Particle-picking in cryo-electron microscopy images with online resources

ABSTRACT

Aims: Use deep learning online resources to identify and pick single particle views in micrographs to enable high quality three-dimensional reconstruction for macromolecular structure determination.
Study design: Using the keyhole limpet hemocyanin dataset, a public cryo electron microscopy (cryo-EM) dataset containing two dimensional projections of the particles in two views (top and side) in several orientations, and a recent deep learning algorithm made available in a GitHub repository, we design the procedure to pick both views with high degree of confidence, using only online resources and running in a standard laptop.
Methodology: The defocus images are subject to a pre-processing stage to increase its contrast and ameliorate its radiometric range. This is followed by a training stage, that needs a few images annotated with examples of both views of interest – top and side view – identified using user-friendly tools available online. The annotated subset of images is divided for train and validation purposes, and the algorithm runs to produce a set of weights, that can be used for inference in any other similar image, locating all the instances of the particles in both views in seconds.
Results: From the 57 images used to evaluate the performance of the algorithm, 63% had both precision and recall better than 90%. The global precision on the test dataset was 91.6% and recall 98.1%. Considering each view separately, top views detections attain a precision of 91.7%, with 100% for recall; side views recall remains on 96.7%, while precision attain 91.5%.
Conclusion: Deep learning methods are a promising tool for extracting large amounts of single particle views as needed for quality 3D structure reconstruction, as they can be implemented with minimal computer skills and trained to achieve state-of-the-art automatic discrimination, as described in this study.

Keywords: particle picking, cryo-EM, deep-learning, 3D reconstruction

1. INTRODUCTION

High resolution cryo-EM reconstruction of a molecular complex requires a huge number of single particles collected from micrographs, because low contrast and contaminated and noisy backgrounds are characteristics of these images inherent to the low electron dosage used to avoid damage to the targeted molecules [1]; the number of particles required for a reconstruction to approximate atomic resolution [2-3] are in the hundreds of thousands.
In the last 20 years several approaches have been developed to overcome this situation and achieve the desirable particle numbers [4-6], from template matching techniques [7-9] in direct or Fourier space to reference-free methodologies requiring only the size range of the particles [10] involved. Since the onset of machine learning techniques, deep learning with its multidimensional layered feature extraction came to the rescue. It's broad and successful application in other domains of known difficulty (speech recognition emotions [11], facial expression [12], specific video surveillance [13]) soon bring it to almost all fields of biology [14-15] and biomedical imagery [16]. Deep learning technology has been used to solve complex problems in microbiology such as predicting drug targets or vaccine candidates, diagnosing microorganisms causing infectious diseases, classifying drug resistance against antimicrobial medicines [17], predicting disease outbreaks, exploring microbial

interactions, and detecting viral plaques, and has been applied to a large number of microorganisms, such as viruses, bacteria, fungi, and parasites [18].

Deep learning is a subset of machine learning (ML) methods that has become popular as the availability of data has greatly increased due to automation and digitization, and standard laptops have become available with high processing power and large storage capacity, as well as graphics processing units (GPU) that allow parallel processing: all these factors have contributed to allowing intensive training of deep learning models on standard computers, even if tens of hours are still involved. The possibility of using cloud processing facilitates the process when the time factor becomes critical.

Our motivation is to show how easy can be nowadays to establish a processing chain to attain results using the last evolution of deep learning algorithms, as both the image datasets and the software become available online and can be used by anyone with a minimum of resources – a standad laptop and, in our case, an image processing environment (Matlab) for which open-source alternatives also exist.

2. MATERIAL AND METHODS

The images are part of the keyhole limpet hemocyanin (KLH) public dataset [19] that can be downloaded online. With a pixel size of 2.2Å, the images were acquired with a Philips CM200 TEM equipped with a 2Kx2K CCD Tietz camera, at a nominal magnification of 66,000x and a voltage of 120 KeV.

The algorithm used is one of the latest developments in one stage algorithms based on convolutional neural networks (CNNs), the 5th version since the introduction of the concept You Only Look Once (YOLO) [20]. YOLO v5 was introduced in 2020 by a different developer and made publicly available in a GitHub repository [21].

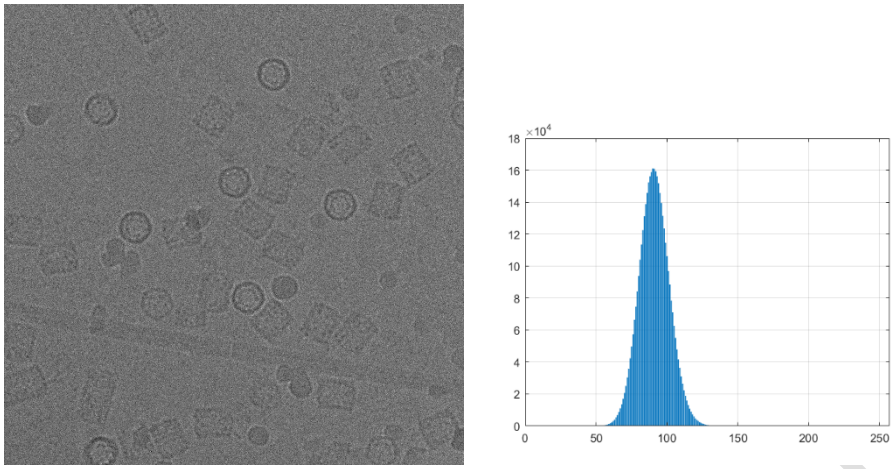
The algorithm has been retrained for the task using a transfer learning technique: since many basic features are common to all detection problems (edges, contrasts, forms, etc.) an already heavily trained network can be used to implement a new problem. The new discriminators will define the last layers of the CNN, tunning the detector according to the details of the specific problem, while the basics defined by the first layers were robustly trained in big data sets like Common Objects in Context (COCO), with 80 classes and more than 200.000 images annotated. Tools available online such as [22] allow to upload our own image data set, annotate each image, and download a set of text files with the corresponding annotations in a user defined format.

The laptop used is equipped with dual Core Intel i7-10750H processor, 16 GB SDRAM and an NVIDIA GeForce RTX 2060 Max-Q 6GB.

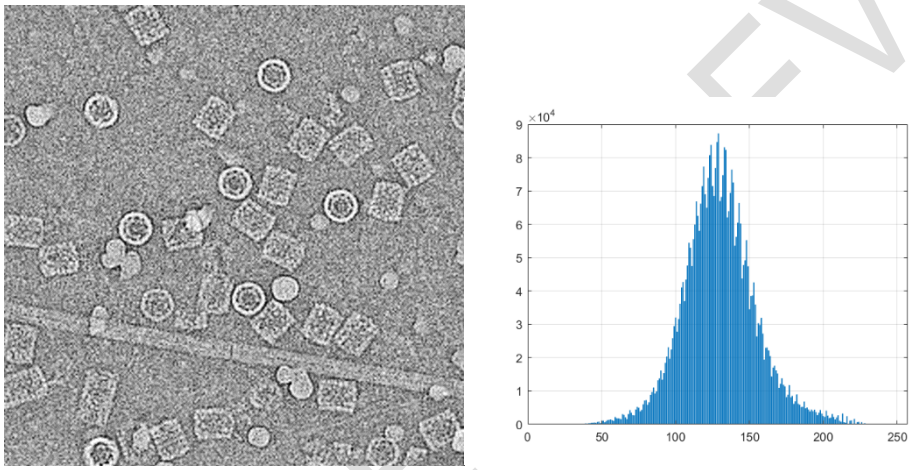
Using a Matlab environment, the 82 defocus images were resized in 1024x1024 tiles with a bicubic interpolation method and a pre-processing stage has been implemented to reduce noise and increase radiometric contrast in each image (Fig. 1), with a 3-step procedure followed by normalization.

The first step consists in a tunned wavelet filter build to suppress details, where a Daubechies wavelet of 15^{th} order was used in two levels of decomposition/reconstruction. A Wiener filter with a kernel of 9x9 pixels were applied to the result, and the last step was a contrast-limited adaptive histogram equalization parametrized with 4x4 tiles and a uniform distribution.

The spread of the radiometry that can be observed in the histogram (figure1, right column) after the pre-processing lead to an increase in contrast which is useful both for building a reference data set to evaluate the results and to the algorithm, as much more features become evident.



(a)



(b)

Fig. 1. Original (a) and pre-processed (b) images, and corresponding histograms..

A subset of 24 images was used to train the detector, splitted in 16 images for train and 8 for validation. An online tool [22] was used to annotate the 24 images, keeping in mind that all objects of interest of each class present in the image should be identified and that the boxes drawn around each object should include minimal background. The algorithm's data augmentation properties extend all the occurrences annotated to many viable possibilities in what concerns scale and orientation. We choose to include two classes in the train, the two views "Top" and "Side".

The 57 remaining images were used to evaluate the performance of the inference using the notions of Precision and Recall. Precision is defined as the percent of particles correctly classified among all the particles identified by the algorithm as positives, given by the quotient True Positives/(True Positives + False Positives). Recall refers to the percentage of positives correctly detected in the sum of all occurrences of real positives, and is given by the quotient True Positives/(True Positives + False Negatives).

Detected particles overlapped/overlaid by any debris, particles with heavily damaged boundaries, aggregates, and side-view particles larger than 1.5 times normal size were considered false positives (figure 2).

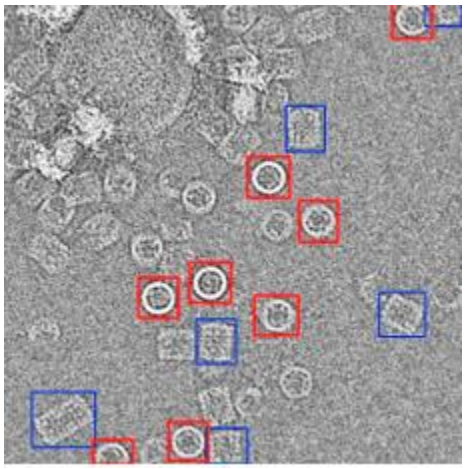


Fig. 2. Example of detections considered false positives - bottom left corner, one side view larger than normal, and on the right side of the image two top and one side views with corrupted boundaries.

Particles whose detection failed, to be counted as false negatives, were particles with the boundary in good condition and without any overlapping exogenous material nor aggregated with any other particle.

4. RESULTS AND DISCUSSION

The train specifications include 1000 epochs and a batch size of 4, for an image size of 1024, and 526 epochs were performed in 32.5 hours, before the "early stop" interrupts the execution, due to stability conditions being reached. Overall precision achieved was 0.916 and recall 0.878, with a mean average precision (mAP@0.5) of 0.930. Individually, the “Top” class has better scores on precision (0.976), recall (0.987) and mAP@0.5 (0.993) than the "Side" class (precision: 0.856, recall: 0.770, mAP@0.5: 0.866), which could be explained by the larger variety included in the side views.
Inference over each image (1024x1024 pixels) with the weights issued by the train is achieved in 2.8 s on the previously described laptop. The outputs consist of a classified image with all occurrences detected by the algorithm differentiated by class (as figure 2) in different colors, and a numerical output, considered for evaluation purposes as the number of true positives founded. Each classified image was analyzed to account for false positives and false negatives with the criteria mentioned above.
From the 57 test images, 63% had both precision and recall better than 90%, and 25% had both metrics above 95% simultaneously. The number of false positives comes mainly from the side views; although we experimented in the sense of reducing to zero the scaling parameter available in the train specifications for data augmentation, the results were no better. The higher contrast of perfect top views account for the better scores of the class.
The global precision on the test dataset was 91.6% and recall 98.1%. Considering each class separately, top views detections attain a precision of 91.7%, with 100% for recall; side views recall remains on 96.7%, while precision attain 91.5%.

4. CONCLUSIONS

Deep learning methods based in heavily trained CNN’s such as the one used in this work, known as YOLOv5, are a promising tool for extracting large amounts of single particle views as needed for 3D macromolecular structure reconstruction. The methodology is simple to implement, and once the algorithm is trained, the inference phase is fast, and can be implemented on a large amount of images with little cost in time and labor, and last but not least, with minimal computer skills.
A major advantage of the large CNN used, the model x of YOLOv5, is the small quantity of false positives and the almost absence of false negatives. These numbers can be tunned easily for a particular application using the parameters available for the inference command, in particular the confidence threshold that reflects the probability associated to the classification of each object detected, that can be displayed in the image output (figure 3).

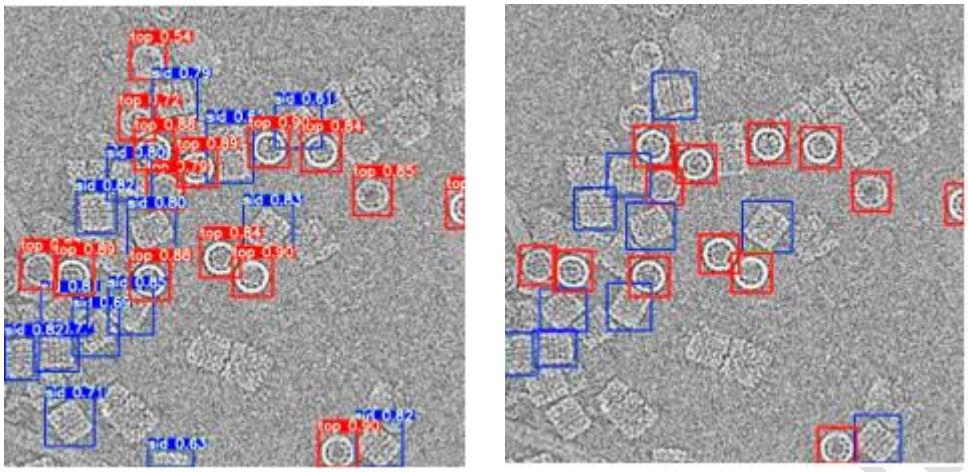


Fig. 3. The image output with the confidence associate to each detection (left) can be used to tune the inference and reduce the number of false positives, using a higher confidence threshold, in the case 0.75 (right)

Obviously, if a higher threshold is selected for the inference process, fewer false positives will be detected, but some true particles that could be used will not be accounted for. If a large dataset is available, more demanding confidence thresholds will be a good option to reduce human curation of the final product.

All the software used is open source, except Matlab which was used for the preprocessing of the images, but there are similar options available online [23] that could be used for the operations described, aiming to improve image contrast and radiometric range.

The availability of online software that demands minimum informatic skills to be used, and the broad application that can be achieved using only limited computational resources and online tools, enables these methodologies to extend to other applications. In future work, we will look for new applications of deep learning methods for other microbiology tasks, focusing those that can be useful in technologically disadvantaged countries, in addition to those that we have already been able to demonstrate, such as counting microalgae [24] with a binocular loupe, detecting polyomaviruses [25] or localize exosomes [26] in transmission electron microscopy images.

This is a tool that will facilitate and speed up some work that was previously imprecise and cumbersome, as well as prone to the subjectivity inherent in all human work.

REFERENCES

1. Zhu Y, Ouyang Q, Mao Y. A deep convolutional neural network approach to single-particle recognition in cryo-electron microscopy. BMC Bioinformatics. 2017; 18(1):348. doi: 10.1186/s12859-017-1757-y.

2.Yan C, Hang J, Wan R, Huang M, Wong CC, Shi Y. Structure of a yeast spliceosome at 3.6-angstrom resolution. Science. 2015; 349(6253):1182-91. doi: 10.1126/science.aac7629

3. Wu J, Yan Z, Li Z, Yan C, Lu S, Dong M, Yan N. Structure of the voltage-gated calcium channel Cav1.1 complex. Science. 2015; 350(6267): aad2395. doi:10.1126/science.aad2395

4. Nicholson WV, Glaeser RM. Review: automatic particle detection in electron microscopy. J Struct Biol. 2001; 133(2-3):90-101. doi: 10.1006/jsbi.2001.4348. PMID: 11472081.

5. Henderson R. Overview and future of single particle electron cryomicroscopy. Arch Biochem Biophys. 2015; 581:19-24. doi: 10.1016/j.abb.2015.02.036.

6. White HE, Ignatiou A, Clare DK, Orlova EV. Structural Study of Heterogeneous Biological Samples by Cryoelectron Microscopy and Image Processing. Biomed Res Int. 2017; 2017:1032432. doi: 10.1155/2017/1032432. PMID: 28191458.

7. Roseman AM. Particle finding in electron micrographs using a fast local correlation algorithm. Ultramicroscopy. 2003; 94(3-4):225-36. doi: 10.1016/s0304-3991(02)00333-9

8. Langlois R, Pallesen J, Ash JT, Nam Ho D, Rubinstein JL, Frank J. Automated particle picking for low-contrast macromolecules in cryo-electron microscopy. J Struct Biol. 2014; 186(1):1-7. doi: 10.1016/j.jsb.2014.03.001. PMID: 24607413

9. Huang Z, Penczek PA. Application of template matching technique to particle detection in electron micrographs. J Struct Biol. 2004; 145(1-2):29-40. doi: 10.1016/j.jsb.2003.11.004.

10. Langlois R, Pallesen J, Frank J. Reference-free particle selection enhanced with semi-supervised machine learning for cryo-electron microscopy. J Struct Biol. 2011; 175(3):353-61. doi: 10.1016/j.jsb.2011.06.004. PMID: 21708269
11. Hadhami A, Yassine BA. Speech Emotion Recognition with deep learning. Procedia Computer Science. 2020; 176:251-260. doi: 10.1016/j.procs.2020.08.027.
12. Saad S, Asghar AS, Muhammad KE, Muhammad RA, Asad M, Teweldebrhan M. Automated Facial Expression Recognition Framework Using Deep Learning. J Healthcare Eng, 2022; vol. 2022. Doi:10.1155/2022/5707930.
13. Hashmi TSS, Haq NU, Fraz MM, Shahzad M. Application of Deep Learning for Weapons Detection in Surveillance Videos. 2021 Intl Conf Digital Futures and Transf Techn (ICoDT2), 2021; 1-6. Doi:10.1109/ICoDT252288.2021.9441523.
14. Webb S. Deep Learning for Biology. Nature. 2018; 554:555-557.
15. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141):20170387. doi: 10.1098/rsif.2017.0387. PMID: 29618526.
16. Xing F, Xie Y, Su H, Liu F, Yang L. Deep Learning in Microscopy Image Analysis: A Survey. IEEE Trans Neural Networks and Learning Systems. 2018; 29(10):4550-4568. doi: 10.1109/TNNLS.2017.2766168. PMID: 29989994.
17. Stephen J G, Barratt JLN, Kennedy PJ, Kaufer A, Calarco L, Ellis JT. Machine learning and applications in microbiology. FEMS Microbiology Reviews, 2021; 45(5). Doi: 10.1093/femsre/fuab015
18. Zhang Y, Jiang H, Ye T, Juhas M. Deep Learning for Imaging and Detection of Microorganisms. Trends in Microbiology, 2021; 29(7):569-572. Doi 10.1016/j.tim.2021.01.006.
19. National Resource for Automated Molecular Microscopy. Downloads (nysbc.org). Accessed 29 March 2022
20. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. IEEE Conf on Comp Vision and Pattern Recognition. 2016;779-788.
21. GitHub. GitHub - ultralytics/yolov5: YOLOv5 🚀 in PyTorch > ONNX > CoreML > TFLite. Accessed 21 June 2021
22. Makesense.AI. https://www.makesense.ai/ . Accessed 25 January 2022
23. Image processing software ImageJ.JS. Accessed 6 June 2023
24. Proença MC, Barbosa M, Amorim A. Counting microalgae cultures with a stereo microscope and a cell phone using deep learning online resources. Bulletin of the National Research Centre. 2022; 46:278. doi:10.1186/s42269-022-00965-z
25. Proença MC. Polyomavirus localized with deep-learning methods. Journal of Advances in Microbiology. 2022; 22(3):52-57. doi: 10.9734/jamb/2022/v22i330445
26. Proença MC, Alves de Matos AP. Case study – apply a deep-learning algorithm to exomes detection with online resources. Annals of Antivirals and Antiretrovirals. 2021; 5(1):33-35. doi: 10.17352/aaa.000014