*Bioinformatics*, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

Manuscript Category

|  |
| --- |
| Subject SectionAutoCryoPicker: An Unsupervised Learning Approach for Fully Automated Single Particle Picking in Cryo-EM ImagesAdil Al-Azzawi1, Anes Ouadou1, David Korasick2**,** John Tanner2, and Jianlin Cheng1\*1Department of Electrical Engineering and Computer Science, 2Department of Biochemistry, University of Missouri, Columbia, MO 65211-2060, USA.\*To whom correspondence should be addressed.Associate Editor: XXXXXXXReceived on XXXXX; revised on XXXXX; accepted on XXXXX Abstract**Motivation:** An important task of macromolecular structural determination by electron cryo-microscopy (cryo-EM) is the single particle picking. Single particle extraction from cryo-EM images is currently laborious, time consuming and labor intensive. Thus, it is important to develop a fully automated framework for single particle picking in cryo-EM images. The supervised machine learning (e.g. deep learning) methods for particle picking often need a large training dataset which requires expensive manual labeling. Other reference-dependent methods rely on low-resolution templates for particle detection, matching and picking, and therefore are not fully automated.**Results:** We design a fully automated approach based on unsupervised clustering for single particle picking in noisy cryo-EM micrographs. Our approach consists of three stages: image preprocessing, particle clustering, and particle picking. The image preprocessing based on image averaging, normalization, contrast transfer correction, histogram equalization, restoration, adaptive histogram equalization, guided image filtering, and morphological operations significantly improve the quality of original cryo-EM images. Our particle clustering method uses an intensity-based distribution model instead of the distance metric. It is much faster and more accurate than traditional K-means and FCM algorithms for single particle picking in cryo-EM images. 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 the bounding box encapsulating the particles. **Availability:** The source code of our method – AutoCryoPicker is available at [https://github.com/jianlin-cheng/AutoCryoPicker](https://mail.umsystem.edu/owa/redir.aspx?REF=FpQUlr4PYZRt6mNWyZlcBgXFbO5Y4fzTqJCpiTd7MO-nvJz0hQPWCAFodHRwczovL2dpdGh1Yi5jb20vamlhbmxpbi1jaGVuZy9BdXRvQ3J5b1BpY2tlcg..)**Contact:** chengji@missouri.edu **Supplementary information:** Supplementary data are available at *Bioinformatics* online. |

# Introduction

For decades X-ray crystallography reigned as the dominant technique for obtaining high-resolution information about macromolecular structure. Single-particle cryo-EM was traditionally used to provide insights into the morphology of large protein complexes that resisted crystallization, albeit at substantially lower resolutions than crystallography. Though the overall strategy has not changed appreciably over the years, very recent technological advances in sample preparation, computation and especially instrumentation are now allowing researchers to use cryo-EM to solve near-atomic-resolution macromolecular structures (Doerr, 2016).

Cryoelectronic microscopy or cryo-electron microscopy (also known as a cryo-EM) is the method originally used to “take photographs” of viruses and other macromolecular complexes. It becomes increasingly popular to study the structures of protein complexes. Through cryo-EM, a number of large molecular complexes, such as ribosome, have been resolved to near atomic resolutions (Z. Liu, 2016). And recent advances in single-particle cryo-EM technique has revolutionized the field of structural biology (K. Al Nasr L. C., 2012.) (K. Al Nasr D. R., 2014) (K. Al Nasr D. R., 2011), and enabled protein complex structure determination at near-atomic resolution (M. L. Baker S. S., 2011) (M. L. Baker T. J., 2007) (A. Biswas, 837–843). An overall single-particle cryo-EM workflow is shown in Fig. S1 in the supplemental document. The first step is sample preparation. Then, the sample is placed on a grid that is rapidly plunged into a cryogen such as liquid ethane (-180°C for liquid nitrogen stages, -269°C for He) for flash-freezing and particle trapping in a thin film of vitreous ice. In addition to capturing the protein structure at the moment of freezing, this process protects the sample to some degree from radiation damage and prevents evaporation of buffer in the high-vacuum conditions of a transmission electron microscope.

Two-dimensional cryo-EM images of the particles are taken by electron microscope, which contain randomly arranged particles along with non-particles—bits of frost, deformed particles, protein aggregates and so on. These images suffer from heavy background noise and low contrast, due to a limited electron dose used in imaging. A large number of single-particle images, extracted from cryo-EM micrographs, are required to perform a reliable 3D reconstruction of the underlying structure. Particle recognition thus represents the first bottleneck in the practice of cryo-EM structure determination.

To address the bottleneck, numerous computational approaches have been proposed to facilitate the particle picking process (Yuanxin Zhu, 2004) (Glaeser., 2001) (P.S Umesh Adiga, 2004) (N.R. Voss, 2009) (Jianhua Zhao, 2013) (Ramin Norousi, 2013). These methods can be basically divided into three categories, generative methods (Grigorieff, 2007) (Patwardhan, 2004) (Penczek, 2004), discriminative classification (Robert Langlois, 2011) (C.O.S. Sorzano, 2009) (Pablo Arbelez, 2011) including the recent deep learning approaches (Feng Wang, 2016) (Yanan Zhu, 2017) and unsupervised learning methods (P.S Umesh Adiga, 2004). The generative approaches measure the similarity to a reference to identify particle candidates from micrographs. A typical generative method employs a template-matching technique (Patwardhan, 2004) (Penczek, 2004) 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 examples, then apply this trained classifier to detect and recognize particle images from micrographs (Robert Langlois, 2011) (C.O.S. Sorzano, 2009) (Pablo Arbelez, 2011). The automated particle picking pipeline based deep learning approached (Feng Wang, 2016) (Yanan Zhu, 2017) consist of two steps: the training step and the test step. In the training step, a set of manually labeled positive and negative samples is used to train a convolutional neural network (CNN) model for particle picking. The trained CNN classifier is then used to select particle images from input micrographs in the test step. In the deep learning particle picking scheme (Feng Wang, 2016), a user is not required to manually pick any particle from micrographs. However, the training dataset has to be manually created to train the model, which is labor intensive.

Although these generative and discriminative approaches have greatly reduced time and effort spent on single-particle data analysis, they 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.

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 (P.S Umesh Adiga, 2004) (N.R. Voss, 2009). But they rarely fully exploit the intrinsic and unique characters of particles to facilitate automated particle picking.

Therefore, the unsupervised approaches are often combined with the template-matching or classification-based approaches to achieve decent picking results. In this paper, we develop a fully automated approach for particle picking (AutoCryoPicker) that is based on advanced image preprocessing, robust unsupervised learning clustering without any human intervention, and sophisticated shape detection. The experimental results demonstrate that the fully automated particle picking scheme can accurately detect and select a sufficient number of particles that are comparable to those picked manually. It is also more accurate than k-means and Fuzzy C-means (FCM) clustering methods. Therefore, our new automated picking approach can significantly reduce time and labor spent on single-particle data analysis and thus greatly relieve a bottleneck in the automated cryo-EM structure determination pipeline.

# Unsupervised Learning Based Fully Automated Particle Picking Framework

Image clustering based unsupervised leering is an essential data analysis in image processing. K-means (MacQueen, 1967) and fuzzy c-means (FCM) (J. C. Dunn, 1973) are the two popular unsupervised clustering algorithms based on the distance or the dissimilarity between data objects. For instance, K-means may use the Euclidean distance between data points. Instead of a binary clustering, FCM has an additional factor called membership. This means that a data point can have partial membership to multiple clusters instead of just single one.

We observed that the distance or dissimilarity metric-based methods in the cryo-EM image clustering has some shortcomings such as: (1) Clustering destabilization: different clustering results may be obtained from different random initializations of cluster centers; and (2) Time consuming: computing distances between any two points or between a data point and a center is time consuming. We design a new clustering approach based on the intensity differences instead of the distance metric for cryo-EM images clustering.

We implemented the AutoCryoPicker framework for automated particle picking in single-particle cryo-EM structure determination as shown in Fig. S2 in the supplementary document. In our automated particle picking scheme, the user is not required to manually pick any particle from the micrographs in consideration. The fully automated approach is divided into three main stages: preprocessing, clustering, and particles picking. In the preprocessing step, several image processing methods are applied to enhance the input cryo-EM images such as image normalization, Contrast Transfer Correction Function (CTF) etc. After that, in the clustering stage a binary mask is generated for particles detecting and picking through clustering. Clustering is done using three different algorithms k-means (MacQueen, 1967), FCM (J. C. Dunn, 1973) a new robustness clustering algorithm that addresses some typical clustering issues such as clusters stabilization and cluster center random initialization.

## Preprocessing Stage

 In standard, the cryo-EM image has become industry in a MRC format where the results of the technique is a three-dimensional grid of [voxels](https://en.wikipedia.org/wiki/Voxel) each with a value corresponding to [electron density](https://en.wikipedia.org/wiki/Electron_density) or [electric potential](https://en.wikipedia.org/wiki/Electric_potential). In this case, the data block of an MRC format file holds a 3D array of data based on different modes. Since our idea is to design such a fully automated particle picking by applying some unsupervised learning clustering algorithms in additional to propose a new, we use to convert all the cryo-EM images dataset from the MRC format to 16-bits PNG images format using EMAN2. The main purpose for doing that is to project the 3D image dimensional to 2D which will be more significantly easier to our implementation model and more response for image preprocessing tools to enhance the high noisy cryo-EM images.

### 2.1.1 Cryo-EM Image Averaging and Normalization

Because different cryo-EM image has different ice thickness and similar pixel value ranges, each individual cryo-EM image should be normalized. We typically normalize cryo-EM images by setting the background mean to zero and background variance to one. With this normalization, the voxel values become equivalent to *Z*-score, i.e., the number of sigma’s above noise level (Abdi, 2010) as it shown in equation (1):

 (1)

Where is the mean of the intensity pixels values, and is the standard deviation. Since each cryo-EM image may have a number of (e.g. 50) frames per image (see Fig. S3 (a), (b), (c), and (d) in the supplemental document for an example) and each single frame image converted from MRC to PNG format using EMAN2 is of bad quality as shown in supplemental Fig. S3 (e), we use the image averaging and normalization function in EMAN2 to average the frames of each image, resulting in the converted cryo-EM of much better quality as shown in supplemental Fig. S3 (f).

### 2.1.2 Cryo-EM Image Contrast Transfer Correction (CTC)

We develop a contrast transfer correct (CTC) procedure to improve signal to noise ratio in cryo-EM images. It has the following three steps.

1. Find Limits to Contrast Stretch: In this step, we specify the low and high image intensity range by detecting the *low\_high* value as it shown in the cryo-EM histogram in supplemental Fig. S4 (top part). By default, values in *low\_high* specify the bottom 2% and the top 2% of all pixel values. The gray values returned can be used by the “*imadjust*” function to increase the contrast of an image as shown at the bottom of supplemental Fig. S4 (The MathWorks, 2018).

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 between 0 and 1 value by dividing the intensity values of each pixel as shown in equation (2) (The MathWorks, 2018).

(2)

where and are the row and column index of cryo-EM image matrix and the column respectively. Supplemental Fig. S5 (a) shows an example of an cryo-EM image after mid-range stretching.

Intensity Adjustment: In this step, the intensity values of the cryo-EM image are adjusted to new values by using MATLAB function “*imadjust*” (The MathWorks, 2018). Fig. S5(b) shows an example of a cryo-EM image with contrast transfer correction.

### 2.1.3 Cryo-EM Histogram Equalization

We use the histogram equalization as an intensive preprocessing step to increase the contrast between the particles objects and the background. Assuming is a cryo-EM image represented by a number of pixels in the range from 0 to *L* pixel intensities in the range [0-1]. denote the normalized histogram of with a bin for each possible intensity as shown in Equation (3) (Woods, 2018).

 (3)

Then, the histogram equalized cryo-EM image is defined by equation (4) (Woods, 2018).

 (4)

where is defined as rounding down to the nearest integer. This is equivalent to transforming the pixel intensities, of by the function in Equation (5) (Woods, 2018). (comments: what does fij mean here?)

 (5)

Supplemental Fig. S6 (a) and (c) an example of the cryo-EM image before and after the histogram equalization.

###  2.1.4 Cryo-EM Image Restoration

We apply Wiener to remove the additive noise and inverts the blurring in cryo-EM images (Woods, 2018). 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 equation (7) (Woods, 2018):

 (7)

where  are respectively power spectra of the original image and the additive noise, and is the blurring filter.

Supplemental Fig. S7 (a) and (b) an example before and after applying Wiener filtering.

### 2.1.5 Cryo-EM Adaptive Histogram Equalization

Global histogram equalization can enhance the local details of the cryo-EM image (Stark, 2000 ). We apply the Adaptive Histogram Equalization (The MathWorks, 2018) 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. Adaptive Histogram Equalization combines neighboring tiles using bilinear interpolation to eliminate artificially induced boundaries. To achieve the uniform distribution or local uniform distribution, assuming that the is the gray value of gray level, the algorithm compares with : if , then map forward; , then map backward based on Equation (8) (Woods, 2018).

 (8)

where . Supplemental Fig. S8, (a) and (c) an example of cryo-EM image before and after applying the Contrast-limited adaptive histogram equalization. The enhanced cryo-EM look darker and particles look brighter. The overall contrast is improved even the histogram looks more smoothly than the previous one.

### 2.1.6 Cryo-EM Guided Image Filtering

We perform guided filter operation on cryo-EM images as an edge-preserving smoothing operator. In this step, the guided filter generates the filtering output by considering the content of a guidance image, which can be the input image itself or another different image. It has a theoretical connection with the matting Laplacian matrix (Woods, 2018) and can better utilize the structures in the guidance image. Let assume that is a guidance image filter, is an input cryo-EM image, and is an output image. Both and are given beforehand and can be identical. The filtering output at a pixel is expressed as a weighted average as shown in Equation (9) (Kaiming He, 2013):

 (9)

where and are pixel indices. The filter kernel is a function of the guidance cryo-EM image and independent of (Kaiming He, 2013)as in Equation (10)*.*

 (10)

To implement guided filtering of image for the cryo-EM, we use a MATLAB function “*imguidedfilter*” (The MathWorks, 2018) that uses the input cryo-EM image under self-guidance. It uses the same cryo-EM image as the guidance image. This step performs edge-preserving smoothing of cryo-EM image in order to reduce the noise while keeping the particles edges (see supplemental Fig. S9 for an example).

### 2.1.7 Morphological Image Operation

Particles region detection and extraction in cryo-EM needs some logical operation to make particles regions (or background regions) similar to each other. Morphological image operation (dilation and erosion) are two fundamental morphological operations. Erosion is a tool to removes some pixel from object boundaries while dilation adds some pixels to the boundaries of objects in an image. We use an opening operation which is dilation followed by erosion with the same structuring element as shown in Equation (11) (The MathWorks, 2018).

 (11)

where is the original cryo-EM image and is the structure element. Supplemental Fig.10, (b) shows an example of cryo-EM image before and after the morphological image operation (image closing with a structural element ). It is the particles are brighter and more isolated after the operation.

**Adjustment of Intensity Levels**

## Particles Clustering Stage

Particles isolation is the most important step in our approach. In this step, a binary mask is constructed based on unsupervised learning clustering approaches. Two standard clustering algorithms K-means (MacQueen, 1967) and FCM) (J. C. Dunn, 1973) as well as our new intensity-based clustering (IBC) algorithm are applied. The description of K-means and FCM can be found in the supplemental document. We describe IBC in details as follows.

This clustering algorithm is based on an intensity distribution model, that relates the intensity difference value (instead of the distance metric), to signed difference intensity values, . Let be a set of images of the same modality containing the same anatomical structure of various subjects (i.e. particles in the CryoEM images), and be all (*L*) the pixels in an image that we want to group into a few consistent “clusters.” Here, is a real intensity value in a specific range, 1 <= *i* <= *L*. Let be the set of the average intensity values of *K* clusters. Let be the index of the cluster whose center (is closest to . The cluster assignment of all pixels is updated iteratively according to the average intensity (of clusters. The centers ( of *K* clusters are initialized as evenly distributed intervals in the intensity range at the equal step size according to Equation (15):

 (15)

where the is the difference between the maximum and the minimum intensity level in each image according.

The procedure of the clustering algorithm for cryo-EM image clustering is shown below.

|  |
| --- |
| **Algorithm: Intensity Based Clustering Algorithm** |
| **Set** number of clusters *K* as the number of adjusted intensity levels |
| **Transform** the 2-D image into 1-D vector of intensity values of all the pixel. *L* is the total number of pixels (L = height × width of image).  |
| **Calculate** the minimum and maximum values of intensity in the image  |
| **Set** the step size of interval as the range of intensity divided by *K*.  |
| **Initialize** cluster centers ( based on the step size |
| **Repeat** |
|  **for** i = 1 to *K* do |
| **for** n = 1 to *L* do |
| **Assign**  to Cluster *k* whose center ( is closest to according to the absolute intensity difference between the two. |
|  |
| **end** |
| **end** |
|  **for** k = 1 to K do |
| **Update** mean of each cluster by calculating the average intensity values of the pixels assigned to the cluster.  |
|  **end**  |
| **Until** there is no change in cluster centers.  |

Supplemental Fig. S13 shows an example of the final clustering result using the intensity-based clustering method. We can notice that the particles are most stably grouped in Cluster 1. Generally, the particles of the different images of the same protein are best identified in the same specific cluster according to our experiments.

# Particles Picking Stage

The las stage of the AutoCryoPicker framework has two main steps. The first step is the binary mask image cleaning and the second step is the particle object detection. In the first step, after the binary mask of the cryo-EM image is generated in the previous stage, some post-processing step such as 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. In the second step, since the regular shape of the particles of the in our test cryo-EM dataset is the common circle shape, a Circular Hough Transform (CHT) is used to detect particles from particle clusters. For another common particle shape in cryo-EM images - square, a square shape detector would be needed here.

### 2.3.1 Cryo-EM Cluster Image Cleaning

We first use a MATLAB function “*imfill*” to performs a flood-fill operation on background pixels of the input binary mask. Then, an image opening is applied to it according to Equation (19) (The MathWorks, 2018).

 (19)

Where is the original input image and is the structure element. This operation is carried out by calling MTALAB function “*imopen*”. Finally, the small objects in the binary mask image is removed by a MATLAB function” *bwareaopen*”. Supplemental Fig.14 (a) and (b) show a cluster image before and after image cleaning.

### 2.3.2 Cryo-EM Particles Detection and Picking

The second step of the particles piking is particle object detection. We measure the properties of the particles rejoins in each cryo-EM image using a MATLAB function “*regionprops*” (The MathWorks, 2018). Supplemental Fig. S15 (c), (f) and (i) show the examples of the particle object counting based ICB, k-means, and FCM, respectively. Although, we extract the boundary box dimensions of each particle object in each single cryo-EM image and shown the detection results. Supplemental Fig. S15, (d), (g), and (j) illustrate the particle detection and picking results of ICB, k-means, and FCM, respectively.

# 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.

## Performance Evaluation

For the preprocessing stage, we use common image preprocessing criteria such as peak signal-to-noise ratio (PSNR), signal-to-noise ratio (SNR), mean squared error (MSE) (Woods, 2018) to evaluate the improvement of the quality of cryo-EM images. The PSNR, SNR, and MSE are defined by equations (20), (21) and (22), respectively (Woods, 2018).

 (20)

where,  is the maximum possible pixel value of the image, and MSE is the mean square error value.

 (21)

 (22)

For the image clustering stage, we use clustering accuracy and misclassification rate which are defined by equations (23) and (24), respectively. Each evaluation metric bases on the representation of the confusion matrix such as the True Positive (TP) which refers to correct detection of positive cases. True Negative (TN) which refers to correct detection of negative cases. False Positive (FP) which refers to incorrect detection of positive cases into negative class. Finally, False Negative (FN) which refers to incorrect detection of negative cases into class positive (Langlois, 2011).

 (23)

 (24)

where TP denotes the number of correctly detected positive cases (real particles), TN the number of correctly detected negative cases, FP the number incorrectly detected positive cases, FN the number of incorrectly detected negative cases. Moreover, Dice Criteria (DIC) is also used for the similarity measure between the cluster image and the Ground Truth (GT). DC is defined by Equation (25) (Rohlfing, 2012):

 (25)

where, is the cluster image and is the ground truth image of . Finally, we use precision, recall, and F1 measure scores (Langlois, 2011) to evaluate the particle picking results in the particle picking stage. The precision, recall, and F measure are defined by Equations (26), (27) and (28), respectively (Langlois, 2011).

 (26)

 (27)

 (28)

## Cryo-EM Datasets and Pre-processing Results

The cryo-EM image dataset used to test AutoCryoPicker is the Apoferritin molecular mass dataset, which is acquired from the cryo-EM Databank (EMPIAR-10146) (Grant T, 2018). The dataset is a multi-frame cryo-EM image dataset that has 20 cryo-EM images in MRC format. Each image has 50 frames, and the image dimension (cryo-EM resolution) is 1240×1200. Each pixel is a 32-bit float number. Beam energy of 300 keV, spherical aberration of 0.0, and the exposure rate of 2 electrons/A^2/frame were used in the cryo-EM experiment to generate the image dataset. Apoferritin molecular mass is 440 kDa and its symmetry is O.

Supplemental Fig. S16 shows the average PSNR, SNR and MSE scores of the cryo-EM images before and after the preprocessing steps. The average PSNR score is increased from 65.76 to 66.25 and the average SNR score from 6.58 to 6.94. The average MSE is reduced from ? to ?. Supplemental Figure S17 illustrates the histogram of PSN, SNR, and MSE scores before and after image preprocessing. The range of PSNR scores has been increased from [66.24-66.25] for the original Cryo-EM images to [69.4-70.8] for the preprocessed ones. The range of the SNR scores is increased from [26-28] for the original images to [30-32] for the preprocessed ones. The range of MSE scores is decreased from [0.994-0.996] to [0.36-0.46] after the preprocessing steps. According to Student’s t test, the p-values of the changes of PSNR, SNR and MSE scores caused by the preprocessing are 5.8e-15, 7.6e-15, and 1.0e-14, respectively, indicating that the preprocessing steps significantly improve the quality of cryo-EM images.

## Particle Clustering and Picking Results

In order to evaluate the performance of automated particle clustering and picking, we randomly select two cryo-EM images from the dataset and manually label the coordinates of the particles in the images, which were as true reference particles in the evaluation. Supplemental Fig. S18 shows the two cryo-EM images that are used to test our method. A typical example of the fully automated picking results is shown in Fig.19 in supplementary data, which demonstrated that most of the particles were correctly picked by AutoCryoPicker. Table. 1 reports the recall, precision, accuracy, F1 score, and running time of AutoCryoPicker based on three clustering algorithms: K-means, FCM, and IBC.

**Table 2.**Fully Automated Particles Picking Performance Results

|  |  |  |  |
| --- | --- | --- | --- |
| Measures  | K-Means | FCM | IBC |
| Average Recall (%) | 77.90 | 73.60 | 99.11 |
| Average Precision (%) | 98.81 | 97.99 | 99.10 |
| Average Accuracy (%) | 77.22 | 72.81 | 98.50 |
| Average F1 Score (%) | 76.78 | 74.97 | 99.24 |
| Average Time (Sec.) | 41.79 | 112.73 | 6.87 |

AutoCryoPicker based on IBC has recall of 99.11%, precision of 99.10%, accuracy of 98.50%, and F1 score of 99.24, which are much higher than its performance based on K-means and FCM. IBC-based particle picking is also much faster than K-means and FCM-based methods. Supplemental Fig. S20 shows the comparison between the particles picked manually by human experts and the AutoCryoPicker based on the three different algorithms: (a) IBC, (b) k-means, and (c) FCM

# Conclusion

The difficulty of the fully automated picking particles from a micrograph is generally associated with the contrast level of the image, which is generally related to its defocus level. Accurate particle picking still requires substantial human intervention and therefore is labor intensive and time consuming. To address the challenge, we develop AutoCryoPicker – the first fully automated particle picking approach based on image preprocessing, unsupervised clustering and shape detection. Our experiment shows that the approach can improve signal to noise ratio in cryo-EM images and rather accurately pick particles. Therefore, the method can free scientists from laborious work on particle picking in cryo-EM images, and substantially improve the precision, recall, consistency and efficiency of particle picking, which can improve the quality and effectiveness of cryo-EM based protein structure determination.

Funding

The work is partially supported by an NIH grant (R01GM093123) to JC and JT.

# References

A. Biswas, D. R. (837–843). A dynamic programming algorithm for finding the optimal placement of a secondary structure topology in cryo-em data. *Journal of Computational Biology*, 2015.

A.K. Jain, M. M. (1999). Data clustering, a review. *ACM Computing Surveys*, 265–323.

Abdi, H. (2010). Normalizing Data. In H. Abdi. The University of Texas at Dallas: In Neil Salkind (Ed.), Encyclopedia of Research Design.

C., A. V. (2010). Chapter 11 Wiener Filtering. In A. V. C., *ntroduction to Communication, Control, and Signal Processing* (pp. 195-210). Verghese: Spring.

C.O.S. Sorzano, E. R.-C.-M. (2009). Automatic particle selection from electron micrographs using machine learning techniques. *Journal of Structural Biology*, 252 – 260.

Doerr, A. (2016). Single-particle cryo-electron microscopy. *Nature Methods*, 23. Retrieved from https://www.nature.com/articles/nmeth.3700

Feng Wang, H. G. (2016). DeepPicker: a Deep Learning Approach for Fully Automated Particle Picking in Cryo-EM. *Journal of Structural Biology*, 325-336.

Glaeser., W. V. (2001). Review: Automatic particle detection in electron. *Journal of Structural Biology*, 90 – 101.

Grant T, R. A. (2018, 03 09). *EMPIAR*. Retrieved from EMPIAR-10146: Apoferritin tutorial dataset for cisTEM: https://www.ebi.ac.uk/pdbe/emdb/empiar/entry/10146/#&gid=1&pid=1

Grigorieff, J. Z. (2007). SIGNATURE: A single-particle selection system for molecular electron microscopy. *Journal of Structural Biology*, 168 – 173.

He, Q. (1999). *A review of clustering algorithms as applied to IR.* University of Illinois at Urbana-Champaign: Technical Report UIUCLIS.

Henderson, T. S. (2012). Electron cryomicroscopy of biological macromolecules. In T. S. Henderson, *Chapter 19.6* (pp. 593–614). International Tables for Crystallography. Retrieved from What is Cryo-EM?: https://www.fei.com

J. C. Dunn. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybern*, 32-57.

Jianhua Zhao, M. A. (2013). TMaCS: A hybrid template matching and classification system for partially-automated particle selection. *Journal of Structural Biology*, 234 – 242.

K. Al Nasr, D. R. (2011). Ranking valid topologies of the secondary structure elements using a constraint graph. *Journal of Bioinformatics and Computational Biology*, 415–430.

K. Al Nasr, D. R. (2014). Solving the secondary structure matching problem in cryo-em de novo modeling using a constrained-shortest path graph algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 419–430.

K. Al Nasr, L. C. (2012.). Building the initial chain of the proteins through de novo modeling of the cryo-electron microscopy volume data at the medium resolutions. *In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 490–497.

Kaiming He, J. S. (2013). Guided Image Filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Langlois, R. e. (2011). A clarification of the terms used in comparing semi-automated particle selection algorithms in Cryo-EM. *J. Struct. Biol*, 348-352.

M. L. Baker, S. S. (2011). Modeling protein structure at near atomic resolutions with gorgon. *Journal of structural biology*, 360–373.

M. L. Baker, T. J. (2007). Identification of secondary structure elements in intermediate-resolution density maps. *Structure*, 7–19.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *in Proc. 5th Berkeley Symp. on Math. Stat. and Probability*, (pp. 281-297). Berkeley, CA.

Mankiewicz, R. (2004). *The Story of Mathematics.* Princeton, NJ: Princeton University Press.

MathWorks. (2018). *Statistics and Machine Learning.* Natick, MA: The MathWorks, Inc.

N.R. Voss, C. Y. (2009). DoG Picker and TiltPicker: Software tools to facilitate particle selection in single particle electron microscopy. *Journal of Structural Biology*, 205 – 213.

P.S Umesh Adiga, R. M. (2004). A binary segmentation approach for boxing ribosome particles in cryo EM micrographs. *Journal of Structural*, 142 – 151.

Pablo Arbelez, B.-G. H. (2011). Experimental evaluation of support vector machine-based and correlation-based approaches to automatic particle selection. *Journal of Structural Biology*, 319 – 328.

Patwardhan, R. J. (2004). A two step approach for semi-automated particle selection from low contrast cryo-electron micrographs. *Journal of Structural Biology*, 19 – 28.

Penczek, Z. H. (2004). Application of template matching technique to particle detection in electron micrographs. *Journal of Structural Biology*, 29 – 40.

Ramin Norousi, S. W. (2013). Automatic post-picking using MAPPOS improves particle image detection from cryo-EM micrographs. *Journal of Structural Biology*, 59–66.

Rebecca F.Thompson, M. W. (2016). An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology. *Methods*, 3-15.

Robert Langlois, J. P. (2011). Reference-free particle selection enhanced hanced with semi-supervised machine learning for cryo-electron microscopy. *Journal of Structural Biology*, 353 – 361.

Rohlfing, T. (2012). Image Similarity and Tissue Overlaps as Surrogates for Image Registration Accuracy: Widely Used but Unreliable. *IEEE Trans Med Imaging*, 153–163.

Rongjian Li, D. S. (2016). Deep Convolutional Neural Networks for Detecting Secondary Structures in Protein Density Maps from Cryo-Electron Microscopy. *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 41-46). IEEE.

Stark, J. A. (2000 ). Adaptive Image Contrast Enhancement Using Generalizations of Histogram Equalization. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 889-869.

The Eva Nogales Lab, H. o. (2018). *The Nogales Lab*. Retrieved from http://cryoem.berkeley.edu/cryoem

The MathWorks, I. (2018). *Image Processing Toolbox™ User's Guide.* Natick, MA: The MathWorks, Inc. Retrieved from MathWorks Accelerating the pace of engineering and science: https://www.mathworks.com/help/images/contrast-adjustment.html

Woods, R. C. (2018). *Digital Image Processi \_4, 4th Edition.* University of Tennessee.

Yanan Zhu, Q. O. (2017). A deep convolutional neural network approach to single-particle recognition in cryo-electron microscopy. *BMC Bioinformatics*, 2-10.

Yuanxin Zhu, B. C. (2004). Automatic particle selection: results of a comparative study. *Journal of Structural Biology*, 3 – 14.

Z. Liu, F. G.-C. (2016). a resolution cryo-em 3d reconstruction of close-packed virus particles. *Structure*, 319–328.