doi:10.1017/S1431927622005207
Microsc. Microanal. 28 (Suppl 1), 2022

CryoDiscovery™: Public Data Based AI/ML Model Enhancements for Cryo-Electron Microscopy Image Analysis

Ryan Dehart[1], Elliot Gray[1], Narasimha Kumar[1*] and Hitha Yeccaluri[1].

[1] HTI Inc., Portland Oregon, United States.
* Corresponding author: kumar@hti.ai

Structural Biology is an emerging critical area for disease research and drug discovery. Cryogenic Microscopy (cryo-EM) is one of the most impactful and vital tools of biological structure analysis today. Due to its importance, Cryo-EM leaders were awarded the 2017 Nobel Prize in Chemistry [1].

We, HTI Inc, have presented our approaches to applying AI/ML for automated 2D class selection in the M&M 2020 first, and improvements and further results in M&M 2021 in our project CryoDiscovery™. AI/ML helps in improving the accuracy, reducing bias, and significantly improving ease of use and hence, productivity.

Our work was supported by NSF Phase I award and the grants from the State of Oregon. We have now received NSF Phase II award to complete the work and productize.

This year, we present our development of a tool to reduce the need for human intervention in Cryo-EM single-particle analysis. Our tool automatically distinguishes "good" from "bad" candidates from among 2D class-average images that are produced by the 2D classification step typically performed during single-particle reconstruction [2]. Ordinarily, human operators can select from among these 2D class-average particles as a way to remove noisy or artifactual data that is not easily removed during prior data processing steps, and in our experience represents a significant labor cost and potential for human bias. Our model is a Convolutional Neural Network and is trained using annotated data comprising three different proteins collected in partnership with the Oregon Health & Science University Cryo-EM center. Our model can distinguish good from bad images with receiver-operating-characteristic and average-precision scores exceeding 0.95, and further analysis of our model's mistakes reveals potential human annotation errors, suggesting the true model performance could be as good or better than human-level. The key learning from the current work is that for the structural analysis of proteins (and other particles), the model must be trained continually with current and diverse data to make it more efficient in detection.

In the last few years, the proven 3-D structures of many of the proteins have been published in public repositories, PDB [3] and EMDB [4]. (There are several thousand entries in PDB and EMDB currently). This trend will continue as protein structures are determined, peer-reviewed and published.

HTI will use these published protein structures to derive 2D projections at multiple angles, using recently available tools. Using those images, we will add a varying amount of noise to simulate the noisy data as seen in microscope images. This will improve the training as the data would cover multiple proteins that cannot be obtained in any one center.

As can be seen in the figures below, there are methods to generate the 2D projections. With those projects, we will add a varying amount of noise to make the image "more real".

For example, a set of tools for 2D extraction has been published in Scipion [5][6] and its extensions, which are open-source projects. The images in Figure 2 were derived using this tool.

With multiple 2D projections of some good set of entries, and applying data augmentation, 100X (of the original Training data) is within reach! With over 2 million training images, an additional challenge would be the computing platform. Computing platforms are being improved and enhanced by leaders such as Intel® and NVIDIA® to address large training datasets. HTI is part of Intel and NVIDIA's program for startups and will have access to the cutting-edge platforms and software to implement this level of training.

To measure the effectiveness of the training, standard AI/ML metrics, such as Precision and Recall, will be used. In addition, the Biochemistry metrics, such as the Fourier Shell Correlation to Resolution, will be used as well.

This talk will detail the methods, tools, and options considered and used.

CryoDiscovery is designed by Health Technology Innovations Inc (HTI) (https://hti.ai/), a startup company in Portland, Oregon [6].

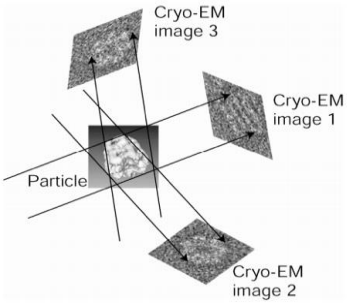

Figure 1. 2D projections from the 3D images in the public repositories. The "Cryo-EM Images" are the 2D images generated by the projection tools at different projection angles.

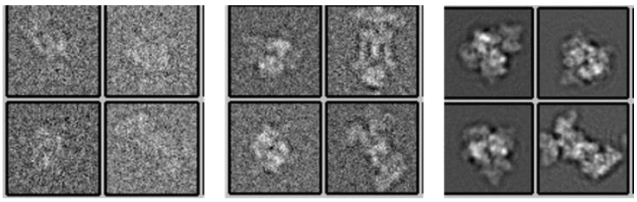

Figure 2**.** An example: Generated 2D projections from the PDB entry *6eti* (*inhibitor-bound ABCG2*) with noise added. The first set of images is generated with an SNR of 0.1 (i.e., lot of noise), the second set with 0.5, and the last set with 50 (i.e., almost no noise).

References:
[1] https://www.nobelprize.org/prizes/chemistry/2017/press-release/
[2] Chung, SC., Lin, HH., Niu, PY. et al. Pre-pro is a fast pre-processor for single-particle cryo-EM by enhancing 2D classification. Commun Biol 3, 508 (2020).
[3] PDB: https://www.rcsb.org/
[4] EMDB: https://www.ebi.ac.uk/emdb/

https://doi.org/10.1017/S1431927622005207 Published online by Cambridge University Press

[5] Scipion: http://scipion.i2pc.es/
[6] The authors acknowledge help and directions from Dr. Slavica Janic, CNRS, Franc

https://doi.org/10.1017/S1431927622005207 Published online by Cambridge University Press