JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2019, 59, 1658−1667

pubs.acs.org/jcim

Clustering Enhancement of Noisy Cryo-Electron Microscopy Single-Particle Images with a Network Structural Similarity Metric

Shuo Yin,[†] Biao Zhang,[†] Yang Yang,[‡] Yan Huang,[§] and Hong-Bin Shen*[,†,‡]

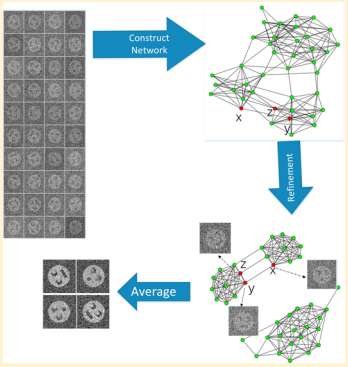
[†]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

[‡]Department of Computer Science, Shanghai Jiao Tong University, and Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai 200240, China

[§]State Key Laboratory of Infrared Physics Shanghai Institute of Technical Physics, Chinese Academy of Sciences, 500 Yutian Road, Shanghai 200083, China

Supporting Information

ABSTRACT: The reconstruction of a three-dimensional model from cryo-electron microscopy (cryo-EM) two-dimensional images is currently a mainstream technology for revealing the structure of biomacromolecules. In this structure solution protocol, an important step is to identify each particle's projection orientation. Because the obtained single-particle images are often too noisy, clustering is an important step to mitigate noise by averaging images within the same class. The core of clustering is to place similar cryo-EM images into the same class; hence, measurement of similarity between data samples is an essential element in any clustering algorithm. As the cryo-EM images are highly noisy, directly measuring the similarity of two images will be easily biased by the hidden noise. In this study, we propose a new network structural similarity metric-based clustering protocol NCEM for clustering the noisy cryo-EM images. We first construct an image complex network for all of the cryo-EM single-particle images, where each image is represented as a node in the network. Then the similarity between two images is refined from the network structural geometry. By extending the similarity measurement from two independent images to their corresponding neighboring sets in the network, this new NCEM has typical advantages over direct measurement of two images for its noise resistance by using the structural information on the network. Our experimental results for both synthetic and real data sets demonstrate the efficacy of the protocol.



1. INTRODUCTION

Because of the technical breakthroughs associated with electron detectors in the past decade, cryo-electron microscopy (cryo-EM) has been a powerful technique for analyzing the structure of a large variety of biological specimens.[1] Unlike X-ray crystallography, which needs crystallization, and nuclear magnetic resonance (NMR), which is difficult with large biomolecules, cryo-EM requires fewer restrictions on samples, making it a more efficient structure solution technique. With the recent progress in data quality and data processing, cryo-EM reconstructions are now able to generate nearly atomic models.[2,3] Moreover, many software tools specialized for cryo-EM image processing have been developed in recent years, which significantly accelerate the handling of high-throughput cryo-EM images. To date, cryo-EM has become a mainstream technology in structural biology and attracted more and more attention in both biology and information science, e.g., guiding drug design.[4]

Upon application of electron microscopy, the cryo-EM images are typically acquired using defocus contrast. The resulting image can be modeled by an ideal phase-contrast image that is modulated by a variable scaling of its Fourier components. The scaling factors are called the CTF.[5] Therefore, the first step of single-particle image analysis is known as the CTF estimation and correction on the acquired micrograph images. Then particle picking is performed to extract single-particle images from micrograph images. Usually, a small electron dose will be used to avoid radiation damage on the targeting molecules. Also, because of the low contrast between the molecule and its surrounding background, the images, i.e., two-dimensional (2D) projections of the single particle, usually exhibit a very low signal-to-noise ratio (SNR) and very poor contrast,[6] which makes it necessary to perform preanalysis on the 2D particle images.

For instance, from a large pool of the obtained 2D images, a necessary step is assigning images from heterogeneous projections into homogeneous classes to separate images from different conformations, which is denoted as "heterogeneous clustering" in this paper. A following processing step will be clustering the images from the same conformation into groups of the same or similar projection orientations to obtain the class mean images for input to the three-dimensional (3D) reconstruction algorithm. This can be called "homogeneous

Received: November 26, 2018
Published: January 24, 2019

DOI: 10.1021/acs.jcim.8b00853
J. Chem. Inf. Model. 2019, 59, 1658−1667

clustering" and is an important step for reducing the bias effects caused by the noise but has inherent difficulty because of the unsupervised clustering setting and an extremely low SNR of the images. The clustering quality will directly affect subsequent reconstruction performance, as the dissimilar angle projection images will dramatically decrease the quality of class mean images, which are reconstruction inputs.

An initial model is generated from class mean images and then refined using all particle images. Finally, the 3D model's resolution will be calculated in a blind manner because no ground truth is available in the real-world applications.[7] In this paper, we will focus on the homogeneous clustering problem to develop a new network-based similarity metric for enhancing the cryo-EM image clustering performance.

As an important step, clustering of single-particle images to find their projection angles is a long-term challenge, which has already been implemented in many common software packages for single-particle image analysis. Many approaches have been proposed. For example, EMAN2[8,9] software combines MSA with multireference alignment (MRA) for the clustering purpose. It first generates translational and rotational invariants for an initial classification. Then, a MSA step is iterated with a MRA step, in which images are aligned to those features and clustered by the *K*-means algorithm, until a predefined number of iterations is reached.

In the XMIPP package, its clustering module is called CL2D,[10] which is also a modified *K*-means method. CL2D uses correntropy[11] as a measurement of similarity of input images and proposes a new clustering criterion to substitute the traditional "nearest" criterion to address the varied SNR issue. Specifically, for the clusters, if their mean images (class averages) have very different SNRs, then it would be improper to simply determine the membership of an image according to the similarity to the class averages. To solve this problem, CL2D proposes a robust criterion for measuring the similarity between an image and a class average compared with other images. Unlike a traditional clustering criterion, which assigns an image to a class average with the highest similarity, CL2D's robust criterion ranks all images with respect to different class averages and recalculates the similarity considering this between-images information. For each class average C_i, it keeps two vectors, V_{in} and V_{out}, where V_{in} keeps correntropy values between images assigned to the class and the class average C_i while V_{out} keeps correntorpy values between images not assigned to this class and the class average C_i. The recalculated similarity between an input image z_i and a class average C_i is calculated by counting how many values in V_{in} and V_{out} are less than the correntorpy between z_i and C_i.

The SIMPLE package uses a greedy *K*-means-based approach to cluster the images, which is named PRIME_CLUSTER.[12] It uses a stochastic hill climbing approach to maximize the correlation between images and their cluster averages (in Fourier space). In the latest improved version of PRIME2D,[13] a simulated annealing approach has been incorporated to optimize the clustering objective functions.

Relion[14] employs an empirical Bayesian approach for 2D classification. It introduces prior information as a regularization term into a traditional likelihood function, which has proven to be useful for both high-resolution reconstruction and 2D and 3D classification in a wide range of experimental studies.[15] Other very popular approaches for 2D classification and averaging that are used include reference-free alignment (RFA) followed by *K*-means clustering as implemented in SPIDER,[16] multireference alignment and multivariate statistical analysis (MRA/MSA) as implemented in IMAGIC,[17] and iterative stable alignment and clustering (ISAC) available in SPARX and SPHIRE.[18]

Although many attempts have been reported, significant challenges remain in this field, for instance, how to measure the similarity of two low-SNR cryo-EM images, which is a fundamental problem in the unsupervised clustering algorithm. Almost all methods described above use "pairwise" similarity, which use only two images to calculate their similarity. In the case of a very low SNR, the similarity between two images derived from a pairwise measurement like correlation or common line would be unreliable because the hidden noise will significantly affect the measurement.

A network or a nearest neighbor-based network have been used in the single-particle image analysis. For example, some works[19,20] use the network partition idea in the heterogeneous image clustering. They first construct a fully connected network by calculating their modified common line similarity and then use the network partition method to generate desired clusters. Common line similarity measures the probability of two images coming from the same conformation. This is different from the homogeneous image clustering, where we need to calculate the probability of two images coming from the same projection orientation. Another recent work incorporated the network idea for noise reduction.[21] They first construct a *K*-nearest neighbor network using their defined image representation, and instead of clustering on the network, they compute the average of *K*-means nearest images for each node, which could be considered as the noise reduction of the images. Also, complex network or graph theory is widely used in molecular simulation to identify and statistically analyze the changes in conformation.[22] For example, graph theory is used to find evidence of a protospacer adjacent motif (PAM)-induced allosteric mechanism.[23]

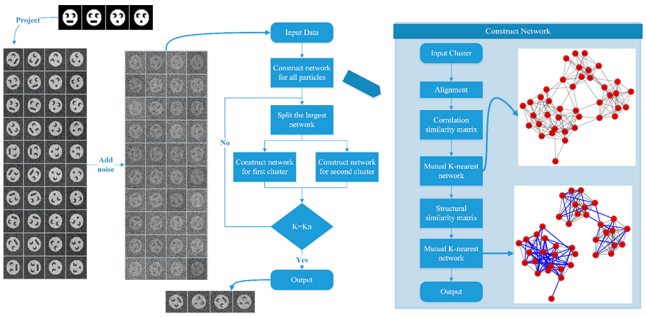
Here we propose a novel network structure metric-based homogeneous clustering approach, NCEM. We first construct an image network for all of the image samples, where each node represents an image. The similarity between two images (nodes) will be recalculated and refined from the network's local or global geometry structure. The merit of such a network-based structural metric is it is robust against perturbations, which are image noise in our case. Experimental results on both synthetic and real-world data sets have demonstrated the efficacy of the proposed NCEM protocol in homogeneous clustering.

2. NCEM ALGORITHM

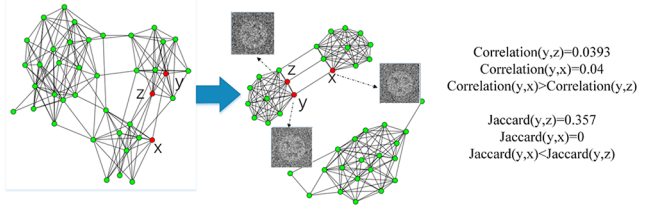
The basic idea of the NCEM algorithm is to cluster the noisy cryo-EM images with a new network-based structural similarity. Three modules are included in NCEM: (1) initial cryo-EM image network construction, (2) initial image network refinement, and (3) hierarchical alignment and spectral clustering.

2.1. Initial Cryo-EM Image Network Construction. Because our input is individual images, we need to construct a network first, where the node represents the image and the edge is determined by a similarity metric. Here we adopt correlation as the similarity measurement to construct the initial network. The correlation between two images (X, Y) is defined as

DOI: 10.1021/acs.jcim.8b00853
J. Chem. Inf. Model. 2019, 59, 1658−1667



(a) Flowchart of NCEM



(b) Case study

Figure 1. (a) Flowchart of the NCEM protocol. On the left is the experimental data generation procedure. The middle panel is a high-level procedure of our algorithm, while the right panel is the detailed workflow of how to construct a network for a given cluster of images. (b) Case study showing the difference in the correlation-based similarity measurement and the network-based Jaccard refined similarity measurement.

High level procedure:
Input: Dataset D, final number of classes K_n
Output: Class information
Do:
Construct network N^R for all images and set current number of class K=1
While current number of class K!=K_n
Find the largest class and its corresponding network net_0;
Split the largest class into two sub classes using spectral clustering;
Construct network net_1, $net2$ for two new class.
Delete net_0 and keep net_1 and $net2$, K+=1
End while
Save the information of classes and exit
Network construction and refinement
Input: a set of images S
Output: image network N^R
Do:
Align images
Compute correlation matrix M_{corr}
Compute mutual K-nearest network N using K= ||S||/3
Compute structural similarity M_{str}
Compute mutual K-nearest network N^R using K= ||S||/3
Output N^R

Figure 2. Pseudocode of the network-based cryo-EM clustering algorithm NCEM.

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{(X - \mu_X)^T (Y - \mu_Y)}{\|X - \mu_X\| \times \|Y - \mu_Y\|} \quad (1)$$

where we treat images X and Y as two one-dimensional vectors and μ_X and μ_Y are the expectations of images X and Y, respectively. Because their expectation is unknown, we use an average image of the data set in our algorithm.

There are many other pairwise similarities, like Fourier-based steerable PCA followed correlation[21] or correntorpy.[10] Considering the computational efficiency, we applied the correlation (eq 1) as our pairwise similarity measure for constructing the initial network. Given a similarity metric between the image samples, there are multiple ways to construct an image network, such as an ϵ-neighborhood network, a k-nearest neighbor (KNN) network, a mutual k-nearest neighbor (mutual KNN) network, and a fully connected network.[24] For the same data set, different types of networks will lead to different clustering results even with the same clustering algorithms due to different network topologies generated by different methods. We have tried different approaches and will use a mutual k-nearest neighbor network in this study (see Discussion for more details).

2.2. Network Structural Similarity Definition for Cryo-EM Images. To achieve accurate clustering, measuring the similarity between two cryo-EM images is critically important. Similarity metrics like correlation, Euclidean distance, or correntropy are pairwise-based measurements; i.e., the value of similarity depends on only the two considered images or nodes. Because of the low SNR of the input images, noise may dominate the computation of similarity and may result in a small correlation for two similar images. Aiming to develop a more robust similarity metric, we adopt network-based structural similarity measurement. This new structural similarity idea will utilize the local or global structure of the network to calculate the similarity between two nodes, which will be more robust compared to traditional methods considering only two nodes.

Because of the development of network science,[25] many structural similarity approaches can be used. Given a constructed network $M = [V, E]$, allowing $\Gamma(x)$ be the set of neighbors of node x, we can define the following structural similarity S_{xy} between x and y:

(1) Jaccard index:

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (2)$$

(2) Sørensen index:[26]

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{k_x + k_y} \quad (3)$$

where k_x and k_y are degrees of nodes x and y, respectively.

(3) HD index (hub depressed index):

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max\{k_x, k_y\}} \quad (4)$$

(4) RA index (resource allocation index):

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z} \quad (5)$$

where k_z is degree of node z.

(5) Katz index:[27]

$$S_{xy} = \sum_{l=1}^{\infty} \beta^l |\mathrm{paths}_{xy}^l| = \sum_{l=1}^{\infty} \beta^l (A^l)_{xy} \quad (6)$$

where β is a free parameter, $|\mathrm{paths}_{xy}^l|$ is the number of the paths with length l between x and y, and A is the adjacent matrix. Note that β must be lower than the reciprocal of the largest eigenvalue of matrix A to ensure the convergence. The Katz index has taken all paths from node x to node y into consideration.

(6) Leicht−Holme−Newman index (LHN2):

DOI: 10.1021/acs.jcim.8b00853
J. Chem. Inf. Model. 2019, 59, 1658−1667

$$S = \lambda_1 D^{-1}\left(I - \frac{\phi A}{\lambda_1}\right)^{-1} D^{-1} \tag{7}$$

where λ_1 is the largest eigenvalue of adjacent matrix A, D is the degree matrix, and ϕ is a free parameter, where a smaller φ will assign higher weights on the shorter paths. The LHN2 index is a variant of the Katz index based on the concept that two nodes are similar if their immediate neighbors are similar.

Jaccard similarity and Sørensen similarity are local structural similarities, which means they consider only two-step-length paths between two nodes. The Katz is a global structural similarity, which means it considers all paths between two nodes. To obtain the refined network N^R from initial network N, we first calculate structural similarity matrix S using the nodes of structural similarity defined above and then construct a new mutual k-nearest network based on S. When using a different network node metric, we will obtain different variants of the proposed NCEM, e.g., NCEM-Jaccard for using the Jaccard index, NCEM-Sørensen for using the Sørensen index, etc.

2.3. Hierarchical Alignment and Spectral Clustering. Then, our purpose now is to partition the image network N^R into meaningful groups or clusters. Many algorithms have been proposed for this task. Here, we adopt the spectral clustering algorithm implementation[28] for the consideration of its efficiency in dealing large-scale data set by standard linear algebra methods as well as its performance. The spectral clustering algorithm aims to find a network partition, which maximizes the ratio of edges within classes and all edges in the whole network:

$$\begin{cases} \text{maximize} \dfrac{1}{K}\displaystyle\sum_{i=1}^{K} \dfrac{\Upsilon_i^T A \Upsilon_i}{\Upsilon_i^T D \Upsilon_i} \\ s.\,t.\ \Upsilon \in \{0, 1\}_{P\times K} \end{cases} \tag{8}$$

where K is the expected number of classes, P is the number of vertices in the network, A is the adjacent matrix of the network, and D is degree matrix of the network. This optimization problem is NP-Complete even for $K = 2$ and even when the network is planar.

As demonstrated in previous studies,[21] any classification algorithm that first attempts to globally align the images would fail whenever there are many different views that cover the sphere. The reason is that in this situation, global alignment will assign very different angles for images with similar projection orientations. Therefore, rather than align all images once and for all, we use a hierarchical clustering approach that will align images within every subcluster. Divisive clustering can work against the local optimal,[29] although the global optimal is not guaranteed. If now we have k classes and we want K classes, $k < K$, then we extract the subnet of the largest class and split it into two parts using spectral clustering algorithm. After this, we will have $k + 1$ classes from the two newly generated classes. Then we align images and construct a new network for these two subnets separately. This process is repeated until the number of classes reaches K.

2.4. Flowchart of NCEM. Figure 1a shows the whole flowchart of our new network-based cryo-EM image clustering algorithm NCEM, and Figure 1b is a case study showing why the network-based similarity metric is better. We adopt a hierarchical divisive clustering approach. In each iteration, we first construct an initial image network, where the structural similarity will be calculated by considering the network topological geometry. On the basis of the new similarity measurement, the refined network is constructed and then a spectral clustering algorithm is applied for the partitioning. Figure 2 shows the pseudocode of the proposed NCEM. Here we use XMIPP's *xmipp_image_align* program for image alignment. It is a template-based alignment algorithm, and if not provided, it generates templates by averaging all images in the data set.

To show the effectiveness of structural similarity, we generated four faces as template images. These faces differ in their eyes (left vs right) and mouth (big vs small). Ten copies of each image with different rotation angles compose a clean data set. By adding noise to the clean data set, we obtain a noisy data set with SNR = 0.2, as shown in the left part of Figure 1a. In the right part of Figure 1a, we use all images to show the network construction procedure. The first network plot is a mutual K-nearest network based on a correlation matrix with K equal to 10. The second network plot is a mutual K-nearest network based on the Jaccard similarity matrix with K equal to 10. Edges existing in the second network and absent in the first network are colored blue. As one can clearly see from the figure, after structural similarity refinement, the new network has a clearer structure, which makes it easier for the clustering algorithm to partition. The Jaccard index is correct because it considers the nodes' neighbors, showing an interesting "many is better than one" phenomenon. To further show the advantage of the Jaccard index, we marked three images (x, y, z) in the networks, shown in Figure 1b, where y and z belong to the same class and z is a rotated image of y. Because of the noise effects and rotation, the correlation between x and y (0.04) is even higher than that between y and z (0.0393), which will lead to a misassignment in this case. However, if we calculate the network Jaccard index of these nodes, we will find that the Jaccard metric between x and y (0) is obviously lower than that between z and y (0.357), which is a reasonable result.

3. EXPERIMENTAL RESULTS

To evaluate the performance of our method, we conducted experiments using both artificial and real data sets. We compared our results with results from CL2D[10] in XMIPP, Prime2D[12] in SIMPLE, EMAN2, and Relion. For CL2D, we use correntropy and robust clustering as recommended in ref 10. For Prime2D, we kept the parameter default set in the program. For EMAN2 and Relion, we set the number of iterations to 10. For our method, we set β in the Katz index equal to half of A's largest eigenvalue and φ in the LHN2 index to 0.4. All of these experiments are conducted on a computer with four Intel Core i7 central processing units and an Ubuntu 16.04 system.

3.1. The Structural Similarity Metric Is Better Than the Pairwise Measurement, Showing the "Many Is Better Than One" Phenomenon. The key point of our new structural similarity approach is extending the measurement of the similarity between two cryo-EM images from a "pair"-based to a "set"-based protocol. Instead of determining whether two cryo-EM images are similar to each other or not, we consider their sets of network neighbors in the new approach. This is expected to be particularly useful in the case of a low SNR of cryo-EM images.

To further intuitively illustrate the effectiveness of structural similarity, we generated two projections from protein EMD-

DOI: 10.1021/acs.jcim.8b00853
J. Chem. Inf. Model. 2019, 59, 1658−1667

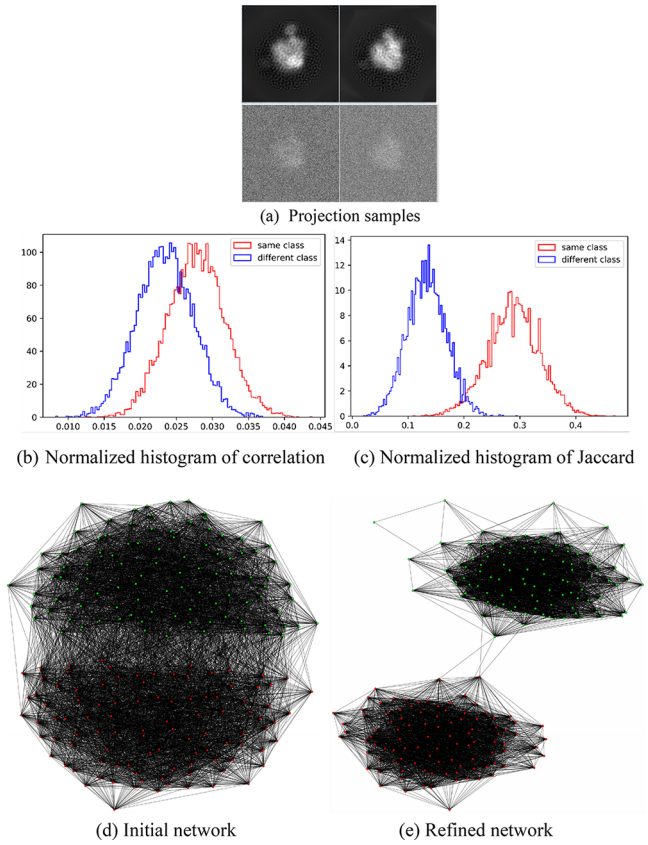


Figure 3. (a) Projections of EMD-5780 and two classes of samples. (b) Histogram of correlation values between two images within and without in the same class. (c) Histogram of network Jaccard values between two images within and without in the same class. (d) Initial mutual KNN network constructed from correlation. (e) Refined network with Jaccard structure similarity.

5780[30] and then rotated the two images and added noise to them to form a small data set containing two classes. Each class contained 100 images; samples and their clean projections are shown in Figure 3a. We then calculated the correlation between images. To evaluate the performance of correlation-based clustering, we plotted the normalized correlation histogram of intra- and interclass image correlation values within and without in the same class, as shown in Figure 3b. We then applied our NCEM approach and constructed a mutual KNN network using NN = 75 and then calculated the Jaccard similarity on this network. The normalized histogram of the Jaccard values is shown in Figure 3c, where we can clearly see that the network Jaccard is a better discriminative value of the two classes. We also constructed a mutual KNN network from the two similarity matrices using NN = 75. Figure 3d shows the initial network, which is constructed from the correlation matrix, and Figure 3e shows a refined network from Jaccard structure similarity. Nodes from different classes are labeled with different colors (green vs red). We can intuitively see that the refined network has fewer connections between the two classes, suggesting a better ability to discriminate.

To numerically show Jaccard's advantages over correlation, we counted the 100 nearest neighbors for each image under the correlation and Jaccard similarity measurement. We found that under correlation measurement, 70.90% images among the 100 nearest neighbors belonged to the same class as the node. While using the Jaccard measurement, the percentage was 97.03%. These results suggested that the network Jaccard metric can help rerank the neighbors and cluster the similar images into the same group.

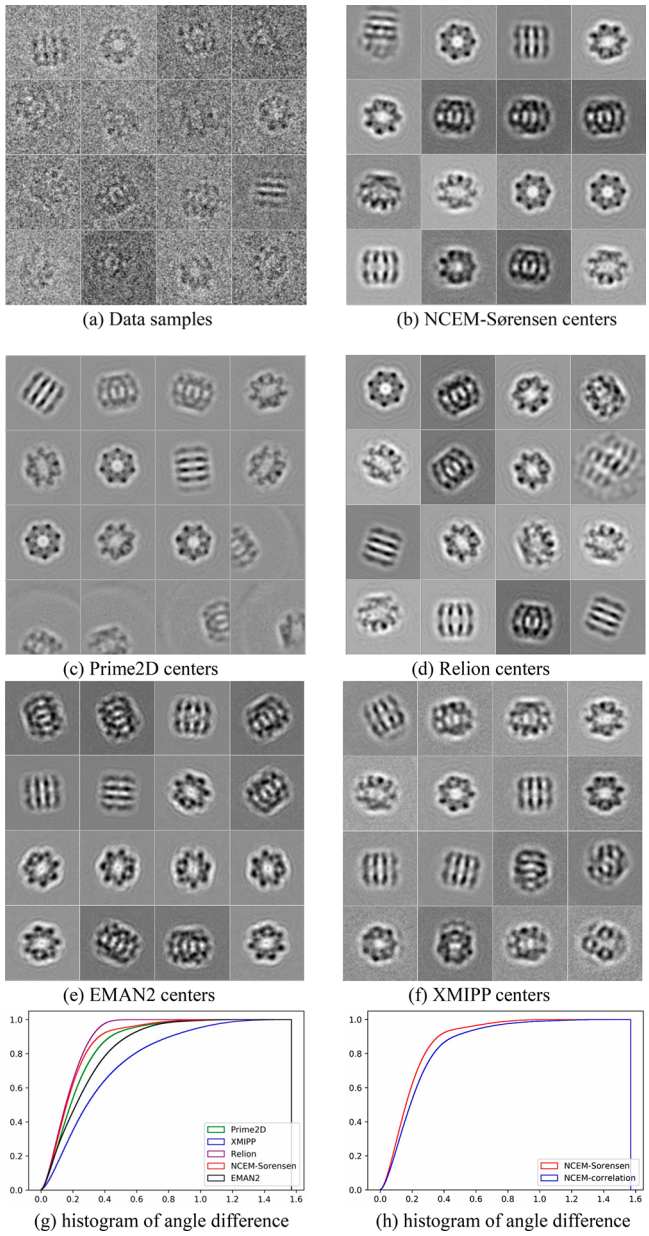


Figure 4. Clustering results on the EMPIAR-10029 data set. (a) Image samples in this data set. (b) Class averages given by the NCEM−Sørensen approach. (c−f) Class averages given by Prime2D, Relion, EMAN2, and XMIPP, respectively. (g) Cumulative histogram of the angle difference between images within the same class using different methods. (h) Cumulative histogram of the angle difference between images within the same class using the refined network (Sørensen) and the initial network (correlation).

Table 1. Area under the Histogram Curve

	Relion	Prime2D	EMAN2	CL2D	NCEM−Sørensen	NCEM−correlation
AUC	1.400	1.345	1.307	1.201	1.374	1.349

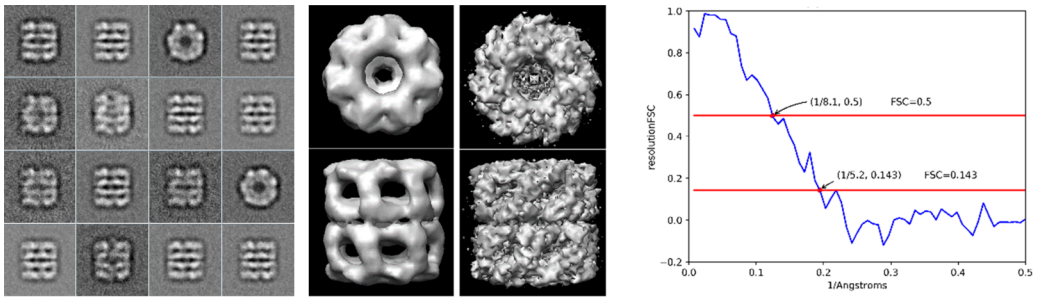
3.2. Clustering Results on a Large Synthetic Data Set.

Here we use the EMPIAR-10029 synthetic benchmark data set

DOI: 10.1021/acs.jcim.8b00853
J. Chem. Inf. Model. 2019, 59, 1658−1667

Table 2. FSC Resolution Values of Different Initial Models on the GroEL Data Set

Initial 3D model resolution	Relion	EMAN2	CL2D	Prime2D	NCEM−Katz
FSC 0.5	12.8 Å	25.6 Å	25.6 Å	21.3 Å	8.05 Å
FSC 0.143	8.0 Å	18.3 Å	18.3 Å	16.0 Å	5.15 Å
reconstruction time	34 min 3 s	21 min 09 s	6 h 34 min	2 h 25 min	37 min 20 s



(a) Katz class averages (b) Initial and real model (c) NCEM-Katz-based FSC Curve

Figure 5. Results of the initial model resolution on the GroEL data set from the NCEM−Katz method. (a) Class averages generated by the NCEM−Katz method. (b) Real model and reconstructed initial model. The right column is from the real model, and the left column is from the reconstructed model. The top row shows top views, and the bottom row shows side views. (c) FSC curve of the NCEM−Katz algorithm.

Table 3. FSC Resolution Values of Different Initial Models on the EMPIAR-10034 Data Set

Initial 3D model resolution	Relion	EMAN2	CL2D	Prime2D	NCEM−Jaccard
FSC 0.5	25.6 Å	42.7 Å	42.7 Å	42.7 Å	21.3 Å
FSC 0.143	21.3 Å	16.0 Å	18.3 Å	16.0 Å	16.0 Å
reconstruction time	10 h 15 min	1 h 30 min	113 h 24 min	8 h 18 min	3 h 36 min

for a clustering performance comparison, which is projected from protein GroEL (Protein Data Bank entry 4HEL), which is a publicly available benchmark data set. This data set contains 10000 images that are 200 × 200 in size. Some samples are shown in Figure 4a. We cluster the images into 16 classes using Prime2D, Relion, EMAN2, CL2D, and our NCEM. Here we set mutual nearest neighbors to 1000 in NCEM considering the large number of samples. The class averages of these methods are shown in panels b−f of Figure 4, respectively. As we did for the evaluation metric of CL2D,[10] we evaluated the results using the distribution of the angle difference of images within the same classes as we knew the projection orientation for every projection in this synthetic data set. We plotted the cumulative histogram of their differences in Figure 4g (see Figure S1 for more results for five other structural similarity metrics). As one can see from this figure, the curve of the NCEM−Sørensen approach is above those of CL2D, Prime2D, and EMAN2, illustrating that classes generated by the NCEM−Sørensen approach are denser (better). Figure 4h compares the clustering results on the refined image network (NCEM−Sørensen) and on the initial image network (NCEM−correlation). The results also show that by using the structural similarity Sørensen, the clustering results will be better, demonstrating the effectiveness of set-based similarity measurement.

Also for a numeric comparison, we calculate the area under the cumulative histogram curve, shown in Table 1. A better clustering result will have a higher value with $\pi/2$ as its upper bound, because the maximum angle difference is $\pi/2$. We can see that, in this data set, our algorithm performs worse than Relion but better than the other three. In addition, we can see that, if we performed the clustering on the initial network using the correlation metric without the network refinement procedure, the AUC will be 1.349, which is below the value from the NCEM−Sørensen approach (see also Figure 4h), indicating quantitatively the necessaritiy of the network refinement. For AUC scores of other structural similarities, see Table S1.

3.3. Experimental Results on a Real-World GroEL Data Set. Additionally, we have used a GroEL real data set[31] in this experiment. This data set contains 26 micrographs 4082 pixels × 6278 pixels in size. The sampling rate is 2.10 Å/pixel, and the microscope voltage is 200 kV. We picked 4694 particles that were 128 × 128 in size from this data set. The corresponding EMDB code of this data set is 1081. We cluster this data set into 16 classes as in the literature.

In terms of the time consumed, EMAN2 runs 21 min, CL2D runs 6 h 34 min, Prime2D takes 2 h 25 min, and Relion takes 30 min, while our NCEM algorithm runs from 33 to 37 min. We then reconstruct an initial volume and use its resolution as a measurement of the clustering performance, following the evaluation method in the study of Prime2D.[12] Here we measure the FSC curve between the EMDB map (EMDB code 1081) and the initial model. We use EMAN2's "e2initialmodel" for this purpose. The "e2initialmodel" program implements a standard projection matching algorithm. It outputs 10 initial models at one time; usually models are sorted according to their qualities (the best is model #1).

Here we pick the best model's FSC value for each method. We evaluate the resolution of each model using an FSC 0.5 value and an FSC 0.143 value (a lower value means a better resolution), as shown in Table 2. We also plot the class averages, initial model, real model, and FSC curve of the NCEM−Katz approach in Figure 5. We found in our experiments that the NCEM-Katz approach shows a better performance on this dataset; for more results of NCEM from

DOI: 10.1021/acs.jcim.8b00853
J. Chem. Inf. Model. 2019, 59, 1658−1667

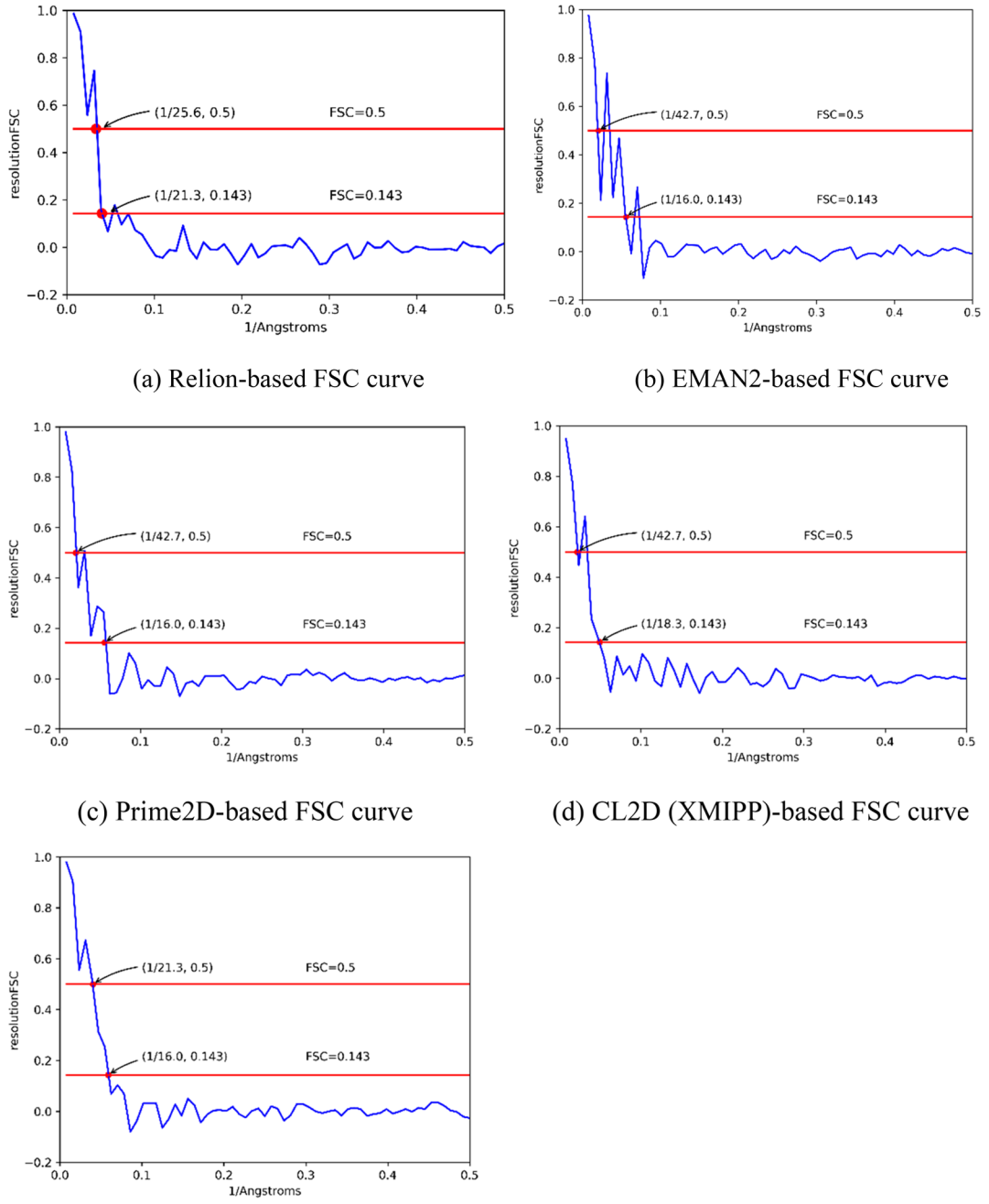


Figure 6. FSC curves of the initial reconstructed models using the cluster averages from different clustering algorithms: (a) Relion clustering algorithm, (b) EMAN2 clustering algorithm, (c) Prime2D clustering algorithm, (d) CL2D (XMIPP) algorithm, and (e) NCEM−Jaccard algorithm.

different similarity metrics, see Table S2. As shown in Table 2, the model from the NCEM−Katz approach has highest resolution, which suggests our model is more similar to the EMD-1081 map. Considering each of these algorithms generates 16 classes and the procedure of the projection matching algorithm, a lower FSC value may indicate more accurate clustering, which means similar images can be grouped into the same classes, and more representative class averages.

3.4. Experimental Results for the EMPIAR-10034 Data Set. To test the algorithm on a larger real-world data set, we used EMPIAR-10034, which is composed of CTF-corrected particle images of tubulin chaperone complexes TBC-DEG Q73L:$\alpha\beta$-tubulin:TBCC.[32] It contains 16801 images, and each image is 128 × 128 in size. It is publicly available. We grouped the images into 128 classes like the others.

In this data set, CL2D runs for more than 4 days, Prime2D takes 8 h 18 min, EMAN2 takes 1 h 30 min, Relion takes 10 h 15 min, while our NCEM−Jaccard algorithm runs from 3 to 6 h (Table 3). We follow the same initial 3D model evaluation method in section 3.3, and the results are listed in Table 3. The FSC curves of the NCEM−Jaccard method and four other methods are shown in Figure 6. For all other variants of NCEM methods, see Table S3 and Figure S3.

DOI: 10.1021/acs.jcim.8b00853
J. Chem. Inf. Model. 2019, 59, 1658−1667

4. DISCUSSION

4.1. Initial Network Construction. Except for correlation, we also explored the correntropy image similarity measure-

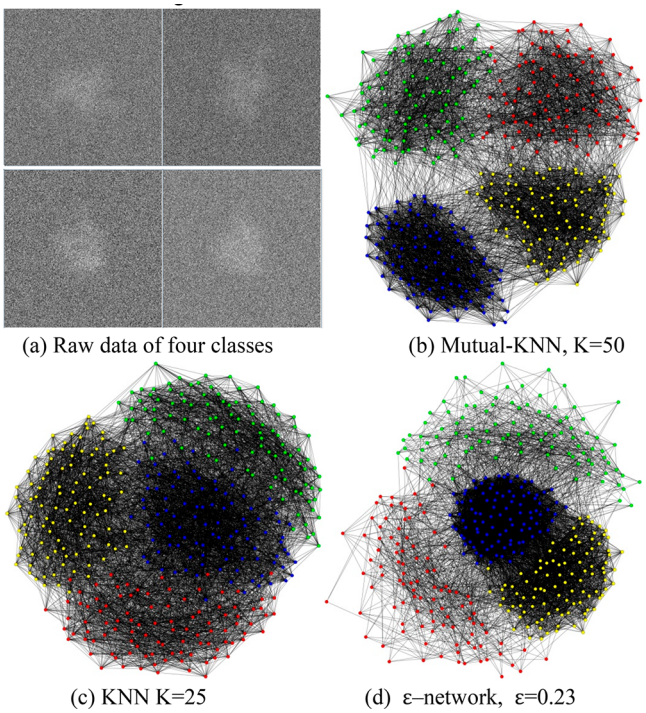


Figure 7. Different intuitive initial networks on the synthetic data set.

ment for constructing the image network, as used in XMIPP's CL2D method. The correntropy between two images (X, Y) is defined as $\mathrm{Cor}(X, Y) = E\{K_\sigma(X - Y)\}$, where $E\{\}$ is the expectation operator and K_σ is a symmetric and non-negative kernel. In practice, the true distribution of $X - Y$ is unknown, resulting in difficulty in calculating the expectation $E\{\}$. In this case, the correntropy can be approximated by its empirical estimate as $\mathrm{Cor}(X, Y) = \frac{1}{M}\sum_{i=1}^{M} K_\sigma(x_i - y_i)$. $K_\sigma = \exp\left[-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right]$, where x_i and y_i are actual pixel values of X and Y, respectively, and M is the total number of pixels in the image. $\sigma = \sqrt{2}\,\sigma_N$, and σ_N is the standard deviation of image noise, the region outside the maximum circle inscribed in images. We reconduct the experiment described in section 3.1 using correntropy instead of correlation, and the final results are not substantially different (see Figure S4). Considering correlation is more computational efficiency and has fewer hyperparameters, we use correlation in the final protocol.

The other important issue for network construction is how to construct the edges in the network. As an empirical result, the ε-neighborhood network tends to connect points within regions of high density and to disconnect points within regions of low density. The K-nearest neighbor network tends to connect points within regions of high density and points in regions of different densities. The mutual k-nearest neighbor network tends to connect points within regions of constant density while separating regions of different densities. Considering that in the cryo-EM images samples are often not taken from evenly spaced angles and SNRs also differ in clusters, leading to varying densities among clusters, we choose to use the mutual KNN in this paper. To further show the advantages of mutual KNN over other kinds of networks, we extend the data set in section 3.1 to four classes as a clearer illustration. Samples from four classes are shown in Figure 7a. We then calculate the correlation matrix and then construct mutual KNN, KNN, and ε-networks, which are shown in panels b–d of Figure 7, respectively. We can see that in this case, among these networks, mutual KNN is sparser and more separable. In the future, we will continue to investigate the effects of the initial network on the final results.

4.2. Clustering Efficiency. Through these experiments, we can see that our new NCEM clustering algorithm is efficient. The clustering module in the XMIPP is essentially a multireference alignment with a divisive K-means method, so it performs alignment $\mathrm{O}(N \times C \times L)$ times and calculates pair-based similarity the same number of times (correlation or correntropy), where N is the number of images in the data set, C is the number of classes, and L is the number of hierarchical levels. In XMIPP, both alignment and calculation of similarity are performed in the pixel space, so it often takes a long time. The clustering algorithm of Prime2D is a K-means-like algorithm, which optimizes the objective function using stochastic hill climbing. Prime2D aligns and calculates similarity in Fourier space. It performs alignment and similarity calculation $\mathrm{O}(N \times C \times \mathrm{iter})$ times, where iter is the number of iterations to reach convergence. The clustering module in Relion also performs alignment and similarity calculation $\mathrm{O}(N \times C \times \mathrm{iter})$ times, so its clustering time cost is similar to that of Prime2D. The clustering of EMAN2 performs MSA to project images into a lower dimension and then performs alignment and calculates similarity, so it runs the fastest of our compared algorithms. In our NCEM clustering algorithm, we use a hierarchal procedure and split the largest cluster into two parts in each hierarchical level. Therefore, it has $\sim\mathrm{O}(N \log N)$ complexity on alignment, and to construct an initial network, we need to compute pair-based similarity (correlation) for $\mathrm{O}(N^2 \log N)$ times. Because alignment takes longer than the calculation of similarity, our algorithm then runs faster than the clustering of XMIPP, Priem2D, and Relion. In the future, we plan to explore a new parallel image network construction method to further accelerate this process.

4.3. Clustering Performance and Trade-off. Our algorithm is essentially a network partitioning problem, while the existing clustering algorithms are often basically modified K-means algorithms. One of the typical merits of our NCEM protocol is it adopts the "set"-based similarity idea, instead of the classical "pair"-based method. It first constructs an initial image network, followed by a network refinement using the structural similarity metric, and then finds the image communities with a hierarchical spectral clustering. Our studies have shown that we can define different network structural similarities, which have shown promising results on both the synthetic and real-world data sets. Among the tested structural similarities, global methods will take more time, especially on the large data sets. A potential price of the network-based NCEM method of the high performance is that a large amount of memory may be required to store the whole network, and the computation of the matrix inverse is involved. The NCEM method has a space complexity of $\mathrm{O}(N^2)$ because we have to keep the whole network in memory during the calculations.

DOI: 10.1021/acs.jcim.8b00853
J. Chem. Inf. Model. 2019, 59, 1658−1667

5. CONCLUSIONS

Clustering of the projected 2D cryo-EM images into groups of similar projection angles is a critical step in generating a high-quality 3D model. This is not an easy task because of the very low SNR of the images. In this paper, to reduce the potential bias caused by the noise effects when measuring the similarity between two images in the clustering, we propose a network-based "set" measurement idea. The essential protocol includes image network construction and refinement, followed by network partitioning in groups of images. Our experimental results on both synthetic and real-world data sets have demonstrated the efficacy of the new NCEM method, which is available at http://www.csbio.sjtu.edu.cn/bioinf/NCEM/.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00853.

Experimental results of different network similarity measurements on different data sets (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: hbshen@sjtu.edu.cn. Telephone: +86-21-34205320. Fax: +86-21-34204022.

ORCID

Hong-Bin Shen: 0000-0002-4029-3325

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2018YFC0910500), the National Natural Science Foundation of China (61725302, 61671288, 91530321, and 61603161), and the Science and Technology Commission of Shanghai Municipality (16JC1404300, 17JC1403500, and 16ZR1448700).

■ REFERENCES

(1) Nogales, E. The development of cryo-EM into a mainstream structural biology technique. *Nat. Methods* **2016**, *13*, 24.

(2) Nogales, E.; Kellogg, E. H. Challenges and Opportunities in the High-Resolution Cryo-EM Visualization of Microtubules and Their Binding Partners. *Curr. Opin. Struct. Biol.* **2017**, *46*, 65−70.

(3) Hurley, J. H.; Nogales, E. Next-Generation Electron Microscopy in Autophagy Research. *Curr. Opin. Struct. Biol.* **2016**, *41*, 211−216.

(4) Kotev, M.; Pascual, R.; Almansa, C.; Guallar, V.; Soliva, R. Pushing the Limits of Computational Structure-Based Drug Design with a Cryo-EM Structure: The Ca2+ Channel $\alpha 2\delta$-1 Subunit as a Test Case. *J. Chem. Inf. Model.* **2018**, *58*, 1707−1715.

(5) Sigworth, F. J. Principles of Cryo-EM Single-Particle Image Processing. *Microscopy* **2016**, *65*, 57−67.

(6) Cheng, Y.; Grigorieff, N.; Penczek, P. A.; Walz, T. A Primer to Single-Particle Cryo-Electron Microscopy. *Cell* **2015**, *161*, 438−449.

(7) Yang, Y.-J.; Wang, S.; Zhang, B.; Shen, H.-B. Resolution Measurement from a Single Reconstructed Cryo-EM Density Map with Multiscale Spectral Analysis. *J. Chem. Inf. Model.* **2018**, *58*, 1303−1311.

(8) Tang, G.; Peng, L.; Baldwin, P. R.; Mann, D. S.; Jiang, W.; Rees, I.; Ludtke, S. J. EMAN2: an Extensible Image Processing Suite for Electron Microscopy. *J. Struct. Biol.* **2007**, *157*, 38−46.

(9) Bell, J. M.; Chen, M.; Baldwin, P. R.; Ludtke, S. J. High Resolution Single Particle Refinement in EMAN2.1. *Methods* **2016**, *100*, 25−34.

(10) Sorzano, C.; Bilbao-Castro, J.; Shkolnisky, Y.; Alcorlo, M.; Melero, R.; Caffarena-Fernández, G.; Li, M.; Xu, G.; Marabini, R.; Carazo, J. A Clustering Approach to Multireference Alignment of Single-Particle Projections in Electron Microscopy. *J. Struct. Biol.* **2010**, *171*, 197−206.

(11) Liu, W.; Pokharel, P. P.; Príncipe, J. C. Correntropy: Properties and Applications in non-Gaussian Signal Processing. *IEEE Trans. Signal Process.* **2007**, *55*, 5286−5298.

(12) Reboul, C. F.; Bonnet, F.; Elmlund, D.; Elmlund, H. A Stochastic Hill Climbing Approach for Simultaneous 2D Alignment and Clustering of Cryogenic Electron Microscopy Images. *Structure* **2016**, *24*, 988−996.

(13) Reboul, C. F.; Eager, M.; Elmlund, D.; Elmlund, H. Single Particle Cryo-EM Improved ab initio 3D Reconstruction with SIMPLE/PRIME. *Protein Sci.* **2018**, *27*, 51.

(14) Scheres, S. H. W. RELION: Implementation of a Bayesian Approach to Cryo-EM Structure Determination. *J. Struct. Biol.* **2012**, *180*, 519−530.

(15) Bai, X. C.; McMullan, G.; Scheres, S. H. How Cryo-EM is Revolutionizing Structural Biology. *Trends Biochem. Sci.* **2015**, *40*, 49−57.

(16) Shaikh, T. R.; Gao, H.; Baxter, W. T.; Asturias, F. J.; Boisset, N.; Leith, A.; Frank, J. SPIDER Image Processing for Single-Particle Reconstruction of Biological Macromolecules from Electron Micrographs. *Nat. Protoc.* **2008**, *3*, 1941−1974.

(17) van Heel, M.; Harauz, G.; Orlova, E. V.; Schmidt, R.; Schatz, M. A new Generation of the IMAGIC Image Processing System. *J. Struct. Biol.* **1996**, *116*, 17−24.

(18) Yang, Z.; Fang, J.; Chittuluru, J.; Asturias, F. J.; Penczek, P. A. Iterative Stable Alignment and Clustering of 2D Transmission Electron Microscope Images. *Structure* **2012**, *20*, 237−247.

(19) Aizenbud, Y.; Shkolnisky, Y. A Max-Cut Approach to Heterogeneity in Cryo-Electron Microscopy. arXiv, 1609.01100, 2016.

(20) Herman, G. T.; Kalinowski, M. Classification of Heterogeneous Electron Microscopic Projections into Homogeneous Subsets. *Ultramicroscopy* **2008**, *108*, 327.

(21) Zhao, Z.; Singer, A. Rotationally Invariant Image Representation for Viewing Direction Classification in Cryo-EM. *J. Struct. Biol.* **2014**, *186*, 153−166.

(22) Bougueroua, S.; Spezia, R.; Pezzotti, S.; Vial, S.; Quessette, F.; Barth, D.; Gaigeot, M.-P. Graph Theory for Automatic Structural Recognition in Molecular Dynamics Simulations. *J. Chem. Phys.* **2018**, *149*, 184102.

(23) Palermo, G.; Ricci, C. G.; Fernando, A.; Basak, R.; Jinek, M.; Rivalta, I.; Batista, V. S.; McCammon, J. A. Protospacer Adjacent Motif-Induced Allostery Activates CRISPR-Cas9. *J. Am. Chem. Soc.* **2017**, *139*, 16028−16031.

(24) Von Luxburg, U. A Tutorial on Spectral Clustering. *Stat. Comput.* **2007**, *17*, 395−416.

(25) Lu, L. Y.; Zhou, T. Link Prediction in Complex Networks: A survey. *Phys. A* **2011**, *390*, 1150−1170.

(26) Sørensen, T. A method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species and its Application to Analyses of the Vegetation on Danish Commons. *Biol. Skr. - K. Dan. Vidensk. Selsk.* **1948**, *5*, 1−34.

(27) Katz, L. A new Status Index Derived from Sociometric Analysis. *Psychometrika* **1953**, *18*, 39−43.

(28) Yu, S. X.; Shi, J. Multiclass Spectral Clustering. In *International Conference on Computer Vision*, 2003; 2003; pp 313−319.

(29) Gray, R. Vector Quantization. *IEEE ASSP Mag.* **1984**, *1*, 4−29.

(30) Lyumkis, D.; Julien, J.-P.; de Val, N.; Cupo, A.; Potter, C. S.; Klasse, P.-J.; Burton, D. R.; Sanders, R. W.; Moore, J. P.; Carragher, B.; Wilson, I. A.; Ward, A. B. Cryo-EM Structure of a Fully Glycosylated Soluble Cleaved HIV-1 Envelope Trimer. *Science* **2013**, *342*, 1484−1490.

(31) Ludtke, S. J.; Chen, D.-H.; Song, J.-L.; Chuang, D. T.; Chiu, W. Seeing GroEL at 6 Å Resolution by Single Particle Electron Cryomicroscopy. *Structure* **2004**, *12*, 1129−1136.
(32) Nithianantham, S.; Le, S.; Seto, E.; Jia, W.; Leary, J.; Corbett, K. D.; Moore, J. K.; Al-Bassam, J. Tubulin Cofactors and Arl2 are Cage-like Chaperones that Regulate the Soluble $\alpha\beta$-Tubulin Pool for Microtubule Dynamics. *eLife* **2015**, *4*, No. e08811.

DOI: 10.1021/acs.jcim.8b00853
J. Chem. Inf. Model. 2019, 59, 1658−1667