PROCEEDINGS OF SCIENCE

FAIR Principles for data and AI models in high energy physics research and education

Avik Roy[1,*]

National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign
1205 Clark St, Urbana, Illinois 61801, United States of America

E-mail: avroy@illinois.edu

In recent years, digital object management practices to support findability, accessibility, interoperability, and reusability (FAIR) have begun to be adopted across a number of data-intensive scientific disciplines. These digital objects include datasets, AI models, software, notebooks, workflows, documentation, etc. With the collective dataset at the Large Hadron Collider scheduled to reach the zettabyte scale by the end of 2032, the experimental particle physics community is looking at unprecedented data management challenges. It is expected that these grand challenges may be addressed by creating end-to-end AI frameworks that combine FAIR and AI-ready datasets, advances in AI, modern computing environments, and scientific data infrastructure. In this work, the FAIR4HEP collaboration explores the interpretation of FAIR principles in the context of data and AI models for experimental high energy physics research. We investigate metrics to quantify the FAIRness of experimental datasets and AI models, and provide open source notebooks to guide new users on the use of FAIR principles in practice.

41st International Conference on High Energy physics - ICHEP2022
6-13 July, 2022
Bologna, Italy

[1]on behalf of the FAIR4HEP collaboration: https://fair4hep.github.io
[*]Speaker

PoS(ICHEP2022)240

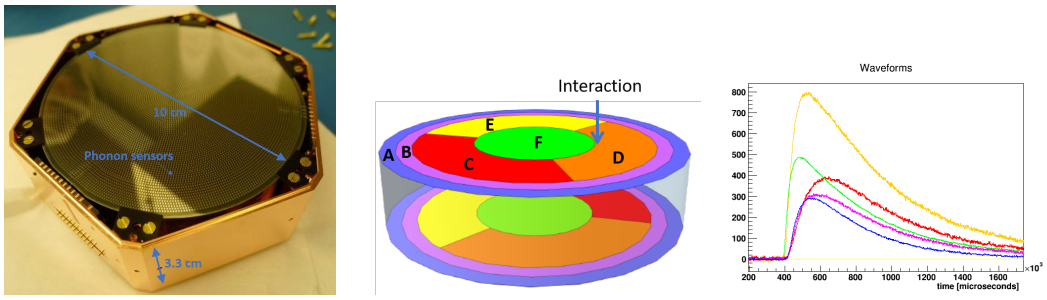
1. Introduction

In order to make datasets findabile, accessibile, interoperable, and reusable (FAIR), a set of data principles have been defined so that scientific datasets could be readily reused by both humans and machines [1]. Originally envisioned for preservation of scientific datasets, the FAIR principles have been interpreted in the context of different kinds of digital objects, including research software [2], notebooks [3], custom digital libraries [4], and machine learning (ML) models [5, 6]. This work summarizes the multifaceted interpretation and applications of the FAIR principles explored by the FAIR4HEP collaboration in the context of High Energy Physics (HEP).

2. FAIR principles for HEP data

Interpretation and application of FAIR principles for HEP datasets has been explored in Ref. [7]. Taking the publicly available $H \to b\bar{b}$ dataset from CMS Open Data portal, this work performs a thorough evaluation of the FAIR-readiness of this dataset using domain-agnostic metrics. Having the established the FAIRness of this dataset, this work demonstrates how adhering to these principles allow this dataset to be AI-ready. This dataset has been made available in formats widely used by the broader ML community. Pedagogical examples of usage of this dataset shows implementation of cutting edge methods in ML for jet classification. Guided by the FAIR principles, the FAIR4HEP collaboration has published multiple datasets for a variety of problems in HEP.

- **Super Cryogeinc Dark Matter Search (Super CDMS) dataset** [8] was obtained from a prototype detector (Figure 1a) for recording phonon scattering from an interacting particle. With six operating channels (Figure 1b), this detector operates at 30 mK and records phonon pulse amplitude as a function of time in each channel (Figure 1c). This dataset consists of timing information about these pulses for 7000 individual measurements taken over 13 different impact locations on the detector. This dataset is being used to reconstruct the impact location based on these timing information and to obtain generative models for simulation of particle intraction.



(a) (b) (c)

Figure 1: (a) A Super CDMS detector prototype and (b) illustration of its different channels. (c) shows the waveforms recorded by different channels for an example particle interaction. These images are taken from https://fair-umn.github.io/FAIR-UMN-CDMS/

- **Laser Response from Electromagnetic Calorimeter (ECal) crystals at CMS** [9] is obtained from the thousands of calibrations performed during the course of Run-2 between 2016 and 2018. The ECal calorimeter at CMS consists of almost 758k lead tungstate ($PbWO_4$)

crystals. Exposure to prolonged radiation during data taking modifies the transparency of these crystals, which is partially recovered when such exposure is taken away (Figure 2). This dataset contains information about the Ecal crystal's laser response and can be used to model this behavior as a function of integrated luminosity to predict future behavior.

3. FAIR Principles for AI Models

Application of the FAIR principles for AI models requires incorporating their dependence on computing environments into the interpretation of FAIR principles. We use the Interaction Network (IN) model [11] for classification of jets from the aforementioned $H \to b\bar{b}$ dataset. Figure 3 shows the IN model architecture and tabulates the default hyperparameters and data dimensions. This network is built to train on graph data structure whose nodes comprise of N_p particle tracks, each with P features, and N_v secondary vertices, each with S features, associated with the jet. The physical description of each feature is given in Appendix C of ref. [11]. It creates a fully connected directed graph with $N_{pp} = N_p(N_p - 1)$ edges for the particle tracks. A separate graph with $N_{vp} = N_v N_p$ generates all possible connections between the particle tracks and the secondary vertices.

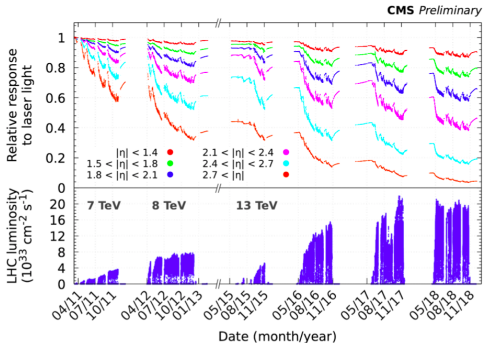
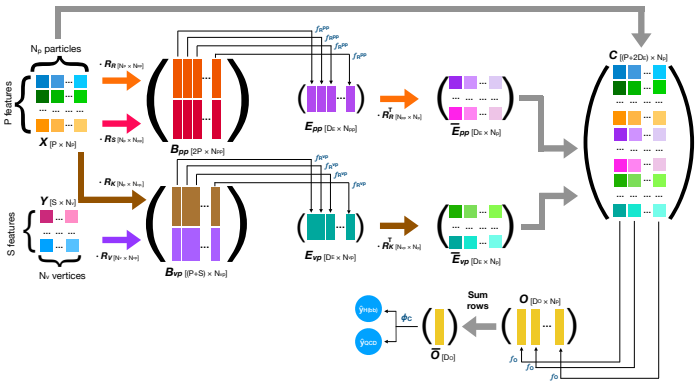


Figure 2: Transparency of CMS ECal crystals in different regions of pseudorapidity (η) during Run-2. This image is taken from Ref. [10]



Default IN hyperparameters	
(P, N_p, S, N_v)	$(30, 60, 14, 5)$
Hidden layers	3
Hidden layer dimension	60
(D_e, D_o)	$(20, 24)$
Activation	ReLU

Figure 3: A Schematic diagram of the network architecture and dataflow in the IN model. This image is taken from Ref. [11]. The choice of model hyperparameters and input data dimensions for the baseline model is given in the accompanying table.

The node level features for the track-track (track-vertex) graph are transformed to edge level features via a couple of interaction matrices, transformed via fully connected NNs, and finally transferred back to the track-level representations by the interaction matrices. The trainable dense MLP creates the post-interaction internal representation that are summed over the tracks and then

linearly combined to produce a two-dimensional, softmax-transformed output to predict individual class probabilities.

3.1 Interoperability of the IN mode

To inspect that the model can infer under multiple deep learning frameworks, we convert the IN model to `ONNX` [12] format and then to `TensorRT` [13] engine from the `ONNX` format model. The `ONNX` format is designed to help the model become portable between the commonly used deep learning formats like `PyTorch, TensorRT, TensorFlow` etc. We also quantitatively evaluate the inference performance before and after the conversion by measuring inference accuracy, running time, and Area under the Curve (AUC) for the Receiver Operating Characteristic (ROC) curve.

We separately infer three test groups using the `PyTorch` and `ONNX` model with a batch size of 128. With all three methods of inference, the model's performance in terms of accurancy and the ROC-AUC score was identical. When comparing the average time needed for inference per batch, the `PyTorch` model took 1.226 ms, while inference with `ONNX` and `TensorRT` models took 12.64 ans 12.31 ms respectively. This order of magnitude difference can be understood from the data transfer between the GPU device hosting these models and the actual host of the data. The running time/epoch refers to the time used to run one batch, including the data transfer between the two sides (device, host) and the inference part in the GPU device. When we increase the batch size from 1 to 128, the running time becomes larger, because the time used for data transfer also increases. The running time difference between the `ONNX` and `TensorRT` models can be attributed to the differences in hardware acceleration.

Here we also test whether different batch size will influence the inference results using the model. We record the performance of the GPU utility and throughput, i.e the number of inferences performed per second, with the change of batch size in `TensorRT` format. The model's performance was stable across batch sizes. Figure 4 shows the results. GPU utility steadily increases up to a batch size of 1000 and saturates at 100% beyind that. The performance of throughput also matches with the GPU utility, it reaches maximum when batch size reaches around 1200, which is the almost same period when the GPU utility is fully occupied.

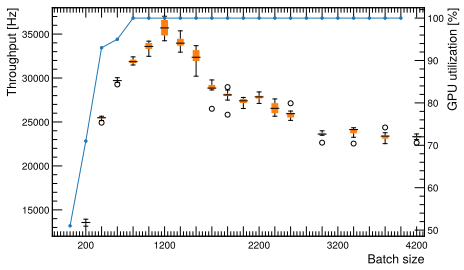


Figure 4: GPU utility and throughput for inference with `TensorRT` engine with different batch size

3.2 Streamlining FAIR AI Development

As AI models are often developed as an intricate software repository, an automation tool to allow a streamlined organization of such repositories has been developed. Cookiecutter Data Science [14] provides one such template, specifically oriented at data science projects. It consists of a logical, reasonably standardized, but flexible project structure hosted on GitHub for performing and sharing data science work. We took inspiration from this and created a fork of this template generator, called `cookiecutter4fair` [15], with additional

features to promote the adoption of our FAIR principles. It allows structural organization of machine learning model repositories, development of containers, and publishing the codebase with persistent identifiers like digital object identifier (DOI). It includes comprehensive instructions to build the necessary environments and allows coherently incorporating notebooks for various studies. This tool has been used to restructure the codebase for the IN model [16] and include notebooks that facilitate studies on its interoperbility and explainability.

4. Pedagogical aspects of FAIR in HEP

FAIR principles can facilitate education in data science and machine learning. Developing classroom contents aligned with the FAIR principles can amplify their reach and effectiveness. On the other hand, introducing students with FAIR principles and their interpretation for data and ML can help standardize modern practices in digital object management.

- **Pedagogical Introduction to FAIR**[1] is a series of notebooks developed based on the previously mentioned Super CDMS dataset. It contains an introduction to FAIR principles, evaluation of FAIRness of datasets, and builds up ML models, both traditional and neural network based, and demonstrates their organization and management in line with the FAIR principles. These notebooks include the analysis of the impact location estimation problem with the CDMS dataset using regularized linear regression, principal component analysis, deep neural netowrks, and variational autoencoders.
- **Data Science for Physics**[2] is a course designed to introduce modern concepts of machine learning via data analysis problems in HEP. This course was originally developed as an online substitute to the junior lab at MIT during the COVID pandemic and will be offered as a full, independent class starting from Spring 2023. It includes as series of lectures on different topics of physics, statistics, and data analysis and requires analysis of physics data from LIGO, CMS, and CHIME experiments. An open source version of this course will also be featured on the MITx[3] platform.

5. Conclusion

With a view to inspiring the modern community-wide standards for preservation and management of digital objects, FAIR4HEP is developing HEP-specific interpretation of FAIR and active implementation by developing FAIR datasets, models, and tools. This work summarizes the ongoing efforts of the project along these ventures.

Acknowledgements

This work was supported by the FAIR Data program of the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research, under contract number DE-SC0021258.

[1] https://github.com/yorkiva/FAIR-Exercises
[2] https://github.com/MIT-8s50/course
[3] https://mitxonline.mit.edu

References

[1] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak et al., *The FAIR guiding principles for scientific data management and stewardship*, *Scientific data* **3** (2016) 1.

[2] M. Barker, N.P. Chue Hong, D.S. Katz, A.-L. Lamprecht, C. Martinez-Ortiz, F. Psomopoulos et al., *Introducing the fair principles for research software*, *Scientific Data* **9** (2022) 1.

[3] R.A. Richardson, R. Celebi, S. Van Der Burg, D. Smits, L. Ridder, M. Dumontier et al., *User-friendly composition of FAIR workflows in a notebook environment*, *Proceedings of the 11th on Knowledge Capture Conference* (2021) 1.

[4] M.S. Neubauer, A. Roy and Z. Wang, *Making digital objects fair in high energy physics: An implementation for universal feynrules output (ufo) models*, *arXiv preprint arXiv:2209.09752* (2022) .

[5] D.S. Katz, F.E. Psomopoulos and L.J. Castro, *Working towards understanding the role of FAIR for machine learning*, *Proceedings of the 2nd Workshop on Data and research objects management for Linked Open Science* (2021) 1.

[6] N. Ravi, P. Chaturvedi, E. Huerta, Z. Liu, R. Chard, A. Scourtas et al., *Fair principles for ai models with a practical application for accelerated high energy diffraction microscopy*, *Scientific Data* **9** (2022) 1.

[7] Y. Chen, E. Huerta, J. Duarte, P. Harris, D.S. Katz, M.S. Neubauer et al., *A FAIR and AI-ready Higgs boson decay dataset*, *Scientific Data* **9** (2022) 1.

[8] M. Fritts and T. Li, *CDMS-dataset*, 2021. 10.34740/kaggle/dsv/2660709.

[9] B. Joshi and R. Rusack, *Laser response in ECAL crystals in CMS detector*, Mar, 2022. 10.5281/zenodo.6394777.

[10] CMS collaboration, *CMS ECAL Response to Laser Light (CERN-CMS-DP-2019-005)*, 2019.

[11] E.A. Moreno, T.Q. Nguyen, J.-R. Vlimant, O. Cerri, H.B. Newman, A. Periwal et al., *Interaction networks for the identification of boosted $h\to b\,b$ decays*, *Physical Review D* **102** (2020) 012010.

[12] J. Bai, F. Lu, K. Zhang et al., "ONNX: Open Neural Network Exchange." `https://github.com/onnx/onnx`, 2019.

[13] NVIDIA, *TensorRT: https://developer.nvidia.com*, 2018.

[14] Driven Data, "Cookiecutter data science." `https://drivendata.github.io/cookiecutter-data-science/`, 2022.

[15] FAIR4HEP, *Cookiecutter4fair: v1.0.0*, 2022. 10.5281/zenodo.7306229.

[16] J.M. Duarte, B. Li, A. Roy and R. Zhu, *Hbb Interaction Network: v0.1.1*, Nov., 2022. 10.5281/zenodo.7305227.