CryoDiscovery (TM): A Machine Learning Platform for Automated Cryo-electron Microscopy 2D Class Selection

Narasimha Kumar[1]

[1]Health Technology Innovations

kumar@hti.ai

Structural Biology is an emerging critical area for disease research and drug discovery. This can be the basis for detecting novel biological threats and hence will help prepare the country's readiness. It should be noted that much of the SARS-CoV-2 virus that causes Covid-19 has been analyzed using structural biology. Cryogenic Microscopy (cryo-EM) is one of the most impactful and vital tools of biological structure analysis today. Single-particle cryo-EM produces images of individual particles, and has the potential to analyze the biological structures at the single-molecule level. However, the images generated by cryo-EM are highly noisy, and hence it takes a significant number of compute-intense steps [1] to process them to build recognizable 3D structures (proteins and others). Some of the steps must be repeated to filter out noisy data. Some of the steps require manual input, especially the 2D and 3D classes, which result in classification errors due to user bias, time waiting, and user fatigue. HTI proposes to address these problems with CryoDiscovery to eliminate or minimize the manual steps to improve productivity and accuracy. The role of CryoDiscovery is highlighted in the simplified workflow diagram (Figure 1). CryoDiscovery is designed by Health Technology Innovations Inc (https://hti.ai/), the recipient of the NSF grant [2]. CryoDiscovery uses AI/ML techniques for automated class selection. To achieve generation and the use of the AI/ML model and to be compatible with the workflow, it has: A training module to generate the AI/ML model. A module that uses that AI/ML model to infer whether images are good or bad during a workflow. A module that translates that inference into generating the data for the subsequent steps. In the figure here, the CryoDiscovery components are inside the dashed box, in the simplified CryoDiscovery diagram (Figure 2). CryoDiscovery uses Convolutional Neural Network (CNN) [3] implementation. CNN, a layered network, iteratively extracting features for classifying the class average images. After training with a few thousand images, the Accuracy and False Negative graph (Figure 3) showed very good results. The network is designed to be retrained, if needed, with few labeled images (often, less than a dozen). In this talk, we will discuss the results. The model was verified quantitatively with the Fourier Shell Correlation-Resolution graph (FSC graph) [4], and qualitatively by comparing against the published structures. We will look at the analysis of datasets from the EMPIAR site [5]. Relion [6] and CryoSPARC™ [7] were used in the flow, along with CryoDiscovery. We will use these examples to demonstrate the viability of CryoDiscovery. 1)Empiar 10204 dataset: beta-galactosidase: CryoDiscovery was used in the Relion example workflow for 2D Class Selection. The FSC graph shows the results from Relion's pre-calculated flow and with CryoDiscovery. As could be noted in Figure 4, it is very close. 2)Empiar 10256: TRPV5: As in the previous test, CryoDiscovery was used in the Relion workflow. The 2D classification was repeated thrice. The image generated in the 3D stage was compared with the published one in the EMDB [7] in Figure 5. 3)Empiar 10295: Clathrin Cages: CryoDiscovery was used in the Relion workflow. In this case, we needed to retrain the model. 2D Classification was repeated thrice. The 3D image was compared against the one in EMDB in Figure 5. 4)Empiar 10025: T20S: CryoDiscovery was used with CryoSPARC workflow, with the 2D Classification step repeated thrice. The FSC graph was used for assessment. Resolution with CryoDiscovery was 2.9Å vs 2.8Å in the data provided by CryoSPARC.(Figure 6) 5)Empiar 10256: Clathrin Cages: This time this dataset was analyzed with CryoDiscovery in the CryoSPARC workflow. The FSC graph showed even a better resolution with CryoDiscovery (3.1Å vs 3.3Å). (Figure 7) Along with the accuracy analysis using the above public datasets, we estimated the productivity improvements. CryoDiscovery eliminates/reduces the wait times (waiting for user input), and with the estimates provided (based on the availability of the user as needed), we could accomplish an ~2X improvement in the processing duration (Figure 8). This work was done with the help of many: Dr. Craig Yoshioka at OHSU, Dr. Justin Kollman at the University of Wshington, and the staff at HTI inc. We acknowledge the grants from the State of Oregon, and the National Science Foundation.

[1]https://academic.oup.com/jmicro/article/65/1/57/2579723

[2] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1939142&HistoricalAwards=false

[3] https://developers.google.com/machine-learning/practica/image-classification/convolutional-neural-networks

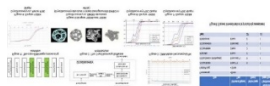[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3165049/

[5] https://www.ebi.ac.uk/pdbe/emdb/empiar/

[6] https://www2.mrc-lmb.cam.ac.uk/groups/scheres/impact.html

[7] https://cryosparc.com,

Figure 1