Al-Azzawi et al. BMC Bioinformatics (2019) 20:326
https://doi.org/10.1186/s12859-019-2926-y

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

AutoCryoPicker: an unsupervised learning approach for fully automated single particle picking in Cryo-EM images

Adil Al-Azzawi[1], Anes Ouadou[1], John J. Tanner[2] and Jianlin Cheng[1,3*]

Abstract

Background: An important task of macromolecular structure determination by cryo-electron microscopy (cryo-EM) is the identification of single particles in micrographs (particle picking). Due to the necessity of human involvement in the process, current particle picking techniques are time consuming and often result in many false positives and negatives. Adjusting the parameters to eliminate false positives often excludes true particles in certain orientations. The supervised machine learning (e.g. deep learning) methods for particle picking often need a large training dataset, which requires extensive manual annotation. Other reference-dependent methods rely on low-resolution templates for particle detection, matching and picking, and therefore, are not fully automated. These issues motivate us to develop a fully automated, unbiased framework for particle picking.

Results: We design a fully automated, unsupervised approach for single particle picking in cryo-EM micrographs. Our approach consists of three stages: image preprocessing, particle clustering, and particle picking. The image preprocessing is based on multiple techniques including: image averaging, normalization, cryo-EM image contrast enhancement correction (CEC), histogram equalization, restoration, adaptive histogram equalization, guided image filtering, and morphological operations. Image preprocessing significantly improves the quality of original cryo-EM images. Our particle clustering method is based on an intensity distribution model which is much faster and more accurate than traditional K-means and Fuzzy C-Means (FCM) algorithms for single particle clustering. Our particle picking method, based on image cleaning and shape detection with a modified Circular Hough Transform algorithm, effectively detects the shape and the center of each particle and creates a bounding box encapsulating the particles.

Conclusions: AutoCryoPicker can automatically and effectively recognize particle-like objects from noisy cryo-EM micrographs without the need of labeled training data or human intervention making it a useful tool for cryo-EM protein structure determination.

Keywords: Clustering, Intensity based clustering (IBC), Micrograph, Cryo-EM, Single particle picking, Protein structure determination

Background

For decades, X-ray crystallography has been the dominant technique for obtaining high-resolution structures of macromolecules. Single-particle cryo-electron microscopy (cryo-EM) was traditionally used to provide low resolution structural information on large protein complexes that resisted crystallization (e.g., highly symmetric particles of viruses). Though the basic workflow of cryo-EM has not changed considerably over the years, recent technological advances in sample preparation, computation, and especially instrumentation, have revolutionized the field of structural biology [1–3], allowing it to solve large protein structures at better than 3 $\mathring{A}$ resolution [4–7].

Cryo-EM micrographs contains two-dimensional projections of the particles in different orientations. Generally, cryo-EM images have low contrast, due to the similarity of the electron density of the protein to that of the surrounding solution, as well as the limited electron

* Correspondence: chengji@missouri.edu
[1]Electrical Engineering and Computer Science Department, University of Missouri, Columbia, MO 65211, USA
[3]Informatics Institute, University of Missouri, Columbia, MO 65211, USA
Full list of author information is available at the end of the article

dose used in data collection. In addition, the micrographs may contain sections of ice, deformed particles, protein aggregates, etc., which can complicate particle picking. Because a large number of single-particle images must be extracted from cryo-EM micrographs to form a reliable 3D reconstruction of the underlying structure, particle recognition, represents a significant bottleneck in cryo-EM structure determination.

To address the bottleneck, numerous computational approaches have been proposed to facilitate the particle picking process [8–14]. These methods can roughly be divided into two categories: generative methods [15–17] and discriminative classification methods [18–20] (e.g. the recent deep learning methods [21, 22]). The generative methods measure the similarity of an image region to a reference to identify particle candidates from micrographs. A typical generative method employs a template-matching technique with a cross-correlation similarity measure to accomplish particle selection. The discriminative methods first train a classifier on a labeled dataset of positive and negative particle examples, then apply it to detecting particle images from micrographs images.

DeepPicker [21] is a deep learning method for semi-automated particle selection and picking. The first part of the method involved the manual creation of training data. The second part was fully automated by learning patterns from the training data to classify particles. DeepEM [22] uses a convolutional neural network (CNN) to recognize particles. The CNN was trained on a manually curated dataset. The training dataset was augmented by adding additional particles images generated by image rotation.

The existing unsupervised approaches distinguish the particle-like objects from background noise in micrographs via an unsupervised learning manner without the need of any labeled training data [10, 11] but, they do not fully exploit the intrinsic and unique characteristics of particles to facilitate automated particle picking. Therefore, the unsupervised approaches are often combined with the reference template matching or classification-based approaches to achieve good picking results. However, in this case, the training dataset has to be manually created to train the model. Although these approaches have greatly reduced time and effort spent on single-particle data analysis, most of them are not fully automated and still require substantial human intervention to initialize the particle selection process. For instance, most methods require users to prepare an initial set of high-quality reference particles used as templates to search for similar particle candidates from micrographs, while the discriminative approaches usually demand the user to manually pick a number of positive and negative samples to train the classifier first.

In this paper, we develop a fully automated approach for particle picking (AutoCryoPicker) that is based on advanced image preprocessing, robust clustering via the intensity distribution, and sophisticated shape detection. The experimental results demonstrate that the fully automated particle picking scheme can accurately detect a number of particles that is comparable to those picked manually. The clustering method is also more accurate than k-means and Fuzzy C-means (FCM) for particle clustering. Therefore, our new automated picking approach can significantly reduce time and labor spent on single-particle data analysis and thus greatly relieves a bottleneck in the automated cryo-EM structure determination pipeline.

Methods

Our AutoCryoPicker framework for automated particle picking is shown in Fig. 1. In this framework, a user is not required to manually pick any particle from the micrographs. The fully automated approach has three main stages: preprocessing, clustering, and particle picking. In the preprocessing stage, several image processing methods are applied to enhance the input cryo-EM images such as image normalization, Contrast Enhancement Correction (CEC), etc. Clustering is done using three different algorithms k-means [23], Fuzzy C-Means (FCM) [24], and a new robustness clustering algorithm, which is the intensity-Based Clustering (IBC) that addresses some typical clustering issues such as cluster destabilization due to random initialization of cluster centers. In the particle picking stage, a final set of particles is selected from clustered particle candidates.

Stage 1: pre-processing

A standard cryo-EM image is stored in the Mixed Raster Content (MRC) format, which defines a three-dimensional grid (array) of voxels each with a value corresponding to electron density or electric potential. In order to apply various image preprocessing techniques to improve the quality of noisy cryo-EM images, we convert cryo-EM images in the MRC format into widely used 16-bits PNG format using EMAN2 [25]. Since our goal is to use the unsupervised learning algorithm to cluster pixels based on the difference in intensity levels in any cryo-EM image, we select a set of advanced preprocessing tools to improve the quality of cryo-EM images. Those tools are tested on two different datasets.

There are two benefits of using the preprocessing. Firstly, those tools improve the contrast of the cryo-EM images by increasing the particle's intensity. Secondly, pre-grouping the pixels inside each particle makes them easier to be isolated by the clustering algorithm. Specifically, the preprocessing tools are selected based on three main objectives: enhancing the global contrast of the cryo-EM, enhancing the local contrast and increasing the intensity level of each particle, and enhancing

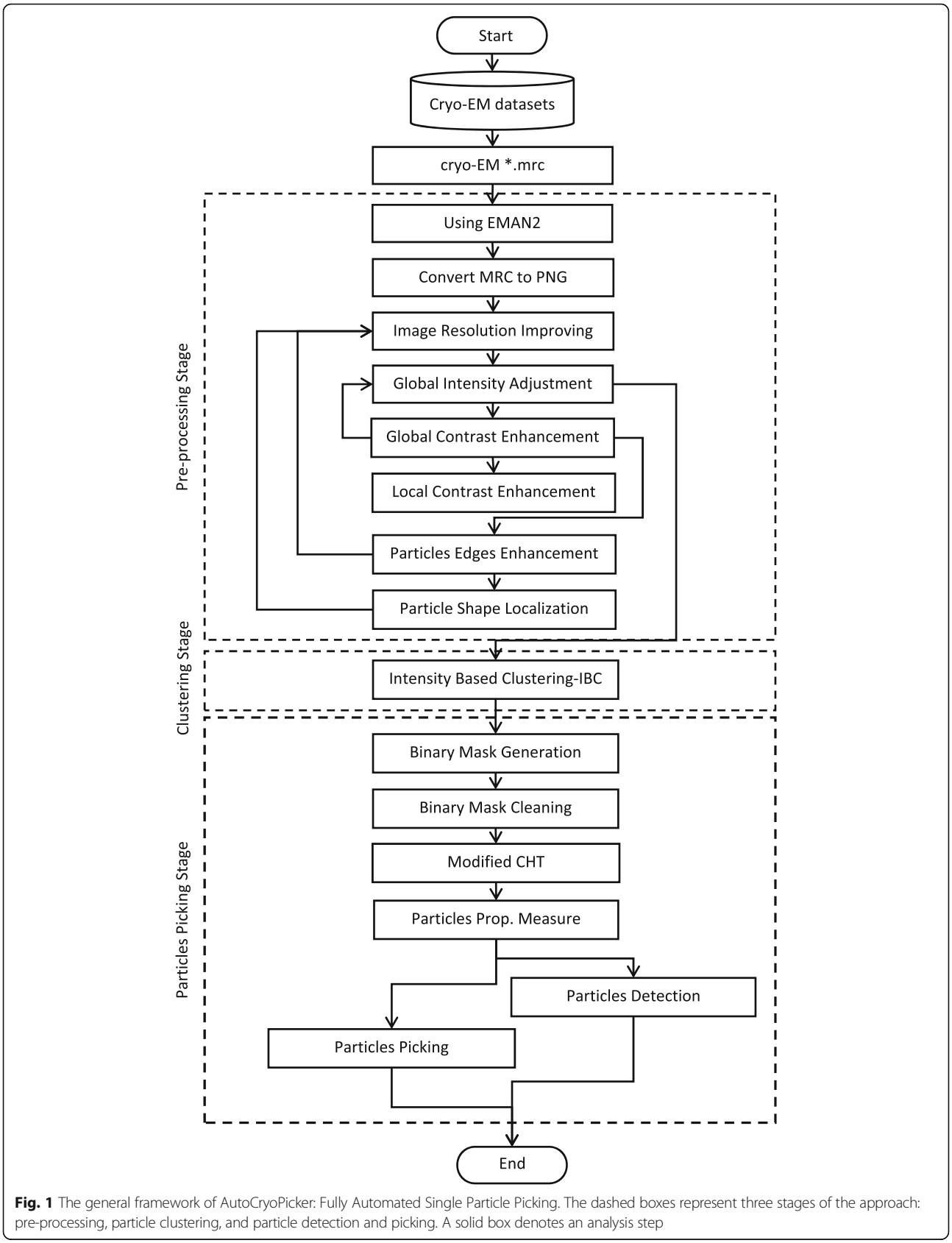

Fig. 1 The general framework of AutoCryoPicker: Fully Automated Single Particle Picking. The dashed boxes represent three stages of the approach: pre-processing, particle clustering, and particle detection and picking. A solid box denotes an analysis step

the particle shapes inside the cryo-EM images. In order to improve the entire contrast between particles and the background, image normalization is used first and then contrast enhancement and correction is applied to increase the global intensity value. To increase the global image contrast, histogram equalization is applied to enhance the pixel intensity level and then image restoration is used to recover and improve the quality of an image. To improve the local contrast and enhancing the definitions of edges in each particle, adaptive histogram equalization is employed. Moreover, guided image filtering is used to perform edge-preserving smoothing of each particle in the cryo-EM image. Finally, morphological image operation is applied to enhance the particle shape and make the particle regions similar to each other and different from the background regions. These preprocessing methods are described in detail in the following steps.

Step 1: Cryo-EM image resolution improving

Cryo-EM images are affected by different factors that either corrupt the micrograph image signal by some gaussian noise or the image resolution. Different cryo-EM images have different artificial objects such as ice, which in some cases, have different thickness and similar ranges of the particle's pixel intensity value. In this case, in a single cryo-EM image, a small number of particles may not have significant difference of scatter power. Technically, the cryo-EM image resolution can be improved using computational image (signal) averaging based on blur motion elimination. This is selected as a main step of the contrast transfer function (CTF) based on the image quality evolution of the single particle cryo-EM and 3D reconstruction tool of viruses [26].

Different cryo-EM images have different intensity value ranges. In order to unify the range values, we renormalize the micrograph by setting the background mean to zero and background variance to one. In this normalization, the pixel values become the Z-score, i.e., the number of sigma's above noise level as shown in Eq. (1) [27]:

$$x' = \frac{x - \bar{x}}{\sigma} \tag{1}$$

Where $\bar{x}$ is the mean of the intensity pixel values, and σ is the standard deviation. For instance, for an image consisting of 50 frames, we used the image averaging and normalization function in EMAN2 [25] to average the 50 frames, resulting in a converted cryo-EM image for further processing and analysis as shown in Fig. 2.

Step 2: global Cryo-EM intensity adjustment

Low-dose micrograph imaging models the exposure to a very low intensity beam in a large defocus area that has both good particle distribution and thin ice. This imaging mode produces very low intensity cryo-EM images. To overcome this problem, intensity adjustment is applied to map the cryo-EM image intensity values to a new range. An Intensity Enhancement Correction (IEC) procedure is used to identify the descent image intensity and improve signal to noise ratio in cryo-EM images. In order to enhance the global intensity adjustment, we apply three different steps.

1) Find Limits to Contrast Stretch: In this step, the range of image intensity is specified by detecting the low and high values via a MATLAB function "*stretchlim*", which returns a two-element vector that consists of the low and upper intensity limits as shown in the cryo-EM histogram in Fig. 3(a). By default, values in low and high intensities specify the bottom 2% and the top 2% of pixel values. In this case, the intensity level of each cryo-EM should be unified. The gray values returned can be used by

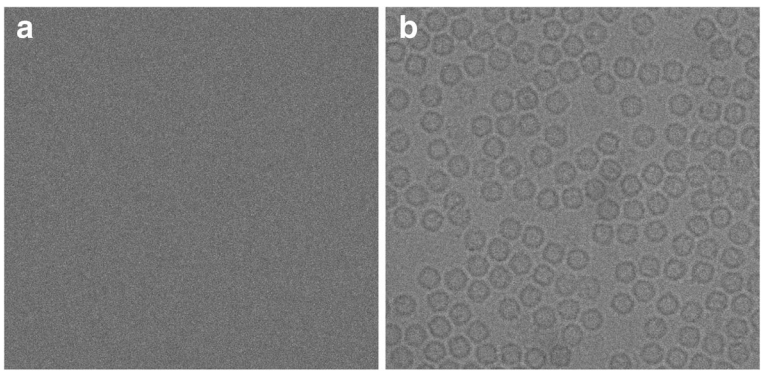

Fig. 2 Cryo-EM image averaging and normalization result using EMAN2. **a** The original cryo-EM image (stack of 50 frame) in the MRC format before the averaging and normalization processing. **b** The cryo-EM image in PNG file format (single frame) after the averaging and normalization processing using EMAN2

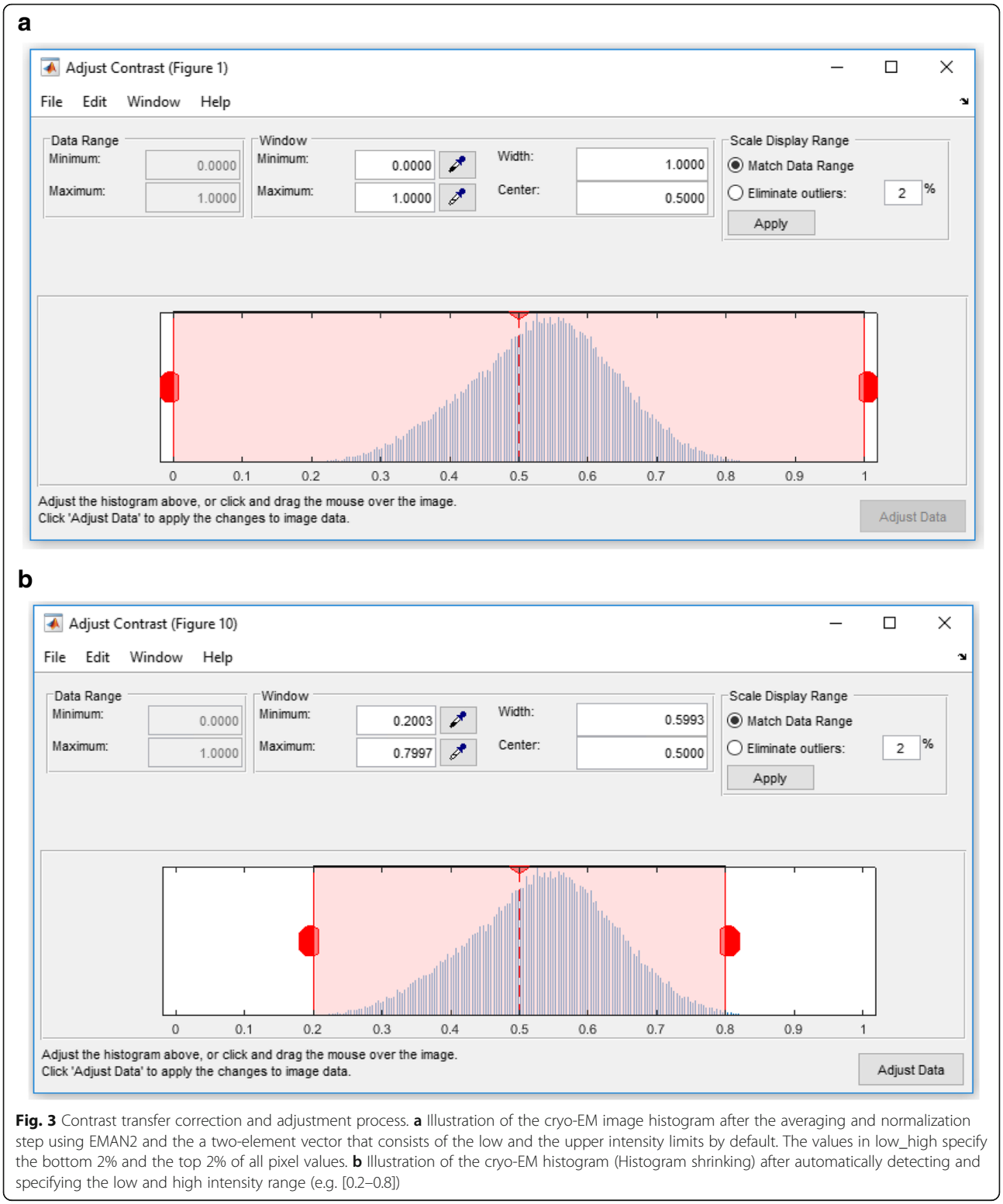

Fig. 3 Contrast transfer correction and adjustment process. **a** Illustration of the cryo-EM image histogram after the averaging and normalization step using EMAN2 and the a two-element vector that consists of the low and the upper intensity limits by default. The values in low_high specify the bottom 2% and the top 2% of all pixel values. **b** Illustration of the cryo-EM histogram (Histogram shrinking) after automatically detecting and specifying the low and high intensity range (e.g. [0.2–0.8])

the "*imadjust*" function [28] to increase the contrast of an image as shown in Fig. 3(b).

2) Mid-Range Stretching: In this step, the cryo-EM image intensity values are stretched to improve their quality. The gray scale image pixels are mapped into the range [0 1] by dividing the intensity values of each pixel as shown in Eq. (2).

$$x_{ij} = \frac{Input\ Image}{High\ Range} \quad (2)$$

where i and j are the row and column index of cryo-EM image matrix respectively and the *High Range* is the highest intensity value in the input image. Figure 4(a) shows an original cryo-EM image, Fig. 4(b) the histogram of the original image, Fig. 4(c) a cryo-EM image after mid-range stretching and Fig. 4(d) the histogram of the stretched image. The histogram in Fig. 4(d) is more stretched than the original one in Fig. 4(b).

1) Intensity Adjustment: The intensity values of the cryo-EM image are adjusted to new values in a condensed smaller range by using the MATLAB function "*imadjust*" [28]. Figure 4(e) shows an

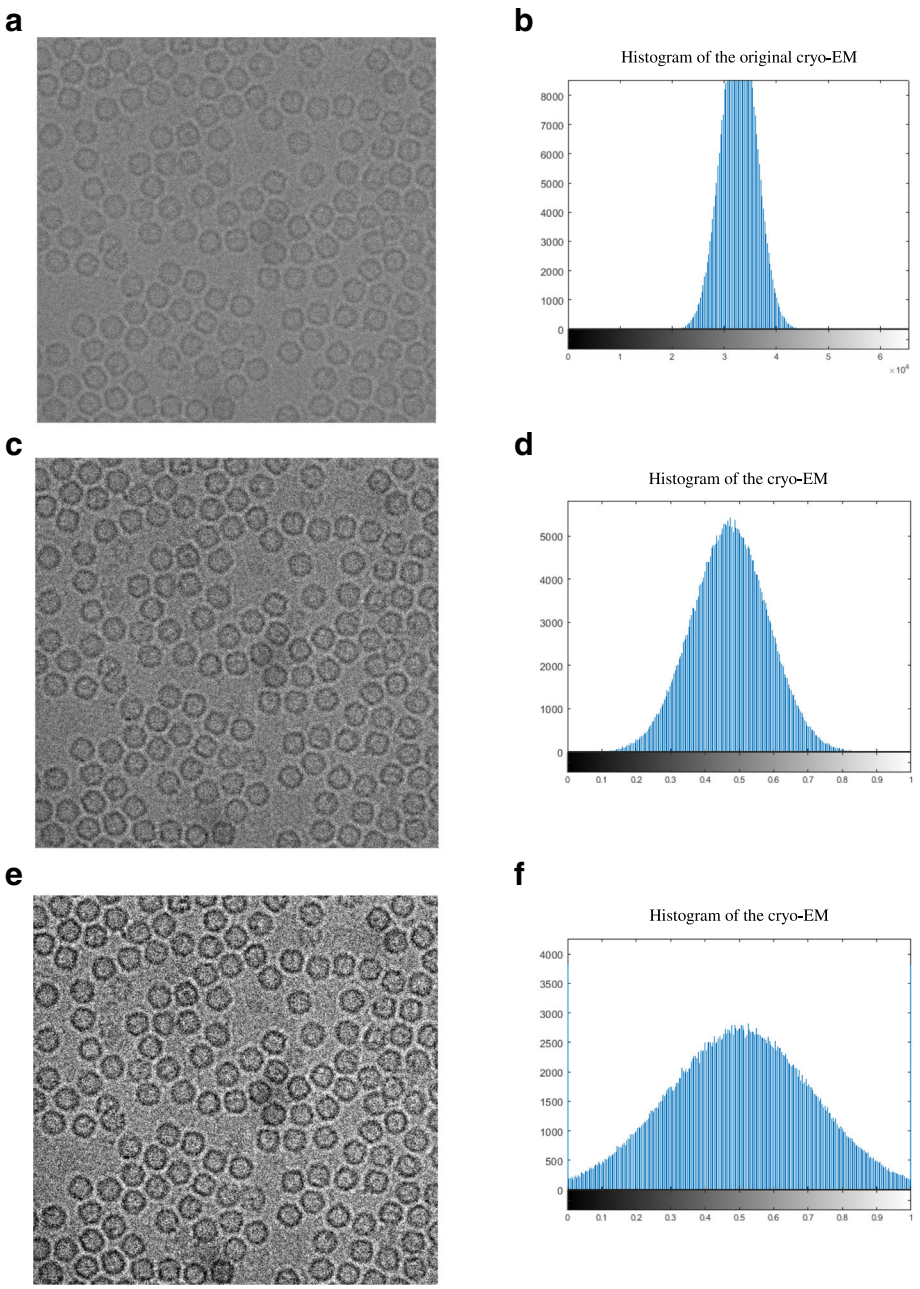

Fig. 4 Cryo-EM Contrast Transfer Correction (CTC) process. **a** The original cryo-EM image after the applying the averaging and normalization process through the EMAN2 software. **b** Histogram of the original cryo-EM image. **c** The cryo-EM image after applying the mid-range stretching based on the low-high intensity range. **d** Histogram of the image in (**c**). **e** The cryo-EM image after applying the contrast enhancement correction (CEC) and image adjustment. **f** The histogram of the cryo-EM image after applying the contrast enhancement correction (CEC)

example of a cryo-EM image with contrast enhancement correction (CEC) and image adjustment, and Fig. 4(f) shows the histogram of Fig. 4(e) where the histogram looks more stretching and the contrast of the cryo-EM is enhanced compared with the original image in Fig. 4(a).

For better demonstrating the effects of the preprocessing steps, we zoom-in one particle image from different datasets. Figure 5(a) and (i) show two original particle images from two different datasets. Figure 5(b) and (j) show the cryo-EM Image resolution being improved by image averaging and normalization. We can notice that image noise has been reduced. Figure 5(c) and (k) illustrates the same single particle images after the global intensity adjustment using Intensity Enhancement Correction (IEC). In comparison with the same particle region in the original micrograph after normalization (Fig. 5(b)), the particles in Fig. 5(c) and (k) has more intensity contrast and are more isolated from the background than the ones in Fig. 5(a) and (b), which will make it easier for clustering algorithms to identify them.

Step3: global Cryo-EM contrast enhancement

Due to the low-dose micrograph imaging mod on a large defocuses particles area, cryo-EM images have low contrast areas where the particles are difficult to detect. Histogram equalization [29] based on a uniform distribution is used to increase and enhance the intensity value of the image pixels. It increases and improves the global image contrast by mapping the original image histogram to a uniform histogram. Figure 5(d) and (l) show an example of a selected particle region in the micrograph after global contrast enhancement-based histogram equalization. Compared with the previous step (e.g. Figure 5(c) and (k)), the particle object regions have more contrast with the background.

Step 4: Cryo-EM noise suppressing

Due to the small electron doses and low contrast between protein and solvent, cryo-EM images tend to be rather noisy [30]. Image restoration is applied to denoise single particle cryo-EM images [31]. Based on the prior knowledge of the degradation process, the image restoration recovers and improves the quality of an image by identifying the type of noise and then removing it. Since the cryo-EM images are often corrupted by typically gaussian noise, the Weiner filter is chosen to model the noise. The Wiener filter is applied to remove additive noise and invert the blurring in cryo-EM images [32]. It minimizes the overall mean square error in the process of inverse filtering and noise smoothing. The Wiener filter in the Fourier domain can be expressed as in Eq. (3).

$$W(f_1,f_2)=\frac{H*(f_1,f_2)S_{xx}(f_1,f_2)}{|H(f_1,f_2)|^2S_{xx}(f_1,f_2)+S_{\eta\eta}(f_1,f_2)} \tag{3}$$

where $S_{xx}(f_1,f_2)+S_{\eta\eta}(f_1,f_2)$ are respectively the power spectra of the original image and the additive noise, and $H(f_1,f_2)$ is the blurring filter. Figure 5(e) and (m) show two different zoom-in particles after applying noise suppressing based image restoration using Wiener filtering. We notice that, in both cases, some background noise is removed, and the structure of the particle object appears more distinctly than the particle object in the previous step (Fig. 5(a)-(d)).

Step 5: local particles contrast enhancement in cryo-EM

In general, the particle picking process depends on the quality of the particles in the cryo-EM. Since there are too many low-quality particle shapes in the cryo-EM images, the local features of the particles such as the contrast, intensity level, and edges, need to be improved and enhanced [26]. Using adaptive histogram equalization (AHE) [32] the particle edges are locally enhanced in the cryo-EM. This is done by improving the local contrast between the particles and background. It provides a sophisticated technique for contrast dynamic range modification (CDRM) based on the intensity histogram shape description. It is applied to small regions of cryo-EM images, called tiles. It enhances the contrast of each tile so that the histogram of the output region approximately matches a specified histogram. The Adaptive Histogram Equalization combines neighboring tiles using bilinear interpolation to eliminate artificially induced boundaries. It is based on a probability model to enhance the contrast condition of each small region (sub-rejoin) using Eq. (4) [32]:

$$p_{rx}(i)=\frac{1}{4}+\left(1-\frac{1}{4}\right)\Phi\left[\left(x-\mu_{ij}\right)\sigma_i^{-1}\right] \tag{4}$$

where $p_{rx}(i)$ is the image contrast-limited adaptive histogram equalization function of pixel value and Φ denotes the cumulative gaussian distribution function for each region, which has a separate location parameter estimate for each region. 1/4 is a constant for the 4-choice task [32]. Figure 5(f) and (n) show two different zoom-in particles after applying local particles contrast enhancement based on contrast-limited adaptive histogram equalization. The particle object intensity (contrast) is significantly improved and enhanced. In both examples, particles look darker and have a higher contrast than the previous particle images (Fig. 5(e) and (m)).

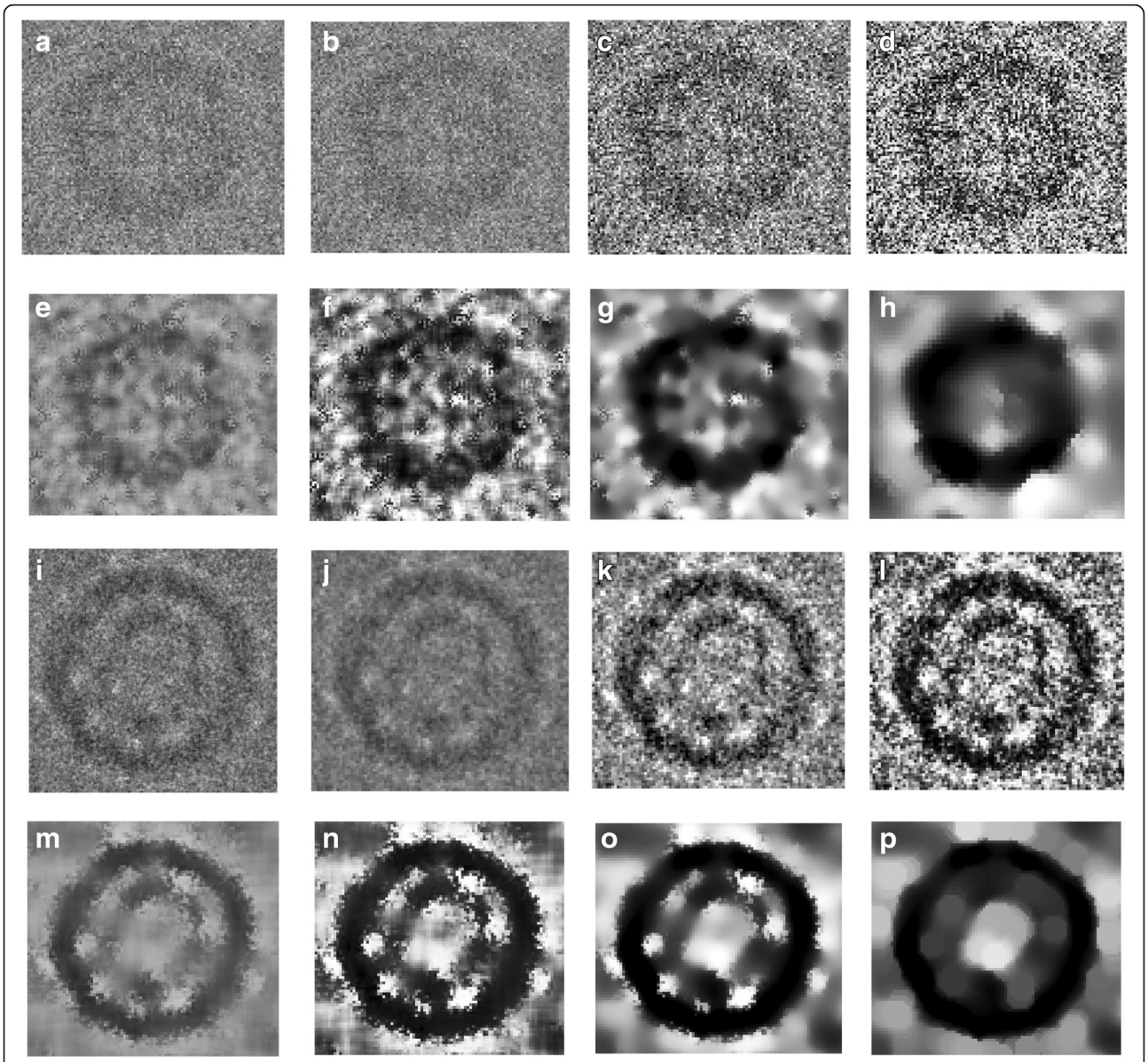

Fig. 5 Illustration of effects of the cryo-EM image analysis on a zoom-in selected particle region using two different examples from two datasets. **a** An original zoom-in selected particle region in the micrograph image in Apoferritin dataset. **b** The normalized single particle image region. **c** The single particle region after applying the contrast enhancement correction (CEC). **d** The single particle region after applying the histogram equalization. **e** The single particle region after applying image resonation with Wiener filtering. **f** The single particle region after applying the contrast-limited adaptive histogram equalization. **g** The single particle region after applying image guided filtering. **h** The single particle region after applying morphological image operation. **i** An original zoom-in selected particle region in a micrograph image in the KLH dataset before the preprocessing steps. **j** The selected particle region in a micrograph image in the KLH dataset after normalization. **k** The selected particle region in a micrograph image in the KLH dataset after applying the contrast enhancement correction (CEC). **l** The selected particle region in a micrograph image in the KLH dataset after applying the histogram equalization. **m** The selected particle region in a micrograph image in the KLH dataset after applying image resonation with Wiener filtering. **n** The selected particle region in a micrograph image in the KLH dataset applying the contrast-limited adaptive histogram equalization. **o** The selected particle region in a micrograph image in the KLH dataset after applying image guided filtering. **p** The selected particle region in a micrograph image in the KLH dataset after applying morphological image operation

Step 6: particle edges enhancement in cryo-EM

In order to localize each particle object in the cryo-EM image, particle edges enhancement is proposed to isolate the particle shapes in the cryo-EM image. Edge-preserving smoothing technique is used to locally smooth and enhance the particle edges in order to localize different particles in any cryo-EM. Guided image filtering [33] is employed to perform edge-preserving and smoothing using the content of a second image, called a guidance image, to influence the filtering. The guided filter generates the filtered output by

considering the content of a guidance image, which can be the input image itself or a different image. It has a theoretical connection with the matting Laplacian matrix [33] and can better utilize the structures in the guidance image. Let us assume that I is a guidance image filter, p is an input cryo-EM image, and q is an output image. Both I and p are given beforehand and can be identical. The filtered output at a pixel i is expressed as a weighted average as shown in Eq. (5) [33]:

$$W_{ij} = \frac{1}{|w|^2} \sum_{k:i \in w_k} \left(1 + \frac{(I_i + \mu_k)(I_j + \mu_k)}{\sigma_k^2 + \epsilon}\right) \tag{5}$$

where i and j are pixel indices. The filter kernel W_{ij} is a

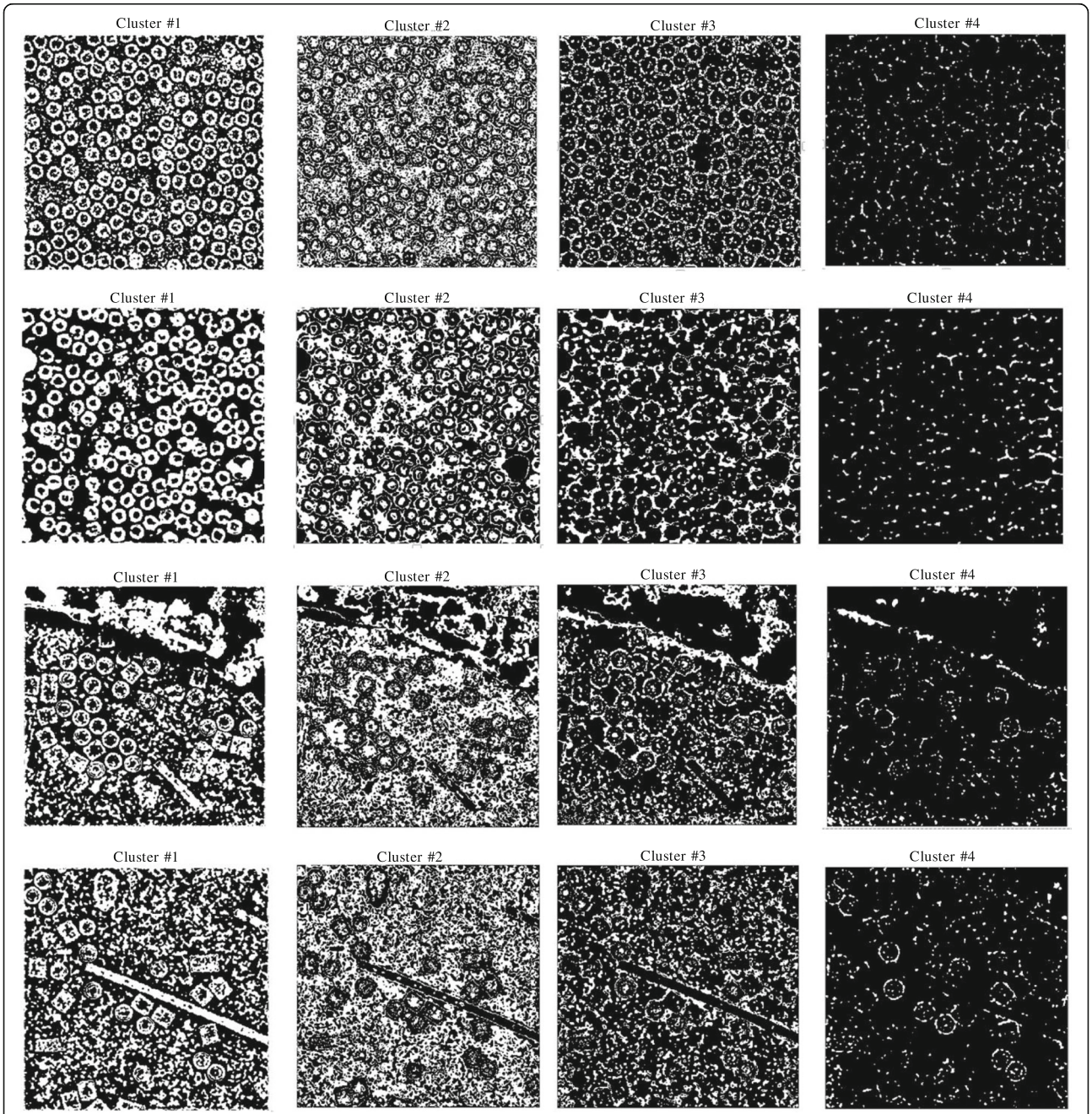

Fig. 6 Different cryo-EM image clustering results using an Intensity-Based Clustering Algorithm (ICB). **a** Two sets of cryo-EM image clustering results (Cluster #1, Cluster #2, Cluster #3 and Cluster #4) on the Apoferritin dataset. Most real particles were always assigned to Cluster 1. **b** Two sets of cryo-EM image clustering results (Cluster #1, Cluster #2, Cluster #3 and Cluster #4) on the KLH dataset. Most real particles were always assigned to Cluster 1

function of the guidance cryo-EM image I and independent of p as in Eq. (6) [33]:

$$q_i = W_{ij}(I)p_j \tag{6}$$

where q_i is the output image after the image guidance filtering and p_j is the input image after the image guidance filtering. A MATLAB function "*imguidedfilter*" is used to implement the guided filtering. It performs the edge-preserving smoothing of the cryo-EM image in order to reduce the noise while keeping the particle edges. Figure 5(g) and (o) show two different zoom-in

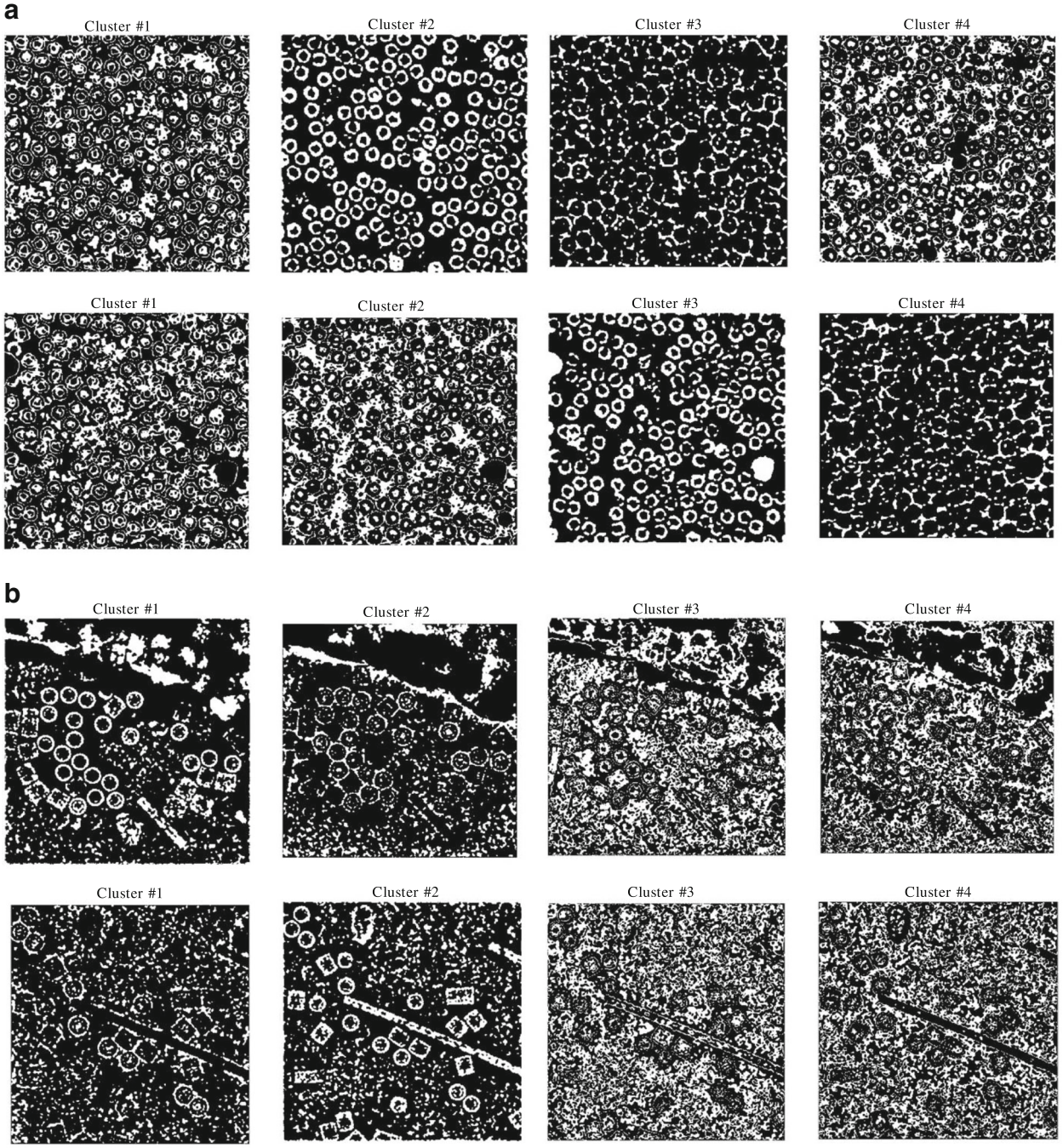

Fig. 7 Different cryo-EM image clustering results using the k-means clustering algorithm. **a** The two sets of cryo-EM images clusters results (Cluster #1, Cluster #2, Cluster #3 and Cluster #4) on the Apoferritin dataset. Most real particles were assigned to Cluster 2 and Cluster 3, respectively. **b** The two sets of cryo-EM image clustering results (Cluster #1, Cluster #2, Cluster #3 and Cluster #4) on the KLH dataset. Most real particles were assigned Cluster 1 and Cluster 2, respectively

particles after applying particle edges enhancement using image guided filtering. The overall contrast of the particle in the cryo-EM image is improved. Compared to the same particle in the previous step (Fig. 5(f) and (n)), particle edges appear more smoothly and some dark spots around the particle object become smoother and brighter while particle object edges become darker. In addition, the particle edges are more connected and have higher contrast than the background.

Step 7: particle shape localization in cryo-EM

The last step of the pre-processing stage is the particle object localization and isolation step. In this step, we use morphological image processing [29], which is a

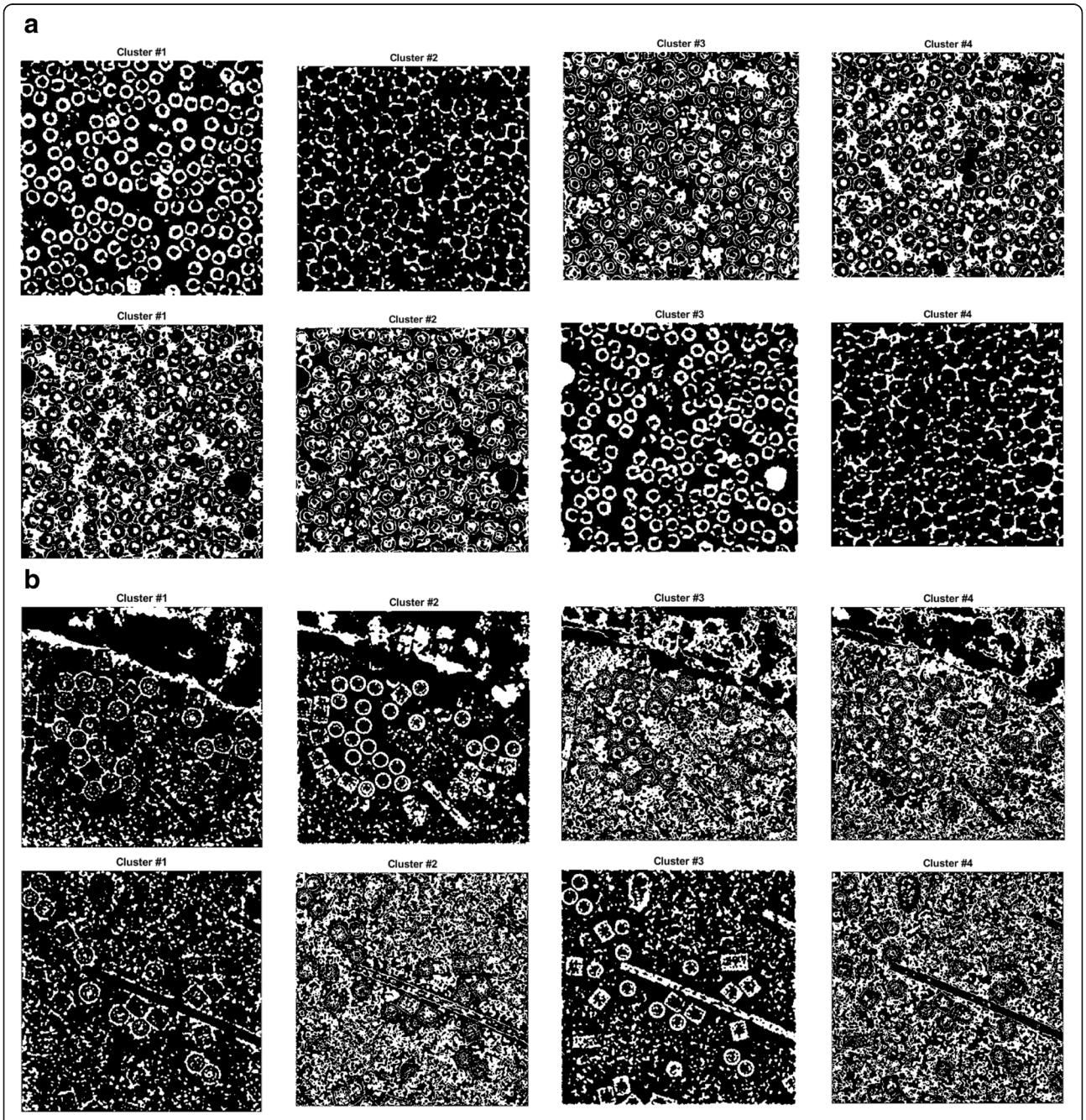

Fig. 8 Different cryo-EM image clustering results using the FCM clustering algorithm. **a** Two sets of cryo-EM images clustering results (Cluster #1, Cluster #2, Cluster #3 and Cluster #4) on Apoferritin dataset. Most real particles were assigned to Cluster 1 and Cluster 3, respectively. **b** Two sets of cryo-EM image clustering results (Cluster #1, Cluster #2, Cluster #3 and Cluster #4) on the KLH dataset. Most real particles were assigned to Cluster 2 and Cluster 3, respectively

collection of non-linear operations related to the shape or morphology of features in an image. Logical operations are applied to make particle regions similar to each other and different from the background regions. We apply an opening dilation operation followed by erosion with the same structuring element as shown in Eq. (7) [29]:

$$A \bullet B = (A \oplus B) \ominus B \tag{7}$$

where A is the original cryo-EM image and B is the structure element. Figure 5 (h) and (p) show two different zoom-in particles after applying shape localization using morphological image operation (image closing with a structural element 5×5). The particle object is

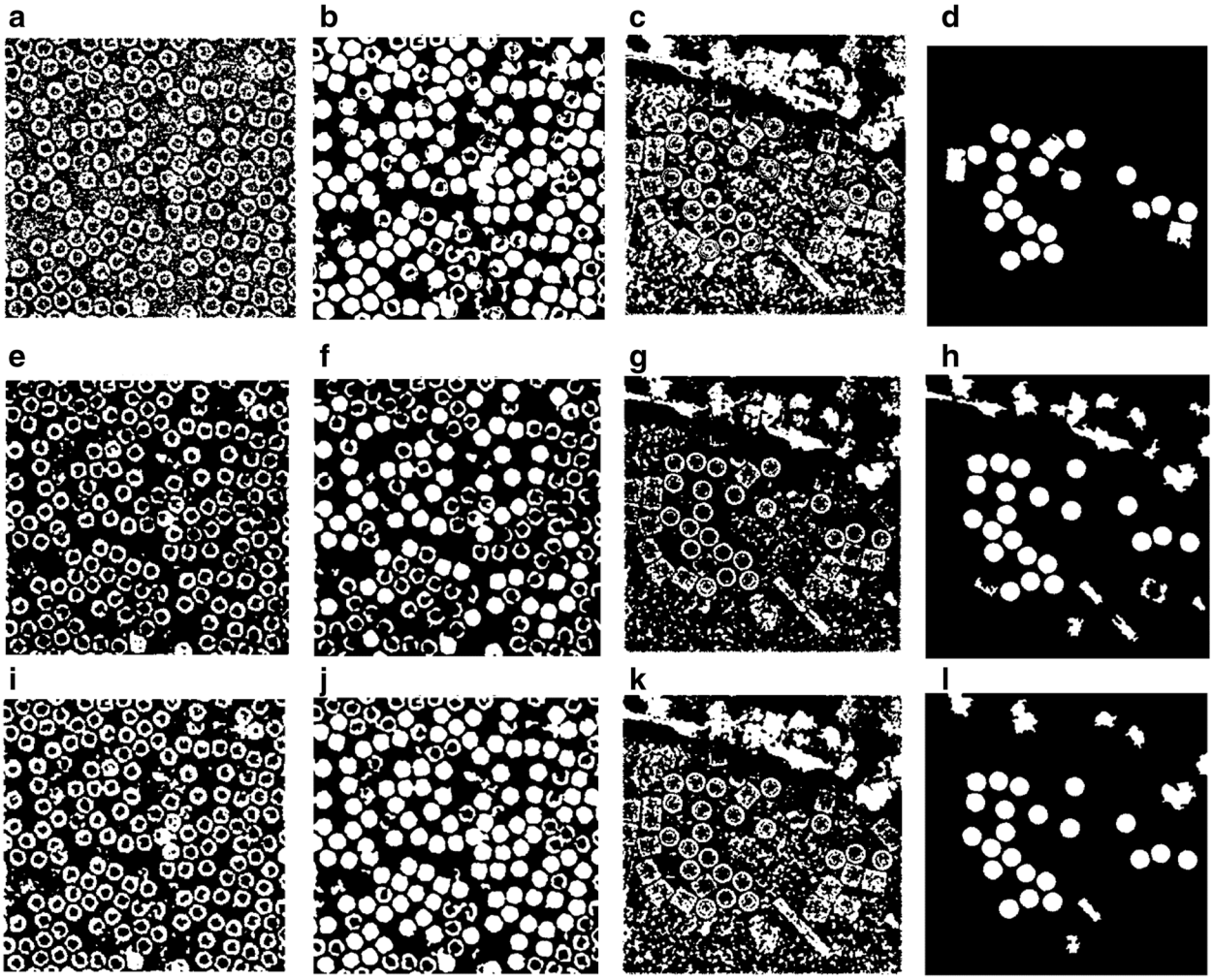

Fig. 9 Cryo-EM Particle Clustering Results after Binary Image Cleaning and Non-Circular Object Removal. **a** The particle clustering image before binary image cleaning and non-circular object removal on the results of ICB clustering of a cryo-EM image from Apoferritin dataset. **b** The particle clustering image after binary image cleaning and non-circular object removal on the results of ICB clustering of a cryo-EM image from Apoferritin dataset. **c** The particle clustering image before binary image cleaning and non-circular object removal on the results of ICB clustering of a cryo-EM image from KLH dataset. **d** The particle clustering image after binary image cleaning and non-circular object removal on the results of ICB clustering of a cryo-EM image from KLH dataset. **e** The particle clustering image before binary image cleaning and non-circular object removal on the results of k-means clustering of a cryo-EM image from Apoferritin dataset. **f** The particle clustering image after binary image cleaning and non-circular object removal on the results of k-means clustering of a cryo-EM image from Apoferritin dataset. **g** The particle clustering image before binary image cleaning and non-circular object removal on the results of k-means clustering of a cryo-EM image from KLH dataset. **h** The particle clustering image after binary image cleaning and non-circular object removal on the results of k-means clustering of a cryo-EM image from KLH dataset. **i** The particles clustering image before binary image cleaning and non-circular object removal on the results of FCM clustering of a cryo-EM image from Apoferritin dataset. **j** The particle clustering image after binary image cleaning and non-circular object removal on the results of FCM clustering of a cryo-EM image from Apoferritin dataset. (k) The particle clustering image before binary image cleaning and non-circular object removal on the results of FCM clustering of a cryo-EM image from KLH dataset. **l** The particle clustering image after binary image cleaning and non-circular object removal on the results of FCM clustering of a cryo-EM image from KLH dataset

significantly improved and more isolated from the background. Also, the particle object structure is fully connected and has a higher contrast. The particle background is smother, compared to the particle background in the previous step Fig. 5(g) and (o).

Stage 2: particle clustering

In this stage, a binary mask is constructed using unsupervised learning clustering methods for particle isolation. Two standard clustering algorithms K-means [23] and FCM) [24] as well as a new intensity-based clustering (IBC) algorithm are applied. This clustering algorithm is based on an intensity distribution model, $P(i; d)$, which relates the intensity difference value d to the signed difference intensity values, i. The detailed description of the Intensity Based Clustering (IBC) algorithm can be found in the Additional file 1: Algorithm 1.

Figure 6(a) and (b) show an example of different cryo-EM clustering results by using the intensity-based clustering method (ICB) with two cryo-EM datasets (Apoferritin [34] and KLH datasets [35]). It is noticed that the particles are most stably grouped in Cluster 1. Generally, the particles of the different images of the same protein can be best identified in the same specific cluster by the ICB method according to our experiments. However, the particles are not most stably grouped in the same cluster by k-means and FCM algorithms due to their random initialization of cluster centers. Figures 7 and 8 show the clustering results of the same cryo-EM images using k-means and FCM respectively. Note that the particles are located in different clusters. For instance, the particles clustering for two cryo-EM images in the first dataset (Apoferritin) using k-means is shown in Fig. 7(a). The particles are grouped in two different clusters (Cluster 2 and 3, respectively). Figure 7(b) shows the same issue for the k-means on the second dataset (KLH). The same problem happens to FCM (Fig. 8).

Stage 3: particle picking

The last stage of the AutoCryoPicker framework has two main steps. The first step is binary mask image cleaning and the second step is particle object detection and picking. In the first step, some post-processing operations (e.g. binary image region and hole filling, morphological image operation using image opening, and small object removal from the binary image) are performed to clean the binary mask produced in the clustering stage. In the second step, a modified Circular Hough Transform algorithm (CHT) [36] is applied to detect particles in the cleaned binary mask.

Step 1: Cryo-EM cluster image cleaning and non-circular object removal

A binary mask of each cryo-EM cluster image is cleaned based on removal of the small and non-circular

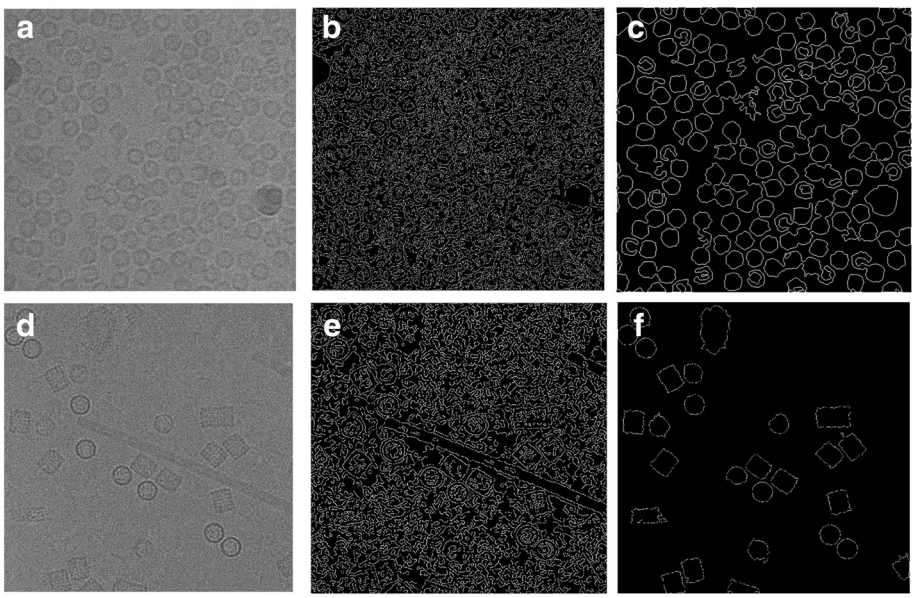

Fig. 10 Modified Circular Hough Transformation (CHT). **a** Original cryo-EM image from the KLH dataset. **b** Edge detection result that will be used later for CHT to detect the center of each circular object in the binary cryo-EM image from the Apoferritin dataset based on using canny edge detection. **c** Edge detection results that will be used later for CHT to detect the center of each circular object in the binary cryo-EM image from the Apoferritin dataset based on using the modified CHT based IBC clustering and boundary pixels list extraction (outline's boundary pixel). **d** Edge detection result that will be used later for CHT to detect the center of each circular object in the binary cryo-EM image from the KLH dataset based on using canny edge detection. **e** Edge detection results that will be used later for CHT to detect the center of each circular object in the binary cryo-EM image from the KLH dataset based on using the modified CHT based IBC clustering and boundary pixels list extraction (outline's boundary pixel)

objects via size filtering and roundness filtering. The detailed description of the image cleaning and non-circular object removal algorithm can be found in the Additional file 1: Algorithm 2. Figure 9 shows the cryo-EM image cluster cleaning results (particles clustering) before and after image cleaning step. Figure 9(b), (f),

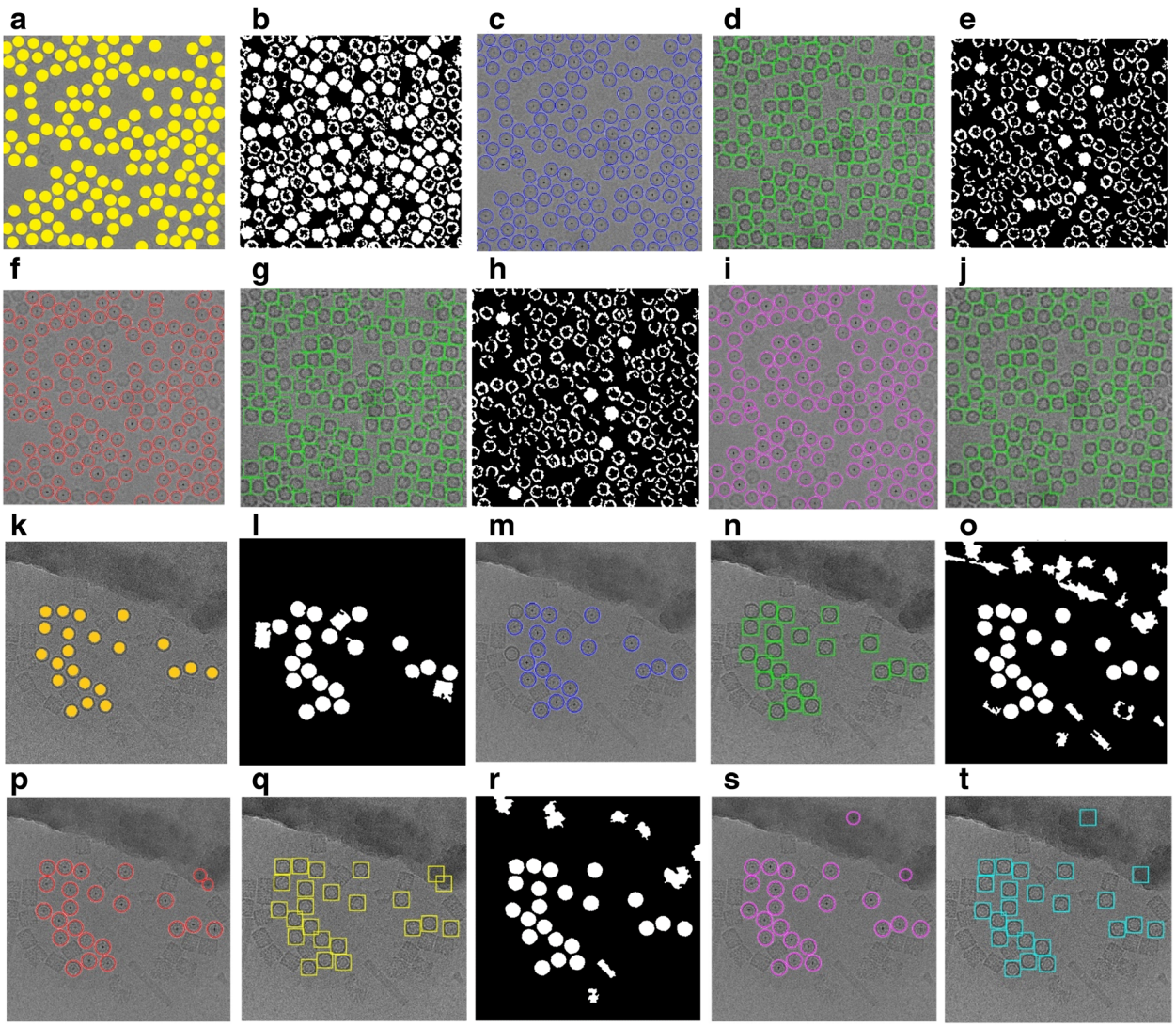

Fig. 11 Top View (Circular) Particles Detection and Picking Results using Modified Circular Hough Transform (CHT). **a** The Ground truth (particles manually labelled) for the cryo-EM image from the Apoferritin dataset. **b** ICB clustering results after the binary image cleaning and non-circular objects removal (Apoferritin dataset). **c** The center of each particle illustrated by the '+' sign and the radius of each particle by the blue circle around each particle (ICB and Apoferritin dataset). **d** The bounding box for each particle object in the original cryo-EM image (ICB and Apoferritin dataset). **e** K-means clustering results after the binary image cleaning and non-circular objects removal (Apoferritin dataset). **f** The center of each particle illustrated by using the '+' sign and the radius of each particle by the blue circle around each particle (k-means results on Apoferritin dataset). **g** The bounding box for each particle (k-means results and Apoferritin dataset). **h** FCM clustering results after the binary image cleaning and non-circular objects removal (Apoferritin dataset). **i** The center of each particle illustrated by the '+' sign and the radius of each particle by the blue circle around each particle (FCM and Apoferritin dataset). **j** The bounding box for each particle in the original cryo-EM image (FCM results and Apoferritin dataset). **k** The ground truth (particles manually labeled) for the cryo-EM image from the KLH dataset. **l** ICB clustering results after the binary image cleaning and non-circular objects removal (KLH dataset). **m** The center of each particle illustrated by the '+' sign and the radius of each particle by the blue circle (ICB and KLH dataset). **n** The bounding box for each particle in the original cryo-EM image (ICB and KLH dataset). **o** K-means clustering results after the binary image cleaning and non-circular objects removal (KLH dataset). **p** Shows the center of each particle illustrated by the '+' sign and the radius of each particle by the blue circle (k-means and KLH dataset). **q** The bounding box for each particle in the original cryo-EM image (k-means and KLH dataset). **r** FCM clustering results after the binary image cleaning and non-circular objects removal (KLH dataset). **s** The center of each particle illustrated by the '+' sign and the radius of each particle by the blue circle (FCM and KLH dataset). **t** The bounding box for each particle in the original cryo-EM image (FCM and KLH dataset)

and (j) show the particles clustering and cleaning results for the cryo-EM images form the Apoferritin dataset using ICB, k-means, and FCM respectively. Figure 9(d), (h), and (I) show the particles clustering and cleaning results for the cryo-EM images form the KLH dataset using ICB, k-means, and FCM respectively. It noticed that the proposed algorithm (ICB) produces significantly cleaner clustering images than the other two standard clustering algorithms.

Step 2: top view (circular) particle detection and picking in Cryo-EM

Since the regular shape of the protein particle in the test cryo-EM dataset is a common shape – circle (top view), a Circular Hough Transform (CHT) [29] is used to detect particles in cluster images. For another common particle shape in cryo-EM images - square, a square shape detector would be needed. The detailed description of the modified Circular Hough Transform (CHT) algorithm can be found in the Additional file 1: Algorithm 3.

The results of replacing the canny edge detection by our IBC algorithm are shown in Fig. 10(c) and (f) using the same images that are used in original CHT (using canny edge detection) in Fig. 10(b) and (e). The detection algorithm returns the center and radius of each particle as is shown in Fig. 11(c), (f), (i), (m), (p), and (s) based on the clustering results of the different clustering algorithms (ICB, k-means, and FCM) respectively for Apoferritin and KLH datasets. For instance, Fig. 11(c) shows the center and radius of each particle illustrated by a '+' sign and a blue circle. A bounding box is drawn around each particle object in the cryo-EM image (Fig. 11(d)). Figure 11(c) and (d) show the results of the particle object detection and picking based on the ICB clustering and the Circular Hough Transform (CHT) on the first dataset (Apoferritin). Figure 10(m) and (n) show the same results on the second dataset (KLH dataset). Figure 11(f) and (g) show the results of the particle object detection and picking based on k-means clustering and the Circular Hough Transform (CHT) on the first dataset. Figure 11(p) and (q) show the same results on the second dataset. Finally, Fig. 11(h) and (j) show the results of the particle object detection and picking based on the FCM clustering and the Circular Hough Transform (CHT) on the first dataset. Figure 11(s) and (t) show the same results on the second dataset.

Step 3: side view (square) particle detection and picking in Cryo-EM

Another common particle shape in the cryo-EM images is a square (side view). In this case, we add another step called circular and non-square object removal form the

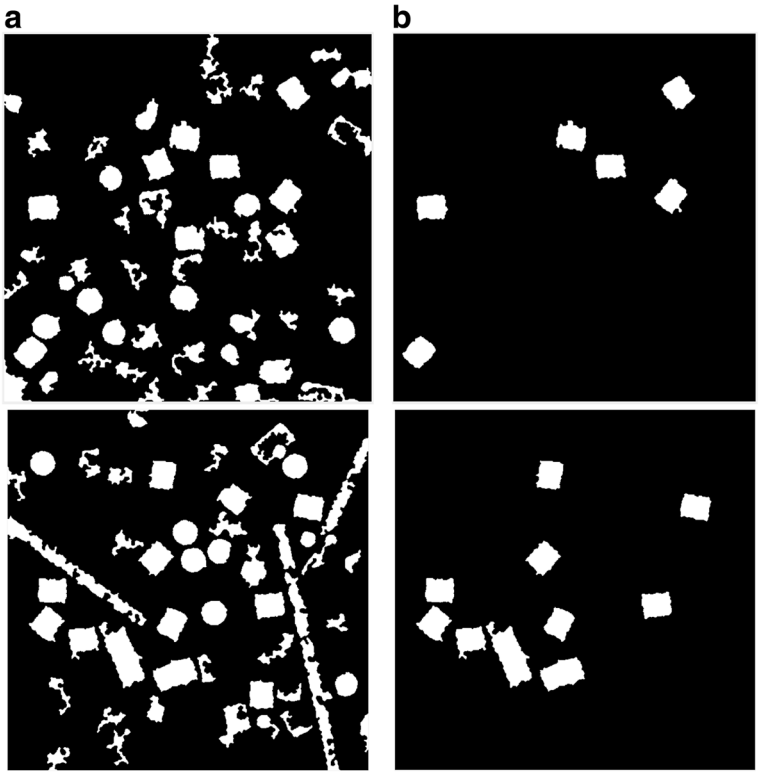

Fig. 12 Cryo-EM clean clustered images after the circular and non-square object removal. **a** The cryo-EM clustered images after image cleaning and small objects removal. **b** The same cryo-EM clustered images after the circular and non-square object removal

ICB clustering image after the cleaning and small object removal step in case of keeping the side view particle shapes (square). The detailed description of the circular and non-square object (particles) removal algorithm can be found in the Additional file 1: Algorithm 4.

Figure 12 shows an example of the cryo-EM clustered images after the circular and non-square object removal. For instance, Fig. 12(a) shows cryo-EM clustered images after image cleaning and small objects removal although, Fig. 12(b) shows the same cryo-EM clustered images after the circular and non-square object removal. After this step, the cleaned image has only the square particle shapes (side view) in the whole cryo-EM images. We can notice that not all the particles (side view) are cleaned after the second post-processing step, but some of them are according to the similarity between the *Max_Allowable_Area* value and the circularities of each square particle object. If the circularity values between each particle shapes (side view-square and top view-circle) are very close, they are eliminated from the cleaned image.

After the circular objects and artifacts have been being removed, the cryo-EM cleaned mask becomes significantly clear for detecting and selecting each square particle. The cleaned binary image has almost only the square objects (particles side view), in this case, we apply the square (side-view) particle detection and picking. The detailed description of the square (side-view) particle detection and picking algorithm can be found in the Additional file 1: Algorithm 5.

The results of the side-view particle shapes detection (square particles) are shown in Fig. 13(c) using different cryo-EM image samples from the KLH dataset.

Step 4: perfect side view (square) particle detection and picking in Cryo-EM

Side-view particle detection (square) and picking is not very accurate. We can notice that some additional objects are attached to the original square particles in addition to some overlapped particles, which are also selected as shown in the final detected results in Fig. 13(b). To

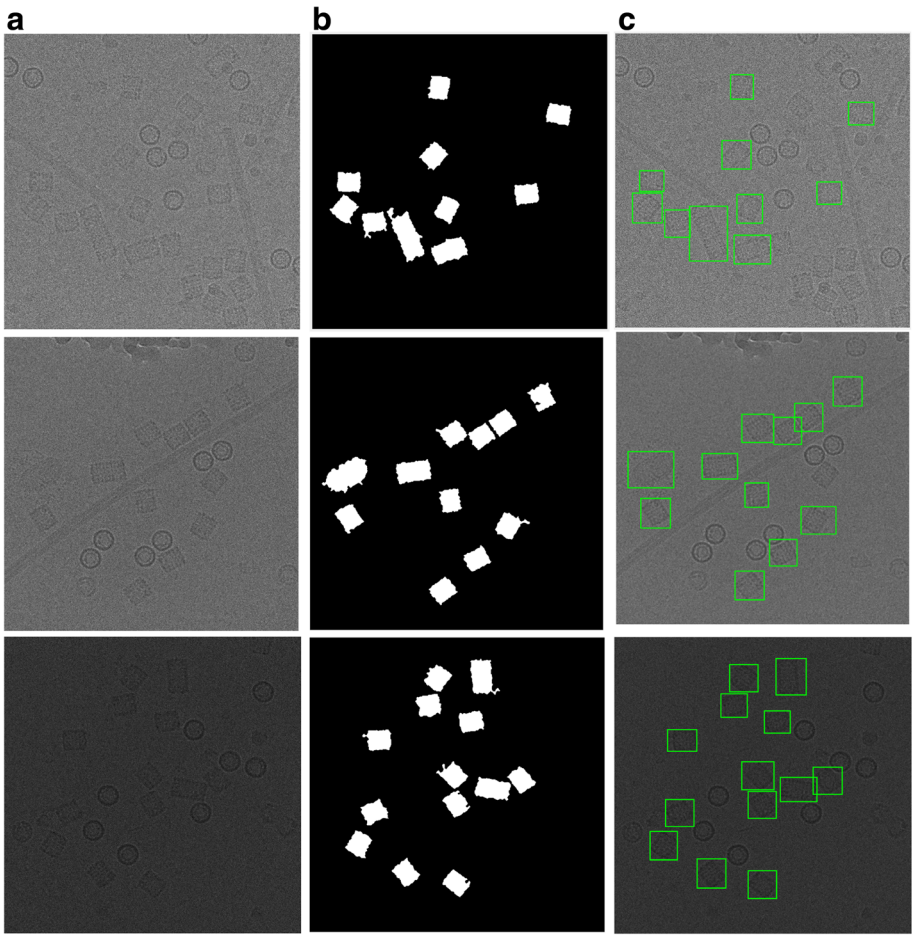

Fig. 13 Side view (square) particles detection and picking results. **a** The original cryo-EM image (KLH dataset). **b** The result after circular and non-square object removal based on the ICB clustering algorithm. **c** Side view (square) particle detection results

overcome this problem, we design another post processing algorithm called perfect square particles shape detection and picking. The detailed description of the perfect square particles shape detection and picking algorithm can be found in the Additional file 1: Algorithm 6.

Figure 14 shows an example of the perfect square particle shapes detection using Feret object diameter. Figure 14(a) shows the square particle shapes in the image after the shapes are smoothed and blurred. Figure 14(b) shows the new boundary box of each particle based on the Feret diameter measures. Figure 14(c) shows the perfect square particle shapes based on the Feret object diameter measurement. Figure 14(d) shows the square particles image after eliminating the outlier objects (overlapped particles). Figure 14(e) shows the square particle detection results (side view) based on the new Feret boundary box. Finally, Fig. 14(f) shows the final results of different particle shape detection and picking (top and side view) based ICB clustering, modified CHT, and perfect square (side view) particle shapes detection using Feret object diameter.

It is noticed that there is almost no true positive (top view particles-circle) missing. In contrast, there are some

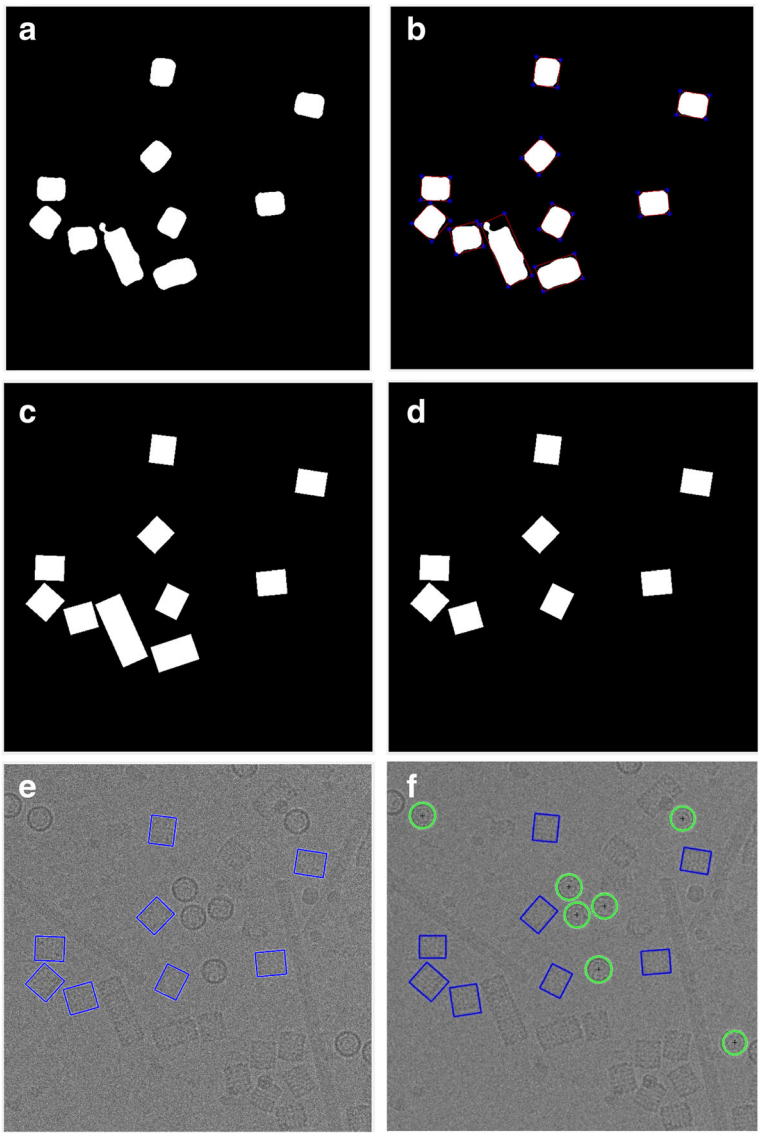

Fig. 14 Perfect square (side view) particle shape detection using the Feret object diameter using (KLH dataset). **a** Square particle image after shapes smoothing and blurring. **b** Boundary boxes (each particle) based on Feret object diameter measurement. **c** Perfect square particle shapes that are generated based on the new boundary box dimension using Feret object diameter measurement. **d** Square particle image after the outlier objects are eliminated. **e** Square particle detection results (side view) based on the new Feret boundary box dimension. **f** The final results of two different particle shape detection and picking (top and side view) based on ICB clustering and modified CHT; and perfect square (side view) particle shapes detection using Feret object diameter

true positive example of square particles (side view) missing. Figure 15 shows some extra cases of the particle detection and picking results for both cases (top and side view) using three different algorithms (ICB, k-means, and FCM). Figure 15(a) shows the original cryo-EM image, while Fig. 15(b), (c), and (d) shows the target detection and picking image using ICB, k-means, and FCM respectively. Those examples have been manually labeled in the case of showing the detection and picking performance. The red dots illustrate hand labeling of the circular particles (top view) while the green squares illustrate hand labeling the squares particles (side view), although, the blue circles showing the particle AutoCryoPicker detection and picking results for the top view particles, and the yellow squares showing the side view particles detection and picking results.

Figure 15 illustrates some cases in which AutoCryoPicker failed to detect and pick in both top and side

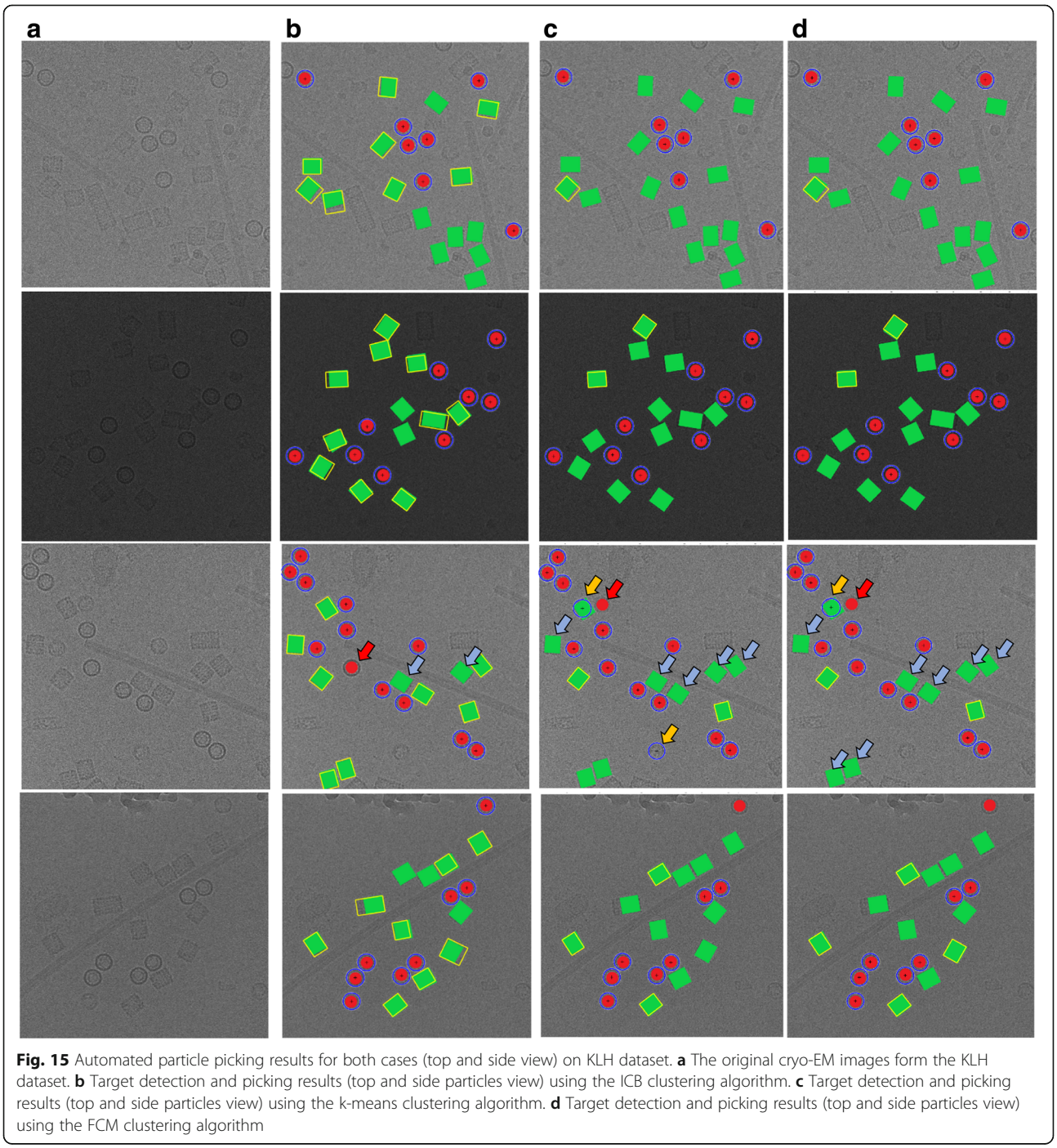

Fig. 15 Automated particle picking results for both cases (top and side view) on KLH dataset. **a** The original cryo-EM images form the KLH dataset. **b** Target detection and picking results (top and side particles view) using the ICB clustering algorithm. **c** Target detection and picking results (top and side particles view) using the k-means clustering algorithm. **d** Target detection and picking results (top and side particles view) using the FCM clustering algorithm

views on the KLH dataset. In the third test example (Fig. 15(b), (c), and (d)), there is one top view circular particle not detected by ICB, k-means, and FCM respectively. Figure 15(b) also shows some side view square particles not recognized by ICB clustering. In both cases (top and side view particles), there are almost no false positive detections by ICB clustering, indicating that AutoCryoPicker rarely picked objects from either the background area or icy area. Figure 15(c) and (d) show some side view particles not detected by k-means and FCM respectively. k-means and FCM missed more particles than ICB clustering. They had some false positives (Fig. 15(c) and (d)). In one case, a side view was mistakenly detected as a top view, and in another case a background area was detected as a top view.

Results

We evaluate the performance of AutoCryoPicker in the three stages according to multiple metrics such as clustering accuracy, particle misclassification (or particles detection) rate, Dice, and time complexity.

Datasets

Images from two datasets (Apoferritin dataset and Keyhole Limpet Hemocyanin (KLH) dataset) are used to evaluate AutoCryoPicker. The particles in the two datasets are regular shapes, which are ideal for testing AutoCryoPicker because it is designed to detect and pick regular (e.g. circular) particle shapes. Two common shapes of protein particles in cryo-EM images are circles and rectangles.

Apoferritin dataset [34] uses a multi-frame MRC image format (32 Bit Float). The size of each micrograph is 1240 by 1200 pixels. It consists of 20 micrographs each having 50 frames at 2 electrons/A^2/frame, where the beam energy is 300 kV. The particle shape in this dataset is circular.

The Keyhole Limpet Hemocyanin (KLH) dataset from US National Resource for Automated Molecular Microscopy [35] uses a single frame image format in a JPG file format. The size of each micrograph is 2048 by 2048 pixels. It consists of 82 micrographs at 2.2 electrons/A^2/pixel, where the beam energy is 300,120 kV. There are two main types of projection views in this dataset: the top view (circular particle shape) and the side view (square particle shape). The KLH dataset [33] is a standard test dataset for particle picking. The KLH dataset is a challenging dataset because of different specimens (different particles) and confounding artifacts (ice contamination, degraded particles, particle aggregates, etc.).

Evaluation metrics

In addition to the proposed clustering algorithm (ICB), we select two popular cluster algorithms (k-means and FCM). We compare them based on three factors. The first one is the running time. K-means and FCM based pairwise distance comparison is more time consuming. The second one is the effectiveness, which includes the clustering accuracy, misclassification rate, dice criteria, precision, recall, and the f1 measure. The third factor is the clustering destabilization. Because K-means and FCM use random selection for cluster initialization, they may group the same points into different clusters in different runs. This requires an extra manual step to select the most appropriate cluster representing particles, which is not fully automated. In contrast, the ICB clustering algorithm is based on computing the interval size to determine the range of the intensity of cluster centers. Therefore, the particles that have the similar intensity values will be grouped into the same cluster.

For the particles clustering stage, we use clustering accuracy and misclassification rate which are defined by Eqs. (20) and (21), respectively. Each evaluation metric is calculated according to the numbers in a confusion matrix such as the True Positive (TP) which refers to the number of correct detections of positive cases, true Negative (TN) the number of correct detections of negative cases, False

Table 1 The results of AutoCryoPicker using the three clustering methods on the first dataset (Apoferritin)

Measures	ICB	k-means	FCM
Sensitivity/Recall (%)	98.11	87.90	83.60
Specificity (%)	97.76	87.97	85.85
Precision (%)	97.11	88.81	87.99
Misclassification Rate (%)	7.784	7.666	15.881
F1 Score (%)	97.61	84.59	83.10
Accuracy (%)	95.36	81.64	78.46
DICE Score (%)	97.76	87.97	85.85
Time consuming (sec.)	1.71	10.29	30.98
Clustering Selection Approach	Fully Automated	Manually	Manually

The table reports the average of the sensitivity or recall, specificity, precision, F1 score, accuracy, DICE score, and the particle clustering time (seconds)

Table 2 The results of AutoCryoPicker using the three clustering methods on the second dataset (KLH)

Measures	ICB	k-means	FCM
Sensitivity/Recall (%)	96.23	93.42	84.67
Specificity (%)	95.095	92.71	94.7925
Precision (%)	95.095	92.71	94.7925
Misclassification Rate (%)	3.77	6.58	15.33
F1 Score (%)	95.595	92.825	88.61
Accuracy (%)	91.8275	87.5025	80.835
DICE Score (%)	95.595	92.825	89.5
Time consuming (sec.)	4.714643	23.8332305	105.676302

The table reports the average of the sensitivity or recall, specificity, precision, F1 score, accuracy, DICE score, and the particle clustering time consuming (seconds)

Positive (FP) the number of incorrect detections of positive cases and False Negative (FN) the number of incorrect detections of negative cases [37].

$$Accuracy = \frac{TP}{TP + TN} * 100 \tag{8}$$

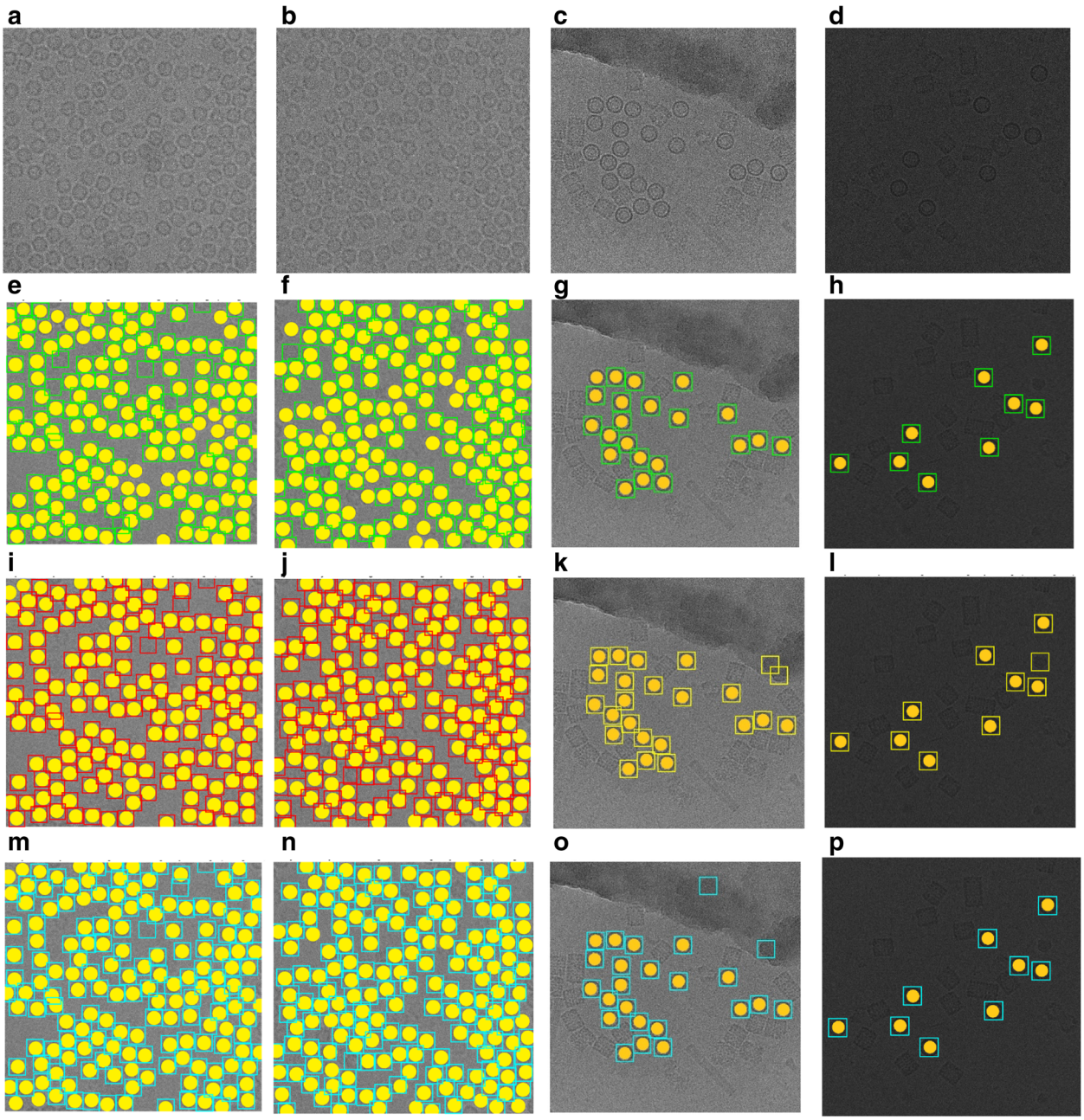

Fig. 16 Automated particle picking results on the two datasets. **a** A cryo-EM image with a high identical particle density and a lack low-frequency from the Apoferritin dataset. **b** A low SNR cryo-EM image from the Apoferritin dataset. **c** A micrograph image from the KLH dataset that includes excessively overlapped particles due to confounding artifacts such as ice contamination, degraded particles, and particle aggregates. **d** A micrograph image from the KLH dataset that has a very low spatial density and different intensity levels. **e** and (**f**) Particle picking results using Intensity Based Clustering Algorithm (ICB) (Apoferritin dataset). **i** and (**j**) Particle picking results using k-means (Apoferritin dataset). **m** and (**n**) Particle picking results using FCM (Apoferritin dataset). **g** and (**h**) Particle picking results using Intensity Based Clustering Algorithm (ICB) (KLH dataset). **k** and (**l**) Particle picking results using k-means (KLH dataset). **o** and (**p**) Particle picking results using FCM (KLH dataset)

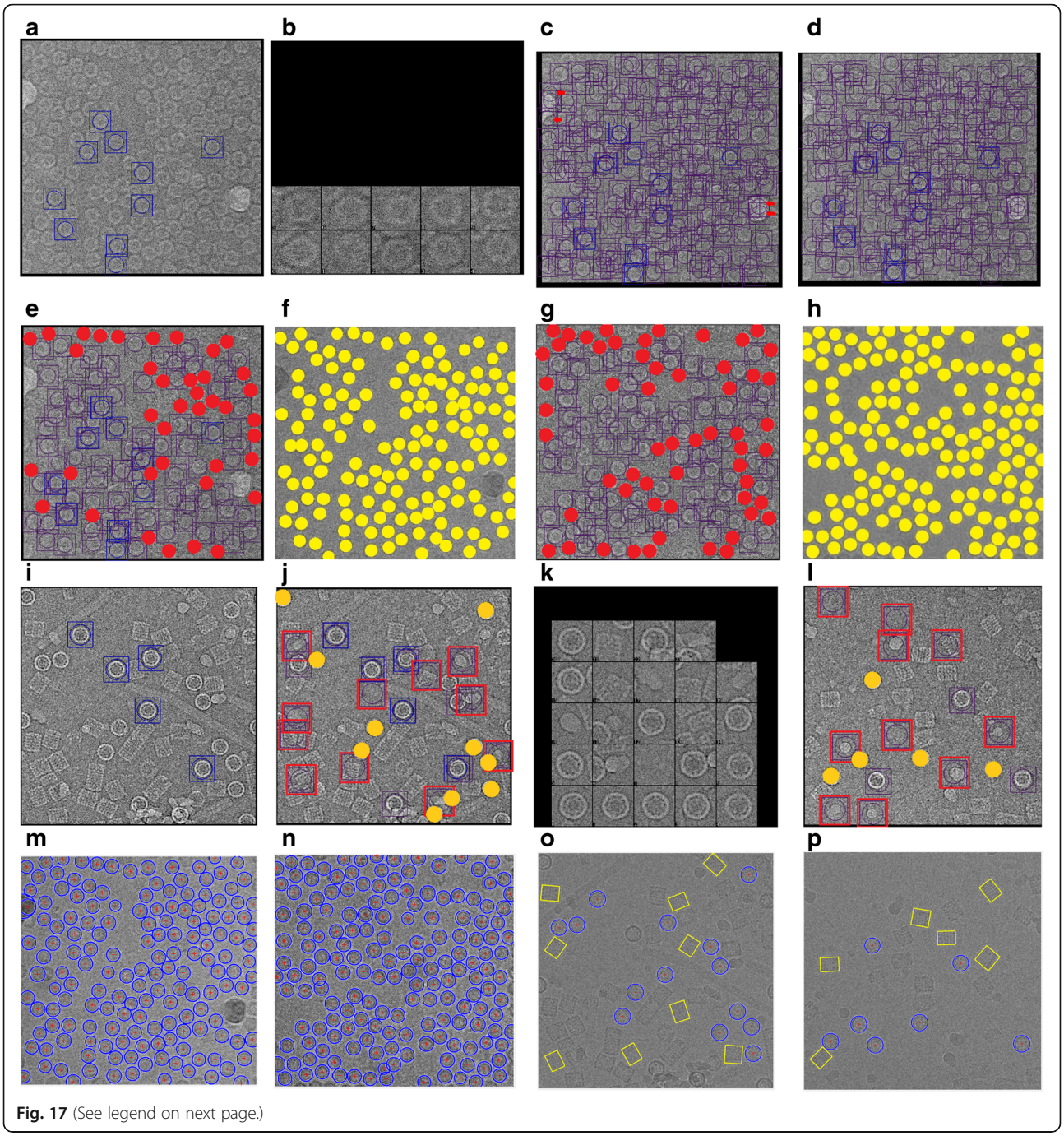

Fig. 17 (See legend on next page.)

(See figure on previous page.)
Fig. 17 Particle picking using EMAN2 and AutoCryoPicker. **a** The manually selected reference particles of the Apoferritin dataset that were used for automated particle picking with EMAN2. **b** Zoomed-in view of the reference particles for the Apoferritin dataset. **c** EMAN2 automatic picking result based on threshold value = 0.0 using the first tested image of the Apoferritin dataset. **d** EMAN2 automatic picking result based on threshold value = 0.5 using the first tested image of the Apoferritin dataset. **e** EMAN2 automatic picking result based the threshold value = 2.3 using the first tested image of the Apoferritin dataset. Red dots mark missed particles). **f** Ground truth of first tested image of the Apoferritin dataset. Yellow dots mark valid particles. **g** EMAN2 automatic picking result based the threshold value = 2.3 using the second tested image of the Apoferritin dataset. Red dots mark missed particles). **h** Ground truth of second tested image of the Apoferritin dataset. Yellow dots mark valid particles. **i** The manually selected reference particles of the KLH dataset that were used for automated picking of top-view (circular) particles with EMAN2. **j** EMAN2 automatic picking result based the threshold value = 0.5 using the first tested image of the KLH dataset. Red squares mark the false positives and the yellow dots the missing particles. **k** Zoomed-in view of the automatically picked particles (threshold value = 0.5) for first tested image of the KLH dataset. **l** EMAN2 automatic picking result based the threshold value = 0.5 using the second tested image of the KLH dataset. Red squares mark the false positives, and the yellow dots mark the missing particles (top-view). **m** Particle picking result from AutoCryoPicker using the first tested image of the Apoferritin dataset. Red '+' mark the center of each particle and blue circles the top-view detected particles in the cryo-EM image. **n** Particle picking result from AutoCryoPicker using the second tested image of the Apoferritin dataset. Red '+' mark the center of each particle and blue circles the top-view detected particles in the cryo-EM image. **o** Particle picking result from AutoCryoPicker using the first tested image of the KLH dataset. Red '+' marks the center of each particle, blue circles the top-view detected particles in the cryo-EM image, and the yellow squares the side-view detected particles in the cryo-EM image. **p** Particle picking result from AutoCryoPicker using the second tested image from the KLH dataset. Red '+' marks the center of each particle, blue circles the top-view detected particles in the cryo-EM image, and the yellow squares the side-view detected particles in the cryo-EM image

$$Misclassification\ Rate = \frac{FP + TN}{Total} * 100 \tag{9}$$

Moreover, Dice Criteria (DIC) is also used for the similarity measure between a cluster image and the Ground Truth (GT). DC is defined by Eq. (10) [38]:

$$Dice = \frac{2(A \cap B)}{A + B} * 100 \tag{10}$$

where, A is the cluster image and B is the ground truth image of A. Finally, we use the precision, recall, and F1 measure scores [37] to evaluate the particle picking results in the particle picking stage. The precision, recall, and F measure are defined by Eqs. (11), (12) and (13), respectively [39]:

$$Precision = \frac{TP}{TP + FP} * 100 \tag{11}$$

$$Recall = \frac{TP}{TP + FN} * 100 \tag{12}$$

$$F1\ measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{13}$$

Particle clustering, detection and picking results

In order to evaluate the performance of automated particle clustering and picking, we generated a true reference by manually picking the particles on the images. Figure 11(a), and (k) show two different cryo-EM images from the two datasets (Apoferritin and KLH), respectively. The results on one image from the Apoferritin dataset are shown in Fig. 10(d), (g), and (j) while the results for KLH dataset are shown in Fig. 11(n), (q), and (t). It was demonstrated that most of the particles were correctly picked by AutoCryoPicker. Table 1 reports the recall, precision, accuracy, F1 score, and the running time of AutoCryoPicker based on three clustering algorithms: K-means, FCM, and IBC. On the Apoferritin dataset the AutoCryoPicker based on ICB clustering achieves a higher accuracy of 95.36% than 84.59 and 78.46% of k-means and FCM respectively. Also, ICB ran significantly faster in particles clustering (average time 1.71 s versus 10.29 s and 30.98 s of k-means and FCM, respectively).

Table 2 shows the results on the KLH dataset. AutoCryoPicker based on ICB achieves a higher accuracy 91.82% than that of k-means and FCM (i.e. 87.50 and 80.83% respectively). The average clustering time of the whole dataset using ICB was 4.7 s on average, faster than the k-means by 23.8 s and 105.8 s of the FCM.

Two different cases from each of the two datasets are illustrated in Fig. 16. Figure 16(a) shows cryo-EM images of a high particle density from the Apoferritin dataset with a low-frequency and Fig. 16(b) a cryo-EM image of low SNR. Figure 16(c) and (d) shows two different micrograph cases from the KLH dataset that consist of excessively overlapped particles and some confounding artifacts such as ice contamination, degraded particles, and particle aggregates. AutoCryoPicker still performed very well on these cases. Figure 16(e)-(p) show the particle picking results using ICB, k-means, and FCM methods on the two datasets, respectively.

Comparison with another particle picking software

EMAN2 was selected as an example of particle picking software for cryo-EM images [25]. The "e2boxer.py" program of EMAN2 was applied to the same images input to AutoCryoPicker.

For the Apoferritin images, a reference set of 10 particles was selected manually (Fig. 17(a), 17(b)) and then automated picking was performed with different threshold values (lower threshold results in more particles picked). For example, use of arbitrarily low threshold values of 0.0 and 0.5 results in most of the valid particles being

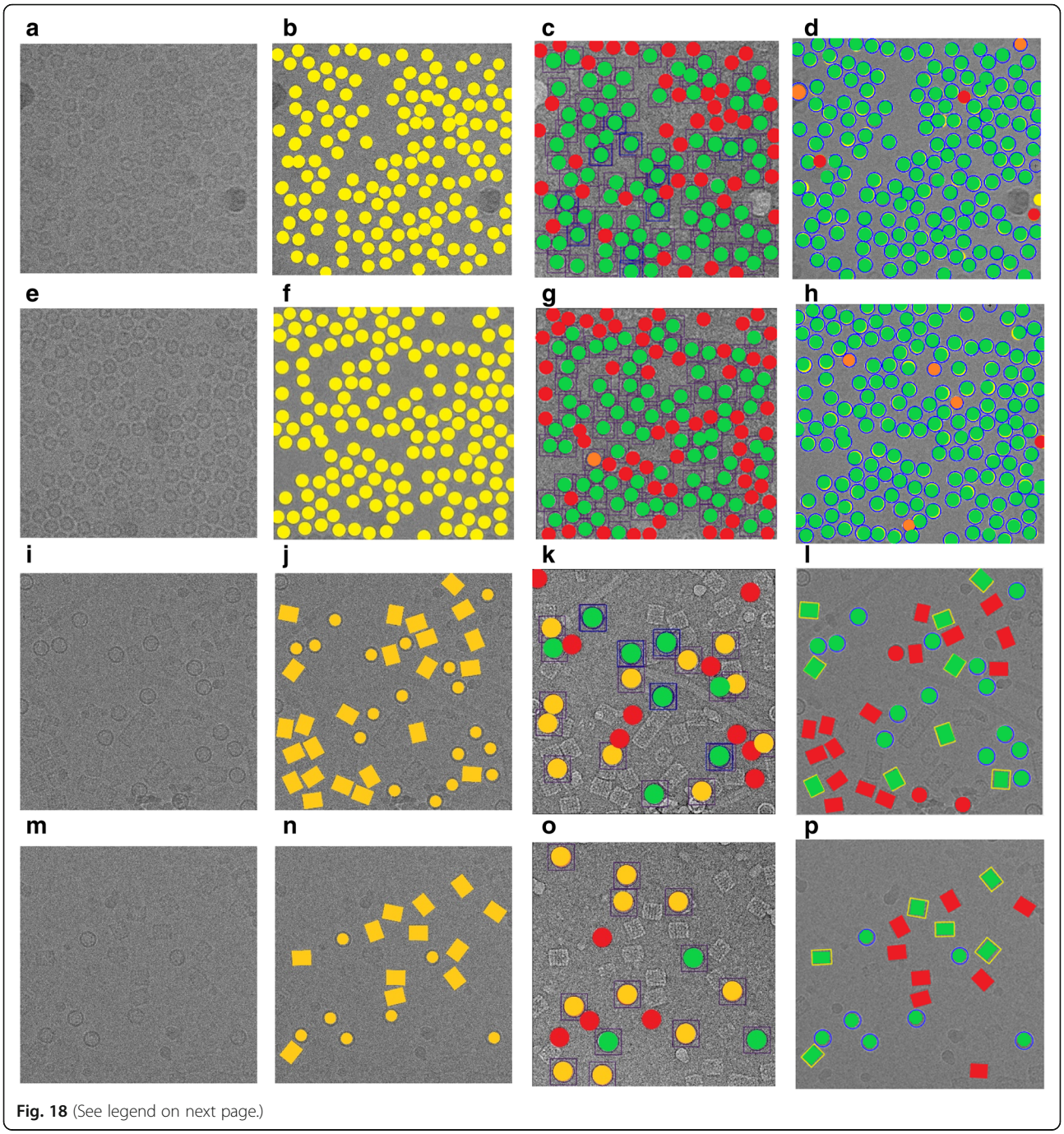

Fig. 18 (See legend on next page.)

(See figure on previous page.)
Fig. 18 Evaluation of particle picking using EMAN2 and AutoCryoPicker. **a** Apoferritin cryo-EM image with top-view particle shapes only. **b** The ground truth (manually particle picking labels) of the first Apoferritin cryo-EM image where each particle is marked by a yellow circle on top of each particle. **c** The particle picking results of the first Apoferritin image using EMAN2. The particles are labeled as follows: Green, True Positive (TP); red, False Negative (FN). **d** The particle picking results of the first Apoferritin cryo-EM image using AutoCryoPicker. The particles are labeled as follows: Green, True Positive (TP); red, False Negative (FN); orange, False Positive (FP). **e** The second original Apoferritin cryo-EM image with top-view particle shapes only. **f** The ground truth (manually particle picking labels) of the second Apoferritin cryo-EM image where each particle is marked by a yellow circle on top of each particle. **g** The particle picking results of the second Apoferritin cryo-EM image using EMAN2. The particles are labeled as follows: Green, True Positive (TP); red, False Negative (FN); orange, False Positive (FP). **h** The particle picking results of the second Apoferritin cryo-EM image using AutoCryoPicker. The particles are labeled as follows: Green, True Positive (TP); red, False Negative (FN); orange, False Positive (FP). **i** The first original KLH cryo-EM image. (j) The ground truth (manually particle picking labels) of the first KLH cryo-EM image where each particle is marked by a yellow circle on top of each particle. **k** The particle picking results of the first KLH image using EMAN2. The particles are labeled as follows: Green, True Positive (TP); red, False Negative (FN). **l** The particle picking results of the first KLH cryo-EM image using AutoCryoPicker. The particles are labeled as follows: Green, True Positive (TP); red, False Negative (FN). **m** The second original KLH cryo-EM image which has top-view particle shapes only. **n** The ground truth (manually particle picking labels) of the second KLH cryo-EM image where each particle is marked by a yellow circle on top of each particle. **o** The particle picking results of the second KLH cryo-EM image using EMAN2. The particles are labeled as follows: Green, True Positive (TP); red, False Negative (FN). **p** The particle picking results of the second KLH cryo-EM image using AutoCryoPicker. The particles are labeled as follows: Green, True Positive (TP); red, False Negative (FN)

selected; however, false positives likely corresponding to thick ice were also selected (Fig. 17(c), (d)). Increasing the threshold to a more reasonable value of 2.3 resulted in no false positives at the expense of leaving several good particles unpicked (Fig. 17(e), (g)). The lack of particle set completeness is evident by comparison to the ground truth result (Fig. 17(f), (h)). In comparison, AutoCryoPicker successfully captured all the valid particles on the images without any false positives (Fig. 17(m), (n)).

Similarly, results of using EMAN2 autopicking with the circular particles in the KLH images yielded incomplete recording of the valid particles and several false positives (Fig. 17(j), (l)). In contrast, AutoCryoPicker was able to identify almost all of the true particles (both the circular and rectangular projections) in the KLH images, without the generation of false positives (Fig. 17(o), (p)).

Quantitative assessment of the comparison is shown in Fig. 18 and Tables 3, 4 and 5. Figure 18(a) and (e) show two original images from the Apoferritin dataset where the images have top-view particle shapes only. Figure 18(b) and (f) show the manually particle picking labels (Ground Truth) where each particle is marked by a yellow circle on top of each particle in the original images. Figure 18(c) and (g) show the particle picking performance results using EMAN2. In terms of evaluating each particle's picking tool in addition to the AutoCryoPicker, three criteria are selected to label and evaluate the particles picking performance results. True Positive (TP) picking where the correct particles are marked by the green circles. False Negative (FN) picking where the missed particles are marked by red circles. False Positive (FP) picking where the incorrectly picked particles are marked by orange circles. Figure 18 (d) and (h) show the same criteria of the particle picking results using AutoCryoPicker. Similarly, two images from the KLH dataset are shown in Fig. 18(i) and (m). Figure 18(j) and (n) show the particles ground truth (hand picking and labeling). Figure 18(k) and (o) illustrate the performance results of the particle picking using EMAN2. Figure 18(l) and (p) show the same performance results using AutoCryoPicker.

Table 3 illustrates the statistical evaluation of the performance results based on the TP, FN, FP for each single particle picking algorithm, as well as the particle shape class and total number of the particles (ground truth) in each image. Note that AutoCryoPicker performed better in detecting two different particle shapes on same images (Table 3).

Table 3 Statistical evaluation AutoCryoPicker and EMAN2 performance using the Apoferritin and KLH images

Cryo-EM images	Particle Shape	Total Particles Number	AutoCryoPicker			EMAN2		
			TP	FN	FP	TP	FN	FP
Apoferritin Image 1	Top-View	151	148	3	2	84	67	0
Apoferritin Image 2	Top-View	160	159	1	5	83	76	1
KLH image 1	Top-View	17	14	3	0	8	9	11
KLH image 2	Top-View	7	7	0	0	3	4	10
KLH image 1	Side-View	24	8	15	0	N/A	N/A	N/A
KLH image 2	Side-View	14	6	8	0	N/A	N/A	N/A

The table reports TP: True Positive picking results where the correct particles are picked, FN: False Negative picking results where some good particles are missed, FP: False Positive picking results where the incorrect particles (other objects such as background or artificial objects) are picked as particles

Table 4 Evaluation of particle picking on Apoferritin images

Measures	AutoCryoPicker	EMAN2
Sensitivity/Recall (%)	98.70	53.92
Precision (%)	97.81	99.41
Misclassification Rate (%)	1.31	46.09
F1 Score (%)	98.25	69.90
Accuracy (%)	96.55	53.76
DICE Score (%)	98.24	69.90

Table 5 Evaluation particle picking on the second KLH image

Measures	AutoCryoPicker	EMAN2
Sensitivity/Recall (%)	90.87	59.44
Precision (%)	98.48	70.46
Misclassification Rate (%)	9.14	40.57
F1 Score (%)	94.24	59.96
Accuracy (%)	89.36	43.33
DICE Score (%)	94.24	37.22

Table 4 illustrates the evaluation of different single particle picking methods by reporting the average performance results using images form the Apoferritin dataset. AutoCryoPicker achieves a higher recall (98.70) and accuracy (96.55) compared to EMAN2 (53.92 and 53.76, respectively). Also, AutoCryoPicker achieved a higher f1 score (98.25) and dice score (98.24), as well as a low false negative rate (1.31).

Finally, Table 5 shows the performance results of different particle picking methods using KLH images. The performance results in Table 5 have been calculated based on the circular particle detection only (top-view particles) since EMAN2 was challenged in detecting two different particle shape in the same image at the same time as shown in Table 3. In this case, AutoCryoPicker achieves higher recall (90.87), precision (98.48), F1 score (94.24), accuracy (89.36), dice score (94.24) and low miss classification rate (9.14).

Conclusions

Accurate particle picking in cryo-EM images still requires substantial human intervention and, therefore, can be labor-intensive and time-consuming. To address this challenge, we develop AutoCryoPicker – a fully automated particle picking approach based on image preprocessing, unsupervised clustering and shape detection. Our experiments show that the approach can significantly improve signal to noise ratio in cryo-EM images and pick particles rather accurately. Therefore, the automated method can relieve scientists from the laborious work of picking cryo-EM particles and help improve the efficiency and effectiveness of cryo-EM based protein structure determination. We conclude that AutoCryoPicker has the potential for being incorporated into the particle picking pipelines of other cryo-EM image processing software.

Additional file

Additional file 1: Supplementary document. (DOCX 100 kb)

Abbreviations

Cryo-EM: cryo-electron microscopy; Micrograph: Digital image taken through a microscope; MRC: Medical Research Council; PNG: Portable Network Graphic

Acknowledgements

Some tests on Apoferritin dataset were carried out by Yuhan Chen. The manuscript has been proofread by Max Highsmith.

Authors' contributions

JC conceived of the project. AA and JC designed the experiment. AA implemented the method and gathered the results. AA, JC, AO and JT analysed the data. AA and JC wrote the manuscript. All authors edited and approved the manuscript.

Funding

Research reported in this publication was supported in part by two NSF grants (DBI 1759934 and IIS1763246) to JC, an NIH grant (R01GM093123) to JC and JT, and an administrative supplement to R01GM065546 (Collaborative Supplements for Cryo-Electron Microscopy Technology Transfer) to JT.
The funding bodies did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets used in this study and the source code of AutoCryoPicker are available at https://github.com/jianlin-cheng/AutoCryoPicker.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare they have no conflict of interest.

Author details

[1]Electrical Engineering and Computer Science Department, University of Missouri, Columbia, MO 65211, USA. [2]Departments of Biochemistry and Chemistry, University of Missouri, Columbia, MO 65211-2060, USA. [3]Informatics Institute, University of Missouri, Columbia, MO 65211, USA.

Received: 19 November 2018 Accepted: 31 May 2019
Published online: 13 June 2019

References

1. Nogales E, Scheres SH. Cryo-EM: a unique tool for the visualization of macromolecular complexity. Mol Cell. 2015;58(4):677–89.
2. Merk A, Bartesaghi A, Banerjee S, Falconieri V, Rao P, Davis MI, Pragani R, Boxer MB, Earl LA, Milne JLS, Subramaniam S. Breaking Cryo-EM resolution barriers to facilitate drug discovery. Cell. 2016;165(7):1698–707.
3. Doerr, Allison. 2016. "Single-particle cryo-electron microscopy." Nat Methods 23. https://www.nature.com/articles/nmeth.3700?draft=collection.
4. Jiang J, Pentelute BL, Collier RJ, Zhou ZH. Atomic structure of anthrax protective antigen pore elucidates toxin translocation. Nature. 2015; 521(7553):545–9.
5. Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Milne JL, Subramaniam S. 2.2 a resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor. Science. 2015;348(6239):1147–51.
6. Campbell, M.G., D. Veesler, A. Cheng, C.S. Potter, B. Carragher. 2015. "2.8 a resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy", Elife 4.
7. Herzik, M.A., Jr., M. Wu, G.C. Lander. 2017. "Achieving better-than-3-a resolution by single-particle cryo-EM at 200 keV.", Nat Methods 14(11):1075–1078.
8. Zhu Y, Carragher B, Robert M, Glaeser D, Fellmann C, Bajaj M. Automatic particle selection: results of a comparative study. J Struct Biol. 2004:3–14.
9. Glaeser RM, Nicholson WV, Robert M. Review: automatic particle detection in electron. J Struct Biol. 2001;133:90–101.
10. Umesh Adiga PS, Malladi R, Baxter W, Glaeser RM. A binary segmentation approach for boxing ribosome particles in cryo EM micrographs. Journal of Structural. 2004;145:142–51.
11. Voss NR, Yoshioka CK, Radermacher M, Potter CS, Carragher B. DoG picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. J Struct Biol. 2009;166:205–13.

12. Zhao J, Brubaker MA, Rubinstein JL. TMaCS: a hybrid template matching and classification system for partially-automated particle selection. J Struct Biol. 2013;181:234–42.
13. Liu Z, Guo F, Wang F, Li T-C, Jiang W. A resolution cryo-em 3d reconstruction of close-packed virus particles. Structure. 2016;24:319–28.
14. Norousi R, Wickles S, Leidig C, Becker T, Schmid VJ, Beckmann R, Tresch A. Automatic post-picking using MAPPOS improves particle image detection from cryo-EM micrographs. J Struct Biol. 2013:59–66.
15. Grigorieff JZ, Chen N. SIGNATURE: a single-particle selection system for molecular electron microscopy. J Struct Biol. 2007;157:168–73.
16. Patwardhan, Richard J Hall, Ardan. 2004. "A two step approach for semi-automated particle selection from low contrast cryo-electron micrographs", J Struct Biol 19–28.
17. Penczek PA, Huang Z, Pawel A. Application of template matching technique to particle detection in electron micrographs. J Struct Biol. 2004;145:29–40.
18. Langlois R, Pallesen J, Frank J. Reference-free particle selection enhanced hanced with semi-supervised machine learning for cryo-electron microscopy. J Struct Biol. 2011:353–61.
19. Sorzano C, Recarte E, Alcorlo M, Bilbao-Castro JR, San-Martîn C, Marabini R, Carazo JM. Automatic particle selection from electron micrographs using machine learning techniques. J Struct Biol. 2009:252–60.
20. Arbelez P, Han B-G, Typke D, Lim J, Glaeser RM, Malik J. Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. J Struct Biol. 2011:319–28.
21. Wang F, Gong H, Liu G, Li M, Yan C, Xia T, Li X, Zeng J. DeepPicker: a deep learning approach for fully automated particle picking in Cryo-EM. J Struct Biol. 2016:325–36.
22. Zhu Y, Ouyang Q, Mao Y. A deep convolutional neural network approach to single-particle recognition in cryo-electron microscopy. BMC Bioinformatics. 2017:2–10.
23. MacQueen, J. 1967. "Some methods for classification and analysis of multivariate observations", in Proc. 5th Berkeley Symp. On math. Stat. And probability. Berkeley, CA. 281–297.
24. Dunn JC. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybern. 1973:32–57.
25. G. Tang, L. Peng, P.R. Baldwin, D.S. Mann, W. Jiang, I. Rees & S.J. Ludtke. n.d. "EMAN2: an extensible image processing suite for electron microscopy", J Struct Biol 157 (PMID: 16859925): 38–46.
26. Guo F., Jiang W. (2014) Single Particle Cryo-electron Microscopy and 3-D Reconstruction of Viruses. In: Kuo J. (eds) Electron Microscopy. Methods in Molecular Biology (Methods and Protocols), vol 1117. Humana Press, Totowa, NJ.
27. Herv'e, Abdi. 2010. "Normalizing Data", By Herv'e Abdi. The University of Texas at Dallas: In Neil Salkind (Ed.), Encyclopedia of Research Design.
28. The MathWorks, Inc. 2018. Image processing toolbox™ User's guide. Natick, MA: The MathWorks, Inc. https://www.mathworks.com/help/images/contrast-adjustment.html.
29. Woods, R. C. Gonzalez, R. E. 2018. "Digital Image Processing", 4th Edition. University of Tennessee.
30. Amit Singer," Mathematics for cryo-electron microscopy", arXiv:1803.06714v1 [physics.comp-ph] 12 Mar 2018.
31. Bhamre T, Zhang T. Singer A. Denoising and Covariance Estimation of Single Particle Cryo-EM Images. J Struct Bio. 2016. p. 195. https://doi.org/10.1016/j.jsb.2016.04.013.
32. Stark JA. Adaptive image contrast enhancement using generalizations of histogram equalization. IEEE Trans Image Process. 2000:889–69.
33. He K, Sun J, Tang X. Guided image filtering. IEEE Trans Pattern Anal Mach Intell. 2013.
34. Grant T, Rohou A , Grigorieff N. 2017. EMPIAR-10146. 07 12. Accessed 03 Sept 2018. https://www.ebi.ac.uk/pdbe/emdb/empiar/entry/10146/#&gid=1&pid=1.
35. N.d.,"KLH Dataset", available Online, http://nramm.nysbc.org/.
36. Atherton, D. Kerbyson, T. 1995. "circle detection using Hough transform filters." proc. 5th Int. Conf. Image process, Appl., U.K. 370–374.
37. Langlois R. A clarification of the terms used in comparing semi-automated particle selection algorithms in Cryo-EM. J Struct Biol. 2011;175:348–52.
38. Rohlfing T. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. IEEE Trans Med Imaging. 2012:153–63.
39. Steve on Image Processing, "Feret Properties – Wrapping Up", concepts, algorithms & MATLAB, https://blogs.mathworks.com/steve/2018/04/17/feret-properties-wrapping-up/.

Publisher's Note