processors and Intel® Xeon Phi™ coprocessors. In analogy to MAP2D in RELION, algo-rithm-level optimizations such as an adaptive E-M approach were adopted in ROME. In addi-tion, the particle image stack was divided based on the unit of individual images across multiple computer nodes in the parallelization of the data flow in the MAP2D module. However, the image stack was divided based on the unit of individual image pixels in the GTM implementa-tion. This allowed the problem to be distributed across multiple cluster nodes and attached coprocessors, while minimizing data transfer and memory requirements. Next, complex data structures were converted from arrays of structures to structures of arrays, making it easier to implement the inner loops of compute-intensive kernels with processor instructions that can be executed on multiple data elements in parallel (vectorization). Loops were also restructured so that independent threads of execution work on data elements that are near one another in memory. This 'cache blocking' attempts to lower the pressure that the algorithms put on the memory subsystem as the number of threads climbs, and tries to make the most effective use of the processor memory caches. To further improve the use of memory and decrease serialization overhead, memory allocation and deallocation were moved from the middle of the algorithm to the initialization phase. Finally, a single-precision data type instead of a double-precision one was used to store the image data in parts of the algorithm where the extra precision is not neces-sary. This treatment can improve performance by about a factor of two.

Results

Classification accuracy with simulated data

To quantify the classification accuracy of our approach, we generated a series of synthetic data-sets with various SNRs, each of which was composed of 50,000 simulated images as described above. Since all the original angles of the simulated data were known, the angular difference between any pair of images classified into each class (hereafter named 'angular distances')

could be computed. We adopted the statistical behavior of the angular distances as a measure of the quality of the corresponding classes [13]. Supposing that n images are classified into one class, we have $\binom{n}{2}$ pairs of angular distances in this class. The histogram distribution of the angular distances from all the classes is a good way to compare the results from different algorithms [13]. A sharper, higher peak at lower angular distances in the distribution curve is expected from a better and more accurate clustering. We also used a second plot that ranks the number of particles over all classes to show the effective classes in the results. With the dataset at a SNR of 0.02, the classification accuracy of our approach is almost the same as that of SPIDER, and was better than that of both EMAN2 and RELION (S3C Fig). For the data at a SNR of 0.01, our approach yielded the sharpest and highest peak at lower angular distances in the distribution plot (Fig 2D), with a total of 495 effective classes (Fig 2E). This feature remained when the SNR was further reduced to 0.005. By contrast, the MAP2D classification implemented in RELION yielded a lower peak (Fig 2D), but only 162 effective classes (Fig 2E). The RELION results obtained with the 'skip_align' option were much worse than those from the standard MAP2D classification procedure (Fig 2D and 2E). Conventional PCA/K-means approaches in EMAN2 and SPIDER resulted in a much lower, wider peak at greater angular distances (Fig 2D and 2E). When the SNR was decreased to 0.005, the classification accuracies of these approaches were further reduced (Fig 2F and 2G and S3E and S3F Fig). These results suggest that a high level of noise significantly compromises the performance of data clustering by conventional PCA/K-means clustering approaches. By contrast, our GTM approach significantly improved the classification accuracy under conditions of reduced SNRs.

We further examined the impact of the dimensionality of the latent space on the classification accuracy in the GTM algorithm. Using the same datasets of simulated images, we compared the measurements of the classification accuracies for the latent spaces assumed in one, two and three dimensions. The experiment with 2D latent space yielded slightly better classification accuracies than those with 1D and 3D latent space in our GTM algorithm for SNRs of 0.01, 0.0067 or 0.005 (Fig 3). This result suggests that the increased dimensionality in manifold learning does not necessarily improve classification accuracy.

We also examined the impact of several control parameters in the GTM algorithm on classification accuracy. First, there are two independent hyper-parameters in GTM: M_{NL} and α. Using the simulated datasets with a SNR of 1/100, we compared the measurements of the classification accuracy for different M_{NL} and α values. The experiment with $M_{NL} = 0.8 \times K$, where K is the class number, yielded better classification accuracy than those with smaller M_{NL} values (S4 Fig). This result suggests that an increased number of nonlinear basis functions can strengthen the classification capability of GTM. For the hyper-parameter α, the experiment with $\alpha = 0.01$ yielded a better classification accuracy than those with smaller or larger α values (S5 Fig). This result suggests that an appropriate α value can penalize the 'peaky' elements in the weight matrix $\mathbf{W}$ to prevent overfitting and to improve the classification ability of GTM. Next, we compared the classification accuracies for different numbers of classes, namely, $K = 100$, 300 and 500. We found that the classification accuracy increases with increasing class number (S6 Fig). However, this was not the case for the other classification methods implemented in RELION and SPIDER (S6 Fig). Finally, we investigated the robustness of our GTM clustering algorithm by varying the initialization. We used different characteristic radii of the Gaussian circles and different values of β, as well as the outputs of PCA, for GTM initialization (S7 Fig). The results indicate that the performance of our GTM clustering is insensitive to the parameters of Gaussian circle inputs, which generally gave rise to better performance than those obtained by the PCA-based initialization (S7 Fig).
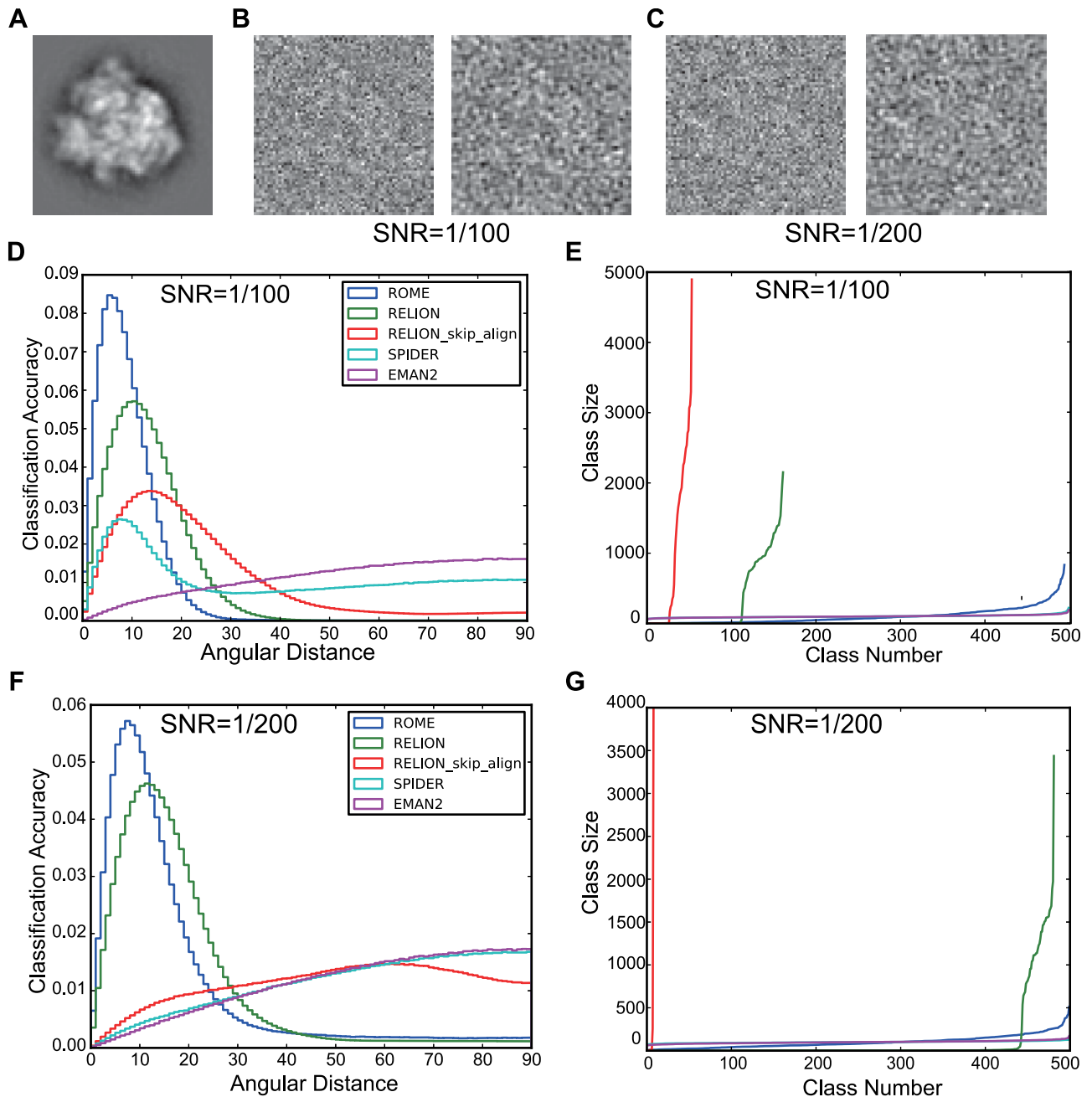
Fig 2. Benchmarking the performance of unsupervised clustering using simulated data. (A) A projection of the 70S ribosome model. **(B and C)** Examples of the simulated images of the 70S ribosome with SNRs of 1/100 **(B)** and 1/200 **(C)**. The right panel in **(B)** and **(C)** shows the low-pass filtered version of each simulated image. **(D and F)** The normalized histogram exhibits the distributions of angular distances resulting from the five classification methods that were applied to the simulated images with SNRs of 1/100 (panel **D**) and 1/200 (panel **F**). **(E and G)** The sizes of classes were ranked for the five classification methods with SNRs of 1/100 (panel **E**) and 1/200 (panel **G**).

https://doi.org/10.1371/journal.pone.0182130.g002

Applications to experimental cryo-EM datasets

To demonstrate the applicability of our clustering approach, we applied our method to several experimental cryo-EM datasets (S1 Fig). First, ROME was used for an unsupervised clustering of a 17,103-particle dataset of the inflammasome complex [8]. The resulting class averages
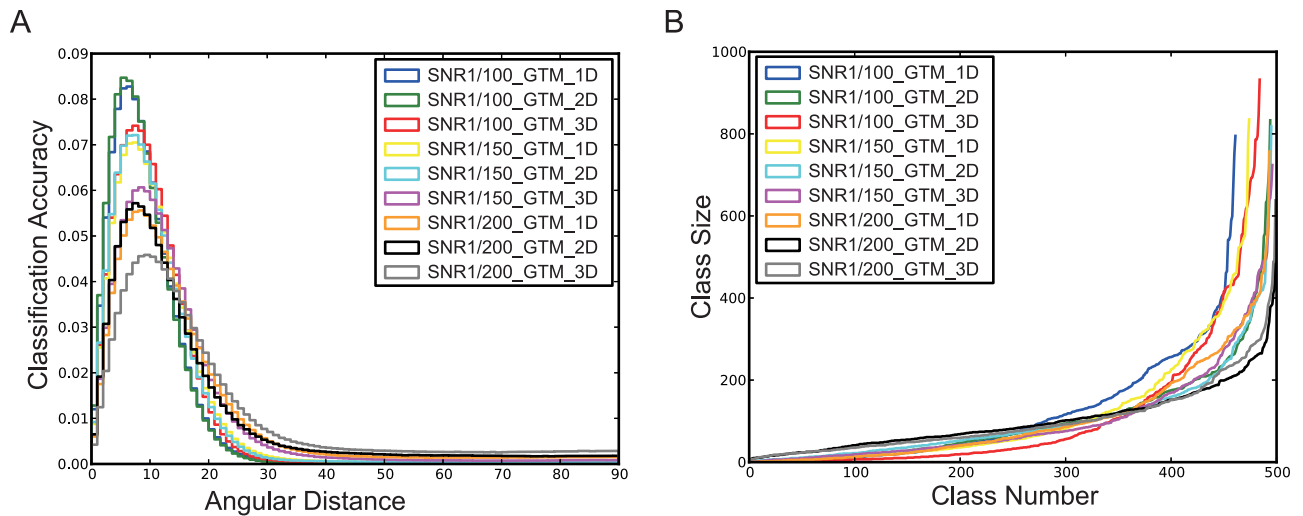
Fig 3. Classification accuracy with one-, two- and three-dimensional latent space in our GTM algorithm. (A) Normalized histograms exhibit the angular distances for the one- and two-dimensional latent space under different SNRs. **(B)** The sizes of classes are for different latent space dimensions with varying SNRs. The label 'GTM_D' in **(A)** and **(B)** represents the number of dimensions. GTM_1D denotes that 500 points in one dimensional latent space were sampled in the GTM algorithm. GTM_2D denotes that 100 points in one dimension and 5 points in the other dimension, a total of 500 points, were sampled by the GTM algorithm. GTM_3D denotes that 20 points in the first dimension and 5 points in each of the other two dimensions, giving a total 500 points, were sampled in the GTM algorithm.

directly presented three different conformations of the complexes corresponding to 10-, 11- and 12-fold symmetry (Fig 4A and S8A Fig). For comparison, 300 classes were produced by reference-free MAP2D classification implemented in RELION [19] (S8B Fig). Second, unsupervised GTM clustering of 96,488 particles of the proteasomal regulatory particle associated with the core particle (RP-CP) sub-complex directly revealed different local features among class averages (Fig 4B) [2,7]. These local differences were verified as corresponding to distinct conformational states through further 3D reconstruction [2,7]. Our results thus indicate that unsupervised GTM clustering can sort out not only the projections of underlying structures into relatively homogenous 2D classes but can also unearth incomplete structures or junk particles, allowing the efficient *in silico* purification of single-particle datasets.

Unsupervised GTM clustering with improved accuracy can help identify hidden heterogeneity in the reference-free classes generated by other methods. For instance, after initial reference-free clustering, particle images of specific classes could be selected for deeper unsupervised clustering by GTM in a hierarchical fashion. Delicate differences between GTM-generated class averages may reveal features corresponding to distinct conformational states. To test this idea, a class of 281 particles resulting from MAP2D-based classification (Fig 4C), whose average visually resembles the side view of the 11-fold inflammasome complex, was further classified into 30 sub-classes by GTM (Fig 5A). Based on the length of the intrinsic structure, the GTM-based reference-free clustering identified the side views of 10-fold and 12-fold inflammasome complexes among the 30 sub-classes (the red boxes in Fig 4D), suggesting that the MAP2D-generated initial class is a mixture of the complexes with 10-, 11- and 12-fold symmetry. Another MAP2D-classified image group with 3961 particles, whose average visually represents a tilted view of the free 19S regulatory particle (RP) that is not associated with a 20S core particle (CP), was further classified by GTM (Fig 4E). This deeper clustering identified 448 single-particle images (11.3%) of RP-CP subcomplexes that were misclassified into the dataset of free RP (Fig 4F).
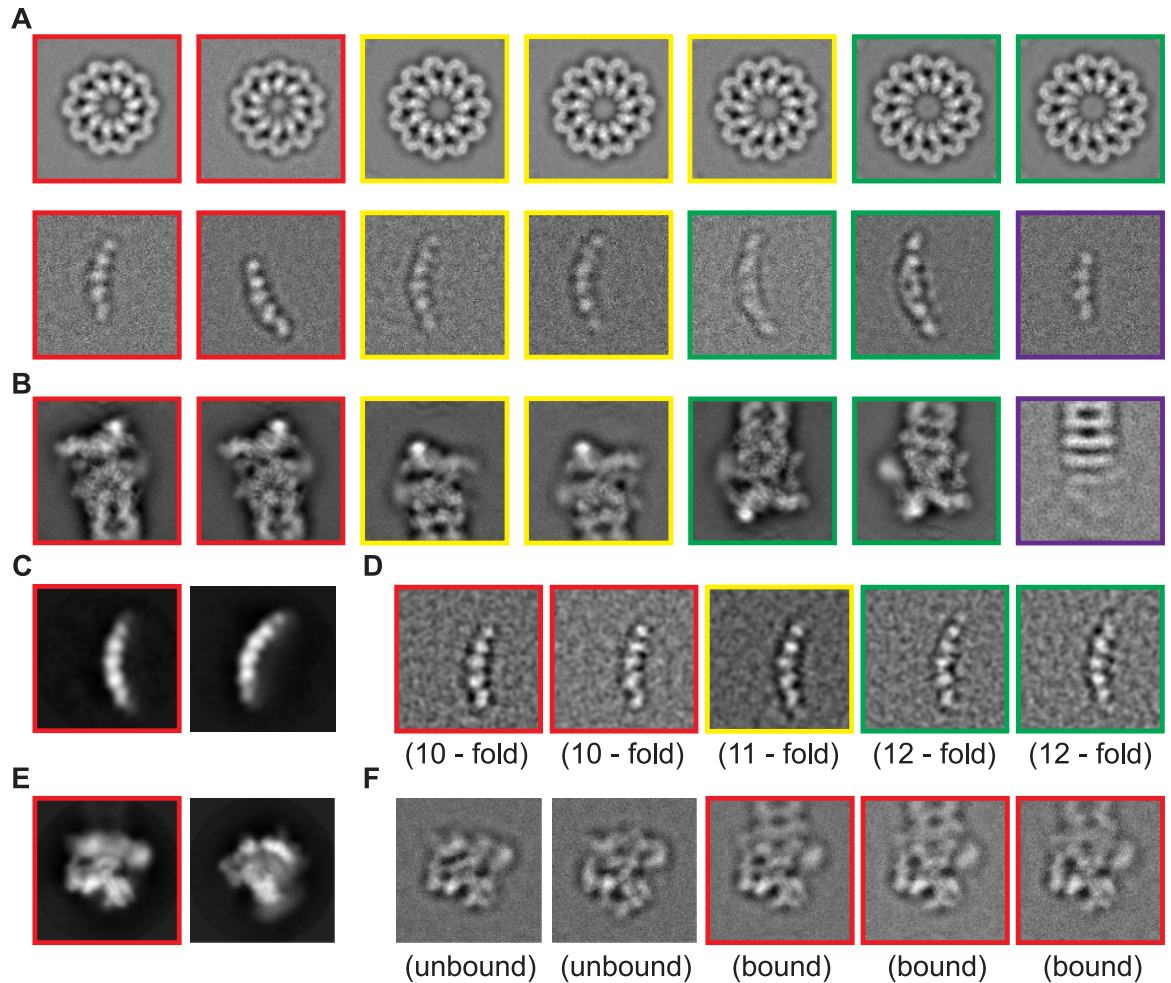
Fig 4. Unsupervised clustering by GTM. (A) Typical class averages of inflammasome particles generated by unsupervised GTM clustering in ROME. Red, yellow and green boxes indicate the top views (first row) and the side views (second row) of 10-, 11-, and 12-fold inflammasome complex, respectively. The side views of the complex structure differ by length. Besides, the purple box denotes the class average of an incomplete inflammasome complex. **(B)** Typical class averages of RP-CP sub-complexes generated by unsupervised GTM in ROME. The red or yellow boxes indicate a pair of class averages showing differences in local features corresponding to the local movement of the Rpn5 subunit of the RP-CP subcomplex [7]. The green box indicates a pair of class averages showing the movement of the Rpn1 subunit of RP-CP subcomplex [7]. The purple box labels the class average of the incomplete RP-CP subcomplex. **(C)** Typical side-view class averages of the inflammasome were initially classified using the MAP2D classifier in a reference-free manner. Two classes among 50 classes visually resemble the 11-fold inflammasome complex particles. **(D)** The class average highlighted by red box in panel (**C**) was further classified by GTM. The red boxes indicate the 11-fold inflammasome particles. The green boxes indicate the 10-fold inflammasome particles that were misclassified by MAP2D into the same class as the rest 11-fold structures. The yellow boxes indicate the 12-fold inflammasome particles that were misclassified by MAP2D into the same class as the rest of the 11-fold structures. **(E)** A 57,001-particle dataset of free RP was initially classified using the MAP2D classifier in a reference-free manner. **(F)** The class marked by the red box in panel (**E**) was further classified by GTM in ROME. Several classes of RP-CP sub-complex particles (red boxes) were found to be misclassified into this free RP class.

Fig 5 presents a further comparison with other algorithms including MAP2D and K-means clustering, which suggests that our GTM outperforms those algorithms in unambiguously identifying hidden heterogeneity with the hierarchical clustering strategy. In this comparison, we selected a MAP2D-classified image group comprising 281 side views of the inflammasome complexes (red boxes in Fig 4C), which was further classified into 30 classes by unsupervised GTM clustering in ROME, by reference-free MAP2D classification in RELION, and by K-
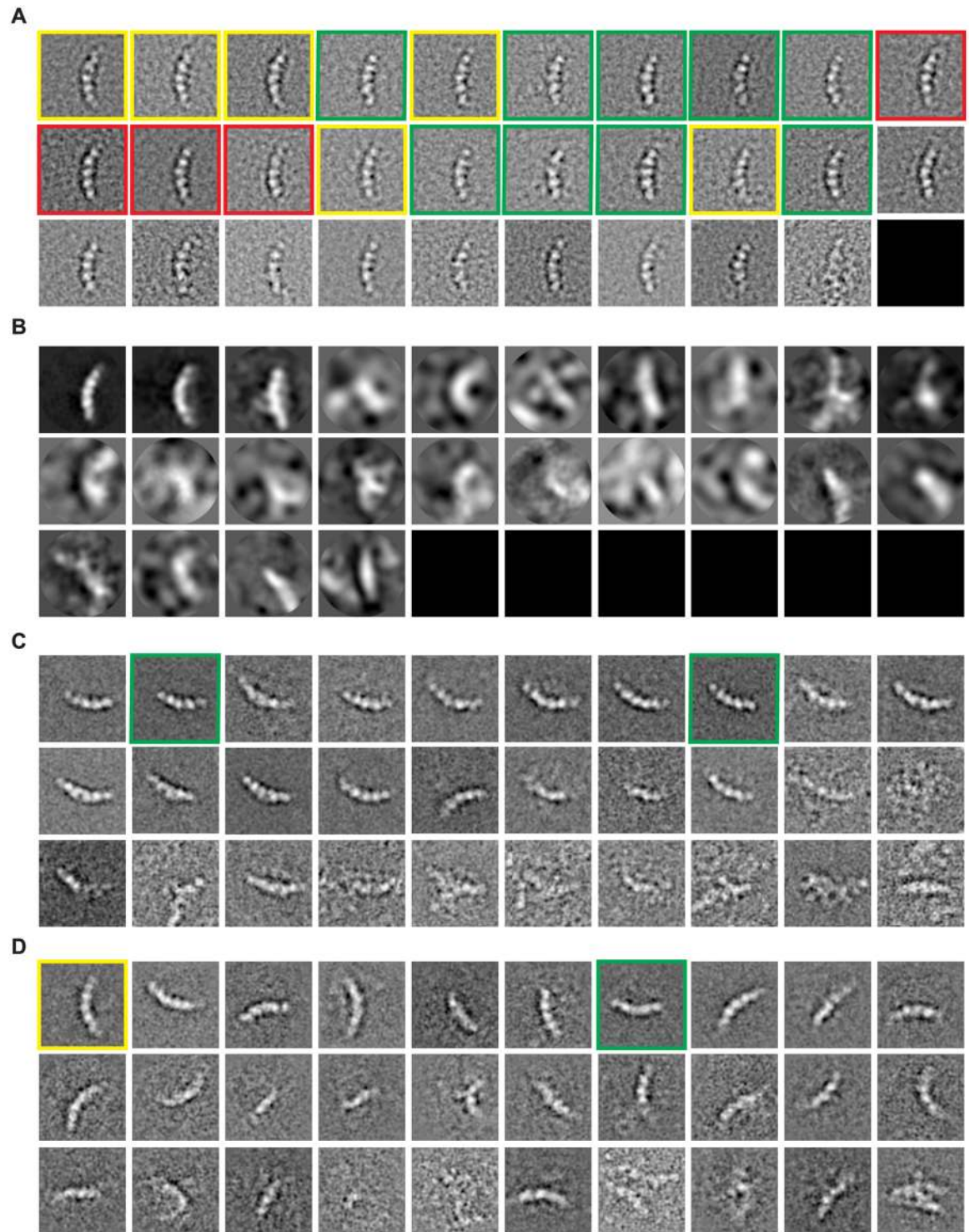
Fig 5. Comparison of hierarchical unsupervised clustering using ROME, RELION, SPIDER and EMAN2. (A)
Unsupervised classification of a MAP2D-generated class into 30 sub-classes using the GTM algorithm in ROME. The red
box marks the side view projection of the 11-fold inflammasome complex. The green box marks the side view projection of the
10-fold inflammasome complex, whose length is smaller than that of the others. The yellow box marks the side view projection
of the 12-fold inflammasome complex, whose length is larger than that of the others. **(B)** Unsupervised classification of the
same MAP2D-generated class into 30 sub-classes in RELION. The major class exhibits the side view projections of 11-fold
inflammasome complex, whereas all the other classes present 'junk' features. **(C)** Unsupervised K-means clustering of the

same MAP2D-generated class into 30 sub-classes in SPIDER. The green boxes highlight the side views of the 10-fold inflammasome complex, whose length is smaller than others. The yellow boxes label the side view projections of the 12-fold inflammasome complex, whose length is larger than that of the others. **(D)** Unsupervised K-means clustering of the same MAP2D-generated class into 30 sub-classes in EMAN2. The yellow boxes label the side view projections of the 12-fold inflammasome complex, whose length is longer than that of the others. The green boxes label the side view projections of 10-fold inflammasome complex, whose length is shorter than that of the others.

https://doi.org/10.1371/journal.pone.0182130.g005

means clustering in SPIDER and EMAN2. The sub-classes resulting from the GTM algorithm all showed clear 2D structural features, indicating a mixture of the side views of 10-, 11- and 12-fold complexes in this dataset, as manifested by their length variation (green box and yellow box in Fig 5A). By contrast, only two meaningful classes were obtained by RELION, while the other sub-classes were either featureless or empty (Fig 5B). The results from SPIDER and EMAN2 appear to be improved over RELION. However, there were still more than 10 sub-classes in both cases that were either featureless or presented aberrant features likely resulting from mis-classification. Our comparative studies suggest that, unlike our GTM-based clustering, other unsupervised clustering approaches, such as MAP2D, PCA and K-means, are either less efficient or ineffective for directly detecting subtle structural variations using the hierarchical clustering strategy.

GTM-based clustering improves *ab initio* reconstruction

To demonstrate the clustering capacity of our approach, 117,471 particles of the free RP complex were classified into 1,000 classes by GTM implemented in ROME in an unsupervised manner. The computation was completed in 228 minutes with 30 rounds of E-M optimization on a computing cluster comprising 512 Intel Xeon E5 CPU cores (S1 Table) and 858 effective classes were obtained (S2 Fig). We used 234 class averages of the highest quality from the 1,000 unsupervised classes to generate an initial 3D reconstruction of the RP complex in EMAN2 (Fig 6A, S2 Fig). As a control, we used the conventional multivariate data analysis (MDA) and K-means clustering to obtain an initial reconstruction of RP complex in EMAN2 from 10,000 raw images of the RP complex (Fig 6B). The FSC curves between the atomic model and the two initial reconstructions using a 0.5-cutoff suggest that the resolution of the initial model using GTM-based class averages was 20.6 Å, which was significantly higher than the resolution, 29.1 Å, of the other initial reconstruction obtained by the conventional approach (Fig 6C). Next, we performed rigid-body fitting of the atomic model of the free RP into the two initial reconstructions. The atomic model fits into the GTM-based initial reconstruction with an excellent agreement (Fig 6A). By contrast, considerable elements in the atomic model were out of the density of the initial reconstruction obtained by the conventional approach (Fig 6B). The contrasting results show that the initial reconstruction built upon our GTM clustering method is of significantly improved quality compared with the conventional one. Taken together, these results suggest that the increased class number and improved classification accuracy of our unsupervised GTM clustering gives rise to a marked improvement of *ab initio* 3D reconstruction with an affordable computational cost.

Optimization of computational performance and efficiency

Our implementation of the GTM-based clustering algorithm in the ROME software has been optimized for modern HPC hardware, such as Intel$^®$ MIC architecture used in the Intel Xeon Phi$^™$ Knights Corner and Knight Landing processors. Our modernized code in ROME outperforms the existing software in both speed and magnitude. First, we compared the computational performance between ROME and RELION using several datasets comprising different
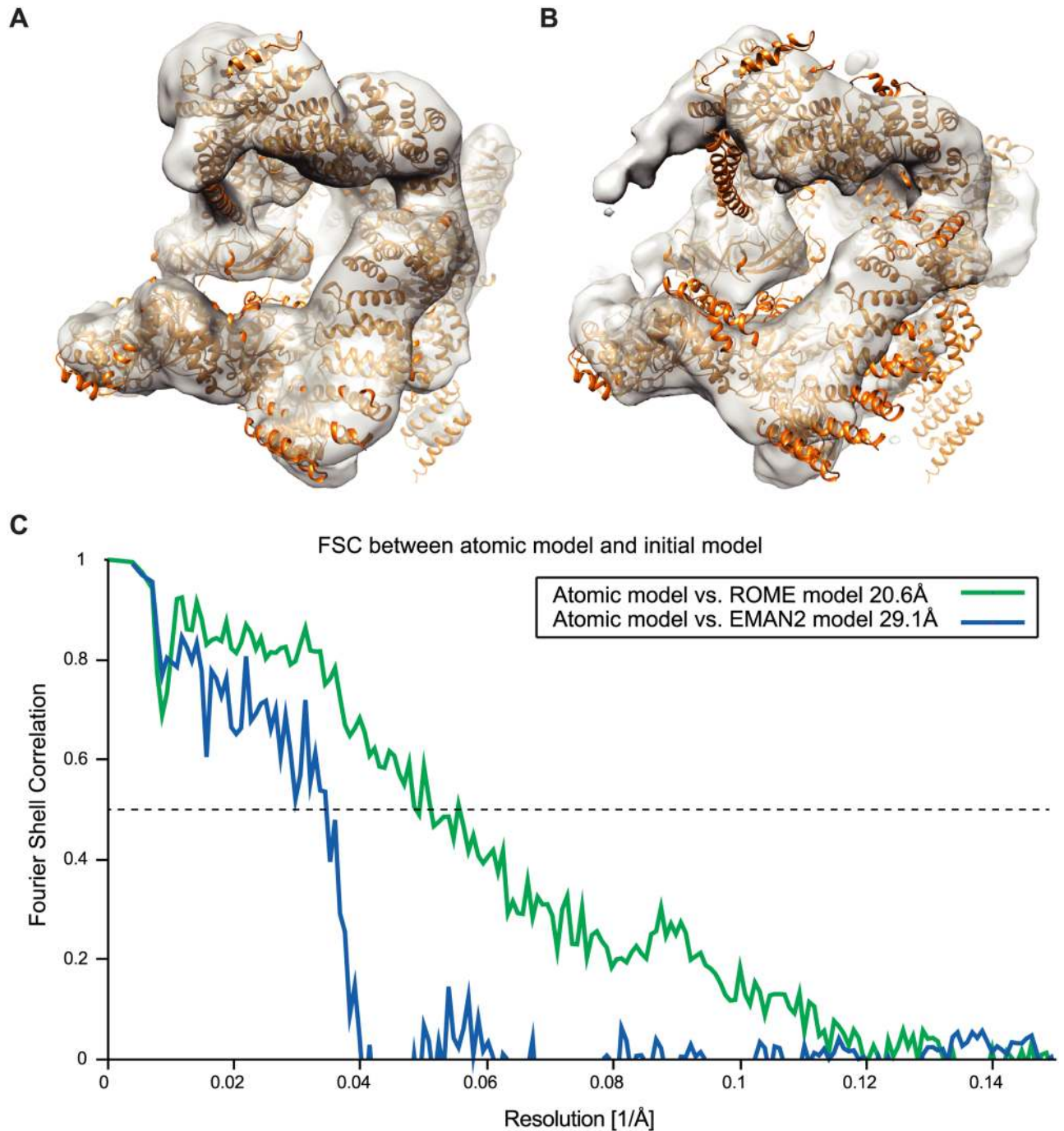
Fig 6. Initial 3D reconstruction from the reference-free class averages of ROME and EMAN2. (A) The initial reconstruction calculated by the ROME-generated class averages is superimposed with the atomic model of free RP shown in a ribbon representation, suggesting that they are highly compatible with each other. **(B)** The initial reconstruction calculated by the EMAN2-generated class averages is superimposed over the atomic model of free RP shown in a ribbon representation. A substantial part of the atomic model is outside of the density of the initial reconstruction, suggesting poor map quality and a large reconstruction error. **(C)** FSC curves between the RP atomic model and the initial reconstructions generated by ROME- and EMAN2-based class averages.

numbers of particles. It took 143 minutes for ROME to classify 57,001 particles of RP into 300 unsupervised classes on a computing cluster comprising 512 Intel Xeon E5 CPU cores ([Fig 7A](#)). By contrast, it took ~40 hours for RELION to classify the same dataset into 300 unsupervised classes on the same cluster. In all cases, the E-M algorithm was run for 30 iterations. Thus, under the same circumstances, ROME was about 10–20 times faster than RELION in unsupervised 2D classification. Second, a 96,488-particle dataset of the RP complex was used to benchmark the performance of ROME in unsupervised clustering with different class numbers on the 512-core computing cluster ([Fig 7B](#)). With increasing class numbers from 100 to 1,000, the running time of GTM clustering was increased in a polynomial trend from 14 to 214 minutes. Specifically, ROME completed the unsupervised clustering of 1,000 classes in 306 minutes with 30 iterations of expectation-maximization ([S1 Table](#)), whereas RELION did not complete the same number of iterations of optimization in a week on the same hardware condition. Furthermore, we also used this dataset to test reference-free classification at the 1,000-class level by MSA and K-means clustering in SPIDER [1,20]. Unfortunately, the program kernel crashed in the reference-free alignment step.

Discussion

Unsupervised data clustering plays an important role in high-resolution structure determination by single-particle cryo-EM [1]. In this study, we introduced a GTM-based unsupervised clustering method incorporating MAP2D-based image alignment, and implemented this approach in the open-source software ROME. One of the key findings of this study is that GTM-based clustering is more tolerant against noise than the other approaches we compared, including maximum-likelihood, PCA and K-means algorithms. This hallmark provides a path towards improved classification accuracy at lower SNRs. Our GTM algorithm built the CTF correction into its mathematical kernel so that the clustering would not be affected by the variation of the defocus value inherent in cryo-EM data. Our software uses the E-M algorithm to maximize the regularized log-likelihood function in statistical nonlinear mapping between the latent and data spaces, which only guarantees the solution of a local optimum. However, the random initialization of the nonlinear mapping ensures the unsupervised nature of data clustering. The class averaging expression under our GTM scheme resembles the Wiener-type filter and gives rise to CTF-corrected, probability-weighted class averages, which inherits the advantage and traits of the maximum-likelihood based approaches [1,25,26].

Utility strategies of unsupervised single-particle clustering by GTM can be highly versatile, problem-oriented, and user-controlled. In a typical scenario, one can pursue a strategy in which all particles are aligned and partitioned into unsupervised classes in single runs of E-M optimization. This gives a glimpse of typical reference-free projection structures of the biomolecules imaged and allows the removal of apparent 'junks'. In other scenarios, particles in specific classes can be further partitioned into deeper sub-classes in a hierarchical fashion. This may allow the user to inspect subtle structural variations within a class to analyze the potential structural heterogeneity within the dataset. This hierarchical clustering strategy can be further interleaved with 3D classification to help verify the quality of 3D classification and clean up 'junk' or poor quality images from 3D classes [2,7,8].

In summary, we developed a statistical manifold learning-based approach for unsupervised single-particle clustering in a massively parallel manner. When optimized for Intel® Xeon® processors and Intel® Xeon Phi™ coprocessors, it exhibited an unprecedented power in obtaining thousands of reference-free class averages with significantly improved clustering accuracy in a massively parallel fashion. Importantly, a greater number of high-quality unsupervised 2D class averages can lead to a more reliable *ab initio* 3D reconstruction. Practical
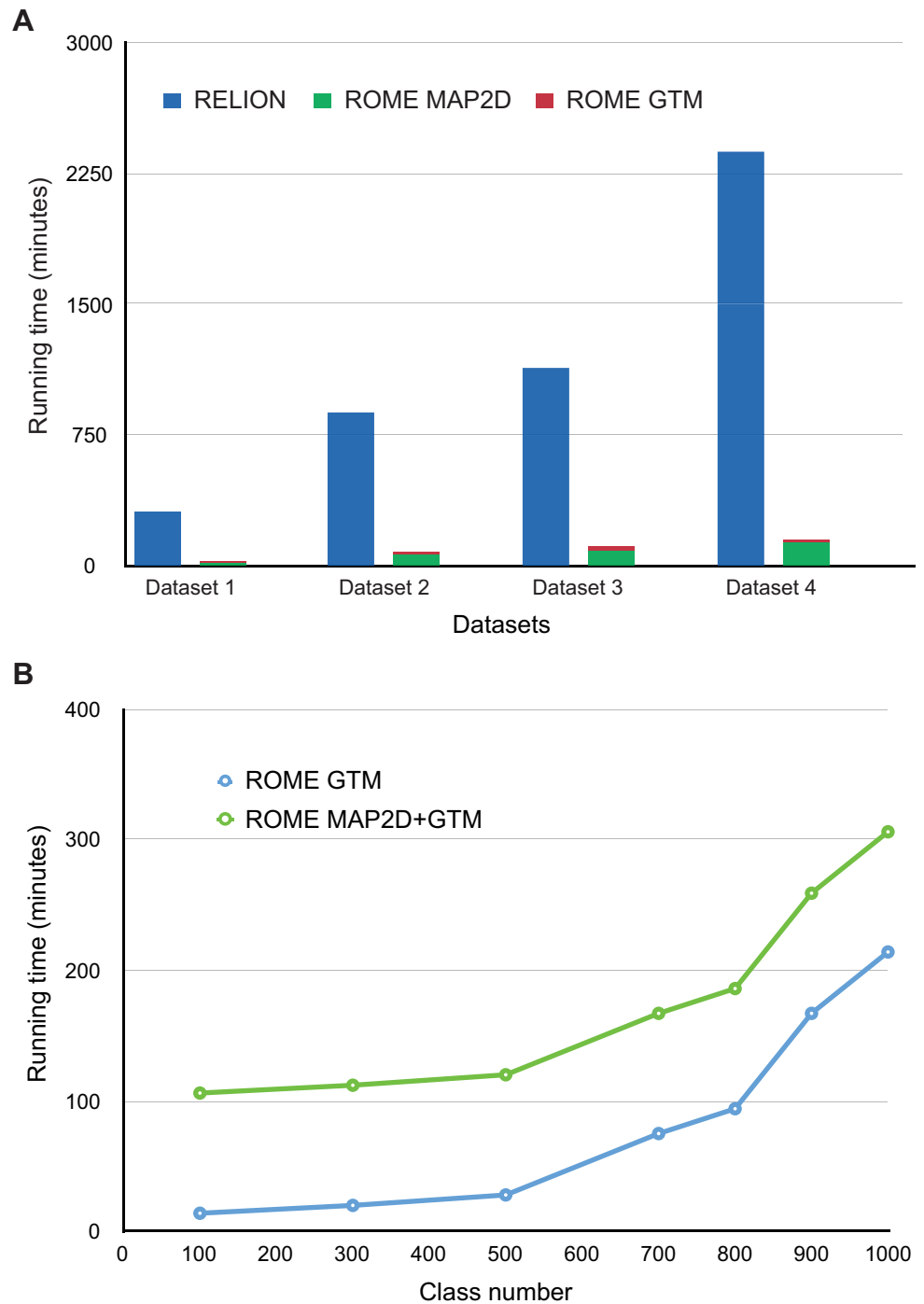
Fig 7. Performance evaluation of unsupervised clustering with ROME. (A) Performance of unsupervised single-particle clustering in ROME versus RELION using different datasets. Unsupervised 2D classification into 300 classes using both software programs were performed on four experimental datasets: Dataset1 refers to the 16,306-particle dataset of the inflammasome with 250×250 box size; dataset2 refers to the 35,407-particle dataset of the free RP complex with 160×160 box size; dataset3 refers to the 96,488-particle dataset of the RP-CP complex with 160×160 box size; dataset4 refers to the 57,001-particle dataset of the free RP complex with 180×180 box size. MAP2D alignment in ROME and GTM clustering for 300 classes wes also performed. The blue, green, and red histograms represent the running time of RELION, MAP2D in ROME, and GTM in ROME, respectively. For more comparison, see S9 Fig and S1 Table. **(B)** The 96,488-particle dataset of the RP-CP subcomplex was used to test the performance of GTM in ROME (blue dots). The green dots represent the total running time including both the MAP2D alignment and GTM clustering in ROME. The running time was polynomially related to the number of classes.

https://doi.org/10.1371/journal.pone.0182130.g007

applications of the algorithm to several experimental datasets demonstrated that unsupervised GTM clustering in ROME is highly effective in distinguishing structural heterogeneity among single-particle datasets. This can in turn, when used iteratively with 3D classification, assist in improving the structural homogeneity of cryo-EM datasets. Our approach has already been used in several studies of high-resolution structural determination from heterogeneous samples [7,8,49].

Supporting information

S1 Fig. Typical results of unsupervised clustering by GTM on three experimental cryo-EM datasets. (**A**) A typical cryo-EM micrograph of the inflammasome. The white boxes mark the particles picked from the micrograph that contributed to the dataset used in testing of our GTM algorithm implemented in ROME. (**B**) Typical reference-free 2D class averages of the 10-fold, 11-fold, and 12-fold inflammasome complex obtained by GTM-based clustering following MAP2D-based image alignment. (**C**) A typical cryo-EM micrograph of the proteasome. We boxed half of the holoenzyme, including half of the CP, in complex with a complete RP, named the RP-CP subcomplex (white boxes). (**D**) Typical reference-free 2D class averages of the RP-CP subcomplex obtained by GTM-based clustering following MAP2D-based image alignment. (**E**) A typical cryo-EM micrograph of the human RP proteasome. We boxed all particles, including the free RP complex and RP-CP subcomplex, whose box center is focused on that of RP (white box). (**F**) Typical reference-free 2D class averages of the RP complex obtained by GTM-based clustering following MAP2D-based image alignment.
(PDF)

S2 Fig. Unsupervised clustering of the RP dataset by ROME. 117,471 particles of the RP complex were classified into 1,000 classes by the unsupervised GTM implemented in ROME. 858 class averages were not blank. These classes comprised different views of the RP complex.
(PDF)

S3 Fig. Comparison of classification results of simulated data at different noise levels. (**A**, **B**) Examples of the corresponding simulated images of the 70S ribosome with SNRs of 1/50 (**A**) and 1/150 (**B**), respectively. The right panel in (**A**) and (**B**) shows the low-pass filtered version of each simulated image. (**C, E**) Normalized histograms show the distributions of angular distances resulting from the five classification methods that were applied to the simulated images with SNRs of 1/50 (panel **C**) and 1/150 (panel **E**). (**D, F**) The sizes of the classes were ranked with respect to the five classification methods for SNRs of 1/50 (panel **C**) and 1/150 (panel **E**).
(PDF)

S4 Fig. Impact of different values of the hyper-parameters M_{NL} in GTM on the unsupervised classification of the simulated data with an SNR of 1/100. M_{NL} was set as $0.8 \times K$, $0.5 \times K$ and $0.2 \times K$, where K is the class number. (**A, C, E**) Normalized histograms showing the distributions of angular distances corresponding to different M_{NL} values. The images were classified into 100 (panel **A**), 300 (panel **C**), and 500 (panel **E**) classes. (**B, D, F**) The sizes of the classes were ranked for different M_{NL} values with $K = 100$ (panel **B**), 300 (panel **D**), and 500 (panel **F**).
(PDF)

S5 Fig. Impact of different values of the hyper-parameter α in GTM on the unsupervised classification of the simulated data with an SNR of 1/100. α is set as 0.0001, 0.01, 1.0 and 100. (**A, C, E**) Normalized histograms showing the distributions of angular distances corresponding to different α values. The images were classified into 100 (panel **A**), 300 (panel **C**),

and 500 (panel **E**) classes. (**B, D, F**) The sizes of the classes were ranked for different α values with $K = 100$ (panel **B**), 300 (panel **D**), and 500 (panel **F**).
(PDF)

S6 Fig. Impact of different class numbers K on the unsupervised classification of the simulated data with an SNR of 1/100. (**A**, **C**, **E**) Normalized histograms showing the distributions of angular distances resulting from the five classification methods. The images were classified into 100 (panel **A**), 300 (panel **C**), and 500 (panel **E**) classes. (**B, D, F**) The sizes of the classes were ranked for the five classification methods for $K = 100$ (panel **B**), 300 (panel **D**), and 500 (panel **F**).
(PDF)

S7 Fig. Impact of different initialization strategies in GTM on the unsupervised classification of the simulated data with an SNR of 1/100 into 500 classes. (**A**, **B**, **C**) Examples of the Gaussian distribution circles corresponding to characteristic radii with 0.1×width (panel **A**), 0.25×width (panel **B**) and 0.4×width (panel **C**). (**D**) A normalized histogram showing the distributions of angular distances resulting from different initialization of the weight matrix **W**. GCI refers to the use of the Gaussian distribution circles to initialize the weight matrix **W**. PCAI refers to the use of the PCA to initialize the weight matrix **W**. (**E**) The sizes of the classes were ranked for the cases from different initializing the weight matrices **W**. (**F**) A normalized histogram showing the distributions of angular distances resulting from different initialization of β. β was set to 1, 0.01 and 100. PCAI refers to the use the PCA to initialize β. (**G**) The sizes of the classes were ranked for different initializing β values.
(PDF)

S8 Fig. Comparison of unsupervised classification of the inflammasome dataset by GTM and MAP2D. 17,103 particles of inflammasome were classified into 300 reference-free classes. Only classes whose particle numbers were greater than 9 were shown. (**A**) Unsupervised clustering using GTM in ROME. Among 300 classes, 49 classes showed the various views of inflammasome complexes with different symmetry. (**B**) Unsupervised classification using the MAP2D procedure in RELION. Only 20 classes exhibited views of inflammasome complexes with different symmetry. The MAP2D in RELION generated significantly fewer effective classes than did the GTM in ROME, indicating that GTM is more efficient for distinguishing structural differences.
(PDF)

S9 Fig. Performance evaluation of MAP2D-based classification algorithms implemented in ROME and RELION. A 96,488-particle dataset of the RP-CP subcomplex particles with a box size of 180×180 pixels was used to test the performance of MAP2D in RELION 1.3 and MAP2D in ROME 1.0. When increasing the class number from 30 to 300, the running time of MAP2D classification in ROME followed a polynomial behavior from 92 to 537 minutes (blue histogram).
(PDF)

S10 Fig. Typical weighted loss function of the unsupervised clustering on six datasets using GTM. (**A-D**) The weighted loss function at each iteration for datasets 1 (panel **A**), 2 (panel **B**), 3 (panel **C**) and 4 (panel **D**). The four datasets were classified into 300 classes by GTM. (**E**) The dataset 5 was classified into 1,000 classes. The results of the weighted loss function at each iteration are shown. (**F**) The simulated data were classified into 100, 300 and 500 classes. The results of the weighted loss function at each iteration are shown.
(PDF)

S1 Table. Summary of computational run time of 2D classification based on ROME and RELION for different classification experiments. All tests were executed on the same computing cluster (32 nodes) consisting of 512 Intel Xeon E5 CPU cores. Based on the convergence criteria using the weighted loss function, all datasets converged within 25 iterations (see S10 Fig). RELION in the table refers to the unsupervised MAP2D-based classification in RELION 1.3. ROME_MAP in the table refers to the unsupervised 2D classification based on MAP2D methods implemented in ROME. ROME_SML in the table refers to the unsupervised 2D classification based on the GTM method implemented in ROME. In all cases, the E-M algorithm was run for 30 iterations.
(PDF)

Author Contributions

Conceptualization: Youdong Mao.

Data curation: Jiayi Wu, Yong-Bei Ma, Shuobing Chen, Yaofang Xu, Youdong Mao.

Formal analysis: Jiayi Wu, Yong-Bei Ma, Yaofang Xu, Youdong Mao.

Funding acquisition: Qi Ouyang, Youdong Mao.

Investigation: Jiayi Wu, Yong-Bei Ma, Youdong Mao.

Methodology: Jiayi Wu, Yong-Bei Ma, Charles Congdon, Bevin Brett, Youdong Mao.

Project administration: Jiayi Wu, Yong-Bei Ma, Qi Ouyang, Youdong Mao.

Resources: Youdong Mao.

Software: Jiayi Wu, Yong-Bei Ma, Charles Congdon, Bevin Brett, Youdong Mao.

Supervision: Youdong Mao.

Validation: Shuobing Chen, Yaofang Xu, Youdong Mao.

Visualization: Jiayi Wu, Yaofang Xu, Youdong Mao.

Writing – original draft: Jiayi Wu, Youdong Mao.

Writing – review & editing: Jiayi Wu, Charles Congdon, Bevin Brett, Qi Ouyang, Youdong Mao.

References

1. Frank J. Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state: Oxford University Press; 2006.

2. Nogales E. The development of cryo-EM into a mainstream structural biology technique. Nature methods. 2016; 13(1):24–7. PMID: 27110629

3. Scheres SH, Gao H, Valle M, Herman GT, Eggermont PP, Frank J, et al. Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nature methods. 2007; 4(1):27–9. https://doi.org/10.1038/nmeth992 PMID: 17179934

4. Katsevich E, Katsevich A, Singer A. Covariance matrix estimation for the cryo-EM heterogeneity problem. SIAM journal on imaging sciences. 2015; 8(1):126–85. https://doi.org/10.1137/130935434 PMID: 25699132

5. Schwander P, Fung R, Ourmazd A. Conformations of macromolecules and their complexes from heterogeneous datasets. Phil Trans R Soc B. 2014; 369(1647):20130567. https://doi.org/10.1098/rstb.2013.0567 PMID: 24914167

6. Tagare HD, Kucukelbir A, Sigworth FJ, Wang H, Rao M. Directly reconstructing principal components of heterogeneous particles from cryo-EM images. Journal of structural biology. 2015; 191(2):245–62. https://doi.org/10.1016/j.jsb.2015.05.007 PMID: 26049077

7. Chen S, Wu J, Lu Y, Ma Y-B, Lee B-H, Yu Z, et al. Structural basis for dynamic regulation of the human 26S proteasome. Proceedings of the National Academy of Sciences. 2016; 113(46):12991–6.

8. Zhang L, Chen S, Ruan J, Wu J, Tong AB, Yin Q, et al. Cryo-EM structure of the activated NAIP2-NLRC4 inflammasome reveals nucleated polymerization. Science. 2015; 350(6259):404–9. https://doi.org/10.1126/science.aac5789 PMID: 26449474

9. Murray SC, Flanagan J, Popova OB, Chiu W, Ludtke SJ, Serysheva II. Validation of cryo-EM structure of IP 3 R1 channel. Structure. 2013; 21(6):900–9. https://doi.org/10.1016/j.str.2013.04.016 PMID: 23707684

10. Zhang J, Minary P, Levitt M. Multiscale natural moves refine macromolecules using single-particle electron microscopy projection images. Proceedings of the National Academy of Sciences. 2012; 109 (25):9845–50.

11. Pascual-Montano A, Donate L, Valle M, Barcena M, Pascual-Marqui R, Carazo J. A novel neural network technique for analysis and classification of EM single-particle images. Journal of structural biology. 2001; 133(2):233–45.

12. Singer A, Zhao Z, Shkolnisky Y, Hadani R. Viewing angle classification of cryo-electron microscopy images using eigenvectors. SIAM Journal on Imaging Sciences. 2011; 4(2):723–59. https://doi.org/10.1137/090778390 PMID: 22506089

13. Sorzano C, Bilbao-Castro J, Shkolnisky Y, Alcorlo M, Melero R, Caffarena-Fernández G, et al. A clustering approach to multireference alignment of single-particle projections in electron microscopy. Journal of structural biology. 2010; 171(2):197–206. https://doi.org/10.1016/j.jsb.2010.03.011 PMID: 20362059

14. Yang Z, Fang J, Chittuluru J, Asturias FJ, Penczek PA. Iterative stable alignment and clustering of 2D transmission electron microscope images. Structure. 2012; 20(2):237–47. https://doi.org/10.1016/j.str.2011.12.007 PMID: 22325773

15. Zhao Z, Singer A. Rotationally invariant image representation for viewing direction classification in cryo-EM. Journal of structural biology. 2014; 186(1):153–66. https://doi.org/10.1016/j.jsb.2014.03.003 PMID: 24631969

16. Baldwin P, Penczek PA. The transform class in SPARX and EMAN2. Journal of structural biology. 2007; 157(1):250–61. https://doi.org/10.1016/j.jsb.2006.06.002 PMID: 16861004

17. de la Rosa-Trevin J, Otón J, Marabini R, Zaldivar A, Vargas J, Carazo J, et al. Xmipp 3.0: an improved software suite for image processing in electron microscopy. Journal of structural biology. 2013; 184 (2):321–8. https://doi.org/10.1016/j.jsb.2013.09.015 PMID: 24075951

18. Hohn M, Tang G, Goodyear G, Baldwin P, Huang Z, Penczek PA, et al. SPARX, a new environment for Cryo-EM image processing. Journal of structural biology. 2007; 157(1):47–55. https://doi.org/10.1016/j.jsb.2006.07.003 PMID: 16931051

19. Scheres SH. RELION: implementation of a Bayesian approach to cryo-EM structure determination. Journal of structural biology. 2012; 180(3):519–30. https://doi.org/10.1016/j.jsb.2012.09.006 PMID: 23000701

20. Shaikh TR, Gao H, Baxter WT, Asturias FJ, Boisset N, Leith A, et al. SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nature protocols. 2008; 3(12):1941–74. https://doi.org/10.1038/nprot.2008.156 PMID: 19180078

21. Sorzano C, Marabini R, Velázquez-Muriel J, Bilbao-Castro JR, Scheres SH, Carazo JM, et al. XMIPP: a new generation of an open-source image processing package for electron microscopy. Journal of structural biology. 2004; 148(2):194–204. https://doi.org/10.1016/j.jsb.2004.06.006 PMID: 15477099

22. Van Heel M, Frank J. Classification of particles in noisy electron micrographs using correspondence analysis. Pattern Recognition in Practice I Gelsema ES, Kanal L (eds) North-Holland Publishing, Amsterdam. 1980:235–43.

23. Van Heel M, Frank J. Use of multivariate statistics in analysing the images of biological macromolecules. Ultramicroscopy. 1981; 6(2):187–94. PMID: 7268930

24. van Heel M, Stöffler-Meilicke M. Characteristic views of E. coli and B. stearothermophilus 30S ribosomal subunits in the electron microscope. The EMBO journal. 1985; 4(9):2389. PMID: 3908096

25. Scheres SH. A Bayesian view on cryo-EM structure determination. Journal of molecular biology. 2012; 415(2):406–18. https://doi.org/10.1016/j.jmb.2011.11.010 PMID: 22100448

26. Scheres SH, Valle M, Nuñez R, Sorzano CO, Marabini R, Herman GT, et al. Maximum-likelihood multireference refinement for electron microscopy images. Journal of molecular biology. 2005; 348(1):139–49. https://doi.org/10.1016/j.jmb.2005.02.031 PMID: 15808859

27. Sigworth F. A maximum-likelihood approach to single-particle image refinement. Journal of structural biology. 1998; 122(3):328–39. https://doi.org/10.1006/jsbi.1998.4014 PMID: 9774537

28. Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. science. 2000; 290(5500):2319–23. https://doi.org/10.1126/science.290.5500.2319 PMID: 11125149

29. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000; 290(5500):2323–6. https://doi.org/10.1126/science.290.5500.2323 PMID: 11125150

30. Silva VD, Tenenbaum JB, editors. Global versus local methods in nonlinear dimensionality reduction. Advances in neural information processing systems; 2002.

31. Law MH, Jain AK. Incremental nonlinear dimensionality reduction by manifold learning. IEEE transactions on pattern analysis and machine intelligence. 2006; 28(3):377–91. https://doi.org/10.1109/TPAMI.2006.56 PMID: 16526424

32. Zhang Z, Chow TW, Zhao M. M-Isomap: Orthogonal constrained marginal isomap for nonlinear dimensionality reduction. IEEE transactions on cybernetics. 2013; 43(1):180–91. https://doi.org/10.1109/TSMCB.2012.2202901 PMID: 22773048

33. Bishop CM, Svensén M, Williams CK. GTM: The generative topographic mapping. Neural computation. 1998; 10(1):215–34.

34. Bishop CM, Svensén M, Williams CK. Developments of the generative topographic mapping. Neurocomputing. 1998; 21(1):203–24.

35. Gorban AN, Kégl B, Wunsch DC, Zinovyev AY. Principal manifolds for data visualization and dimension reduction: Springer; 2008.

36. Saul LK, Roweis ST. Think globally, fit locally: unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research. 2003; 4(Jun):119–55.

37. Zhang Z, Chow TW, Zhao M. Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization. IEEE Transactions on Knowledge and Data Engineering. 2013; 25(5):1148–61.

38. Weinberger KQ, Saul LK. Unsupervised learning of image manifolds by semidefinite programming. International Journal of Computer Vision. 2006; 70(1):77–90.

39. Yan S, Hu Y, Xu D, Zhang H-J, Zhang B, Cheng Q. Nonlinear discriminant analysis on embedded manifold. IEEE Transactions on Circuits and Systems for Video Technology. 2007; 17(4):468–77.

40. Zhang Z, Li F, Zhao M, Zhang L, Yan S. Robust neighborhood preserving projection by nuclear/l2, 1-norm regularization for image feature extraction. IEEE Transactions on Image Processing. 2017; 26(4):1607–22. https://doi.org/10.1109/TIP.2017.2654163 PMID: 28103554

41. Yan S, Wang H, Fu Y, Yan J, Tang X, Huang TS. Synchronized submanifold embedding for person-independent pose estimation and beyond. IEEE Transactions on Image Processing. 2009; 18(1):202–10. https://doi.org/10.1109/TIP.2008.2006400 PMID: 19068429

42. Dashti A, Schwander P, Langlois R, Fung R, Li W, Hosseinizadeh A, et al. Trajectories of the ribosome as a Brownian nanomachine. Proceedings of the National Academy of Sciences. 2014; 111(49):17492–7.

43. Jeffers J, Reinders J. Intel Xeon Phi coprocessor high-performance programming: Newnes; 2013.

44. Van Heel M. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy. 1987; 21(2):111–23. PMID: 12425301.

45. Allinson N, Yin H, Allinson L, Slack J. Advances in self-organising maps: Springer; 2001.

46. Tagare HD, Barthel A, Sigworth FJ. An adaptive Expectation–Maximization algorithm with GPU implementation for electron cryomicroscopy. Journal of structural biology. 2010; 171(3):256–65. https://doi.org/10.1016/j.jsb.2010.06.004 PMID: 20538058

47. Tang G, Peng L, Baldwin PR, Mann DS, Jiang W, Rees I, et al. EMAN2: an extensible image processing suite for electron microscopy. Journal of structural biology. 2007; 157(1):38–46. https://doi.org/10.1016/j.jsb.2006.05.009 PMID: 16859925

48. Penczek P, Radermacher M, Frank J. Three-dimensional reconstruction of single particles embedded in ice. Ultramicroscopy. 1992; 40(1):33–53. PMID: 1580010

49. Lu Y, Wu J, Dong Y, Chen S, Sun S, Ma Y-B, et al. Conformational Landscape of the p28-Bound Human Proteasome Regulatory Particle. Molecular Cell. 2017. https://doi.org/10.1016/j.molcel.2017.06.007 PMID: 28689658

50. Suloway C, Pulokas J, Fellmann D, Cheng A, Guerra F, Quispe J, et al. Automated molecular microscopy: the new Leginon system. J Struct Biol. 2005; 151(1):41–60. https://doi.org/10.1016/j.jsb.2005.03.010 PMID: 15890530.