doi:10.1017/S1431927622005335

CryoDiscovery™: Federated Learning as a Key Tool for Cryo-electron Microscopy Image Analysis

Elliot Gray[1], and Narasimha Kumar[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 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 State of Oregon. We have now received NSF Phase II award to complete the work and productize.

The key learning is that the for the structural analysis of proteins (and other particles), the model must be trained continually with more data to make it more efficient in detecting the structures of the diverse set. The bulk of the data is held privately by Pharma, Research Hospitals and Universities. Intellectual Property considerations and privacy laws prevent the availability to a central development site such as ours. Besides, the data is emerging at a rapid rate.

As it would not be practical to obtain the data in a central place for AI/ML Model development, HTI will use Federated Learning methods to continually train the model. CryoDiscovery will be among the earliest implementations of Federated Learning in the cryo-EM space.

Federated Learning: The concept of Federated Learning has emerged to address the privacy and security constraints and offer a distributed collaborative environment for training AI/ML models [2] [3] [4]. The aim is to have the different centers/sites participate in the collaborative training, without having to share the data.

This idea has been put into practice in hand-writing recognition algorithms for phones and keypad devices successfully [5] [6]. HTI will extend this to cryo-EM processing. Besides, the data for cryo-EM is huge and that makes it ideal for a distributed environment.

The goal is to collate the knowledge, not the data, from various sites/servers. This method, in addition to providing a means to coalesce the ML methods from multiple sessions, provides an established pathway for data isolation, security and privacy as shown in Figure 1. Only the model's parameters are sent. This allows the core model to be updated from the data owners/sites/servers continually and provide an updated model to all data owners with an improved model. To implement this architecture, and to guarantee data privacy and protection, models would have to be run within secure execution environments as shown in the Figure.

https://doi.org/10.1017/S1431927622005335 Published online by Cambridge University Press

The computer industry leaders, Intel® and NVIDIA®, are investing significantly to make platforms for Federated Learning widely available. These platforms provide support for native secure environments to implement this architecture. Both NVIDIA's FLARE [7] and Intel's OpenFL [8] are open-source Federated learning frameworks and gathering significant support. Both are supporting HTI and CryoDiscovery for this project.

This talk will detail the methods, tools, and options for model aggregation. CryoDiscovery will be among the first to adopt Federated Learning in the cryo-EM space.

CryoDiscovery is designed by Health Technology Innovations Inc (HTI) (https://hti.ai/), a startup company in Portland, Oregon [9].

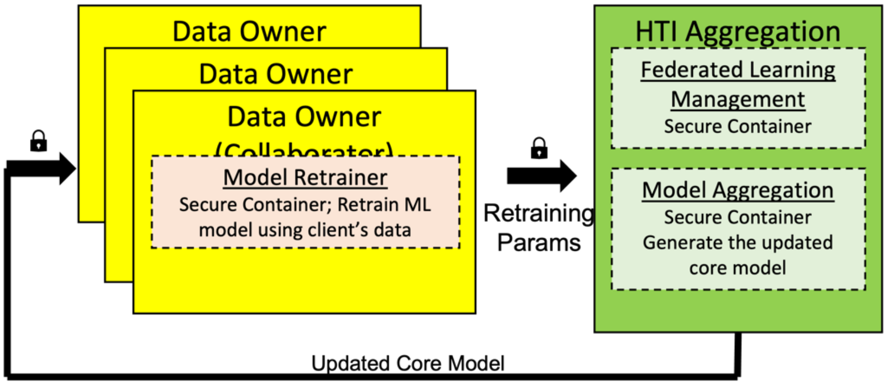

Figure 1. Federated Learning Example with 3 Collaborators. Dashed boxes represent secure execution containers. Data between Collaborators and Aggregation Server encrypted

References:

[1] https://www.nobelprize.org/prizes/chemistry/2017/press-release/
[2] Federated Learning: A survey. https://arxiv.org/pdf/1907.09693.pdf
[3] Early paper on Federated Learning: https://arxiv.org/pdf/1602.05629.pdf
[4] Detailed FL Paper: https://arxiv.org/pdf/1912.04977.pdf
[5] Google Handwriting Recognition Blog: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
[6] Google Keyboard Prediction: https://arxiv.org/pdf/1811.03604v2.pdf
[7] NVIDIA FLARE: https://developer.nvidia.com/flare
[8] G. Anthony Reina et al. OpenFL: https://arxiv.org/pdf/2105.06413.pdf
[9] The authors acknowledge help from and discussions with the NVIDIA FLARE team and the Intel FL team

https://doi.org/10.1017/S1431927622005335 Published online by Cambridge University Press