DeepCryoPicker

An unsupervised learning approach for fully automated single particle picking in Cryo-EM images

Zachary M Lipperd

Dept. Electrical Engineering and Computer Science

University of Missouri, Columbia, Mo

zmlmcb@umsystem.edu

Abstract

Cryogenic electron microscopy (cryo-EM) has revolutionized the determination of protein structures. However, the computational reconstruction of these protein structures remains a challenging task, hindered by the diversity of particle shapes and the extremely low signal-to-noise ratio of micrographs.. Human intervention is often required to create a high-quality set of particles, a process which is time-consuming, error-prone, and requires extensive manual annotation.

We propose DeepCryoPicker, a fully automated, unsupervised approach for single particle picking in cryo-EM micrographs. Our solution involves three stages: image preprocessing, particle clustering, and particle picking. Image preprocessing techniques, including image averaging, normalization, contrast enhancement correction, histogram equalization, restoration, adaptive histogram equalization, guided image filtering, and morphological operations. These steps have shown to significantly improve the cryo-EM image quality, allowing for more accurate clustering and picking. The particle clustering stage leverages an intensity distribution model, which has outperformed traditional methods like K-means and Fuzzy C-Means in speed and accuracy for our task. Particle picking employs image cleaning and a region-based method that identifies connected regions in the processed image and encapsulates the particles within bounding boxes. This approach, aided by earlier preprocessing steps, effectively distinguishes between protein particles and background noise, providing a more precise selection of particles.

DeepCryoPicker will autonomously identify protein particles in cryo-EM micrographs, eliminating the need for human intervention or labeled training data. This breakthrough has the potential to greatly expedite and improve the accuracy of cryo-EM protein structure determination.

Introduction

Background

For decades, X-ray crystallography has been the primary method for obtaining high-resolution structures of macromolecules. However, single-particle cryo-electron microscopy (cryo-EM), once used for providing low-resolution structural information on large protein complexes, has undergone significant advancements. With improvements in sample preparation, computation, and instrumentation, cryo-EM has revolutionized structural biology, enabling the resolution of large protein structures.

Despite these advancements, cryo-EM micrographs, which contain two-dimensional projections of particles in different orientations, pose several challenges. The images often have low contrast due to the similarity in electron density between the protein and the surrounding solution. The limited electron dose used in data collection, as well as the presence of ice sections, deformed particles, protein aggregates, among other artifacts, further complicate the micrographs. This complexity, coupled with the need for a large number of single-particle images for a reliable 3D reconstruction, makes particle picking a significant bottleneck in cryo-EM structure determination.

Numerous computational approaches have been proposed to facilitate the particle picking process. These can be divided broadly into generative and discriminative classification methods. Generative methods measure the similarity of an image region to a reference, while discriminative methods train a classifier on a labeled dataset before applying it to detect particle images from micrographs. Both methods, however, have their limitations. Generative methods are dependent on high-quality references, and discriminative methods require a large labeled dataset, which necessitates extensive manual annotation.

Our project, DeepCryoPicker, aims to overcome these challenges by designing a fully automated, unsupervised approach for particle picking. The approach involves image preprocessing, particle clustering, and particle picking. Image preprocessing includes a series of techniques designed to enhance the quality of cryo-EM images, thereby improving subsequent clustering and picking. Particle clustering is based on an intensity distribution model, which is faster and more accurate than traditional K-means and Fuzzy C-Means algorithms. Finally, particle picking leverages image cleaning and region selection to effectively detect particles' shape and center and encapsulate them within bounding boxes.

This paper will outline our methods, present our results, and discuss the implications of our findings in the context of cryo-EM protein structure determination. The proposed approach has the potential to substantially improve the efficiency and accuracy of the particle picking process.

Related work

Several computational approaches have been proposed to facilitate the particle picking process in cryo-EM images, which can generally be divided into generative methods and discriminative classification methods. Generative methods identify particle candidates from micrographs by measuring the similarity of an image region to a reference. A typical technique used in this approach is template-matching with a cross-correlation similarity measure. On the other hand, discriminative methods involve training a classifier on a labeled dataset of positive and negative particle examples, which is then used to detect particle images from micrographs.

Deep learning methods such as DeepPicker and DeepEM have recently been introduced for semi-automated particle selection and picking. These methods involve a combination of manual creation of training data and automated learning of patterns from the training data to classify particles. While these methods have been useful in reducing the time and effort spent on single-particle data analysis, they are not fully automated, and they still require a significant level of human intervention.

Existing unsupervised approaches aim to distinguish particle-like objects from background noise in an unsupervised learning manner. However, these approaches do not fully leverage the intrinsic characteristics of particles for automated particle picking. While these methods have been combined with reference template matching or classification-based approaches for better picking results, they often require a manually curated training dataset, which is time-consuming and labor-intensive.

Furthermore, previous methods have struggled with issues such as insufficient and undiversified training datasets, high false-positive particle detection rates due to the sliding window technique, and challenges associated with dealing with various low-SNR micrographs.

In contrast, our approach with DeepCryoPicker aims to address these limitations by proposing a fully automated method for particle picking. It incorporates advanced image preprocessing, robust clustering via the intensity distribution model, and sophisticated region selection. This approach aims to significantly reduce the time and effort spent on single-particle data analysis, thereby addressing a significant bottleneck in cryo-EM structure determination.

Methods

Our DeepCryoPicker framework for automated particle picking is shown in Fig. 1. This framework is fully automated and does not require the user to manually pick any particles from the micrographs. The approach has three main stages: preprocessing, clustering, and particle picking.

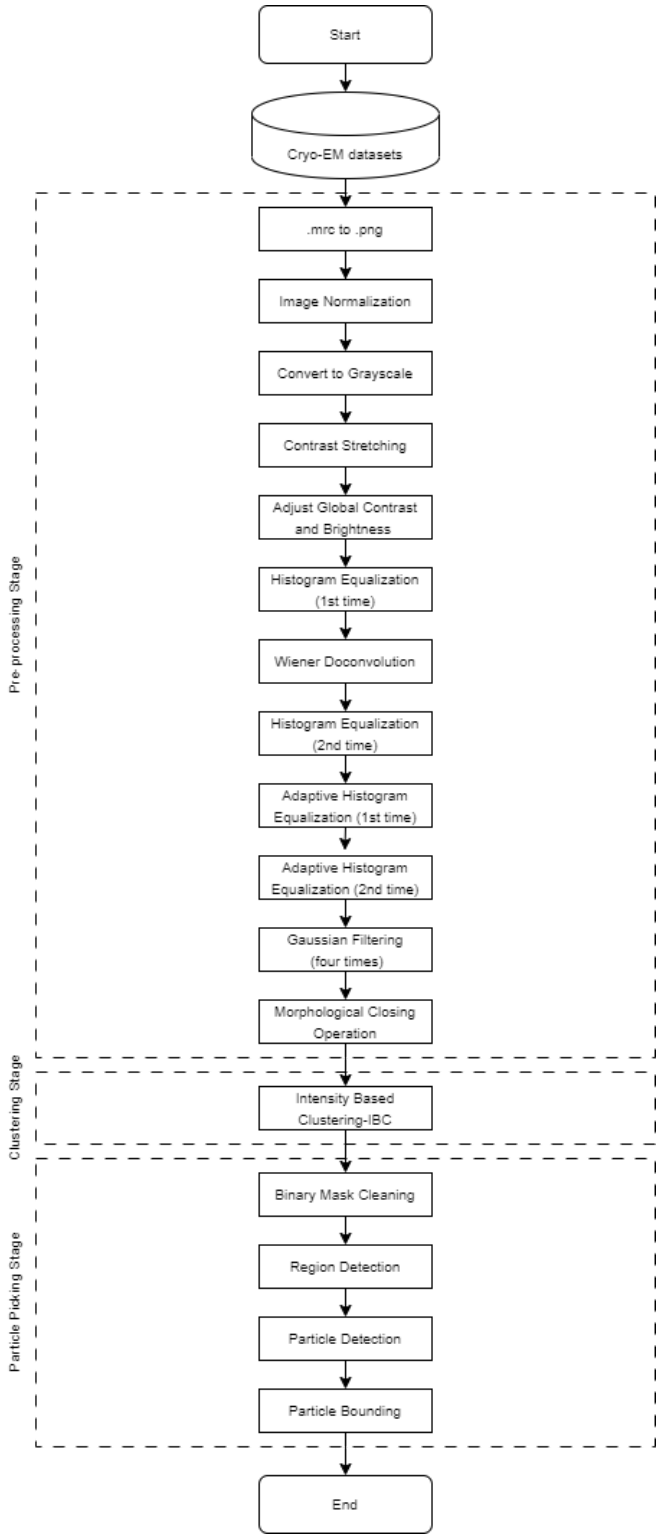
In the preprocessing stage, several image processing methods are applied to enhance the input cryo-EM images. These methods include image normalization, contrast enhancement correction (CEC), histogram equalization, restoration, adaptive histogram equalization, and guided image filtering. We also utilize morphological operations to further improve the quality of the cryo-EM image.

For the clustering stage, we employ a method based on an intensity distribution model. This method has been shown to be faster and more accurate than traditional clustering algorithms such as k-means and Fuzzy C-Means (FCM) for particle clustering, and it avoids issues such as cluster destabilization due to random initialization of cluster centers.

In the particle picking stage, particles are identified from the clustered particle candidates. Rather than using a traditional shape detection algorithm, we use a region detection algorithm that we designed. This algorithm efficiently identifies the location and shape of each particle and creates a bounding box to encapsulate the particles.

These stages combined provide a comprehensive and efficient approach to automatically identify protein particles in cryo-EM micrographs, thus greatly improving the process of protein structure determination.

Fig 1. General Framework for DeepCryoPicker. The dashed boxes represent the three stages of the approach: pre-processing, particle clustering, and particle detection. A solid box denotes an analysis step



Datasets

In our study, we have considered four distinct protein shapes in micrographs, sourced from a variety of micrograph datasets, as illustrated in Fig. 2. The first protein shape is circular, as exemplified by the apoferritin (Fig. 2a, e, f). The second protein shape is square, as seen in the side-view of KLH (Fig. 2b, g). There are two main types of projection views in this dataset. The top view is circular, while the side view is square. The third protein shape that is considered is the general case of an irregularly shaped protein such as the β-galactosidase (Fig. 2c, h) and 80S ribosome (Fig. 2d, i).

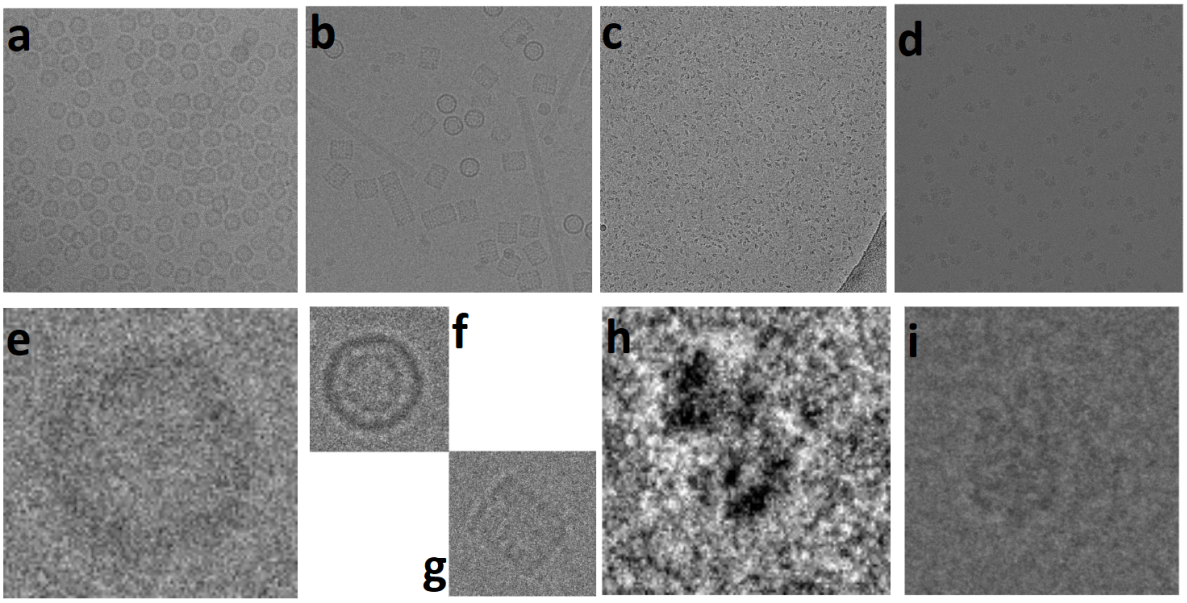


Fig 2. Micrograph datasets used for this project. **a.** Apoferritin micrograph. **b.** KHL micrograph. **c.** β-galactosidase micrograph. **d.** 80S ribosome micrograph. **e.** Picked particle from an Apoferritin micrograph **f.** Picked particle from an KLH micrograph (top view) **g.** Picked particle from an KLH micrograph (side view) **h.** Picked particle from an β-galactosidase micrograph **j.** 80S ribosome

Stage 1: pre-processing

The preprocessing stage serves two key purposes. Primarily, it bolsters the contrast of the cryo-EM images by amplifying the intensity of the particles. In addition, it clusters pixels within each particle, simplifying their isolation by the clustering algorithm. The selection of preprocessing tools is grounded in three key goals: enhancing the overall contrast of the cryo-EM, amplifying the local contrast and the intensity level of each particle, and refining the particle shapes within the cryo-EM images.

To elevate the overall contrast between particles and their background, we initially employ image normalization followed by contrast enhancement and correction to boost the global intensity value. To amplify the global image contrast, we apply histogram equalization to enhance pixel intensity levels, and subsequently use image restoration to restore and improve the image quality.

To enhance local contrast and refine the definition of edges in each particle, we utilize adaptive histogram equalization. Moreover, we employ guided image filtering for edge-preserving smoothing of each particle in the cryo-EM image. Lastly, morphological image operations are performed to enhance particle shape and ensure particle regions bear similarity to each other and exhibit distinctness from the background regions. Detailed descriptions of these preprocessing methods are provided in the subsequent steps.

Step 1: Mixed Raster Content (MRC) to Portable Network Graphic (PNG) conversion

A typical cryo-EM image is saved in the Mixed Raster Content (MRC) format. This format establishes a three-dimensional grid or array comprised of voxels, each assigned a value that corresponds to the electron density or electric potential. To enhance the quality of these cryo-EM images, which often contain a significant amount of noise, we employ several image preprocessing techniques. As a part of this process, we transform the cryo-EM images from the MRC format into the more universally accepted 16-bit PNG format.

Step 2: Image Normalization

The initial step in our preprocessing pipeline is image normalization. This technique is a fundamental and crucial step in image processing, as it brings the pixel intensities of the image into a specific range. In our case, the normalization process transforms the pixel intensities so that they fall within the range [0,1].

Normalization is performed using the formula:

```
normalized_data = image_data/np.max(image_data)
```

This means that each pixel intensity in the image is divided by the maximum pixel intensity value found in the image. This operation scales all pixel values between 0 (representing black) and 1 (representing white).

The primary reason for performing image normalization is to standardize the dynamic range of pixel intensities across different images. This is particularly important when dealing with cryo-EM images, which can come from different sources and could therefore have different intensity ranges.

Cryo-EM images often have their signal corrupted by various factors, including Gaussian noise and issues with image resolution. Furthermore, they often contain artificial objects, such as ice, which can sometimes exhibit similar ranges of pixel intensity values as the particles themselves due to variations in thickness. This can lead to situations where a subset of particles within a single cryo-EM image might not exhibit significant differences in scatter power.

By normalizing the images, we ensure that subsequent image processing steps, such as contrast enhancement or histogram equalization, work consistently across all images, regardless of their original intensity ranges. Furthermore, image normalization improves the contrast of the image, which allows for better distinction between particles and the background. This improves the effectiveness of the subsequent clustering algorithm, as it is easier to distinguish and isolate individual particles. The normalization process thus plays a crucial role in enhancing the quality of the cryo-EM images and setting a solid foundation for the subsequent preprocessing steps.

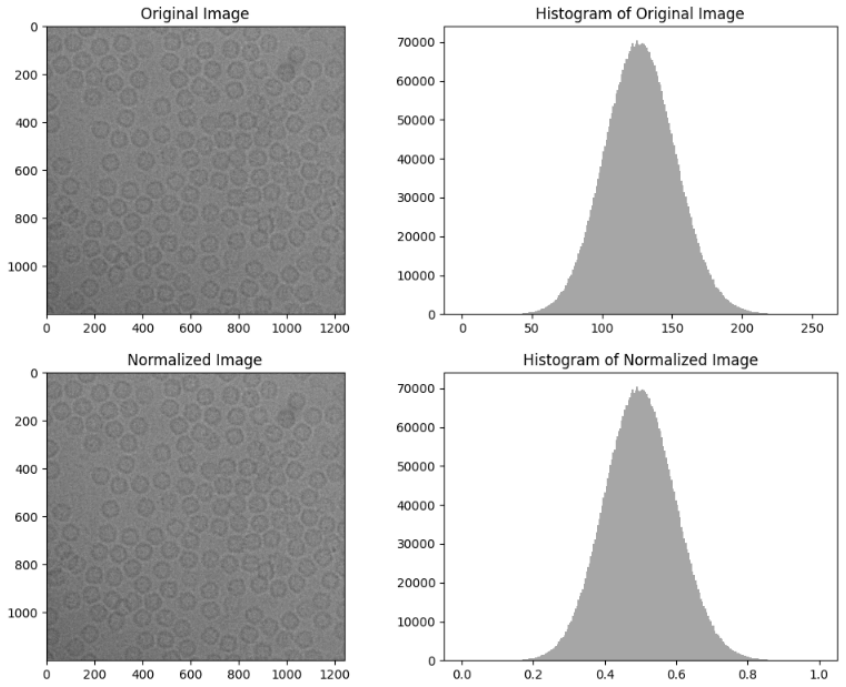


Fig 3. Image Normalization and Its Impact on Pixel Intensity Distribution. The left panel shows the original cryo-EM image, and the corresponding histogram is displayed in the second panel, indicating the distribution of pixel intensities. The third panel illustrates the effect of image normalization on the original image, highlighting the scaling of pixel values to the range [0,1]. The final panel presents the histogram after normalization, revealing a modified pixel intensity distribution that enhances the contrast and visibility of particles in the cryo-EM image.

Step 2: Image to Grayscale

Cryo-EM images inherently carry information in grayscale, where pixel intensities reflect electron densities rather than color information. Thus, the transformation of the image to grayscale simplifies the subsequent processing steps by reducing the data complexity without losing significant information.

This conversion is achieved by transforming the three-color channels (Red, Green, and Blue) of the normalized image into a single channel representing different shades of gray. The gray value of each pixel is calculated as a weighted sum of the corresponding Red, Green, and Blue values. This step results in a single-layered image that still retains the crucial intensity variations essential for identifying particles.

By moving to grayscale, we reduce the computational load for the subsequent steps, allowing us to focus on the significant features of the images, such as contrast and particle shapes. It also sets the stage for further contrast adjustments, restoration techniques, and edge detection, which are more effectively applied to single-channel grayscale images.

Step 3: Contrast Stretching

The third stage in the preprocessing of cryo-EM images is contrast stretching, a crucial step in enhancing the visibility of the features within the images. Contrast stretching, as the name suggests, involves rescaling the intensity range of the image, which ultimately improves the contrast and allows for better visualization of the particles present in the image.

The specific method of contrast stretching we implemented in this study is based on the percentile values of pixel intensities in the grayscale image. We opted to rescale the intensity values of our image such that the 5th percentile and the 95th percentile of pixel intensities mapped to the full intensity range of the image. Essentially, this method

trims the extreme 5% of darkest and brightest pixels and expands the remaining range, which includes most of the useful information, to cover the full grayscale spectrum.

This specific approach aids in reducing the influence of outliers in the data (i.e., extremely bright or dark spots), which could be due to noise or other artifacts. By stretching the contrast in this manner, we ensure that the significant features of the image are well distributed across the entire grayscale range, thus increasing the contrast and enhancing the visibility of particles against the background. This improved contrast is especially beneficial in the subsequent stages of particle picking and image analysis.

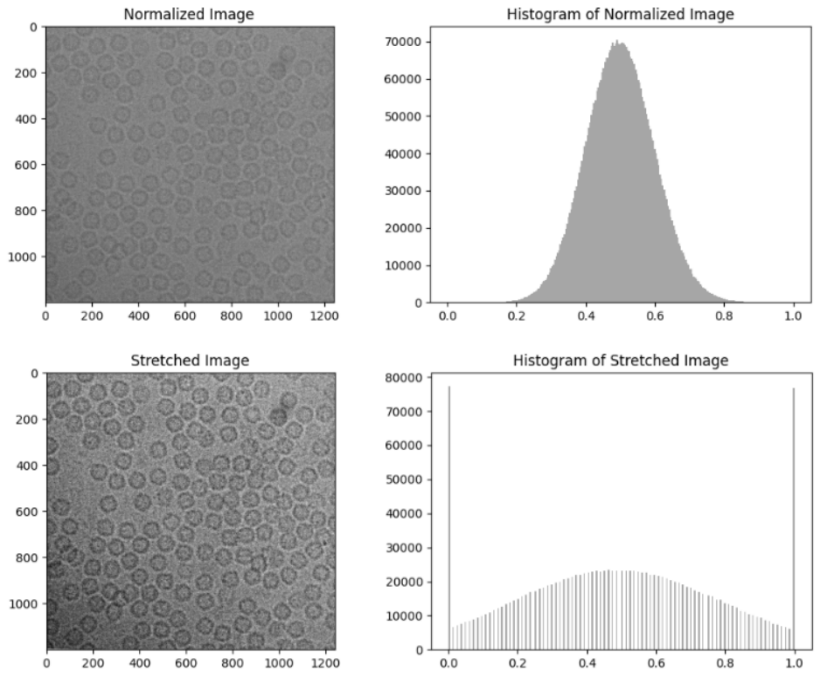


Fig 4. Contrast Stretching and Its Impact on Pixel Intensity Distribution. The left panel presents the normalized cryo-EM image, with the corresponding histogram displayed in the second panel, indicating the distribution of pixel intensities after normalization. The third panel demonstrates the effect of contrast stretching on the normalized image, emphasizing the redistribution of pixel values based on the 5th and 95th percentile of the pixel intensity. The final panel showcases the histogram after contrast stretching, revealing a further modified pixel intensity distribution that increases the contrast and enhances the visibility of particles in the cryo-EM image

Step 4: Global contrast and brightness adjustment

The nexstep in our preprocessing involves adjusting the global contrast and brightness of the cryo-EM image. This operation is performed using a gamma correction technique. Gamma correction is a nonlinear operation often used to control the brightness and contrast of an image. In the context of our application, the aim is to amplify the visibility of particles and reduce the prominence of background noise.

In specific terms, the process involves applying a gamma correction of 0.5 to the stretched image data. This value is chosen because it is empirically observed to improve the visibility of particles in our cryo-EM images. This is mathematically expressed as $I' = I^{\gamma}$, where I is the original pixel intensity, I' is the adjusted intensity, and γ is the gamma value. By using a gamma value of 0.5, we effectively increase the contrast of darker regions (those with lower pixel intensity) without overly brightening the already bright regions. This adjustment can help to further separate particles from the background, thereby facilitating subsequent processing and analysis steps.

It should be noted that while this approach can enhance the visibility of particles, it may also amplify certain image artifacts or increase the prominence of noise in some cases. However, subsequent steps in our preprocessing pipeline are designed to mitigate these potential issues.

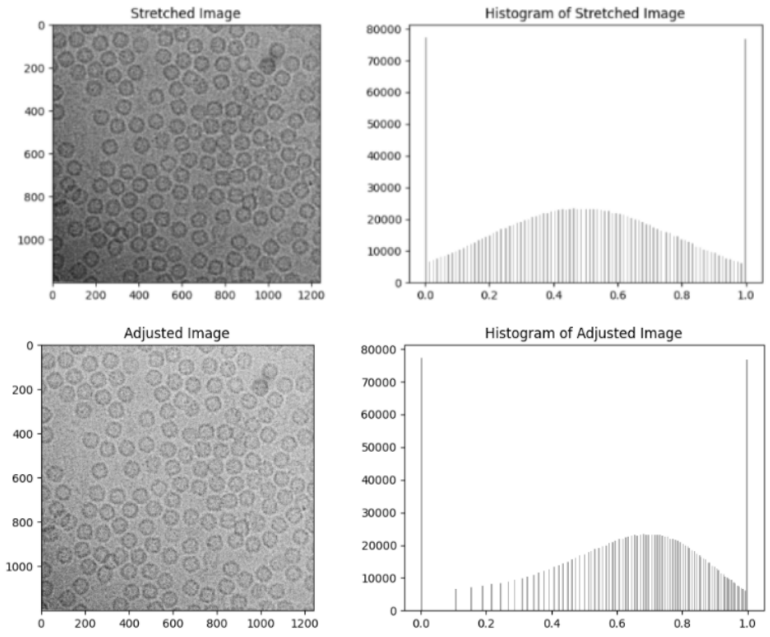


Fig 5. Adjusting Global Contrast and Brightness and Its Impact on Pixel Intensity Distribution. The left panel displays the cryo-EM image post contrast stretching, and its corresponding histogram is shown in the second panel, reflecting the distribution of pixel intensities. The third panel demonstrates the effect of gamma correction on the image, accentuating the visibility of particles by adjusting the global contrast and brightness. The final panel presents the histogram post gamma correction, revealing a transformed pixel intensity distribution that further enhances the contrast between particles and the background in the cryo-EM image.

Step 5: Histogram Equalization (1st time)

The fifth step in our preprocessing workflow is Histogram Equalization, a technique that aims to adjust the pixel intensity distribution of the cryo-EM image in order to enhance the contrast. Histogram Equalization operates by flattening the pixel intensity distribution, effectively spreading out the most frequent intensity values. The goal of this operation is to maximize the image's contrast and improve the visibility of details that were previously obscured due to poor contrast.

In our specific implementation, we employed the exposure.equalize_hist function from the skimage library. This function applies a non-linear stretching to the intensity values in the image, redistributing the pixel intensities in the image so that all intensity values are equally likely to occur, resulting in a uniform histogram. This transformation allows for better visualization and processing of the cryo-EM images, particularly in the context of particle picking where distinguishing between particles and background is critical.

This step is particularly beneficial in scenarios where the contrast within an image is poor due to the limitations in the dynamic range of the recording system, or when a substantial portion of the grayscale is not utilized, which is a common occurrence in cryo-EM images. By redistributing the pixel intensities, we are able to utilize the entire grayscale range, improving the visibility and distinguishability of the particles in the image.

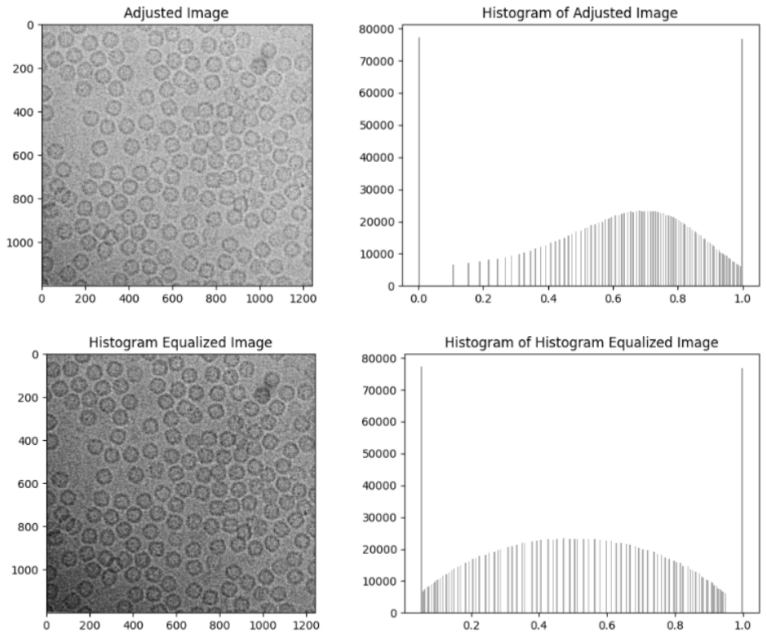


Fig 6. Application of Histogram Equalization and Its Impact on Pixel Intensity Distribution. The left panel shows the cryo-EM image after global contrast and brightness adjustment, with its corresponding histogram displayed in the second panel. The third panel reveals the effect of the first application of histogram equalization on the image, demonstrating the redistribution of pixel intensities to maximize contrast. The final panel presents the histogram post-equalization, illustrating the flattened, uniform distribution of pixel intensities, effectively utilizing the entire grayscale range to enhance image contrast and visibility of particles in the cryo-EM image.

Step 6: Weiner deconvolution

Wiener Deconvolution constitutes the sixth step of our image preprocessing pipeline. This technique seeks to deblur the image by reversing the effects of a known blurring function, in our case, approximated by a Point Spread Function (PSF). The deconvolution process is vital for our preprocessing as it helps in recovering the original image details that may have been blurred or distorted during the image acquisition process. In our implementation, we used a simple 3x3 PSF where all elements are equal, i.e., an isotropic blur. However, this is a generalized approximation, and for better results, a more accurate PSF could be estimated based on the specifics of our dataset, marking an area for potential improvement.

After defining the PSF, we applied a convolution operation to our image data that had undergone histogram equalization. The convolution operation, performed with 'same' mode and 'symm' boundary conditions, resulted in a blurred version of the image data. The Wiener deconvolution was then carried out on the blurred image data using the same PSF and a balance parameter of 0.1. The balance parameter is used to control the trade-off between inverse filtering and noise smoothing, a critical aspect of Wiener deconvolution.

The outcome of the Wiener deconvolution is an image where the blurring effects have been mitigated, leading to a sharper representation of the protein structures in the cryo-EM image. The details within each protein structure become more distinguishable, which facilitates the subsequent steps of image preprocessing and particle identification.

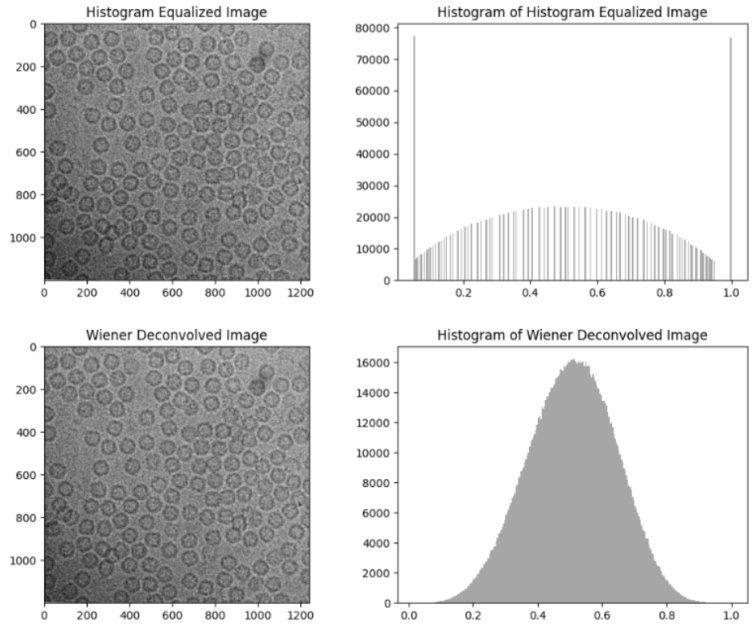


Fig 7. Impact of Wiener Deconvolution on Image Quality and Pixel Intensity Distribution. The first panel presents the cryo-EM image post histogram equalization, and the corresponding histogram is shown in the second panel, illustrating the distribution of pixel intensities. The third panel displays the image after Wiener deconvolution, revealing a sharper and more defined representation of the protein structures. The histogram in the last panel provides the distribution of pixel intensities post Wiener deconvolution, indicating a redistribution of intensities that aids in improving image details and contrast

Step 7: Histogram Equalization (2nd time)

The seventh step in the preprocessing is a second round of histogram equalization. This step is critical to further enhance the global contrast of the image, allowing better visibility of the protein structures. Given that cryo-EM images often contain non-uniform illumination and noise, this second round of histogram equalization can further improve the contrast by redistributing pixel intensities across the image, particularly those that have been altered during the Wiener deconvolution process.

In this specific application, we use the function equalize_hist from the exposure module in the skimage library. This function works by spreading out the most frequent intensity values in the image, i.e., it flattens the intensity distribution of the image, enhancing the contrast globally.

This step is particularly important as it helps to ensure that the micrograph images have uniform contrast and brightness, which is necessary for the subsequent image processing steps, particularly for the application of machine learning algorithms. By ensuring a uniform distribution of pixel intensities, this step aids in reducing bias in the feature extraction stage of the processing pipeline.

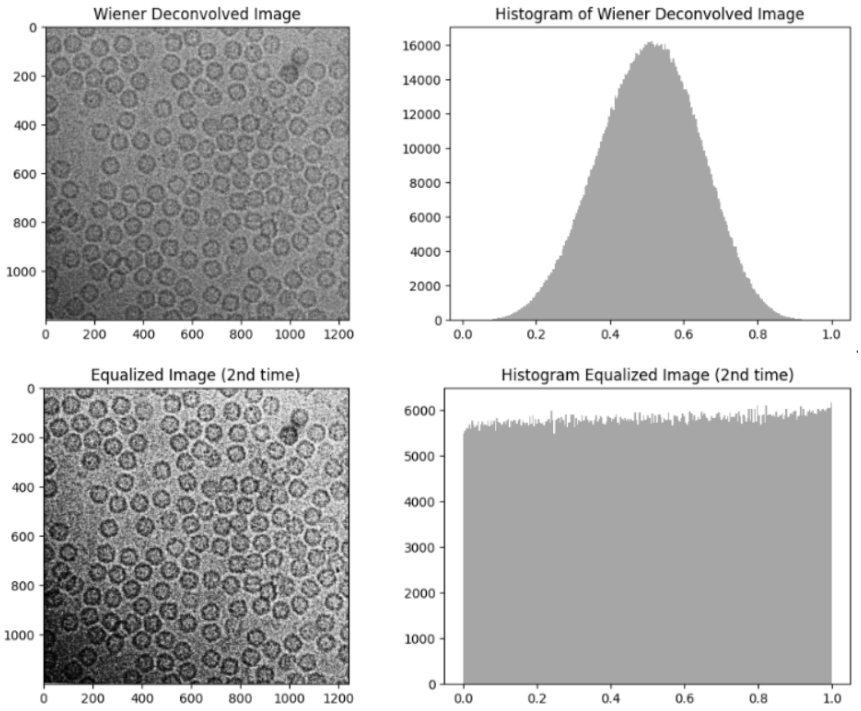


Fig 8. Second Round of Histogram Equalization and Its Impact on Pixel Intensity Distribution. The first panel displays the cryo-EM image after Weiner deconvolution, with the corresponding histogram in the second panel showing the pixel intensity distribution at this stage. The third panel showcases the image after the second round of histogram equalization, indicating further enhancement in the global contrast. The final panel presents the histogram after the second histogram equalization, demonstrating a further spread in the pixel intensity distribution.

Step 8 and 9: Adaptive Histogram Equalization

Adaptive Histogram Equalization, also known as Contrast Limited Adaptive Histogram Equalization (CLAHE), is an advanced form of histogram equalization that performs the contrast enhancement operation locally. This process can be more effective than global techniques as it adapts to local changes in contrast, which can be particularly beneficial in images where the lighting conditions vary across different regions.

Initially, the CLAHE method is applied with a low clip limit of 0.02. The clip limit refers to the threshold for contrast amplification. By setting a low value, the algorithm limits the amount of contrast enhancement applied to the image, thereby preventing excessive enhancement of noise. The first round of CLAHE serves to correct for small-scale contrast variations without significantly impacting the global contrast of the image.

Following the initial round of CLAHE, the process is repeated with a higher clip limit of 0.99. This second round is intended to further enhance the contrast in regions of the image that were not sufficiently improved during the first application. By using a higher clip limit, the algorithm can increase the contrast in these regions without over-enhancing the already improved sections of the image.

The double application of CLAHE serves to fine-tune the contrast enhancement process, ensuring that all regions of the image are sufficiently enhanced without causing over-enhancement of certain sections. This procedure is particularly beneficial for cryo-EM images, which often exhibit wide variations in contrast across different regions.

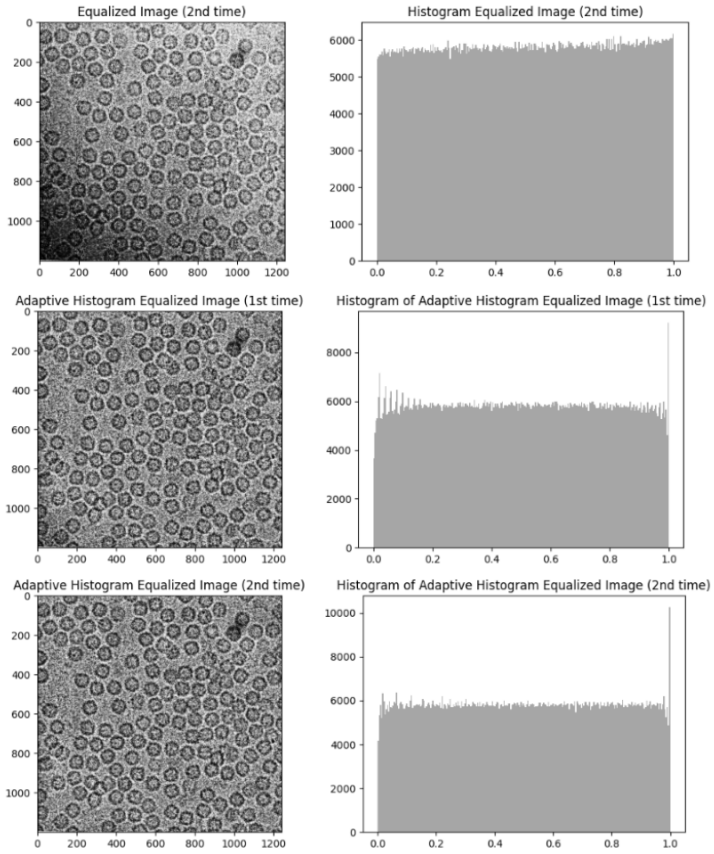


Figure 9: Contrast Enhancement through Adaptive Histogram Equalization. The leftmost panel displays the image following the second round of histogram equalization, and the second panel shows the corresponding histogram, reflecting the pixel intensity distribution. The middle panels depict the image and histogram after the first round of adaptive histogram equalization (with a clip limit of 0.02), demonstrating a local contrast enhancement while preserving global contrast. The rightmost panels showcase the image and histogram after the second round of adaptive histogram equalization (with a clip limit of 0.99), further enhancing the contrast in regions that were not adequately improved during the first round. The progressive enhancement of contrast through adaptive histogram equalization optimizes the visibility of particles within the cryo-EM image.

Step 10: Gaussian Filtering

Gaussian Filtering, often employed in image processing, is a smoothing technique that blurs an image to reduce noise and detail. It uses a Gaussian function, which is a function of space with radial symmetry, meaning that the filtering effect is identical in any direction. Gaussian filtering is linear, and it can be applied to an image with the convolution operation. In the context of cryo-EM images, this operation is necessary to reduce the high-frequency noise that can interfere with downstream processes like particle detection and classification.

In our preprocessing pipeline, we perform Gaussian Filtering four times in succession, each time with a sigma (standard deviation for the Gaussian kernel) of 1. This parameter controls the extent of blur: a smaller sigma results in less blur, and a larger one leads to more. By applying the filter multiple times, we progressively blur the image, effectively reducing high-frequency noise while retaining the general structure of particles.

Simultaneously, we employ a denoising step using Total Variation (TV) regularization, a popular method for image denoising. This technique operates by minimizing the total variation of the image, thus reducing noise while preserving edges. The balance parameter (weight=0.1 in our case) controls the extent of denoising, with higher values producing more denoising (at the potential cost of losing more details).

In combination, Gaussian Filtering and Total Variation denoising help us obtain a cleaner image that facilitates particle identification and interpretation, preparing the image for the final morphological operation.

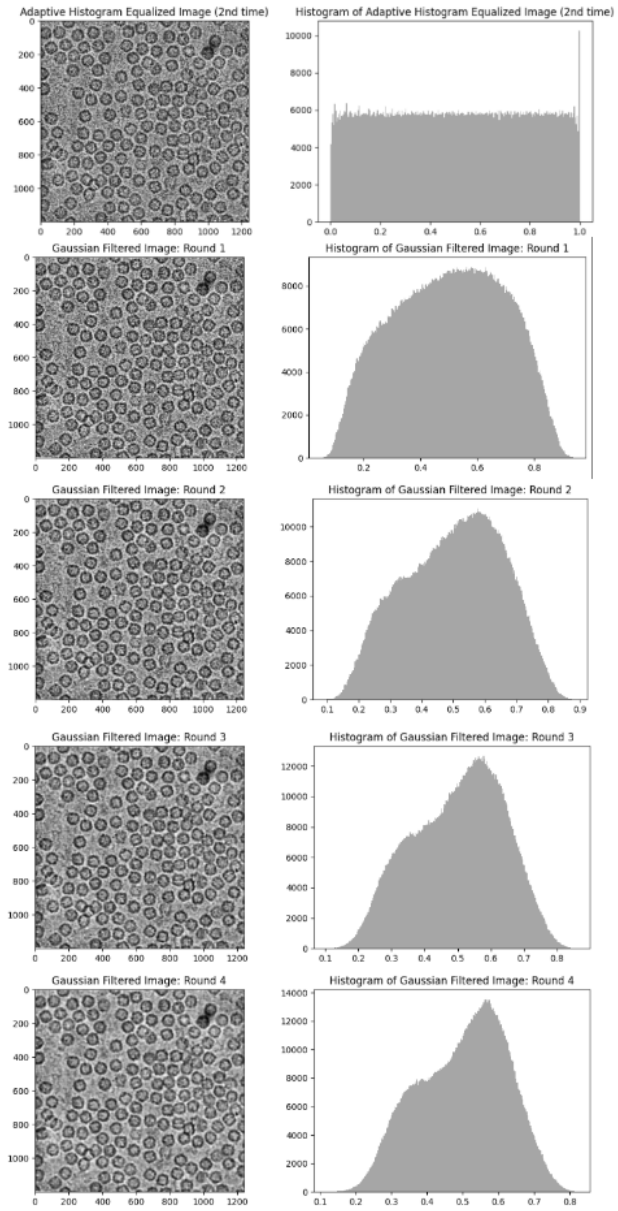


Fig 10. Adaptive Histogram Equalization and Subsequent Gaussian Filtering Effects on Cryo-EM Image and Pixel Intensity Distribution. The first two panels show the image after the second Adaptive Histogram Equalization and its corresponding histogram, respectively. Panels 3 & 4, 5 & 6, 7 & 8, and 9 & 10 present the effects of each subsequent round of Gaussian Filtering on the image and its pixel intensity distribution. As we progress from the initial image through four rounds of Gaussian Filtering (from left to right),

Step 11: Morphological Closing Operation

The final step of our preprocessing pipeline applies a Morphological Closing Operation to the image. Morphological operations are a set of operations that process an image based on its shape, and the closing operation is specifically designed to close small holes in the foreground.

This operation is particularly beneficial for cryo-EM image processing as it aids in removing the small dark spots within the brighter particles, which could otherwise lead to misinterpretations. It works by first performing a dilation operation (which generally expands the bright regions of the image) followed by an erosion operation (which shrinks the bright regions). This sequence allows the closing operation to maintain the size of the larger particles while filling in the small holes.

In our case, we use a disk-shaped structuring element with a radius of 5 pixels. This structuring element is moved across the entire image and at each pixel, the maximum value within the reach of the structuring element replaces the current pixel. As a result, small holes (i.e., dark regions surrounded by brighter ones) get filled in, improving the consistency of individual particles and thereby facilitating easier identification and analysis.

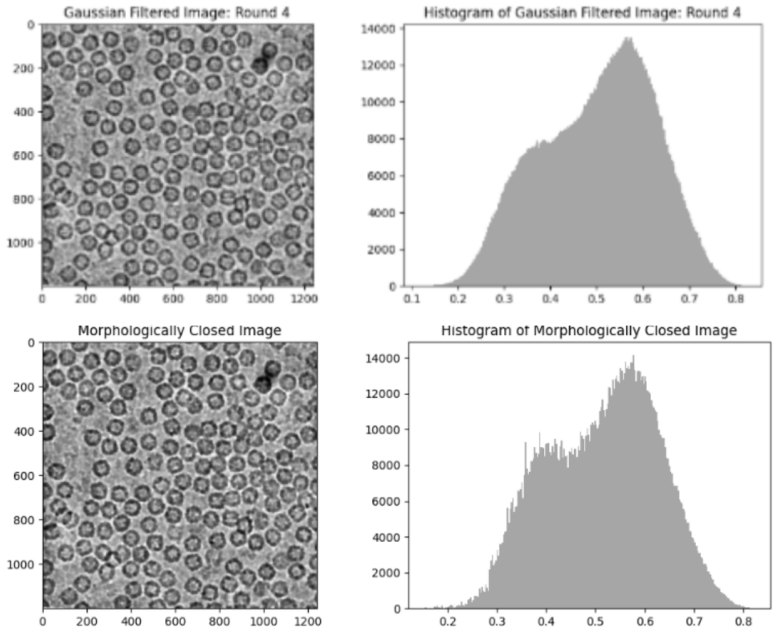


Fig 11. The Effect of Morphological Closing Operation on the Cryo-EM Image and Pixel Intensity Distribution. The first panel shows the cryo-EM image after the final step of Gaussian Filtering, and the corresponding histogram is displayed in the second panel, demonstrating the pixel intensity distribution after the filtering process. The third panel illustrates the effect of the Morphological Closing Operation on the image, specifically highlighting the filled small holes and smoother particles. The final panel presents the histogram after the morphological closing operation, revealing a modified pixel intensity distribution that further enhances the consistency and contrast of particles in the cryo-EM image.

Preprocessing conclusion

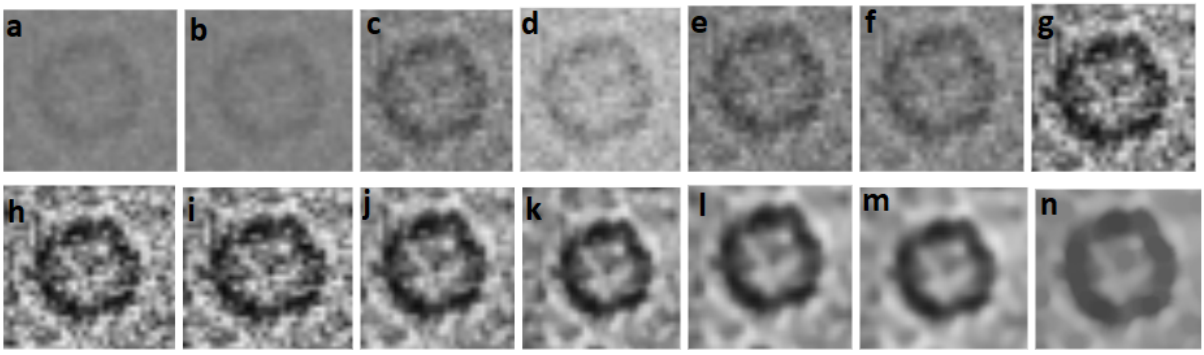
In order to more effectively showcase the impacts of the preprocessing steps, we have chosen to focus on one particle image derived from the dataset. Fig 12 presents a single particle image sourced from the apoferritin dataset. The images in Fig 12 illustrate how the resolution of the cryo-EM image is enhanced by the preprocessing processes. It is apparent that these processes lead to a significant reduction in image noise.

Fig 12 (a) presents our original particle image. Fig 12(b) illustrates how the resolution of the cryo-EM image is enhanced by image averaging and normalization processes. Fig 12(c) presents the same single particle post contrast stretching. The effects of our global brightness and contrast can be seen in Fig 12 (d). When compared to the particle region in the original micrograph after normalization Fig 12(b), the particles in Fig 12(d) exhibit a heightened intensity contrast and are more distinguishable from the background than those in Fig 12 (b). In Fig 12(e), we observe the influence of our first histogram equalization application. This step enhances the overall contrast of the image, making the particle more distinguishable from its background.Subsequently, Fig 12(f) displays the particle after Weiner deconvolution. This technique aids in reducing blur and noise, thereby refining the clarity of the particle's structure. The second round of histogram equalization is demonstrated in Fig 12(g), further escalating the contrast and aiding in the differentiation of the particle from its surroundings.

Fig 12(h) and 12(i) depict the results of two sequential applications of adaptive histogram equalization. These steps adaptively modify the contrast, further emphasizing the particle against its backdrop, and enhancing the details of its structure. The impact of our four-stage Gaussian filtering is presented in Fig 12(j) through 12(m). With each successive application, the smoothing effect of the filter progressively reduces high-frequency noise while preserving the particle's shape and features. Lastly, Fig 12(n) reveals the effect of the morphological closing operation. This step completes our preprocessing by further enhancing the particle's visibility and shape, making it more accessible and identifiable for subsequent analysis stages.

This step-by-step visualization effectively demonstrates how each component of our preprocessing pipeline contributes cumulatively to improving the quality of the cryo-EM images and setting the stage for successful downstream analysis.

Fig 12. Detailed representation of each preprocessing step applied to a single particle image. (a) Original image, (b) Image after normalization and averaging, enhancing resolution and reducing noise, (c) Image post contrast stretching, (d) Image with adjusted global brightness and contrast, increasing particle-background distinction, (e) Image after first histogram equalization, enhancing overall contrast, (f) Image after Weiner deconvolution, reducing blur and noise, (g) Image following the second histogram equalization, further escalating contrast, (h-i) Images after two rounds of adaptive histogram equalization, enhancing particle structure, (j-m) Images demonstrating the four-stage Gaussian filtering, progressively reducing noise, (n) Final image after the morphological closing operation, enhancing particle visibility and shape for subsequent analysis.



By comparing the complete cryo-EM images before and after preprocessing, one can observe the remarkable impact of our pipeline on overall image quality. This comparison is illustrated in Fig 13, where the top panel shows the original cryo-EM image and its histogram distribution, and the bottom panel exhibits the same image after undergoing the complete preprocessing procedure.

In the original image, we can see that the noise, irregular illumination, and other distortions make it challenging to discern the individual particles. However, in the preprocessed image, these challenges are substantially mitigated. The noise is significantly reduced, and the contrast between the particles and the background is markedly increased.

This enhancement in image quality directly impacts the success of downstream analysis. It significantly improves the performance of particle detection and classification algorithms, leading to more accurate and reliable results. This demonstrates the essential role of the preprocessing steps in the analysis of cryo-EM data.

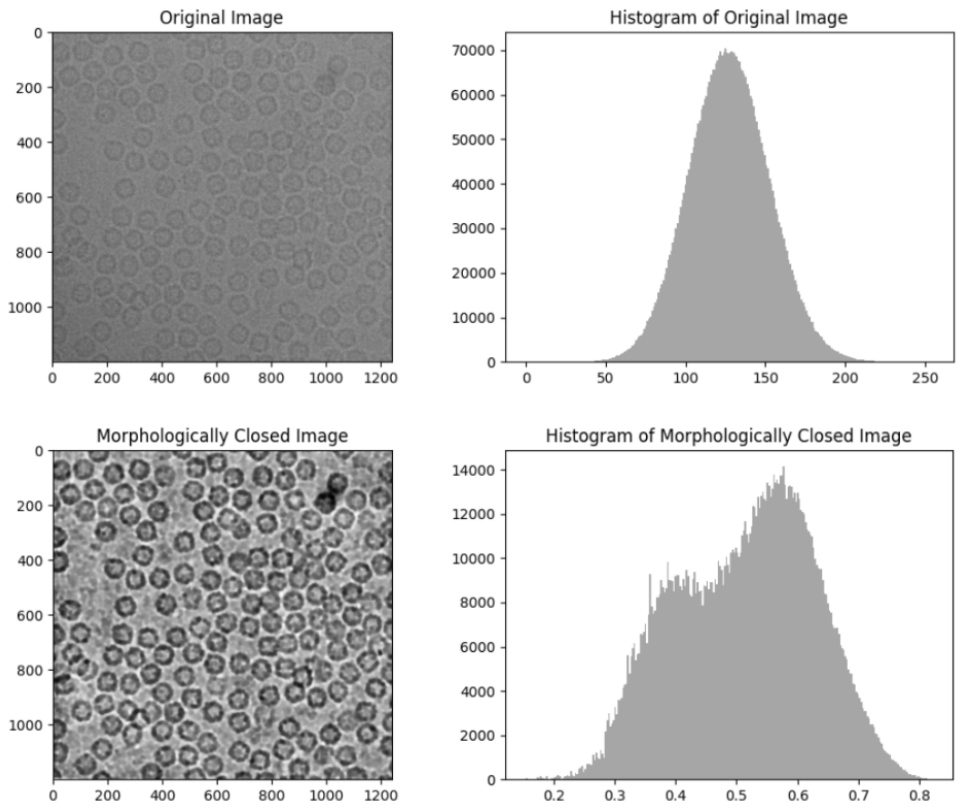


Fig 13: Comparison of the original and preprocessed cryo-EM images. The top panel displays the original image and its corresponding histogram, demonstrating the challenges posed by noise, uneven illumination, and other distortions. The bottom panel presents the same image post-preprocessing, with a markedly enhanced contrast between particles and background, significantly reduced noise, and an altered histogram indicative of the improved image quality.

Stage 2: Clustering Stage

The Clustering stage of our cryo-EM data analysis involves creating a binary mask using unsupervised learning clustering methods to isolate particles. In our approach, we considered three clustering methods: Fuzzy C-Means (FCM), K-means, and our custom Intensity-Based Clustering (IBC). The objective of this step is to assign each pixel in the cryo-EM image to a specific cluster, thus facilitating the identification and isolation of particles.

FCM and K-means are well-established clustering methods, but their application to cryo-EM data poses certain challenges. Specifically, their reliance on random initialization of cluster centers can lead to inconsistent grouping of particles across different images, as illustrated in Fig 14 and Fig 15. For example, in our experimentation with these methods, we noticed that the particles within the same protein images were not consistently assigned to the same cluster.

Fig 14. Comparison of Fuzzy C-Means (FCM) clustering results for Apoferritin (top set) and HLH (bottom set) cryo-EM images. Each row illustrates the clusters generated by the FCM algorithm on the respective image set. Note the inconsistent grouping of particles across different images, indicative of FCM's limitations in this application.

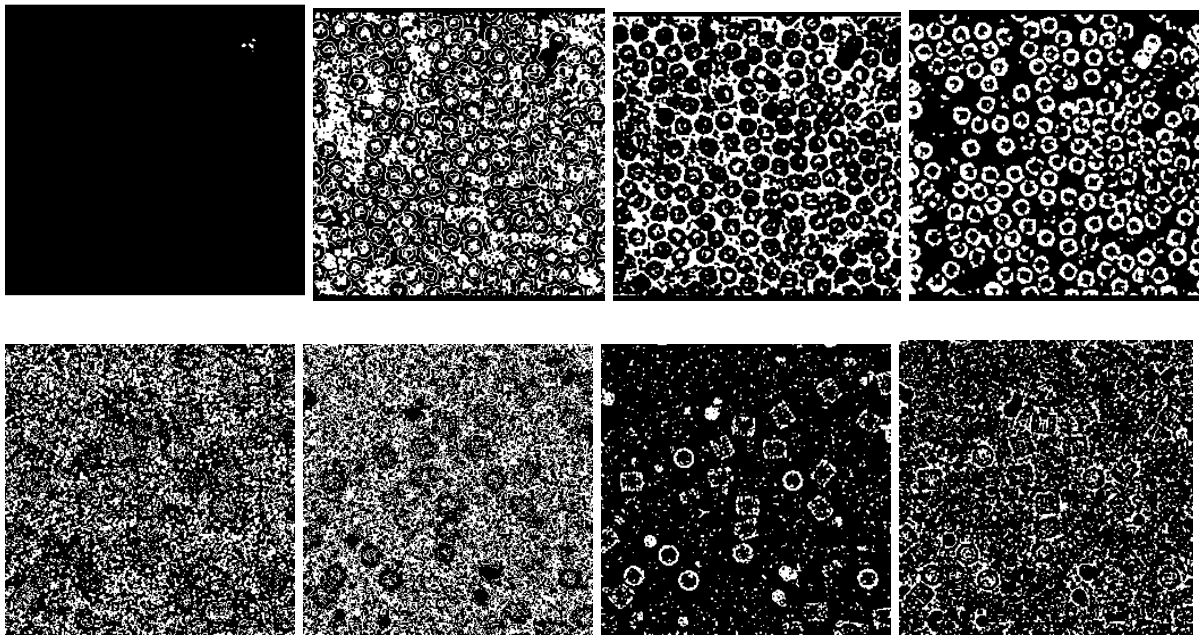
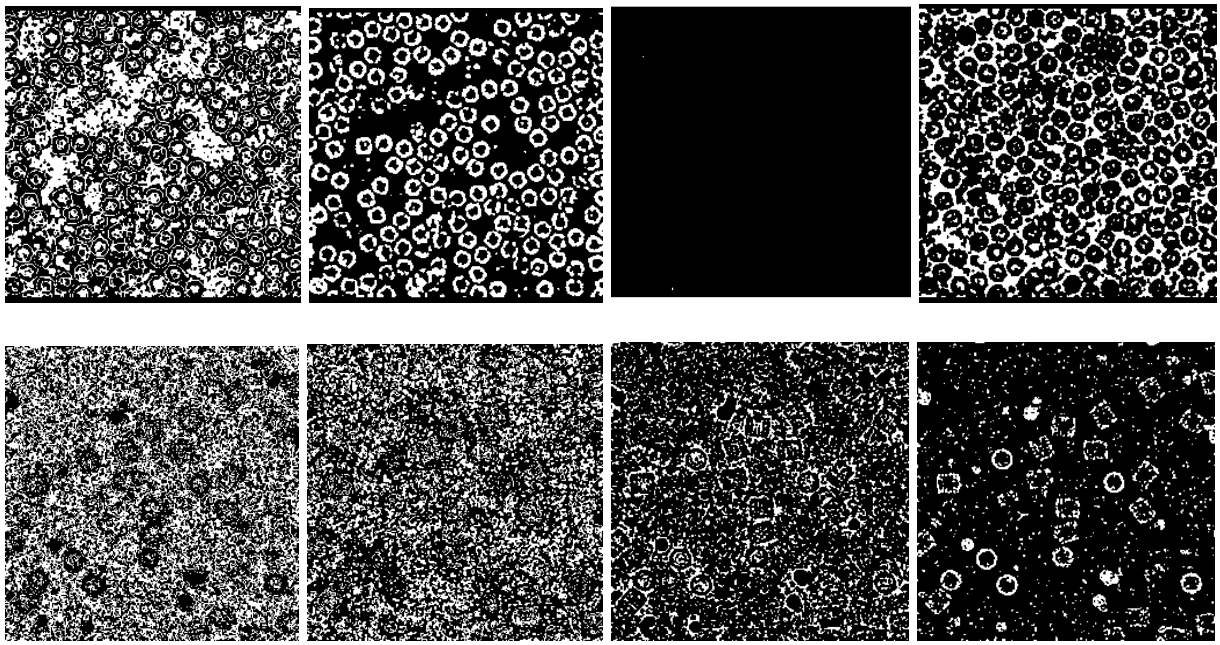
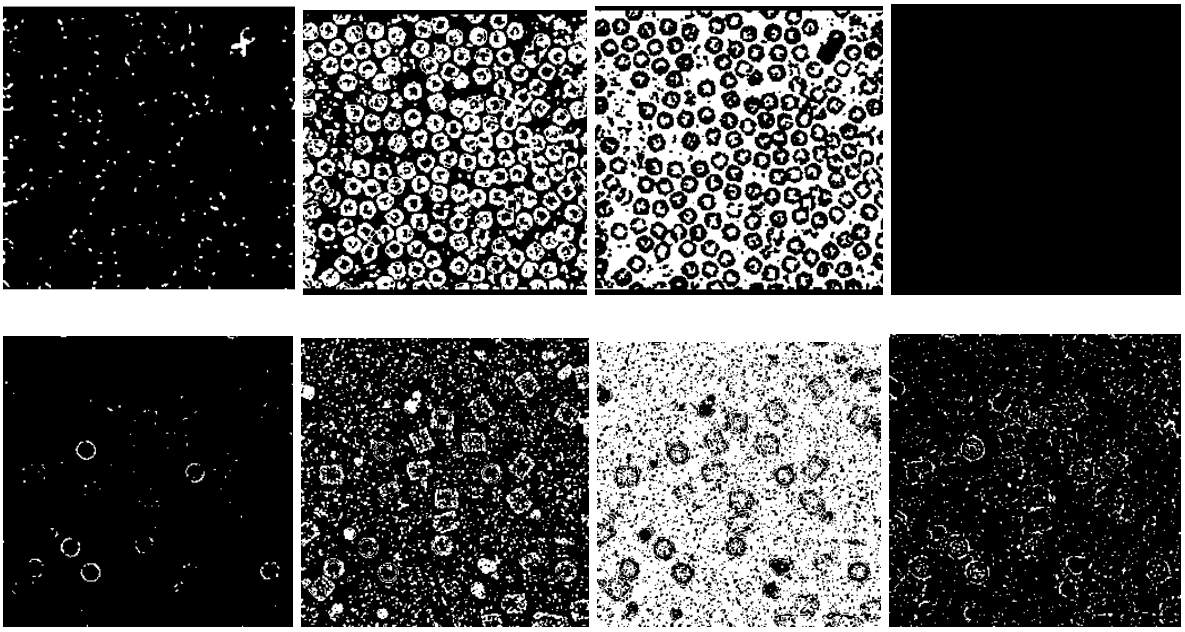


Fig 15. Comparison of K-means clustering results for Apoferritin (top set) and HLH (bottom set) cryo-EM images. Each row illustrates the clusters generated by the K-means algorithm on the respective image set. Similar to FCM, K-means also exhibits inconsistent particle grouping across different images due to the random initialization of cluster centers.



In contrast, our custom IBC algorithm bases the clustering process on an intensity distribution model which correlates the intensity difference value 'd' to the signed difference intensity values, 'i'. This approach ensures a more stable and consistent particle grouping across different images of the same protein, as shown in Fig 16 for the Apoferritin and KLH datasets respectively. The particles were most consistently grouped in Cluster 2, underscoring the stability and reliability of our IBC method.

Fig 16. Clustering results of our custom Intensity-Based Clustering (IBC) method for Apoferritin (top set) and HLH (bottom set) cryo-EM images. Each row demonstrates the clusters generated by the IBC algorithm on the respective image set. Note the consistent particle grouping across different images, highlighting the stability and effectiveness of our IBC method in particle identification and isolation.



Ultimately, we decided to adopt our custom IBC method for our cryo-EM data analysis. This decision was motivated by IBC's ability to provide stable and consistent clustering results, which is crucial for subsequent analysis stages. By using IBC, we can ensure that particles are consistently grouped and isolated, thereby facilitating more accurate particle detection and classification.

Custom IBC

The custom clustering algorithm we developed initiates by flattening the input image into a one-dimensional vector ‘img_vector’. This vector contains the intensity values of all the pixels in the image.

Next, we initialize two lists, ‘clusters’ and ‘cluster_indices’, to store the pixel values and their respective indices for each cluster. Each list is initialized to the size of ‘num_clusters’, which is the number of clusters we want to create.

We then calculate the range of intensity values in the image by subtracting the minimum value from the maximum value. This range is divided by the number of clusters to determine the step value ‘stepv’. This step value is the distance between each cluster center in the intensity spectrum. The cluster centers ‘K’ are then initialized as an array starting from ‘stepv’ and ending at the maximum intensity value, with increments of ‘stepv’.

Following the initialization, the algorithm loops through each pixel value in ‘img_vector’. For each pixel, it computes the absolute difference between the pixel's intensity value and all the cluster centers ‘K’. The ‘np.argmin’ function is used to find the index of the cluster center that is closest to the pixel's intensity value. The pixel's intensity value and its index are then added to the corresponding cluster and ‘cluster_indices’.

Finally, for each cluster, we create a binary mask that represents the distribution of the cluster's pixels in the image. This is done by initializing an array ‘cluster_img’ of zeros with the same shape as ‘img_vector’, then setting the values at the indices corresponding to the current cluster to ‘i + 1’. The resulting array is reshaped back to the original image's shape and added to the ‘binary_masks’ list.

In summary, this algorithm partitions the pixels in the image into a specified number of clusters based on their intensity values, and produces a binary mask for each cluster to facilitate the identification and extraction of particles in the image.

Stage 3: Particle Picking Stage

In the final stage of our cryo-EM data analysis pipeline, we focused on particle picking, which involves two primary steps: binary mask image cleaning and particle object detection and picking.

In the first step, binary mask image cleaning, we performed several post-processing operations to refine the binary mask produced during the clustering stage. These operations included binary image region and hole filling, the application of morphological image opening, and the removal of small objects from the binary image. These procedures served to eliminate noise and irrelevant components, thereby enhancing the accuracy of subsequent particle picking.

The second step entails the detection and picking of particles. Unlike the previous versions that used a modified Circular Hough Transform algorithm, our approach utilizes a region selection method. This method works by scanning the cleaned binary mask and selecting the regions that fit our criteria for being a particle, based on properties such as size, intensity, and shape. This approach enables us to efficiently and effectively identify and isolate particles from the cryo-EM images.

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

In our approach, the binary mask for each cryo-EM clustered image undergoes a cleaning process, where small and non-circular objects are removed via size filtering and roundness filtering. This operation aims to refine the mask by removing artifacts and noise, thereby ensuring a more accurate subsequent particle picking process.

Our implementation involves defining specific parameters tailored to each dataset, namely Apoferritin, Beta-galactosidase, KLH, and Ribosome. These parameters include structuring elements for morphological operations and thresholds for the area and minimum size of the objects.

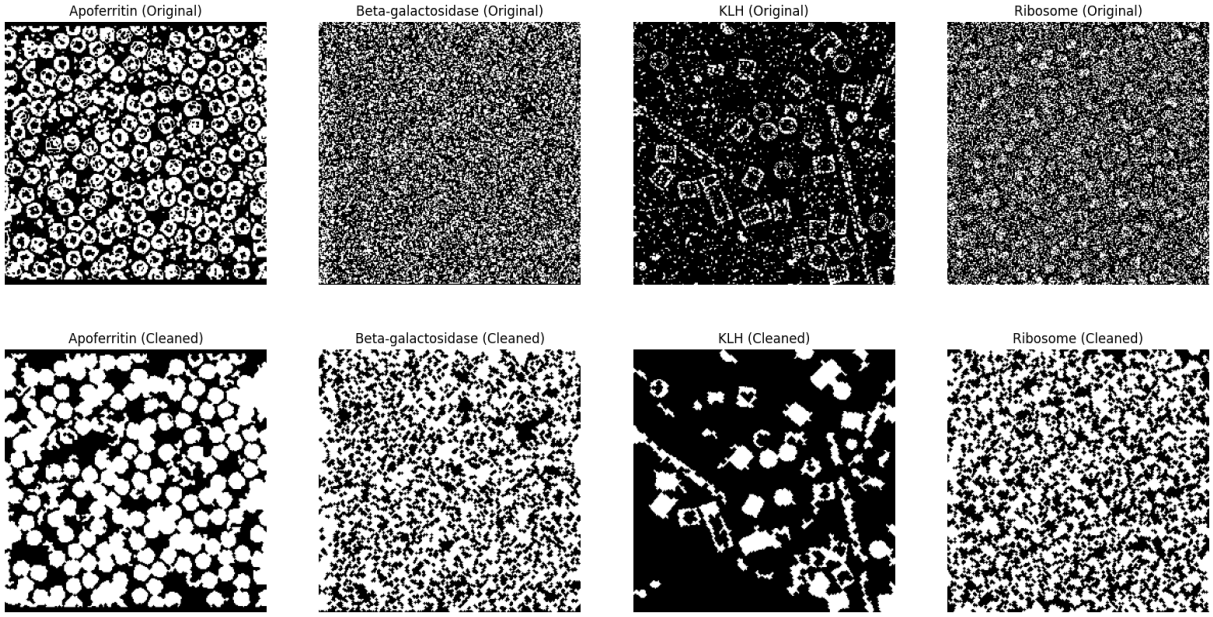
The cleaning process begins with the conversion of the grayscale image to a binary image, followed by a morphological closing operation using the dataset-specific structuring element. Morphological closing is particularly effective in closing small holes in the foreground and connecting disjoint parts of an object.

Next, we fill the holes in the binary image using the dataset-specific area threshold and remove small objects with a defined minimum size. This operation effectively removes irrelevant artifacts and noise in the image while preserving the relevant particles.

Finally, the cleaned binary mask images are stored for the subsequent stage of particle detection and picking.

Please note that additional steps such as morphological opening, roundness filtering, and labeling of connected components were contemplated and can be added based on the needs of the dataset and the specific criteria for the particle shape and size.

Fig 17. Comparative display of binary mask application on diverse datasets. The top row presents a single representative image from each dataset: Apoferritin, Beta-galactosidase, KLH, and Ribosome, respectively. The bottom row illustrates the corresponding cleaned images following the binary mask cleaning procedure. The resulting images highlight the effectiveness of our binary masking in isolating protein particles against background noise. However, it also underscores the need for further fine-tuning, especially evident in the case of Beta-galactosidase and Ribosome datasets.



The binary masking process is crucial for effectively distinguishing the protein particles from the background noise in cryo-EM images. However, the performance of the binary masking can vary based on the inherent characteristics of different datasets, and some may require more meticulous fine-tuning than others. For instance, the Beta-galactosidase and Ribosome datasets have presented more challenges in achieving satisfactory binary mask results.

Beta-galactosidase and Ribosome protein particles possess intricate structures, and their images contain a greater degree of noise or variations in particle size and shape. As such, the current binary masking process, which applies a generic approach to all datasets, may not fully capture the unique features of these particles.

To improve the binary masking for these specific datasets, and indeed for all cryo-EM datasets, the following approaches could be considered:

Adaptive Thresholding: Current thresholding may not be sufficient for images with varying illumination conditions or high-frequency noise. Adaptive thresholding could provide more robust binary masks by setting local thresholds for different regions of the image based on their specific intensity distributions.

Shape-Based Filtering: In addition to size filtering, shape-based filtering could be useful in further refining the binary masks. This could involve more sophisticated measurements of particle shape, such as eccentricity, solidity, or convexity, to better distinguish particles from noise and artifacts.

Multi-Level Clustering: The clustering stage could be expanded to include multi-level or hierarchical clustering, which could help to better distinguish between particles and noise, especially in more complex images.

Machine Learning Approaches: Deep learning models could be trained to perform binary masking by learning the specific features of different types of particles. This could lead to more accurate and robust binary masks, especially for more complex or challenging datasets.

User Interactive Parameter Tuning: An interactive tool could be developed to allow users to manually adjust parameters for the binary masking process, such as size and roundness thresholds. This could provide more flexibility and control for researchers working with particularly challenging datasets.

This is not an exhaustive and the suitability of each approach would need to be evaluated based on the specifics of each dataset and the requirements of the subsequent analysis stages.

Step 2-4: Region Detection

For our region detection, we utilize a two-step process to identify and filter regions of interest in the cryo-EM images.

First, we apply a labeling technique on the connected components within the cleaned images. This is achieved by applying labels to regions within the image. These labels take into consideration the image's connectivity and assign unique labels to separate regions. Subsequently, we compute properties for the labeled regions generating valuable information about each region's characteristics. These properties include information on the labels area, bounding box, centroids, convex area, convex image, coordinates, eccentricity, equivalent diameter, euler number, extent, filled area, label, major and minor axis length, orientation, perimeter and solidity. In parallel, we calculate the roundness of these regions as a ratio of the area and perimeter squared, which is a useful metric for filtering based on particle shape.

Once the properties and roundness for each region have been determined, we proceed to filter out the regions that do not meet our predefined criteria. In our current implementation, we focus on filtering regions based on their area, with a min_area and max_area threshold to exclude regions that are too large or too small. Additionally, we consider the roundness threshold to eliminate regions with irregular shapes that are unlikely to be particles of interest.

It is worth mentioning that future improvements to the filtering process could be explored. For instance, adopting a percentile-based filtering method, where regions falling within the smallest and largest 10% of the identified areas are removed, could help refine the region detection process.

Region detection in cryo-EM images poses several unique challenges, which might have led to the sub-optimal performance of our current method. One of the main issues lies in distinguishing between the actual protein particles and various forms of noise present in the images, such as ice, carbon, or other artifacts. These noise elements can often be mistaken for particles due to their similar sizes and shapes, leading to false positives in the detection process. Conversely, real particles that exhibit irregular shapes or are partially obscured may go undetected, resulting in false negatives.

Another critical challenge lies in the inherent variability of the particles themselves. The particles in our datasets, such as KLH, Beta-galactosidase, and Ribosomes, can vary significantly in their appearance, even within a single dataset. This variability makes it difficult to establish a consistent and reliable set of criteria for detecting particles across all images.

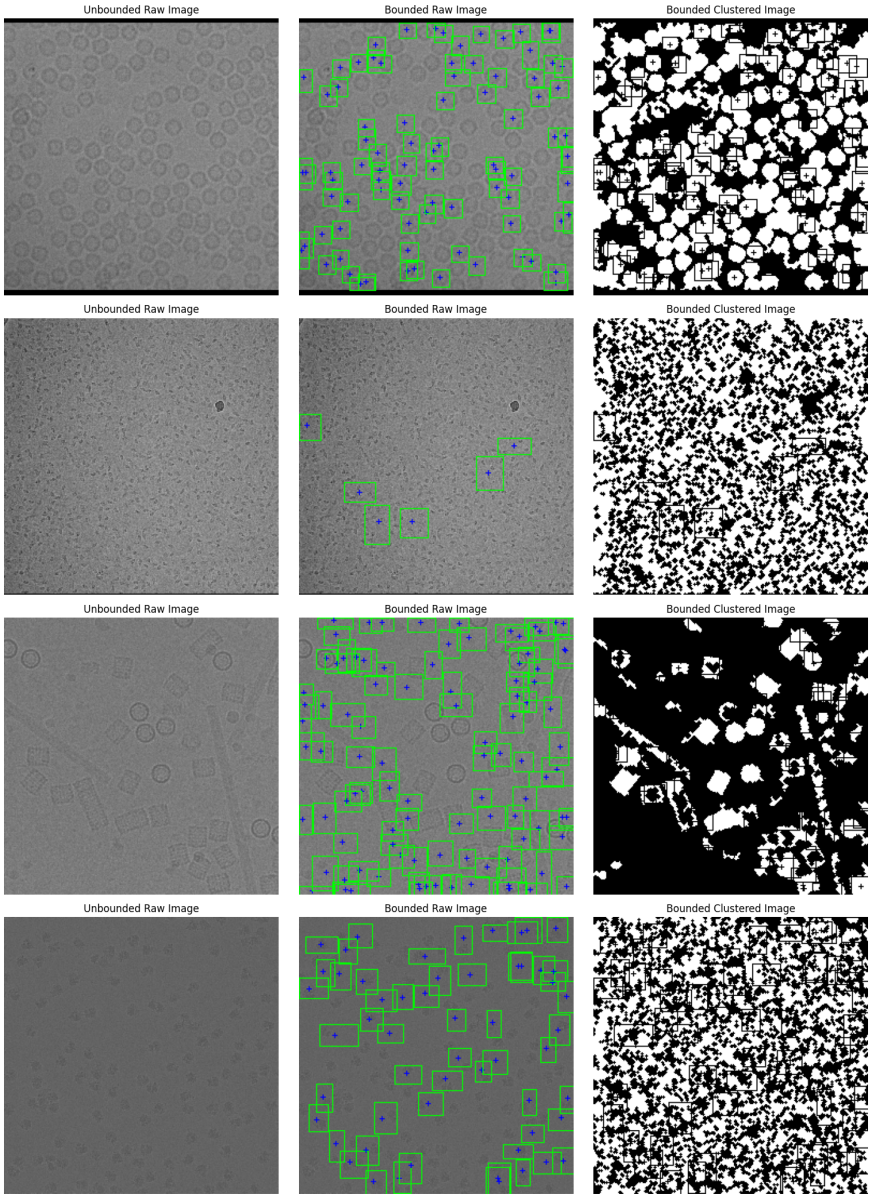
Furthermore, the high level of noise and low contrast typically present in cryo-EM images can exacerbate these issues, making it even more difficult to accurately identify particles. While our post-processing steps in the clustering stage have helped to mitigate this somewhat, it remains a significant challenge that can impact the efficacy of the region detection phase.

Fig 18. illustrates these issues quite starkly, with many particles not being detected and marked, and other areas marked that do not actually contain particles. While disappointing, this outcome also provides valuable insights into the areas we need to focus on for improving our region detection method.

The sub-optimal performance of our current region detection approach underscores the need for further improvements and adaptations. By refining our filtering techniques, incorporating additional region properties, and

exploring alternative detection methods, we can aim to enhance our detection accuracy, and better handle the challenges posed by cryo-EM image analysis. It's essential that our method is designed to be robust against the specific challenges and characteristics of our data, in order to maximize the potential for accurate and reliable particle detection.

Fig 18. Comparison of unbounded, bounded raw, and bounded clustered cryo-EM images for four protein data sets. The first column displays the original unbounded cryo-EM images for Apoferritin, Beta-galactosidase, KLH, and Ribosome, respectively. The second column presents the corresponding bounded raw images post-region detection, showcasing the regions identified as potential particles. The third column exhibits the bounded clustered images after the custom intensity-based clustering (IBC) and subsequent binary mask cleaning. This visual representation provides a clear comparison of the transformation each image undergoes through the preprocessing, region detection, and clustering stages of our data analysis pipeline. The discrepancies between the bounded raw and bounded clustered images highlight the areas of potential improvement in our current region detection approach.



Results

It's important to remember that all stages of the image analysis pipeline are interconnected, and improvements in one stage can often yield significant benefits in subsequent stages. This is particularly relevant in our context where preprocessing, clustering, and binary mask cleaning all play significant roles in shaping the inputs for the region detection stage.

Better preprocessing can help enhance the signal-to-noise ratio and contrast in our cryo-EM images, making the particles more distinguishable from the background and from each other. Advanced techniques such as denoising, histogram equalization, or contrast-limited adaptive histogram equalization (CLAHE) might be explored further to improve the quality of the input images.

Refining our clustering algorithm is another area that can positively impact region detection. By developing more robust and accurate clustering methods, we can ensure that our intensity-based clustering accurately separates particles from the background, leading to more effective and precise isolation of particles. Enhancements to our custom Intensity-Based Clustering (IBC) method could involve adapting the algorithm to better accommodate the intensity variation across different datasets or implementing more sophisticated measures of intensity difference.

Binary mask cleaning is another crucial step in our pipeline, as it helps eliminate noise and other unwanted elements from the binary masks produced during clustering. This process could potentially be enhanced through more rigorous and effective filtering techniques that better account for the shape and size variability of the particles in our datasets. For instance, we could use a more refined approach to removing small and non-circular objects, taking into account the shape variations in our particle datasets.

By improving these stages of our pipeline, we can provide our region detection algorithm with cleaner, better-separated inputs, which can, in turn, help it more accurately and consistently identify the regions of interest. As always, the key lies in balancing the need for accuracy and consistency with the computational efficiency and practicality of our approach.

In the region detection phase of our framework, we utilized an approach based on labeling and property analysis of regions in the binary masks of cryo-EM images. While this method has shown that it can be effective in our context, it's important to acknowledge that there is much room for improvement or alternative approaches.

One potential improvement could be the refinement of the filtering process. The current filtering method relies on setting an area range to accept or reject identified regions. However, these values are currently static and may not optimally handle variations across different data sets or images. A more dynamic approach could involve calculating area thresholds based on statistical properties of the region sizes within each image or dataset, such as percentiles, to more effectively account for this variation.

Another enhancement could be to incorporate additional or more complex region properties into the filtering process. While we currently utilize 'area' and 'roundness' as our primary region properties, the use of other characteristics like 'eccentricity', 'solidity', or 'extent' might offer more nuanced filtering, enabling us to better discriminate between particle-containing regions and noise.

In terms of alternative approaches, one might consider methods such as the Circular Hough Transform (CHT). The CHT is a well-known technique for detecting circular objects within an image. It would seem a good fit for identifying particles in cryo-EM images, which are often roughly circular. However, in practice, the particles within our datasets do not always appear as clean circles, especially in the case of particles like KLH, which have side views showing up as squares, or Beta-galactosidase and Ribosomes, whose particles show irregularities in shape. In these instances, a method based on the CHT could actually underperform, as it may not accurately identify these non-circular particle representations.

Overall, while the Circular Hough Transform has proven valuable in many applications, it is not always the best tool for every job. The choice of region detection method will always be context-dependent, and it's essential to select a technique that aligns with the specific characteristics and challenges of the data at hand. For the task of identifying

particles in cryo-EM images, our current approach, focusing on region properties and specifically designed filtering techniques, offers a promising and effective solution.

Conclusion

In this study, we have presented a novel pipeline for automated particle detection in cryo-EM images that incorporates advanced techniques such as custom intensity-based clustering (IBC) and region-based particle identification. Our method has the potential to streamline and automate the labor-intensive process of particle picking, significantly enhancing the efficiency of protein structure determination via cryo-EM.

However, our results have highlighted areas for improvement. Particularly, our approach revealed challenges in consistently identifying particles across varying protein structures, with the region detection stage often failing to capture all particles, and erroneously marking non-particle regions. Moreover, our binary mask cleaning stage may require more fine-tuning to achieve better results, particularly for the Beta-galactosidase and Ribosome datasets.

Future work will involve improving our preprocessing and clustering stages, as well as exploring more sophisticated techniques for region detection, possibly incorporating machine learning algorithms. We also aim to refine our binary mask cleaning process, ensuring that it is more adaptive to the unique characteristics of each protein structure.

Despite these areas of improvement, we believe that the potential of our approach is promising. The incorporation of our IBC method, along with our unique region-based particle identification, adds a new perspective to the particle picking process, pushing the boundaries of current methods. With further refinement, our method may prove to be a valuable addition to the particle picking pipelines of cryo-EM image processing software.

References

Adil Al-Azzawi, Anes Ouadou, Highsmith Max R, John J. Tanner, Jianlin Cheng. DeepCryoPicker: Fully Automated Deep Neural Network for Single Protein Particle Picking in cryo-EM. https://doi.org/10.1101/763839

Al-Azzawi, A., Ouadou, A., Tanner, J.J. et al. AutoCryoPicker: an unsupervised learning approach for fully automated single particle picking in Cryo-EM images. BMC Bioinformatics 20, 326 (2019). https://doi.org/10.1186/s12859-019-2926-y

Al-Azzawi, A., Ouadou, A., Max, H. et al. DeepCryoPicker: fully automated deep neural network for single protein particle picking in cryo-EM. BMC Bioinformatics 21, 509 (2020). https://doi.org/10.1186/s12859-020-03809-7

Nogales E, Scheres SH. Cryo-EM: a unique tool for the visualization of macromolecular complexity. Mol Cell. 2015;58(4):677–89.

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.

Jiang J, Pentelute BL, Collier RJ, Zhou ZH. Atomic structure of anthrax protective antigen pore elucidates toxin translocation. Nature. 2015;521(7553):545–9.

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.

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.

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.

Glaeser RM, Nicholson WV, Robert M. Review: automatic particle detection in electron. J Struct Biol. 2001;133:90–101.

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.

Frank, J. (2006). Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State (2nd ed.). Oxford University Press.

Henderson, R. et al. (2012). Tilt-pair analysis of images from a range of different specimens in single-particle electron cryomicroscopy. Journal of Molecular Biology, 413(5), 1028–1046.

Sigworth, F. J. (1998). A maximum-likelihood approach to single-particle image refinement. Journal of Structural Biology, 122(3), 328–339.

Scheres, S. H. W. (2012). RELION: Implementation of a Bayesian approach to cryo-EM structure determination. Journal of Structural Biology, 180(3), 519–530.

Chen, S., McMullan, G., Faruqi, A. R., Murshudov, G. N., Short, J. M., Scheres, S. H. W., & Henderson, R. (2013). High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. Ultramicroscopy, 135, 24–35.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE International Conference on Computer Vision, 2015 Inter, 1026–1034.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv.