FAIR Principles for data and AI models in HEP Research and Education

Avik Roy
University of Illinois at Urbana-Champaign

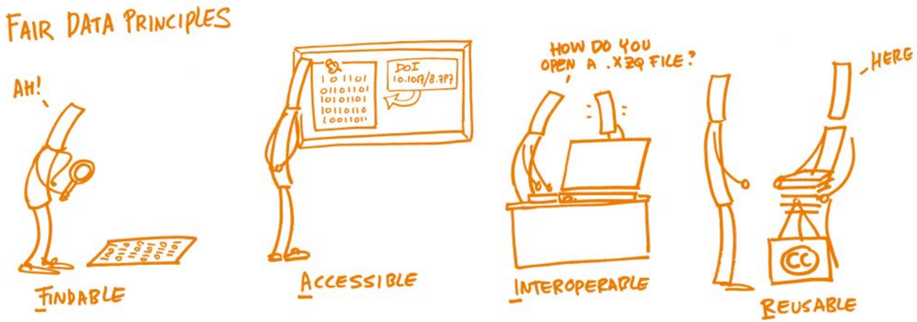
on behalf of the **FAIR4HEP** *project*

July 9, 2022

The FAIR Principles

- To inspire scientific data management for reproducibility and maximal reusability[1]
- Originally proposed for scientific data
- Can be interpreted as guidelines to manage and preserve other Digital Objects (DOs) e.g. research software[2], tutorials and notebooks[3], AI and ML models[4]
- Different working groups working on FAIR guidelines for different DOs (e.g FAIR4RS, FAIR workflows, FAIR VREs)

Findable:	locating DOs in a failsafe fashion
Accessible:	obtaining DOs along with their context, content, and format
Interoperable:	being usable across multiple computing platforms
Reusable:	specifying the context and extent of reusing DOs



FAIR4HEP: FAIR data and AI for HEP

- Multi-disciplinary, multi-institute team for learning how data-intensive HEP research can benefit from FAIR principles and vice versa
- Develop community standards to implement FAIR principles and tools to implement them
- Develop and share benchmark FAIR data and models
- Explore interplay between data and models to explore interpretability and model robustness

What makes data and AI FAIR for physicists?

How FAIR principles facilitate today's physics research?

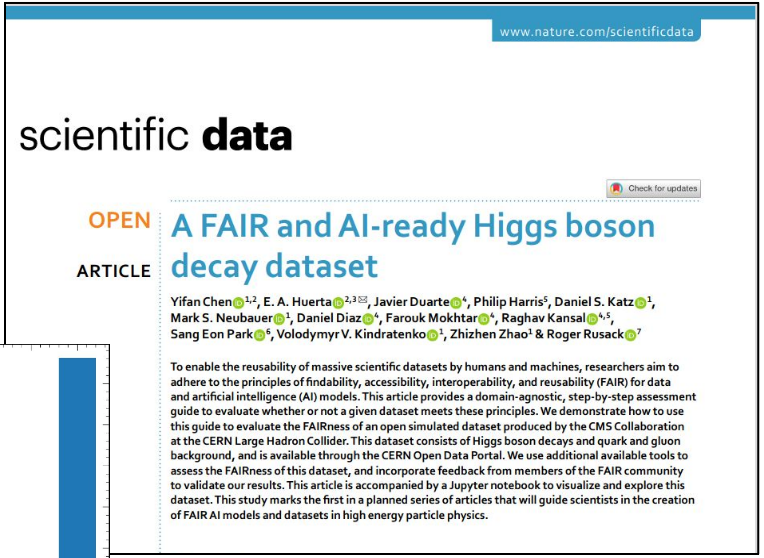
Are AI models in HEP robust?

How well do we understand AI models and their relationship with data?

visit us: https://fair4hep.github.io

Exploring the FAIR Principles for datasets in HEP[5]

- Explored the FAIR principles for the Hbb tagging dataset[6] from CMS open data
- Demonstrate a domain-agnostic evaluation of FAIR-readiness using community-standard tools
- Investigates AI-readiness of the dataset
  - Format of the data and its compatibility with popular ML libraries
  - Pedagogical examples of AI usage via notebooks[7]


www.nature.com/scientificdata

scientific **data**

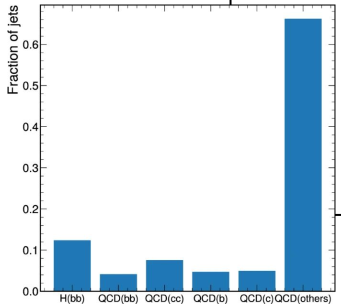
Check for updates

OPEN ARTICLE

A FAIR and AI-ready Higgs boson decay dataset

Yifan Chen[1,2], E. A. Huerta[2,3] ✉, Javier Duarte[4], Philip Harris[5], Daniel S. Katz[1], Mark S. Neubauer[1], Daniel Diaz[4], Farouk Mokhtar[4], Raghav Kansal[4,5], Sang Eon Park[6], Volodymyr V. Kindratenko[1], Zhizhen Zhao[1] & Roger Rusack[7]

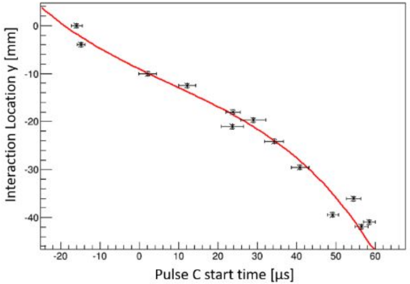
To enable the reusability of massive scientific datasets by humans and machines, researchers aim to adhere to the principles of findability, accessibility, interoperability, and reusability (FAIR) for data and artificial intelligence (AI) models. This article provides a domain-agnostic, step-by-step assessment guide to evaluate whether or not a given dataset meets these principles. We demonstrate how to use this guide to evaluate the FAIRness of an open simulated dataset produced by the CMS Collaboration at the CERN Large Hadron Collider. This dataset consists of Higgs boson decays and quark and gluon background, and is available through the CERN Open Data Portal. We use additional available tools to assess the FAIRness of this dataset, and incorporate feedback from members of the FAIR community to validate our results. This article is accompanied by a Jupyter notebook to visualize and explore this dataset. This study marks the first in a planned series of articles that will guide scientists in the creation of FAIR AI models and datasets in high energy particle physics.

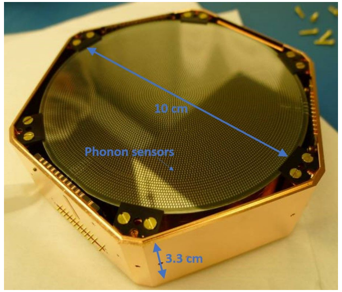


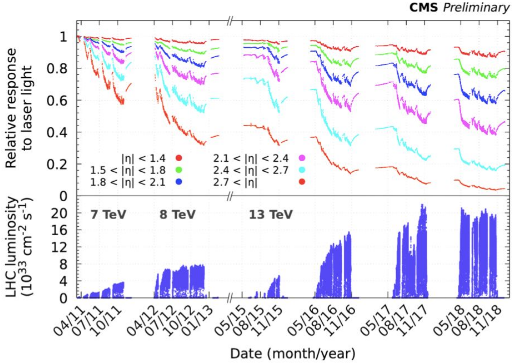
DOI: 10.1038/s41597-021-01109-0

FAIR Datasets from the FAIR4HEP Team

- Dataset from Super Cryogenic Dark Matter Search (SuperCDMS) detector prototype[8]
  - Detector operates at 30 mK and collects response to phonon flux via multiple channels
  - Dataset consists of timing information from detector's response to radioactive source excitation
- Laser Response from Electromagnetic Calorimeter (ECal) crystals at CMS[9]
  - Includes information about radiation damage to ECal crystals
  - Contains data from ~10k/year calibrations during Run 2

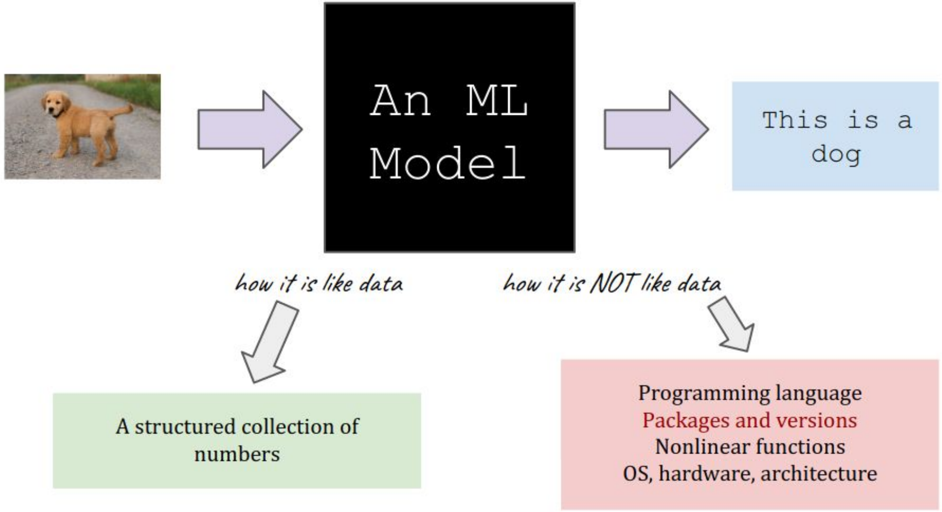






FAIR Principles for AI Models

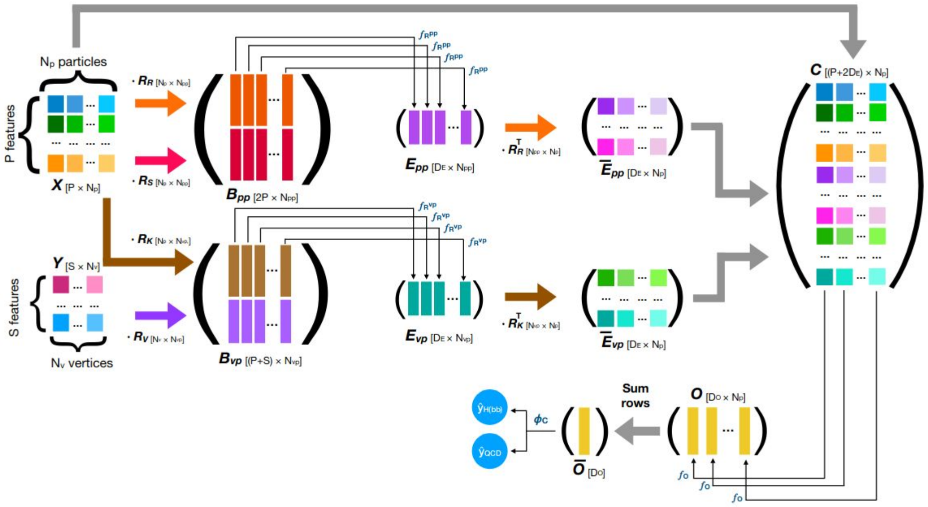
- Technical: Specific OS/software/package dependencies, availability of dataset and its provenance
- Analytical:
  - Using model inference for new/curtailed/similar datasets
  - Retraining model
  - Feature Engineering
  - Reoptimizing model hyperparameters



A Benchmark AI Model: The Interaction Network

- State of the art Graph Neural Network (GNN) Model trained to distinguish $H \rightarrow bb$ jets from QCD background
- Input to the model:
  - 60 particle tracks, 30 features per track
  - 5 secondary vertices, 14 features per vertex
  - Particle-particle and particle-vertex interaction matrices create an interaction network
  - Three MLP as transformation networks:
    - f_r : particle interaction
    - f_r^{pv} : particle-vertex interaction
    - f_o : pre-aggregator

Image from: [10.1103/PhysRevD.102.012010](https://doi.org/10.1103/PhysRevD.102.012010)

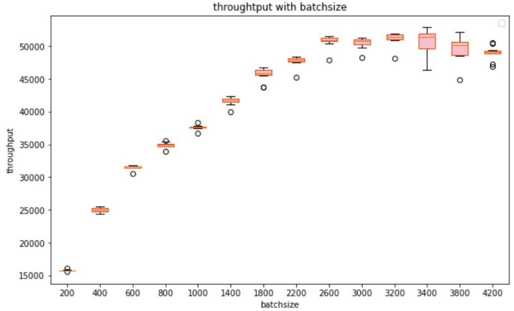


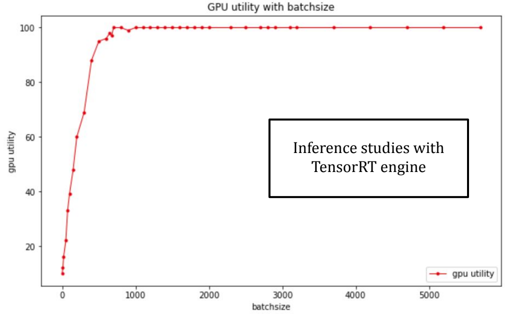
Study of Model Interoperability

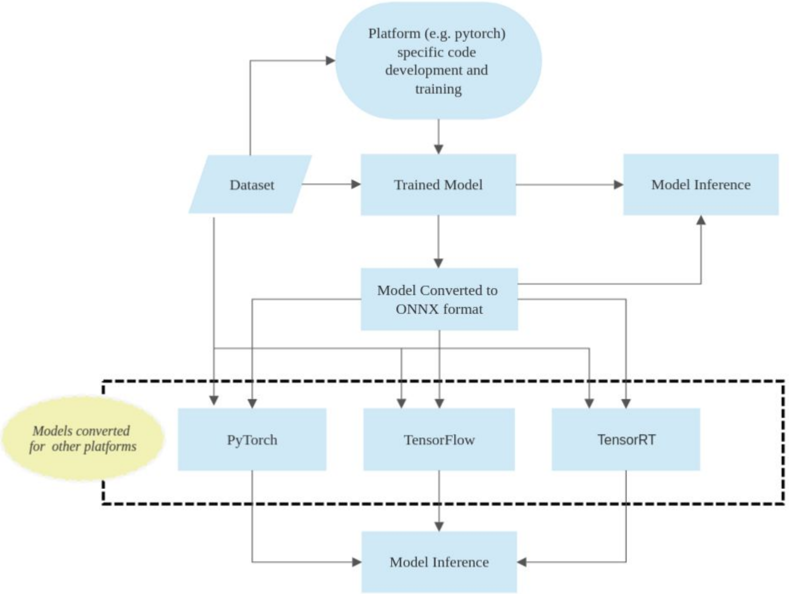
- How can we use a model with different machine architectures, OS, ML libraries?

ML Library/Metric	Accuracy (%)	time / epoch (ms)	AUC score (%)
PyTorch	97.4	1.1	99.1
Onnx	97.4	0.8	99.1
TensorRT	97.4	9.9	99.1

Studies with 10k events with a batchsize of 1







Streamlining FAIR AI Development: The Cookiecutter

- Automation tool to implement best practices in organizing the development of AI models
- Based on Cookiecutter Data Science: adapted for HEP
- Creates an organized template for code and data organization
  - Automated download of data
  - Incorporate notebooks for studies
  - Containerization tools
  - Comprehensive list of project dependencies to build environment- compatible with tools like `conda` and `pip`
- A FAIRified repository for the Interaction Network model is being developed (link)

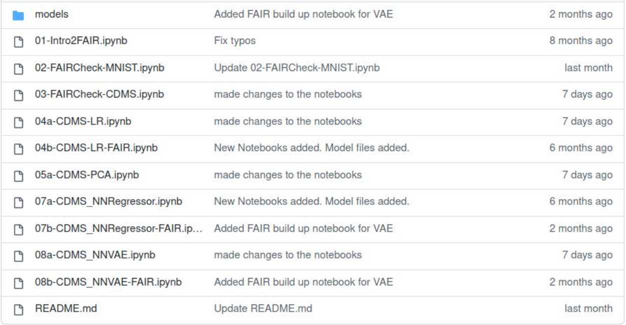
```
├── LICENSE
├── Makefile           <- Makefile with commands like `make data` or `make train`
├── README.md          <- The top-level README for developers using this project.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
├── docs               <- A default Sphinx project; see sphinx-doc.org for details
│
├── models             <- Trained and serialized models, model predictions, or model summar
│
├── notebooks          <- Jupyter notebooks. Naming convention is a number (for ordering),
│                         the creator's initials, and a short `-` delimited description, e
│                         `1.0-jqp-initial-data-exploration`.
│
├── references         <- Data dictionaries, manuals, and all other explanatory materials.
│
├── reports            <- Generated analysis as HTML, PDF, LaTeX, etc.
│   └── figures        <- Generated graphics and figures to be used in reporting
│
├── requirements.txt   <- The requirements file for reproducing the analysis environment, e
│                         generated with `pip freeze > requirements.txt`
│
├── setup.py           <- makes project pip installable (pip install -e .) so src can be ir
├── src                <- Source code for use in this project.
│   ├── __init__.py    <- Makes src a Python module
│   │
│   ├── data           <- Scripts to download or generate data
│   │   └── make_dataset.py
│   │
│   ├── features       <- Scripts to turn raw data into features for modeling
│   │   └── build_features.py
│   │
│   ├── models         <- Scripts to train models and then use trained models to make
│   │   │                 predictions
│   │   ├── predict_model.py
│   │   └── train_model.py
│   │
│   └── visualization  <- Scripts to create exploratory and results oriented visualizations
│       └── visualize.py
│
└── tox.ini            <- tox file with settings for running tox; see tox.readthedocs.io
```

Studies performed by Avik @ UIUC

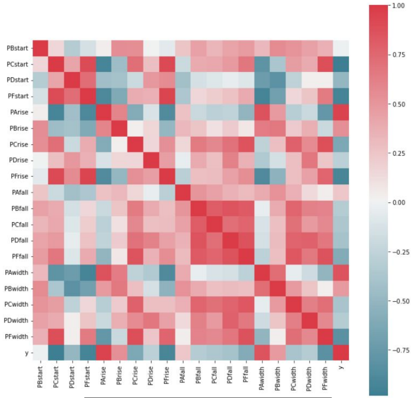
Pedagogical Introduction to FAIR

- Pedagogical introduction for FAIR principles and FAIR model development
- Dedicated tutorials for FAIR and ML in the context of HEP
- Uses the SuperCDMS dataset[8] as benchmark data
- Explores a wide array of machine learning models- from linear regression to generative models

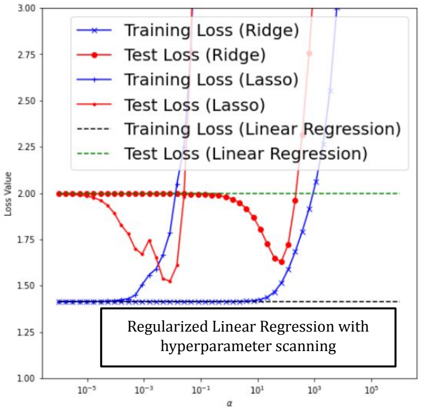
Link to repo: https://github.com/yorkiva/FAIR-Exercises



models	Added FAIR build up notebook for VAE	2 months ago
01-Intro2FAIR.ipynb	Fix typos	8 months ago
02-FAIRCheck-MNIST.ipynb	Update 02-FAIRCheck-MNIST.ipynb	last month
03-FAIRCheck-CDMS.ipynb	made changes to the notebooks	7 days ago
04a-CDMS-LR.ipynb	made changes to the notebooks	7 days ago
04b-CDMS-LR-FAIR.ipynb	New Notebooks added. Model files added.	6 months ago
05a-CDMS-PCA.ipynb	made changes to the notebooks	7 days ago
07a-CDMS_NNRegressor.ipynb	New Notebooks added. Model files added.	6 months ago
07b-CDMS_NNRegressor-FAIR.ip...	Added FAIR build up notebook for VAE	2 months ago
08a-CDMS_NNVAE.ipynb	made changes to the notebooks	7 days ago
08b-CDMS_NNVAE-FAIR.ipynb	Added FAIR build up notebook for VAE	2 months ago
README.md	Update README.md	last month



Feature correlation matrix



Regularized Linear Regression with hyperparameter scanning

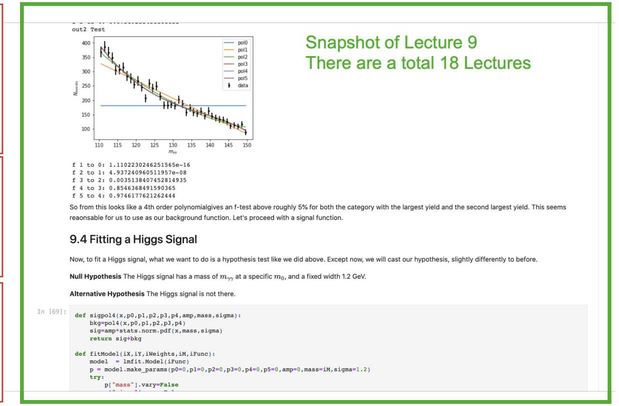
Data Science for Physics Class

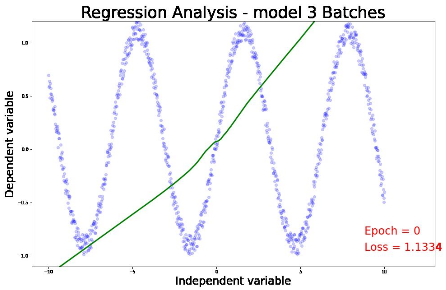
- Initially designed as an online replacement of Junior lab at MIT during COVID pandemic
- Designed around real data analysis from different research frontiers
- To be launched as a full, independent course in Spring 2023 (material already available [here]())
- Dataset and project materials developed with FAIR guidelines
- An online version will be made available via the MITx platform

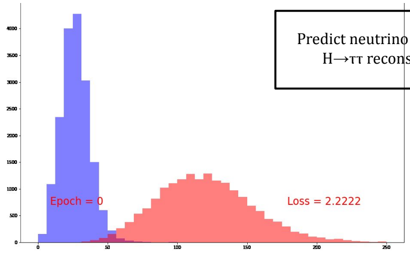
Project 1 : Gravitational Wave Data From LIGO

Project 2 : Collider Physics Data from the Compact Muon Solenoid on the Large Hadron Collider

Project 3 : Cosmic Microwave Background (simulated) Data



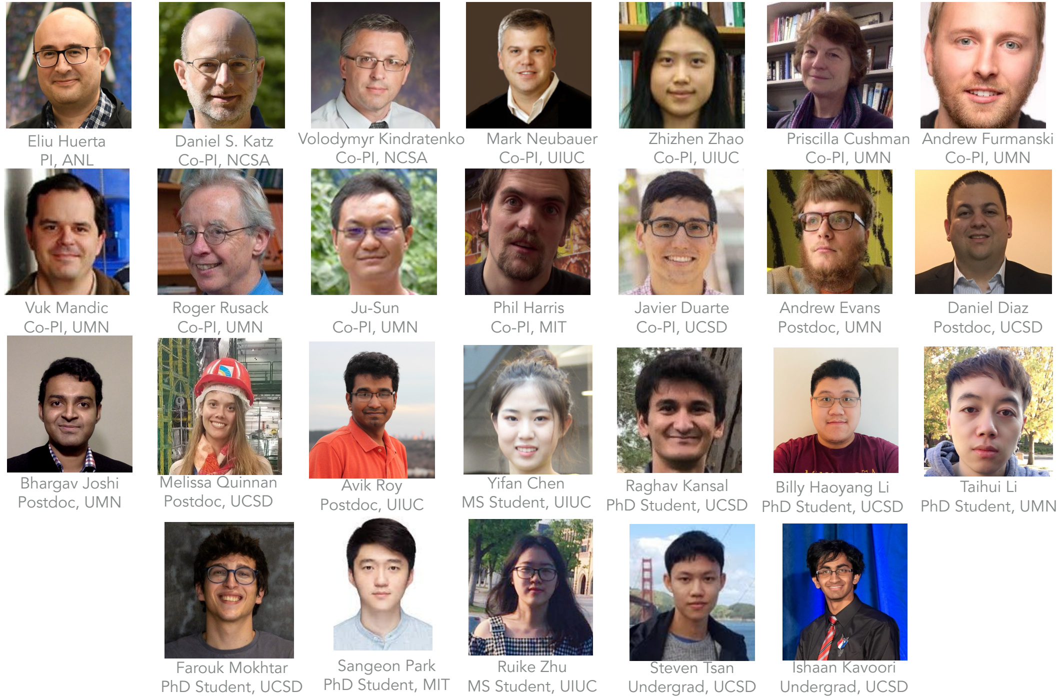




Predict neutrino 4-vector for H→ττ reconstruction

Summary

- FAIR principles set a standard guideline for curation and preservation of digital content for scientific research
- FAIR4HEP is developing HEP-specific interpretation of FAIR and active implementation by developing FAIR datasets, models, and tools



References

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
2. Chue Hong, Neil P., Katz, Daniel S., Barker, Michelle et al. RDA FAIR4RS WG. (2022). FAIR Principles for Research Software (FAIR4RS Principles) (1.0). https://doi.org/10.15497/RDA00068
3. Richardson, R. A., et al. "User-friendly Composition of FAIR Workflows in a Notebook Environment." Proceedings of the 11th on Knowledge Capture Conference. 2021. https://doi.org/10.1145/3460210.3493546
4. Katz, D. S., Psomopoulos, F. E., and Castro, L. J. "Working towards understanding the role of FAIR for machine learning." DaMaLOS@ ISWC (2021): 1-7. https://doi.org/10.4126/FRL01-006429415
5. Chen, Y., Huerta, E.A., Duarte, J. *et al.* A FAIR and AI-ready Higgs boson decay dataset. *Sci Data* **9,** 31 (2022). https://doi.org/10.1038/s41597-021-01109-0
6. Duarte, J; (2019). Sample with jet, track and secondary vertex properties for Hbb tagging ML studies HiggsToBBNTuple_HiggsToBB_QCD_RunII_13TeV_MC. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.JGJX.MS7Q
7. Duarte, J., Rao, A. & Würthwein, F. Code for jmduarte/capstone-particle-physics-domain. *Zenodo.* https://doi.org/10.5281/zenodo.5594610 (2021)
   Chen, Y. & Duarte, J. Code for FAIR4HEP/FAIR4HEP-Toolkit. *Zenodo*, https://doi.org/10.5281/zenodo.5146623 (2021)
8. Fritts, M & Li, T. (2021). *CDMS-Dataset*. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2660709
   Additional documentation and details: https://github.com/FAIR-UMN/FAIR-UMN-CDMS, https://fair-umn.github.io/FAIR-UMN-CDMS/
9. Joshi, B & Rusack R. (2022). *Laser Response in ECAL Crystals in CMS Detector (Version v1)*. Zenodo. https://doi.org/10.5281/zenodo.6394778
   Additional details: https://fair-umn.github.io/fair_ecal_monitoring