

Notice: This manuscript has been authored by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE522 AC05-00OR22725. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for the United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<https://www.energy.gov/doi-public-access-plan>).

Towards FAIR Workflows for Federated Experimental Sciences

Gayathri Saranathan
Hewlett Packard Labs
Hewlett Packard Enterprise
Singapore
gayathri.saranathan@hpe.com

Foltin Martin
Hewlett Packard Labs
Hewlett Packard Enterprise
Fort Collins, USA
martin.foltin@hpe.com

Aalap Tripathy
Hewlett Packard Labs
Hewlett Packard Enterprise
Austin, USA
aalap.tripathy@hpe.com

Annmary Justine
Hewlett Packard Labs
Hewlett Packard Enterprise
Fort Collins, USA
annmary.roy@hpe.com

Maxim Ziatdinov⁺
Physical Sciences Division,
Pacific Northwest National Lab,
Richland, WA 99352, USA
maxim.ziatdinov@pnnl.gov

Ayana Ghosh⁺
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
ghosha@ornl.gov

Kevin Roccapriore⁺
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
roccapriorkm@ornl.gov

Suparna Bhattacharya
Hewlett Packard Labs
Hewlett Packard Enterprise
Bangalore, India
suparna.bhattacharya@hpe.com

Paolo Faraboschi
Hewlett Packard Labs
Hewlett Packard Enterprise
San Jose, USA
paolo.faraboschi@hpe.com

Abstract—A de-centralized, peer-to-peer AI metadata framework is demonstrated which can enable end-to-end metadata & lineage tracking for distributed Machine Learning pipelines spanning edge, High Performance Computing, and cloud environments. With a specific example of end-to-end microscopy algorithm and datasets, the proposed method shows how to enable reproducibility, audit trail, provenance of metadata artifacts. The emerging needs of automation in experimental sciences, ML-centric workflows, and FAIR metadata management across federated compute environments is addressed.

Keywords—HPC, FAIR, AI, ML, DKL

I. INTRODUCTION

Modern AI enabled experimental science research involves intricate workflows that span various experimental and computational facilities. In microscopy, Researchers no longer operate within the confines of a single microscope; instead, they collaborate within extensive instrumentation networks across user facilities. The challenge lies not in the isolated creation or deployment of these workflows but in the seamless integration and optimization. This challenge is underscored by the requirements outlined in RFPs [1], where some vendors are called upon to provide software tools that adhere to FAIR principles (Findable, Accessible, Interoperable, and Reusable) for the management and preservation of scientific data [2,3].

While the data acquisition may be done in ~milliseconds, data transfer is performed ad-hoc in minutes and the scientist/researcher spends weeks-months in iterating over experimental parameters at the instrument and in the algorithm used in their AI surrogate or physical model. In this paper describes how end-to-end metadata & lineage tracking is performed for distributed Machine Learning pipelines spanning instrumentation edge, High Performance Computing & cloud environments using a framework such as Federated CMF [4]. This paper addresses the challenges in real-world scanning-transmission electron microscopy data and workflow involving Deep Kernel learning algorithm [5-8].

II. AUTONOMOUS MICROSCOPY WORKFLOW WITH COMMON METADATA FRAMEWORK

In [8], the study presents the Deep Kernel Learning (DKL) algorithm, which combines Gaussian process-based Bayesian optimization with a feedforward neural network. Its primary goal is to uncover material structure-property relationships using microscope data. The DKL model efficiently trains on a limited dataset from material samples, enabling it to predict values and uncertainties for unexplored points. This aids in selecting high-uncertainty points by active learning, for further examination, traditionally done by experienced microscopy experts. This approach saves time and avoids sample deterioration. Previous research confirms that even sparse random sampling, as low as ~1% of points, suffices for accurate predictions by DKL. This closed loop workflow (in Fig. 1) involves experiment, pre-processing, selection, training, inference, and steering stages. While requiring a GPU, this loop is computationally appropriate to execute on accelerator-equipped edge compute systems.

Figure 1: Federated CMF applied to an Active Learning Workflow

However, during development the experimentation for feasibility and correctness is typically done in a HPC system (with a cluster of compute nodes). The roadmap for future evolution of this workflow involves the coupling of a deterministic molecular dynamics simulation (run in an HPC system) to make real-time predictions of materials properties [5] which is an additional input to the experimental steering algorithm (not discussed further in this submission).

Federated CMF involves a python client-side library to integrate with the workflow. Any participating compute system can stand-up a CMF server to enable syncing of metadata from multiple CMF clients. CMF indexes a unique

identifier (UUID) of a metadata artifact to the version of code it is produced from (GitHub commit-id reference). This enables the reconstruction of lineage and version history of any artifact produced from a pipeline. Since pipeline names are unique, it is possible to merge lineage and version histories on any CMF Server instance from subscribing clients (with cmf push/pull semantics). This also enables merging of execution and artifact history between server instances. The paper incorporated the CMF client-side into the DKL workflow which enabled capturing the instrumentation parameters, subset of data used during data selection, model artifacts produced and hyperparameters used during training, and uncertainty/prediction captured during the inference.

III. LINEAGE TRACKING IN DKL EXPERIMENT STAGES

Federated CMF is instrumented with the DKL - Spectral Reconstruction and Active Learning method. This method includes three experiments: Full Dataset Training, Partially Sampled Data Training (1%, 5%, and 10% of the full dataset), and Active Learning. Each experiment has Data Preprocessing, Training, and Inference stages. Unlike standard model training, Active Learning is an iterative process, and its metadata includes data selection, labeling, training metrics, inference model snapshots at each stage, and the chosen data for the next exploration step. Tracking this metadata enables data-driven decisions, effective progress monitoring, and reusability. Fig. 2 depicts the metadata lineage tracking for the DKL workflow in these experiments.



Figure 2: Lineage Diagram for DKL workflow generated by CMF

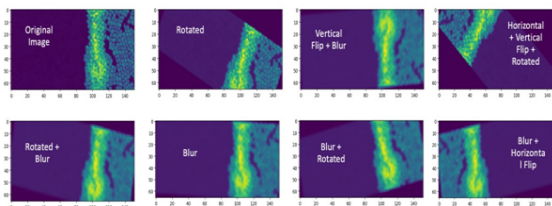


Figure 3: Task Augmentation [9]

A. Full Dataset Training

It begins with the input dataset, which includes plasmonic microscope images and their known spectra. The data preprocessing stage converts the input dataset into patches and their associated spectra. The data is then split into training and testing sets. Training is performed on the train set, generating a train loss and a trained model. The model is reused in the inference stage with the test dataset, producing model predictions and a test loss. The CMF tracks input to output at each stage as in the highlighted Full training experiment. The augmented Dataset is used to perform the training for

robustness as there are only minimal samples available, as given in Figure 3.

B. 10% Data Training

The processed input undergo sampling of random 10% selection to validate the model performance with minimal data. After sampling, the training and inference step is similar to the full dataset training.

C. Active Learning

1% of input patches are randomly selected for initial training. The model then iteratively selects (Data Selection) and labels data to improve performance, which feeds into the Model Re-training stage. During multiple exploration steps, the model is trained, and at each step, a trained model is produced. CMF systematically captures metadata, enabling HPC to edge systems accessibility. This tracking of model stages supports continuous training from different edge. The model is also reusable for direct inference or fine-tuning. The system's interoperability allows input data from various experiment sites, with HPC handling model training, and active learning executed at microscopic locations, offering an efficient real-time solution.

IV. NEXT STEPS

This work is further extended to demonstrate the real-time use of the artefacts and metadata tracked with CMF on simulated data and use it to guide the simulation model without having to retrain from scratch.

V. ACKNOWLEDGMENTS

This research (A.G.) is sponsored by the INTERSECT Initiative as part of the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

REFERENCES

- [1] NERSC-10 Technical RFP: <https://www.nersc.gov/systems/nersc-10/draft-tech-req/>
- [2] Carole Goble, Sarah Cohen-Boulakia, Stian Soiland-Reyes, Daniel Garijo, Yolanda Gil, Michael R. Crusoe, Kristian Peters, Daniel Schober; FAIR Computational Workflows. *Data Intelligence* 2020; 2 (1-2): 108–121. doi: https://doi.org/10.1162/dint_a_00033 [2]
- [3] FAIR Principles: <https://www.go-fair.org/fair-principles/>
- [4] A. J. Koomthanam, S. Serebryakov, A. Tripathy, G. Nayak, M. Foltin and S. Bhattacharya, "Common Metadata Framework: Integrated Framework for Trustworthy AI Pipelines," in *IEEE Internet Computing*, doi: 10.1109/MIC.2024.3377170.
- [5] Bhowmik, Debsindhu, Mukherjee, Debangshu, Oxley, Mark, Ziatdinov, Maxim, Jesse, Stephen, Kalinin, Sergei V., and Ovchinnikova, Olga. 2021. "Building an edge computing infrastructure for rapid multi-dimensional electron microscopy". United States. <https://www.osti.gov/servlets/purl/1813209>.
- [6] Mukherjee, Debangshu, Kevin M. Roccapriore, Anees Al-Najjar, Ayana Ghosh, Jacob Hinkle, Andrew R. Lupini, Rama K. Vasudevan, Sergei V. Kalinin, Olga S. Ovchinnikova, Maxim A. Ziatdinov and Nageswara S. V. Rao. "A Roadmap for Edge Computing Enabled Automated Multidimensional Transmission Electron Microscopy." *Microscopy Today* 30 (2022): 10 - 19.
- [7] Al-Najjar, Anees, Nageswara S. V. Rao, Ramanan Sankaran, Maxim A. Ziatdinov, Debangshu Mukherjee, Olga S. Ovchinnikova, Kevin M. Roccapriore, Andrew R. Lupini and Sergei V. Kalinin. "Enabling Autonomous Electron Microscopy for Networked Computation and Steering." 2022 IEEE 18th International Conference on e-Science (e-Science) (2022): 267-277.
- [8] Roccapriore, Kevin M., Sergei V. Kalinin and Maxim A. Ziatdinov. "Physics Discovery in Nanoplasmonic Systems via Autonomous

Experiments in Scanning Transmission Electron Microscopy.”
Advanced Science 9 (2021): n. pag.

[9] Saranathan, Gayathri & Foltin, Martin & Tripathy, Aalap & Ziatdinov,
Maxim & Koomthanam, Ann & Bhattacharya, Suparna & Ghosh,

Ayana & Roccapriore, Kevin & Sukumar, Sreenivas Rangan &
Faraboschi, Paolo. (2023). Towards Rapid Autonomous Electron
Microscopy with Active Meta-Learning. 81-87.
10.1145/3624062.36260.