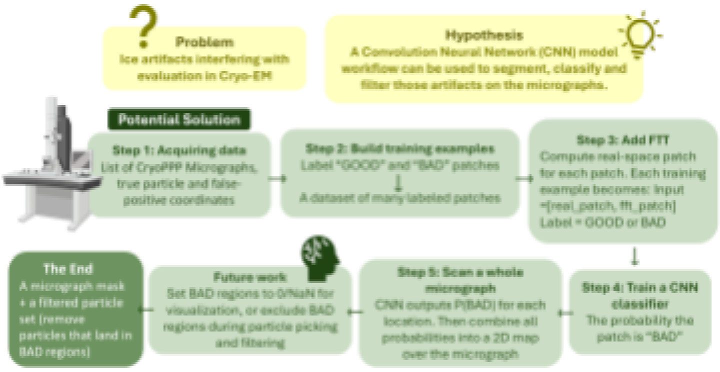
Automated Detection of Ice Artifacts in Cryo-EM Using Machine Learning

Tharushi Rajaguru[1], Shehani Kahawatte[1], Diana Oliveira[2], Elif Dursun[3]

1. *Department of Chemistry, University of Cincinnati, Cincinnati, Ohio 45221, United States*
2. *International Iberian Nanotechnology Laboratory (INL), Braga, Portugal*
3. *Department of Mechanical and Materials Engineering, University of Cincinnati, Cincinnati, Ohio 45221, United States*

INTRODUCTION

Cryogenic electron microscopy (Cryo-EM) is an advanced technique for analyzing the structures of biological macromolecules such as large protein complexes. In Cryo-EM, biomacromolecules are rapidly frozen and fixed in clear ice which allows acquisition of two-dimensional (2D) images in all directions. A computer is then used to process 2D images and reconstruct the three-dimensional (3D) structures of the studied macromolecules [1]. For sample preparation, purified proteins are dropped casted in transmission electron support grids and rapidly frozen in a thin-layer of solution, forming thin-like ice, and then imaged at cryogenic temperatures. Nevertheless, analysing Cryo-EM micrographs poses a significant challenge due to the low signal-to-noise ratio, poor image contrast, and the presence of contaminants such as ice artifacts. Thereafter, making the distinction between true signal and noise is extremely difficult. Low signal-to-noise ratio and poor image contrast is difficult to avoid as biological samples are sensitive to radiation. Those problems can only be improved by decreasing the voltage, which results in a very low signal-to-noise ratio [1]. Therefore it is crucial to improve those with advanced processing techniques. We hypothesize that an automated artifacts detection tool that implements a deep convolutional neural network (CNN) model workflow can segment, classify and filter misleading ice artifacts on the micrographs.



Problem
Ice artifacts interfering with evaluation in Cryo-EM

Hypothesis
A Convolution Neural Network (CNN) model workflow can be used to segment, classify and filter those artifacts on the micrographs.

Potential Solution

Step 1: Acquiring data
List of CryoPPP Micrographs, true particle and false-positive coordinates

Step 2: Build training examples
Label "GOOD" and "BAD" patches
A dataset of many labeled patches

Step 3: Add FTT
Compute real-space patch for each patch. Each training example becomes: Input =[real_patch, fft_patch]
Label = GOOD or BAD

Step 4: Train a CNN classifier
The probability the patch is "BAD"

Step 5: Scan a whole micrograph
CNN outputs P(BAD) for each location. Then combine all probabilities into a 2D map over the micrograph

Future work
Set BAD regions to 0/NaN for visualization, or exclude BAD regions during particle picking and filtering

The End
A micrograph mask + a filtered particle set (remove particles that land in BAD regions)

Figure 1. Schematic overview of this study

MATERIALS & METHODS

Data Preparation

In this study, Cryo-EM micrographs and corresponding annotations were obtained from the cryoPPP dataset [3]. Specifically, data associated with EMPIAR accession 10061 were used in this study. The Electron Microscopy Public Image Archive (EMPIAR) is an open public resource for raw image data used in the generation of 3D Cryo-EM maps [2]. Micrographs were employed as input images, while expert-validated particle coordinates and false positive annotations were used to generate ground truth labels. To construct a supervised learning dataset, fixed square image patches (64x64 px) were extracted from the micrographs. Patches centered on false-positive coordinates were labeled as ‘BAD’ and patches centered on validated particle coordinates were labeled as ‘GOOD’. In addition to using raw image patches as the input, we computed Fast Fourier Transform (FFT) of each patch that reveals periodic and frequency domain signatures of ice, such as crystalline rings and high-frequency attenuation that are difficult to detect reliably in real space alone [4].

Model Building and Training

A supervised machine learning model based on CNN was developed to classify Cryo-EM patches as ‘BAD’ or ‘GOOD’ [5]. The dataset was first divided as training and validation sets. To improve model generalization and robustness, data augmentation was applied on the training patches. This ensures the model learns features invariant to orientation and slight variations in the micrographs. The network consists of multiple convolutional layers to extract spatial features from the input patches, followed by fully connected layers for classification.Hyperparameters, including learning rate, batch size, network depth and regularization parameters, tuned to optimize model performance. The model was optimized using the Adam optimizer and a cross-entropy loss function. Finally, model performance was evaluated on the validation set using various parameters.

RESULTS

Figure 2.A) shows an image of a micrograph where we overlaid the identified particles and false positives. Figure 2.B) shows the patches labeled False positives with the **highest high-frequency scores**. Each image tile corresponds to a spatial patch, and the value above each patch indicates its computed high-frequency score. Despite having high scores, these patches appear visually homogeneous and noise-dominated, lacking clear particle-like structure. There was no visual difference between False positives and particles, which we account for the poor performance of the machine learning model as revealed by score metrics in Table 1.

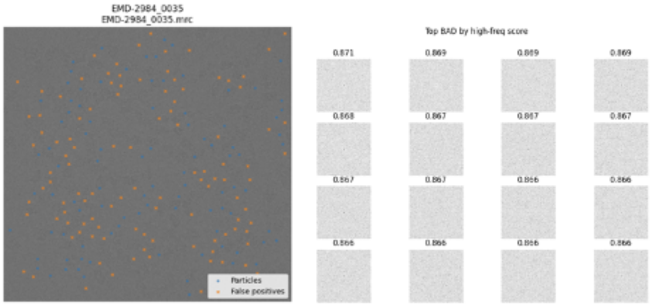


Figure 2. A) An image of a micrograph that shows the Particles (Blue) and False positives (Orange) B). Images of patches of False positives after computing FFT.

FUTURE WORK

Future work will explore larger patch sizes and more target use of information as extended ice signatures such as crystalline ice rings are only partially captured at smaller spatial scales (64-128 px). We also plan to include micrographs from different proteins to make the model more generalizable and to improve robustness to noisy labels . The final goal of the project is to extend the patch-based classifier into a spatial masking approach for automated ice region exclusion at the micrograph level.

Metric	Value
Accuracy	0.56
F1	0.33
Training loss	0.69

Table1. Score metrics for the evaluation of the model performance based on Test set of the CNN model

CONCLUSION

In this study, we investigated an automated approach to detect misleading ice artifacts in Cryo-EM using a CNN-based model. Although FFT-based features were included to highlight frequency domain ice signatures, the model showed limited performance due to the strong visual similarity between true particles and false positives. This work highlights the challenges and provides a foundation for improving model design in future studies.

REFERENCES

[1] B. Xu and L. Liu, Protein Science, vol. 29, no. 4, pp. 872–882, Apr. 2020, doi: 10.1002/pro.3805.

[2] Iudin A, Korir PK, Somasundharam S, Weyand S, Cattavitello C, 2023. “EMPIAR: the Electron Microscopy Public Image Archive.” Nucleic Acids Res., 51, D1503-D1511. https://doi.org/10.1093/nar/gkac1062.

[3] Dhakal A, Gyawali R, Wang L, Cheng J., 2023 doi: 10.1101/2023.02.21.529443.

[4] Amat F, Castaño-Diez D, Lawrence A, Moussavi F, Winkler H, Horowitz M. Enzymol. 2010 doi: 10.1016/S0076-6879(10)82014-2.

[5]R.J. Pally, S. Samadi. Environmental Modelling & Software. Volume 148, 2022 ,105285, https://doi.org/10.1016/j.envsoft.2021.105285.