doi:10.1017/S1431927620021145
Microsc. Microanal. 26 (Suppl 2), 2020

CryoDiscovery^{TM}: A Machine Learning Platform for Automated Cryo-electron Microscopy Particle Classification

Narasimha Kumar[1], John Harkness[2], Craig Yoshioka[3], Shiva Aditham[1], Tuan Phamdo[1] and Kennedy Brown[1]

[1]Health Technology Innovations, Portland, Oregon, United States, [2]Rewire Neuroscience, Portland, Oregon, United States, [3]Oregon Health & Science University, Portland, Oregon, United States

Cryogenic electron microscopy (Cryo-EM) produces high-resolution 3D images at angstrom levels used by researchers across a broad range of fields including structural biology, life Science, materials science, nanotechnology, semiconductors, energy, environmental science, and food science. Advancements in microscopy hardware enable production of 2D and 3D micrographs with near sub-Angstrom resolution, but require exponentially increasing data processing and storage capability. Images generated by cryo-EM are visually noisy, and each project can produce more than 100,000 images and take weeks to arrive at one viewable 3D structure. Many steps in the cryo-EM workflow require manual intervention and analysis that can take several weeks and result in errors due to user bias, time waiting and user fatigue. Current image processing and data analysis solutions are not well-integrated, requiring extensive manual user involvement and long wait times before assessing image quality. Here we describe our development of machine learning models for automation of single particle classification during cryo-EM image processing with repeatable accuracy levels and integrated into the cryo-EM workflow for easy deployment with a new machine learning platform, called CryoDiscovery.

We tested several Convolution Neural Network (CNN) designs for ML training and inference using a private set of over 20,000 images and metadata files. CNN architectural considerations include network depth, activation function and hyperparameters. Our CNN processed image data via a layered approach, iteratively through repeated transformations (in the "hidden" layers) to extract features before classifying them (in the "output" layer) 2-D and 3-D class selection. CNN models were trained using image data found in *mrcs* files, non-image metadata found in *star* files, and image annotations (*ground truth)* found in *selection* files using a computer with a dual socket 2nd gen Intel Xeon® CPU (8 cores each) with 4 NVIDIA 2070-Ti GPUs and 96GB of physical memory. Data preparation was conducted by trained researchers prior to ML training, and consisted of image retrieval, resolution normalization, image augmentation, and non-image data selection. Verification of our models was done by analyzing maximum prediction accuracy with low variance, and false negatives to minimize misclassification of good data, and the impact of using non-image data to improve model prediction accuracy. Model boosting was used to generate strong prediction algorithms and more consistent results from multiple simple models [1]. Three models were trained sequentially and used for inferencing, as shown in Figure 1. The third model was used as a tie-breaker when results from the first two models disagreed on the classification of the results.

Fourier Shell Correlation vs. Resolution (1/Aº) [2] was used to verify that the resolution (at threshold) meets published results. Secondly, we calculated the Mean Square Error (MSE) of 2D predicted images vs. ground truth images to provide a leading indicator of 3D model differences. Lastly, we examine Structural Similarity Index (SSIM) for structure level comparison [3 & 4]. Our prediction results reached over 90% accuracy with only a 3% false negative rate (Figure 2). *These image processing steps (2D classification, 3D Init model and classification, & 3D Refinement) took only hours to complete with our*

system. In order to verify the model with larger datasets, we verified our ML inference results using publicly available datasets, such as EMPIAR, and other private and public datasets.

The objective of this research is to produce a software tool that consistently classifies particles with a high-level of accuracy and is easily integrated into the cryo-EM workflow. The approach will be to increase the training and validation datasets from a wide range of users and particle types (research labs, proteins, etc.), utilize existing convolutional neural network frameworks and develop new techniques running experiments to optimize the models, integrate the prototype into established cryo-EM workflows for end-to-end processing, and produce a delivery method for easy deployment. The expected results will improve accuracy and productivity reducing the time to produce cryo-EM 3D structures from weeks to hours.

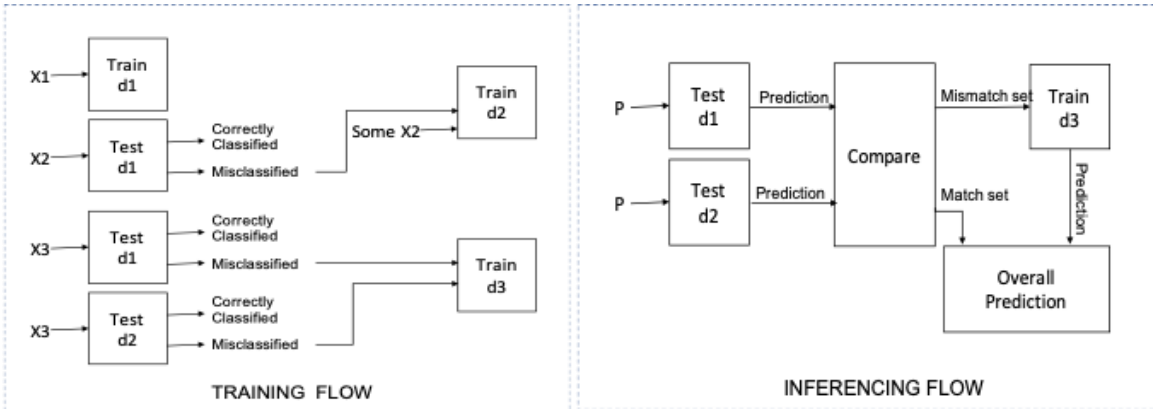

Figure 1. Boosting allows CryoDiscovery to use multiple models to arrive at more consistent classification results. Models are trained sequentially, but only two (d1 and d2) are used for initial classification inference. The third model, d3, is used for tie-breaking when d1 and d2 disagree on classification results.

Figure 2. Preliminary results from training ML models indicate high accuracy of classification over 60 training epochs.

https://doi.org/10.1017/S1431927620021145 Published online by Cambridge University Press

References

1. Schapire, R., *A Brief Introduction to Boosting.* Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999.

2. Liao, H.Y. and J. Frank, *Definition and estimation of resolution in single-particle reconstructions.* Structure, 2010. **18**(7): p. 768-775.

3. Ndajah, P., et al. *SSIM image quality metric for denoised images*. in *Proc. 3rd WSEAS Int. Conf. on Visualization, Imaging and Simulation*. 2010.

4. Brunet, D., *A study of the structural similarity image quality measure with applications to image processing.* 2012.

https://doi.org/10.1017/S1431927620021145 Published online by Cambridge University Press