BILLY LI

CHEP

MAY 11, 2023

FAIR4HEP: FAIR AI MODELS IN HIGH ENERGY PHYSICS

UCSD

FAIR4HEP MISSION

- DOE ASCR-funded collaboration (3-year project: 2020–2023)
  - To advance our understanding of the relationship between our data and AI models by empowering scientists to explore both through the development of frameworks adhering to the principles of **f**indability, **a**ccessibility, **i**nteroperability, and **r**eusability (FAIR)
  - Using HEP as the science use-case
    - Investigate FAIR ways to share AI models and data
    - Create an environment where novel approaches to AI can be explored and applied to new data
    - Enable new insights for applying AI techniques
- Collaborate with partners: CERN Open Data Portal, Zenodo, DLHub
- Operate within larger community: Australian Research Data Commons (ARDC), Research Data Alliance (RDA)

OUTLINE

- Motivation
- FAIR Principles and Datasets in HEP
- FAIR AI models in HEP
  - Cookiecutter4FAIR
- Projects implementing FAIR princples
- Vision & Outlook

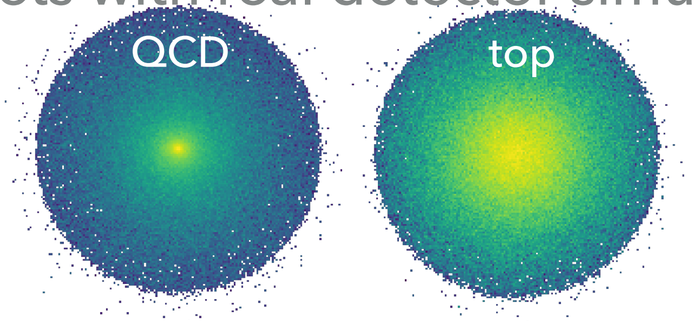
MOTIVATION 1: CREATING/SHARING REFERENCE DATA SETS FOR HEP

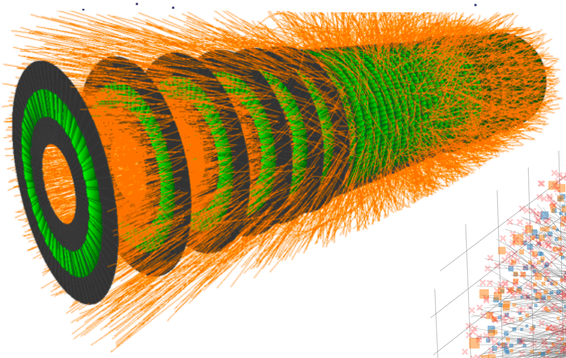
MOTIVATION 1: CREATING/SHARING REFERENCE DATA SETS FOR HEP

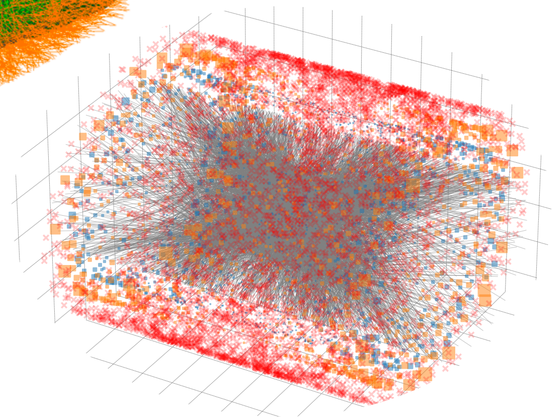
- Engage ML community for interesting, realistic tasks in experimental HEP
  - As ImageNet (an image dataset organized according to the WorldNet hierarchy) accelerated advances in computer vision, do the same for HEP

MOTIVATION 1: CREATING/SHARING REFERENCE DATA SETS FOR HEP

- Engage ML community for interesting, realistic tasks in experimental HEP
  - As ImageNet (an image dataset organized according to the WorldNet hierarchy) accelerated advances in computer vision, do the same for HEP
- Calls at many workshops for more public HEP data sets with real detector simulation for ML applications
  - Example: dataset for top tagging based on Pythia+Delphes
  - Example: dataset for tracking based on ACTS (kaggle TrackML challenge)
  - Example: dataset for H(bb) tagging based on CMS open simulation
  - Example: dataset for particle-flow based on Pythia+Delphes

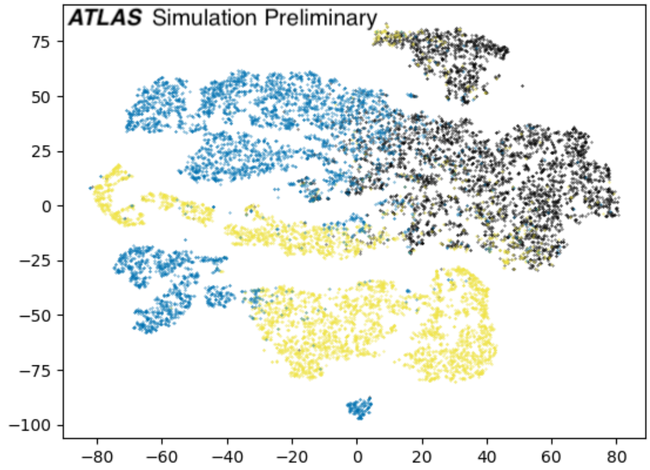






MOTIVATION 2: SHARING/REUSING AI MODELS ACROSS HEP

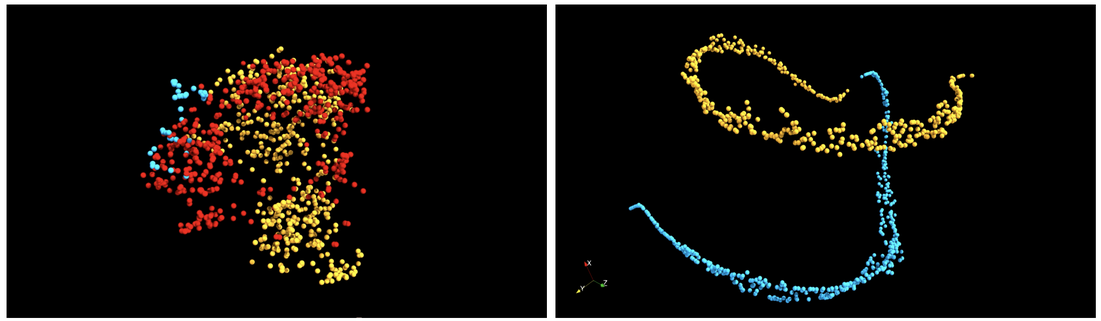
- Allow AI models developed for one experiment to be (re-)trained and (re-)used easily in another experiment
  - Example 1: ATLAS studied GravNet developed by CMS collaborators [https://cds.cern.ch/record/2753414] for physics object localization using point cloud segmentation
  - Example 2: CMS collaborators are using SPANet developed by ATLAS collaborators



MOTIVATION 3: EXPLORING RELATIONSHIPS BETWEEN DATA AND AI MODELS

MOTIVATION 3: EXPLORING RELATIONSHIPS BETWEEN DATA AND AI MODELS

- Easier to build upon existing work (e.g. through transfer learning)

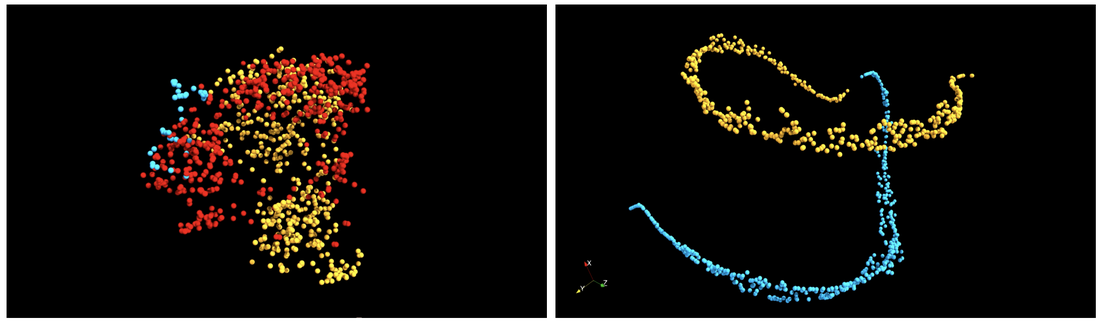


Left: Xception pre-trained on ImageNet applied to galaxies

Right: After fine-tuning on galaxy data, two galaxy clusters can be clearly identified

MOTIVATION 3: EXPLORING RELATIONSHIPS BETWEEN DATA AND AI MODELS

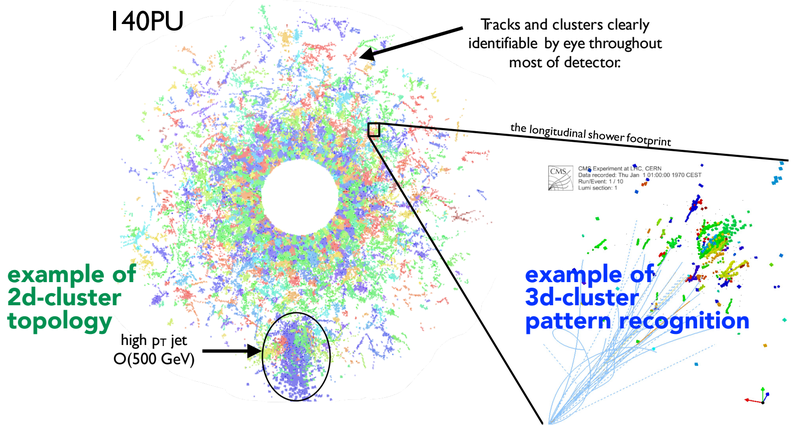
- Easier to build upon existing work (e.g. through transfer learning)

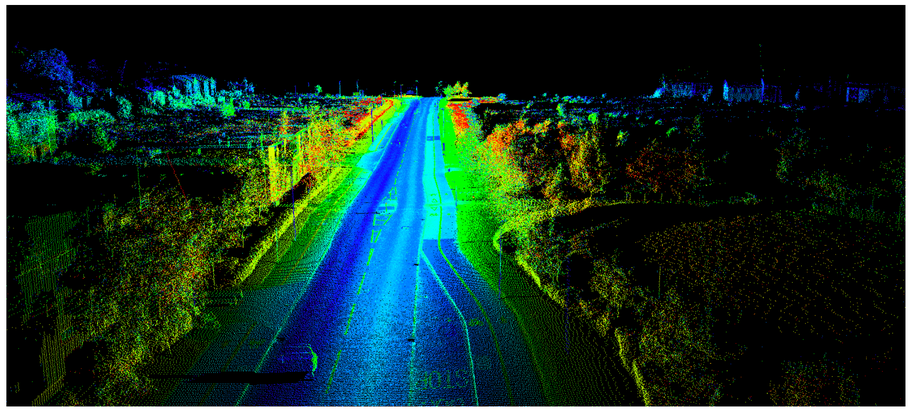


Left: Xception pre-trained on ImageNet applied to galaxies

Right: After fine-tuning on galaxy data, two galaxy clusters can be clearly identified

- Share work beyond HEP
  - AI models developed for HEP-specific tasks may be useful in other domains (e.g. LiDAR point cloud data)





FAIR PRINCIPLES & DATASETS IN HEP

UCSD

FAIR DATA PRINCIPLES

Learn more: https://www.go-fair.org/fair-principles/

Image: book.fosteropenscience.eu

FAIR DATA PRINCIPLES

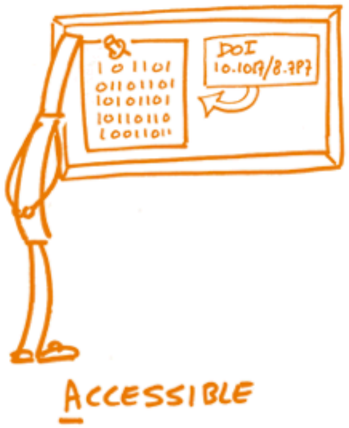
Learn more: https://www.go-fair.org/fair-principles/

- F1. (meta)data have **unique** and **persistent** identifier
  F2. data are described with rich metadata
  F3. metadata specify the data identifier
  F4. (meta)data are registered or indexed in a searchable resource

Image: book.fosteropenscience.eu

FAIR DATA PRINCIPLES

Learn more: https://www.go-fair.org/fair-principles/

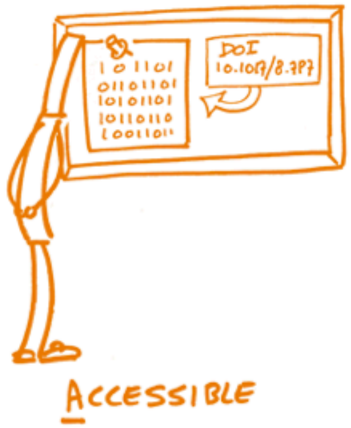


- F1. (meta)data have **unique** and **persistent** identifier
  F2. data are described with rich metadata
  F3. metadata specify the data identifier
  F4. (meta)data are registered or indexed in a searchable resource

Image: book.fosteropenscience.eu

FAIR DATA PRINCIPLES

Learn more: https://www.go-fair.org/fair-principles/

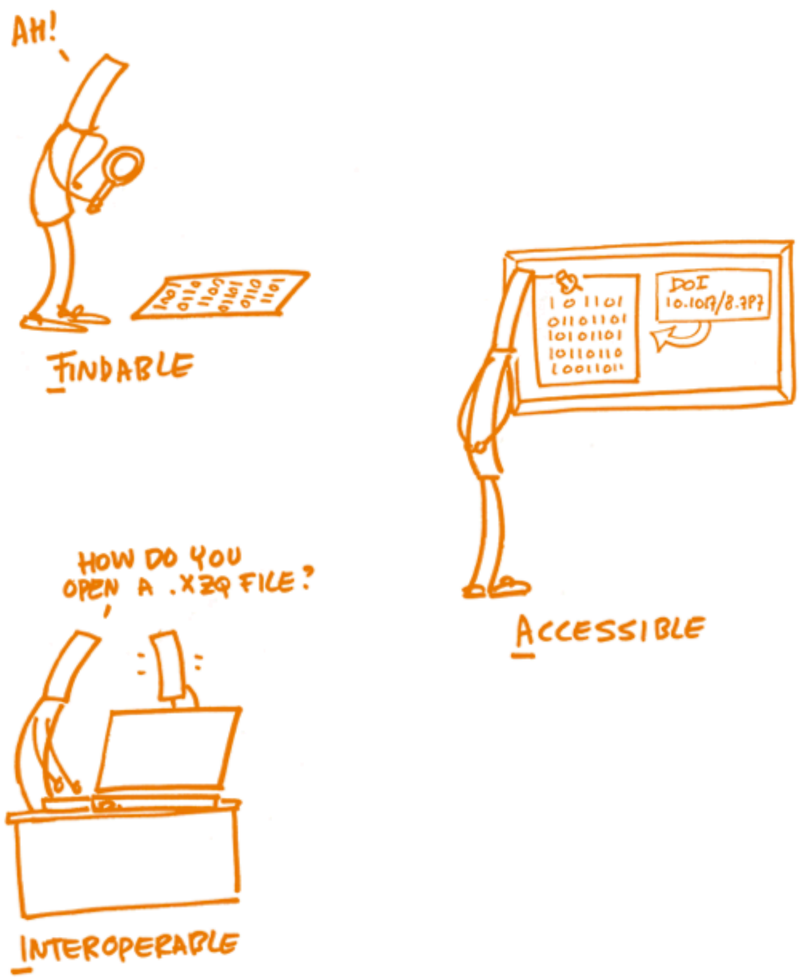


- F1. (meta)data have **unique** and **persistent** identifier
  F2. data are described with rich metadata
  F3. metadata specify the data identifier
  F4. (meta)data are registered or indexed in a searchable resource
- A1. (meta)data are retrievable using standardized protocol
  A1.1 protocol is open, free, and universally implementable
  A1.2 protocol allows for authentication and authorization
  A2. metadata are accessible, even when the data is not

Image: book.fosteropenscience.eu

FAIR DATA PRINCIPLES

Learn more: https://www.go-fair.org/fair-principles/

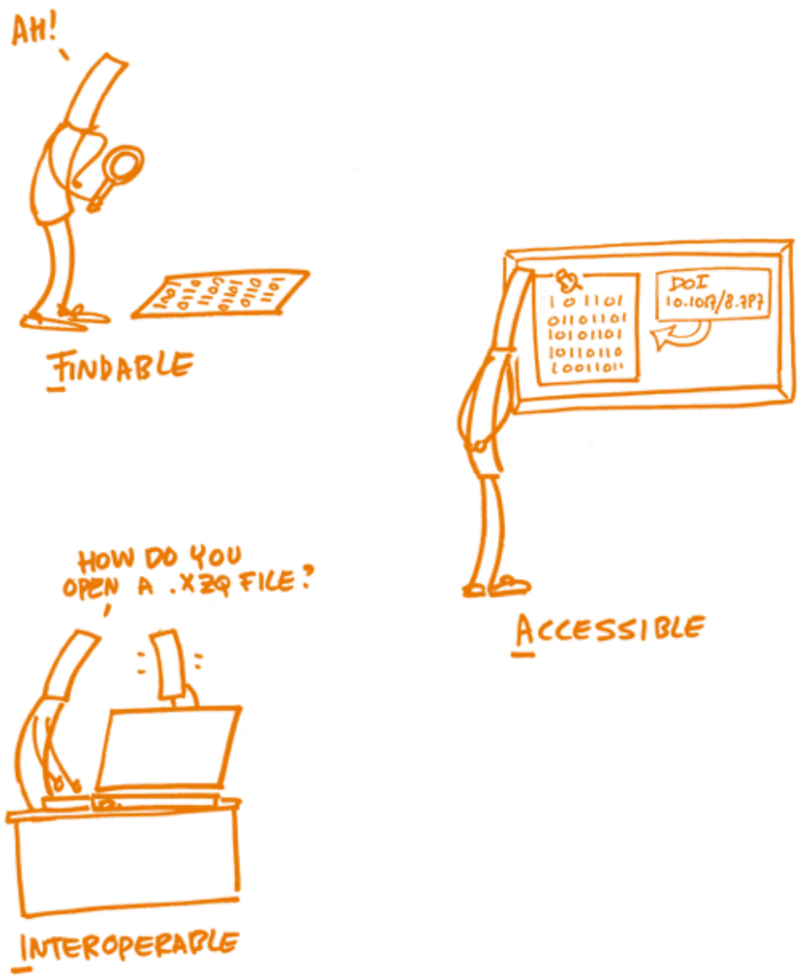


- F1. (meta)data have **unique** and **persistent** identifier
  F2. data are described with rich metadata
  F3. metadata specify the data identifier
  F4. (meta)data are registered or indexed in a searchable resource
- A1. (meta)data are retrievable using standardized protocol
  A1.1 protocol is open, free, and universally implementable
  A1.2 protocol allows for authentication and authorization
  A2. metadata are accessible, even when the data is not

Image: book.fosteropenscience.eu

FAIR DATA PRINCIPLES

Learn more: https://www.go-fair.org/fair-principles/



- F1. (meta)data have **unique** and **persistent** identifier
  F2. data are described with rich metadata
  F3. metadata specify the data identifier
  F4. (meta)data are registered or indexed in a searchable resource
- A1. (meta)data are retrievable using standardized protocol
  A1.1 protocol is open, free, and universally implementable
  A1.2 protocol allows for authentication and authorization
  A2. metadata are accessible, even when the data is not
- I1. (meta)data use a formal, shared, and broadly applicable language for knowledge representation
  I2. (meta)data use **vocabularies** that follow FAIR principles
  I3. (meta)data include qualified references to other (meta)data

Image: book.fosteropenscience.eu

FAIR DATA PRINCIPLES

Learn more: https://www.go-fair.org/fair-principles/

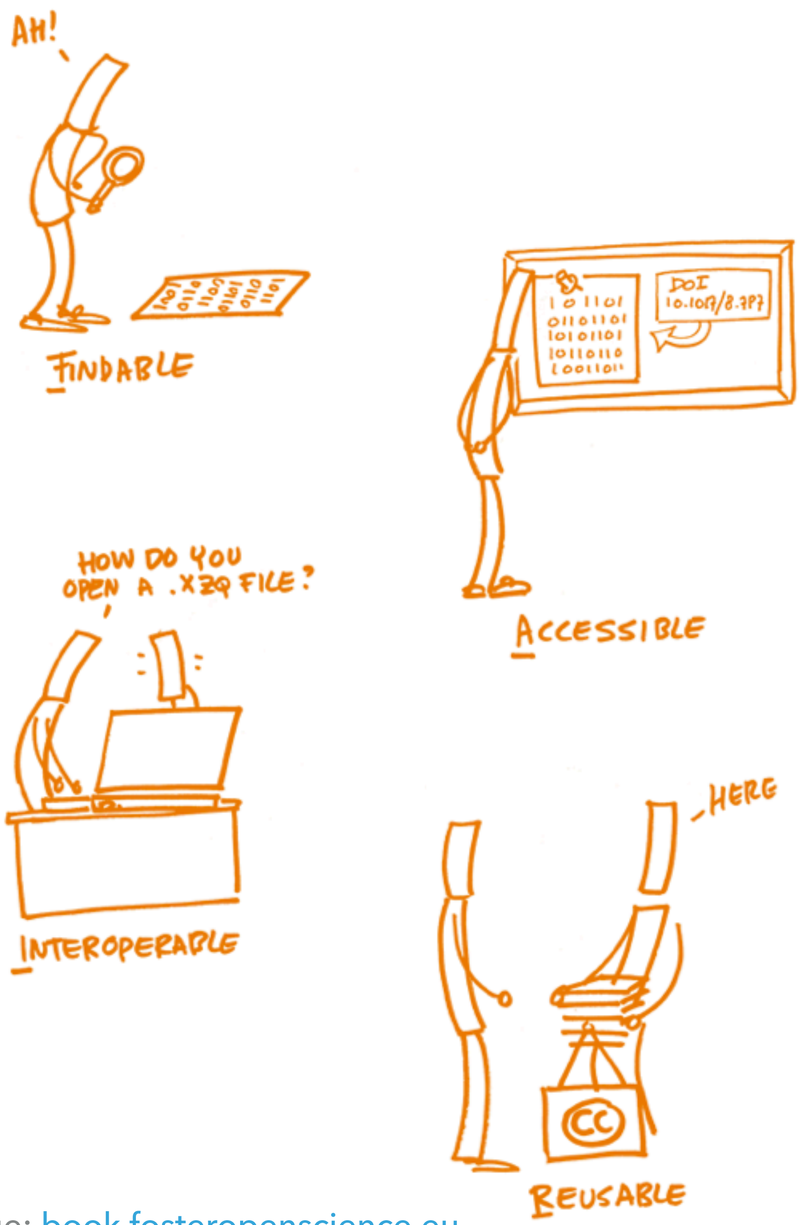


Image: book.fosteropenscience.eu

- F1. (meta)data have **unique** and **persistent** identifier
  F2. data are described with rich metadata
  F3. metadata specify the data identifier
  F4. (meta)data are registered or indexed in a searchable resource
- A1. (meta)data are retrievable using standardized protocol
  A1.1 protocol is open, free, and universally implementable
  A1.2 protocol allows for authentication and authorization
  A2. metadata are accessible, even when the data is not
- I1. (meta)data use a formal, shared, and broadly applicable language for knowledge representation
  I2. (meta)data use **vocabularies** that follow FAIR principles
  I3. (meta)data include qualified references to other (meta)data

FAIR DATA PRINCIPLES

Learn more: https://www.go-fair.org/fair-principles/

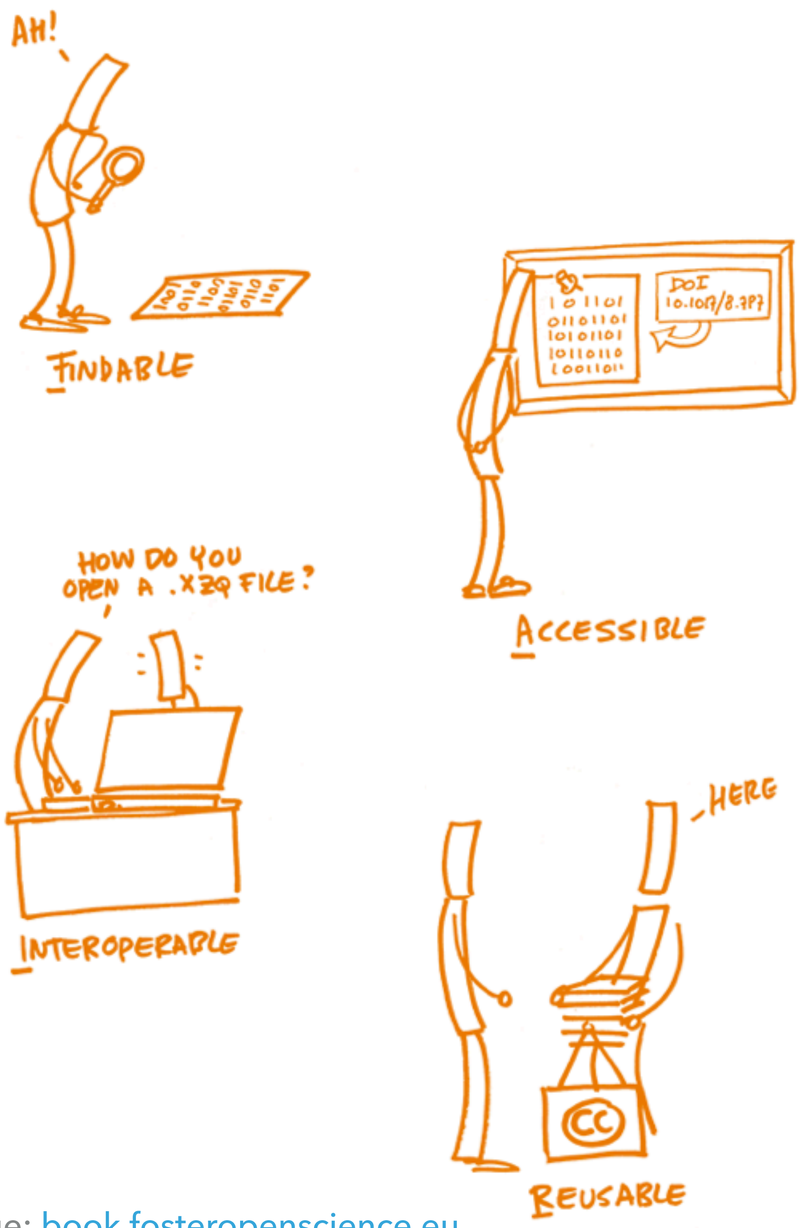


Image: book.fosteropenscience.eu

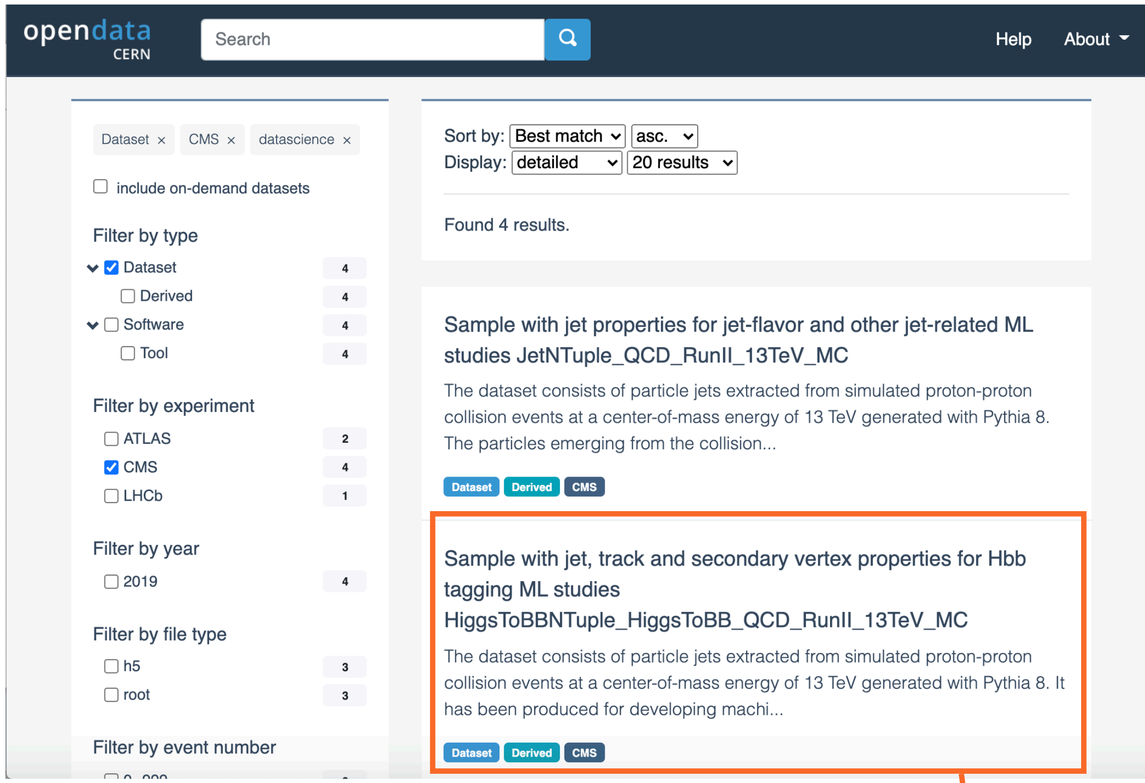
- F1. (meta)data have **unique** and **persistent** identifier
  F2. data are described with rich metadata
  F3. metadata specify the data identifier
  F4. (meta)data are registered or indexed in a searchable resource
- A1. (meta)data are retrievable using standardized protocol
  A1.1 protocol is open, free, and universally implementable
  A1.2 protocol allows for authentication and authorization
  A2. metadata are accessible, even when the data is not
- I1. (meta)data use a formal, shared, and broadly applicable language for knowledge representation
  I2. (meta)data use **vocabularies** that follow FAIR principles
  I3. (meta)data include qualified references to other (meta)data
- R1. (meta)data have a plurality of accurate and relevant attributes
  R1.1. (meta)data have clear and accessible data usage license
  R1.2. (meta)data are associated with their provenance
  R1.3. (meta)data meet domain-relevant community standards

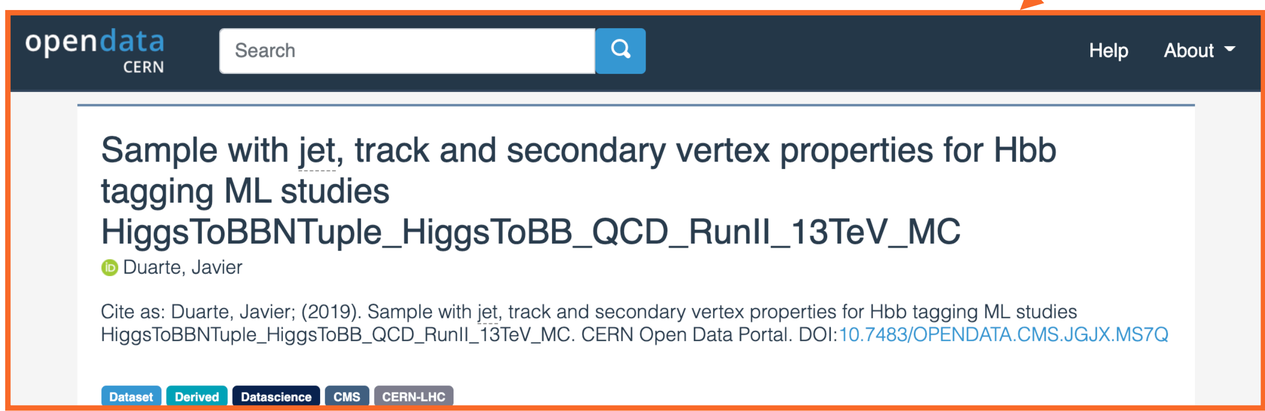
EXAMPLE DATASETS AND AI MODELS

- Advance important tasks in HEP with reference datasets and AI models to explore FAIRness criteria for both
  - **H(bb) jet tagging**
  - Jet generation/simulation
  - Particle-flow reconstruction
  - ECAL crystal calibration
  - Level-1 trigger jet reconstruction
  - Charged particle tracking
  - …

H(BB) JET TAGGING DATASET

- Hosted on CERN Open Data Portal
  - Collaborative effort between CERN IT-CDA and RCS-SIS groups, LHC and OPERA experiments
  - Built with Invenio library management software
  - Products (i.e. data, software, documentation, provenance) shared under open licenses and issued DOIs
  - EOS data storage; access via XRootD, HTTP
- H(bb) dataset [10.7483/OPENDATA.CMS.JGJX.MS7Q]
  - 182 files, 245 GB, 18 million total entries (jets)
    - event features, e.g. MET, ρ (average density)
    - jet features, e.g. mass, p_T, N-subjettiness variables
    - particle candidate features, e.g. p_T, η, Φ
    - charged particle / track features, e.g. impact parameter
    - secondary vertex features, e.g. flight distance





IS THIS H(BB) JET TAGGING DATASET FAIR?

- Evaluated the FAIRness of this dataset in 10.1038/s41597-021-01109-0
- Lessons learned: Difficult to satisfy "**Use FAIR Vocabularies**": requires the metadata values and qualified relations should be FAIR themselves, that is, terms should be findable from open, community-accepted vocabularies (i.e. jargon should be avoided or clearly defined)

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.

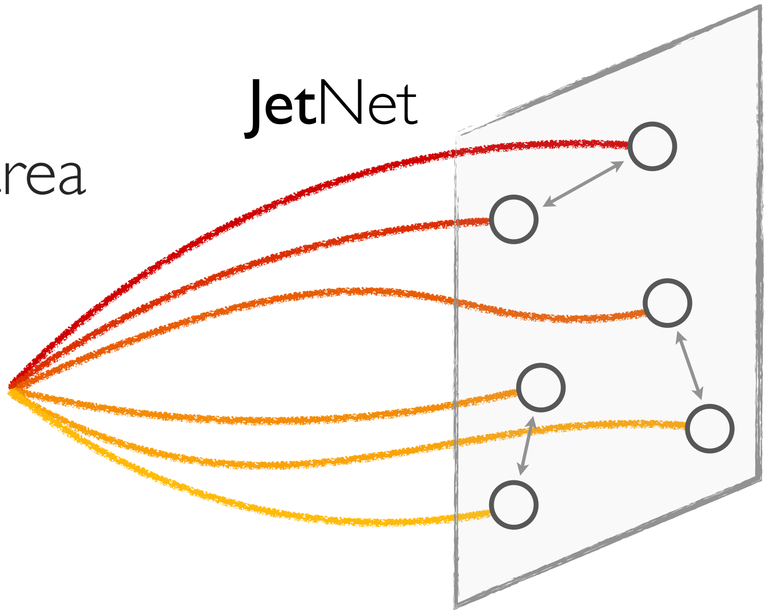
Metric	Evaluation
F1. (Meta)data are assigned globally unique and persistent identifiers.	
Identifier Uniqueness: this metric measures whether there is a scheme to uniquely identify the digital resource.	**Pass**. The DOI for the data (which resolves to a URL[29]) follows a registered identifier scheme.
Identifier Persistence: this measures whether there is a policy that describes what the provider will do in the event an identifier scheme becomes deprecated.	**Pass**. The use of a DOI provide a persistent interoperable identifier.
F2. Data are described with rich metadata.	
Machine-readability of Metadata: to meet this metric, a URL to a document containing machine-readable metadata for the digital resource must be provided.	**Pass**. The URL for the metadata[57] in JSON Schema with REST API is available. The use of JSON Schema provides clear human and machine readable documentation. Also, running the URL through the Rich Result Test shows the data page contains rich results.
Richness of Metadata: data are described with rich metadata	**Partially pass**. Reviewing the DataCite metadata for the DOI shows a fairly sparse record. The metadata can be improved with richer fields.
F3. Metadata clearly and explicitly include the identifier of the data they describe.	
Resource Identifier in Metadata: this measures if the metadata document contains the identifier for the digital resource that meets F1 principle.	**Pass**. The association between the metadata and the dataset is made explicit because the dataset's globally unique and persistent identifier can be found in the metadata. Specifically, the DOI is a top-level and a mandatory field in the metadata record.
F4. (Meta)data are registered or indexed in a searchable resource	
Index in a searchable resource: this measures the degree to which the digital resource can be found using web-based search engines	**Pass**. The dataset is indexed by Google Dataset Search engine.
A1. (Meta)data are retrievable by their identifier using a standardized communications protocol	
A1.1: The protocol is open, free and universally implementable	
Access Protocol: it measures whether the URL is open access and free.	**Pass**. HTTP get on the identifier's URL returns a valid document
A1.2. The protocol allows for an authentication and authorization where necessary	
Access Authorization: it requires specification of a protocol to access restricted content.	**Pass**. This is an open dataset, accessible to everyone on the internet. The data is non-profit and privacy-unrelated, so no access authorization is needed.
A2. Metadata should be accessible even when the data is no longer available	
Metadata Longevity: it requires metadata to be present even in the absence of data	**Pass**. Metadata is stored separately in the CERN Open Data server. As per FAIR Principle F3, this metadata remains discoverable, even in the absence of the data, because it contains an explicit reference to the DOI of the data. Data and metadata will be retained for the lifetime of the repository. The host laboratory CERN, currently plans to support the repository for at least the next 20 years.

Table 1. Findable and Accessible principle assessment checks for the CMS H($b\bar{b}$) Open Dataset.

Metric	Evaluation
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	
Use a Knowledge Representation (programming) Language: use a formal, accessible, shared, and broadly applicable language for knowledge representation	**Pass**. As described in Section 3, this dataset is represented based on the ROOT framework with Python interface. The notebook we release with this manuscript provides the required tools to handle this dataset using HDF5. The metadata is represented following the JSON Schema draft 4. Both are widely used formats in Physics.
Provide Human-readable descriptions	**Pass**. The description and data semantics of this dataset provides rich information on how to use the dataset.
I2. (Meta)data use vocabularies that follow FAIR principles.	
Use FAIR Vocabularies: it requires the metadata values and qualified relations should be FAIR themselves, that is, terms should be findable from open, community-accepted vocabularies.	**Partially pass**. I2 requires the controlled vocabulary used to describe datasets to be documented and resolvable using globally unique and persistent identifiers. For domain-specific terms, we leverage a vocabulary PhySH (Physics Subject Headings), a physics classification scheme developed by American Physical Society (APS). Some terms in dataset descriptions and semantics are registered in PhySH. However, since PhySH is still under development, there is not very good coverage of the narrower experimental concepts. For the terms not covered, references and hover definitions are provided. For general terms, the metadata follows the vocabulary from JSON Schema and a minimal set of FAIR terms are used.
I3. (Meta)data include qualified references to other (meta)data.	
Use Qualified References: The goal is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data.	**Partially pass**. There are connections with other datasets. A list of derived datasets is available at the dataset site [27]. Each referenced external piece of dataset is qualified by a resolvable URL and a unique CERN data identifier in metadata. To improve, the papers of these related data can be provided, from which more information about methods and workflow used to derive this dataset can be retrieved, and external datasets should be references by permanent identifiers rather than URLs.
R1.1. (Meta)data are released with a clear and accessible data usage license.	
Accessible Usage License: the existence of license document for (meta)data are being measured	**Pass**. This dataset is released under Creative Commons CC0 dedication. The license field is present in the metadata.
R1.2. (Meta)data are associated with detailed provenance.	
Detailed Provenance: Who / What / When produced the data? Why / How was the data produced?	**Pass**. The dataset is derived from other data, e.g.[58,59], using public software[60] that was made public to process and reduce it. We are able to track the original authors and data sources. But ideally, this workflow would be described in a machine-readable format.
R1.3. (Meta)data meet domain-relevant community standards.	
Meet Community Standards: it measures whether a certification of the resource meeting community standards exists.	**Pass**. Both metadata and data meet the CERN Open Data community standards and thus have been released on the CERN Open Data repository.

Table 2. Interoperable and Reusable principle assessment checks for CMS H($b\bar{b}$) Open Dataset.

- **Problem:** ML + HEP development right now is generally not very accessible, standardized, or reproducible
- **Our solution:** JetNet Python library to facilitate research in this area, with:
  - Easily accessible datasets/interfaces
  - Standardized evaluation metrics
  - More conveniences for helping researchers in this area



JetNet

DOI 10.5281/zenodo.5598104 | pypi package 0.1.1 | docs passing | downloads 74/month | code style black

A library for developing and reproducing jet-based machine learning (ML) projects.

JetNet provides common standardized PyTorch-based datasets, evaluation metrics, and loss functions for working with jets using ML. Currently supports the flagship JetNet dataset, and the Fréchet ParticleNet Distance (FPND), Wasserstein-1 (W1), coverage and minimum matching distance (MMD) metrics all introduced in Ref. [1], as well as jet utilities and differentiable implementation of the energy mover's distance [2] for use as a loss function. Additional functionality is currently under development.

Installation

JetNet can be installed with pip:

```
pip install jetnet
```

Shout out to Carlos who presented JetNet 1 hour ago!

FAIR AI MODELS IN HEP

UCSD

FAIR AI MODELS IN HEP

- We have FAIR principles for HEP dataset, but what can FAIR mean for AI models?
- Paper exploring defining FAIR AI models in HEP: arXiv:2212.05081
- Proposes a working definition of a FAIR AI model
- Develops a template: Cookiecutter4FAIR to encourage these principles

FAIR AI Models in High Energy Physics

Javier Duarte[1], Haoyang Li[1], Avik Roy[2], Ruike Zhu[2,3], E. A. Huerta[3,4], Daniel Diaz[1], Philip Harris[5], Raghav Kansal[1], Daniel S. Katz[2], Ishaan H. Kavoori[1], Volodymyr V. Kindratenko[2], Farouk Mokhtar[1,6], Mark S. Neubauer[2], Sang Eon Park[5], Melissa Quinnan[1], Roger Rusack[7], and Zhizhen Zhao[2]

[1]University of California San Diego, La Jolla, California 92093, USA
[2]University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA
[3]Argonne National Laboratory, Lemont, Illinois 60439, USA
[4]The University of Chicago, Chicago, Illinois 60637, USA
[5]Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[6]Halıcıoğlu Data Science Institute, La Jolla, California 92093, USA
[7]The University of Minnesota, Minneapolis, Minnesota 55405, USA

PROPOSED FAIR PRINCIPLES FOR AI MODELS

Table 1. Proposed FAIR principles for fully trained AI models used for AI-inference only. These principles may be extended for retraining use cases by amending our proposed definition for the 'Reusability' principle.

F: The AI model, and its associated metadata, are easy to find for both humans and machines.
F1. The AI model is assigned a globally unique and persistent identifier. F2. The AI model is described with rich metadata. F3. Metadata clearly and explicitly include the identifier of the AI model they describe. F4. Metadata and the AI model are registered or indexed in a searchable resource.
A: The AI model, and its metadata, are retrievable via standardized protocols.
A1. The AI model is retrievable by its identifier using a standardized communications protocol. A1.1. The protocol is open, free, and universally implementable. A1.2. The protocol allows for an authentication and authorization procedure, where necessary. A2. Metadata are accessible, even when the AI model is no longer available.

I: The AI model interoperates with other models, data, and/or software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.
I1. The AI model reads, writes and exchanges data in a way that meets domain-relevant community standards. I2. The AI model includes qualified references to other objects, including the (FAIR) data used to train the model.
R: The AI model is both usable (for inference) and reusable (can be understood, built upon, or incorporated into other models and/or software).
R1. The AI model is described with a plurality of accurate and relevant attributes. R1.1. The AI model is given a clear and accessible license. R1.2. The AI model is associated with detailed provenance, such as information about the input data preparation and training process. R2. The AI model includes qualified references to other models and/or software, such as dependencies. R3. The AI model meets domain-relevant community standards.

COOKIECUTTER4FAIR

- GitHub: https://github.com/FAIR4HEP/cookiecutter4fair

- A solid start of developing your FAIR AI projects

Table 2. Map between existing capabilities of the `coookiecutter4fair` AI project template and our proposed FAIR principles for AI models. The $*$ symbol indicates that the process is not yet fully automated and requires additional manual steps.

Principle	GitHub repository	Zenodo upload	DLHub upload	Docker or Apptainer image	License
Findable	✓				
Accessible		✓	*		
Interoperable				✓	
Reusable			*	✓	✓

```
├── LICENSE            <- License for reusing code
├── Makefile           <- Makefile with commands like `make data` or `make train`
├── CITATION.cff       <- Standardized citation metadata
├── README.md          <- The top-level README for developers using this project
├── data
│   ├── processed      <- The final, canonical data sets for modeling
│   └── raw            <- The original, FAIR, and immutable data dump
│
├── Dockerfile         <- For building a containerized environment
│
├── docs               <- A default Sphinx project for documentation; see sphinx-doc.org for details
│
├── models             <- Trained and serialized models, model predictions, or model summaries
│
├── notebooks          <- Jupyter notebooks. Naming convention is a number (for ordering),
│                         the creator's initials, and a short `-` delimited description, e.g.
│                         `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.g.
│                         generated with `pip freeze > requirements.txt`
│
├── setup.py           <- Makes project pip installable (`pip install -e .`) so src can be imported
├── 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
```

COOKIECUTTER4FAIR DEMO

https://asciinema.org/a/554117

```
Restored session: Tue Jan 24 21:07:10 PST 2023
(⎈ |nautilus:cms-ml)➜  ~ conda activate cookiecutter
(cookiecutter) (⎈ |nautilus:cms-ml)➜  ~ cookiecutter https://github.com/FAIR4HEP/cookiecutter4fair
```

COOKIECUTTER4FAIR DEMO

https://asciinema.org/a/554117

```
Restored session: Tue Jan 24 21:07:10 PST 2023
(⎈ |nautilus:cms-ml)➜  ~ conda activate cookiecutter
(cookiecutter) (⎈ |nautilus:cms-ml)➜  ~ cookiecutter https://github.com/FAIR4HEP/cookiecutter4fair
You've downloaded /Users/jduarte/.cookiecutters/cookiecutter4fair before. Is it okay to delete and re-download it? [yes]: yes
project_name [project_name]:
repo_name [project_name]:
author_name [Your name (or your organization/company/team)]:
orcid [Your ORCID]:
description [A short description of the project.]:
Select open_source_license:
1 - MIT
2 - BSD-3-Clause
3 - Apache-2
4 - GPLv3
5 - No license file
Choose from 1, 2, 3, 4, 5 [1]:
data_doi [10.XXXX/XXXX]:
code_doi [10.XXXX/XXXX]:
provide_dockerfile [yes]:
```

COOKIECUTTER4FAIR DEMO

https://asciinema.org/a/554117

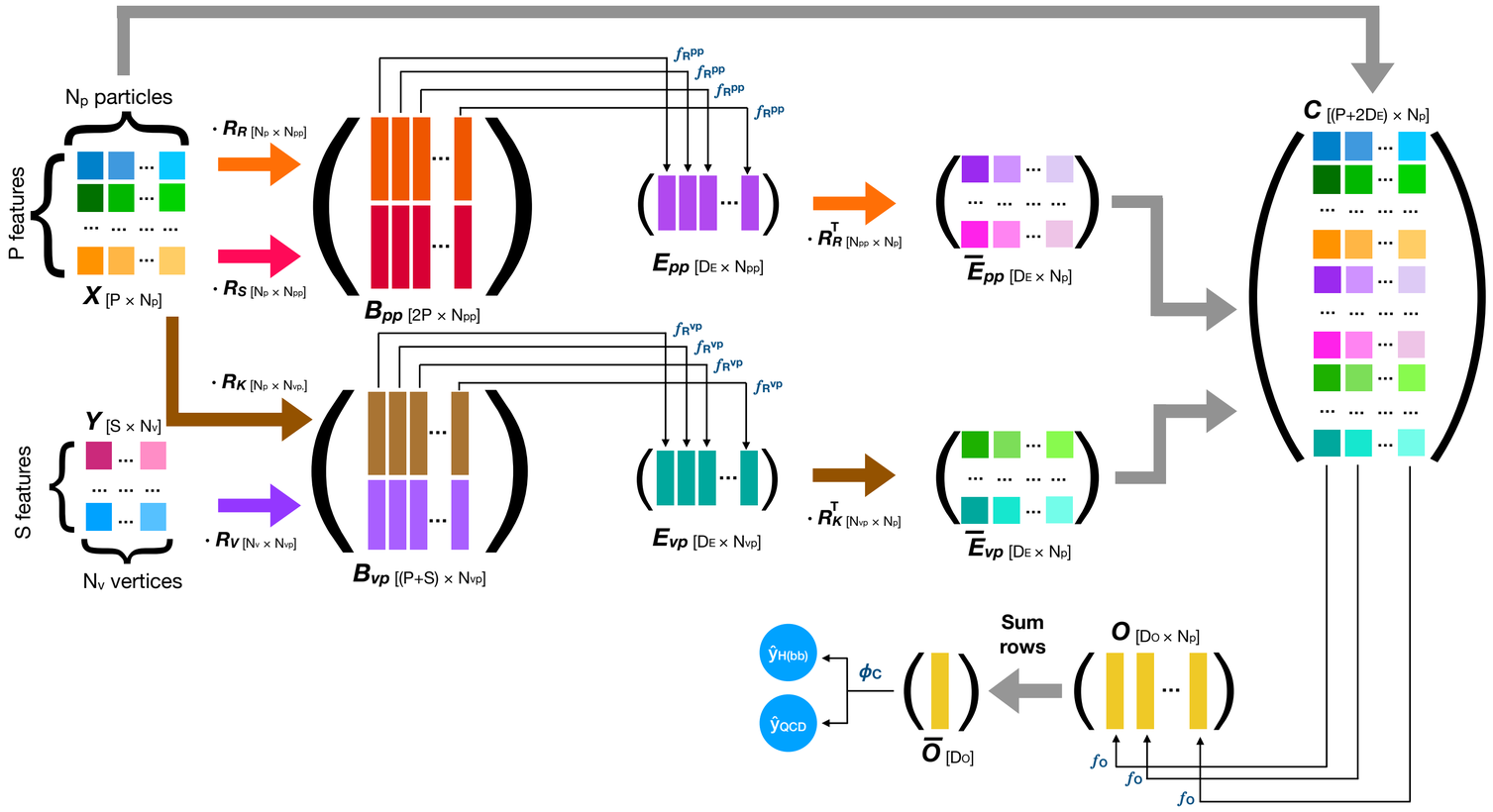
```
4 - GPLv3
5 - No license file
Choose from 1, 2, 3, 4, 5 [1]:
data_doi [10.XXXX/XXXX]:
code_doi [10.XXXX/XXXX]:
provide_dockerfile [yes]:
(cookiecutter) (⎈ |nautilus:cms-ml)➜  ~ ls -l project_name
total 80
-rw-r--r--  1 jduarte  staff   279 Jan 24 21:07 CITATION.cff
-rw-r--r--  1 jduarte  staff   143 Jan 24 21:07 Dockerfile
-rw-r--r--  1 jduarte  staff  1114 Jan 24 21:07 LICENSE
-rw-r--r--  1 jduarte  staff  4125 Jan 24 21:07 Makefile
-rw-r--r--  1 jduarte  staff   229 Jan 24 21:07 README.md
drwxr-xr-x  4 jduarte  staff   128 Jan 24 21:07 data
drwxr-xr-x  8 jduarte  staff   256 Jan 24 21:07 docs
drwxr-xr-x  3 jduarte  staff    96 Jan 24 21:07 models
drwxr-xr-x  3 jduarte  staff    96 Jan 24 21:07 notebooks
drwxr-xr-x  3 jduarte  staff    96 Jan 24 21:07 references
drwxr-xr-x  4 jduarte  staff   128 Jan 24 21:07 reports
-rw-r--r--  1 jduarte  staff   103 Jan 24 21:07 requirements.txt
-rw-r--r--  1 jduarte  staff   255 Jan 24 21:07 setup.py
drwxr-xr-x  7 jduarte  staff   224 Jan 24 21:07 src
-rw-r--r--  1 jduarte  staff   408 Jan 24 21:07 test_environment.py
-rw-r--r--  1 jduarte  staff   234 Jan 24 21:07 tox.ini
(cookiecutter) (⎈ |nautilus:cms-ml)➜  ~
```

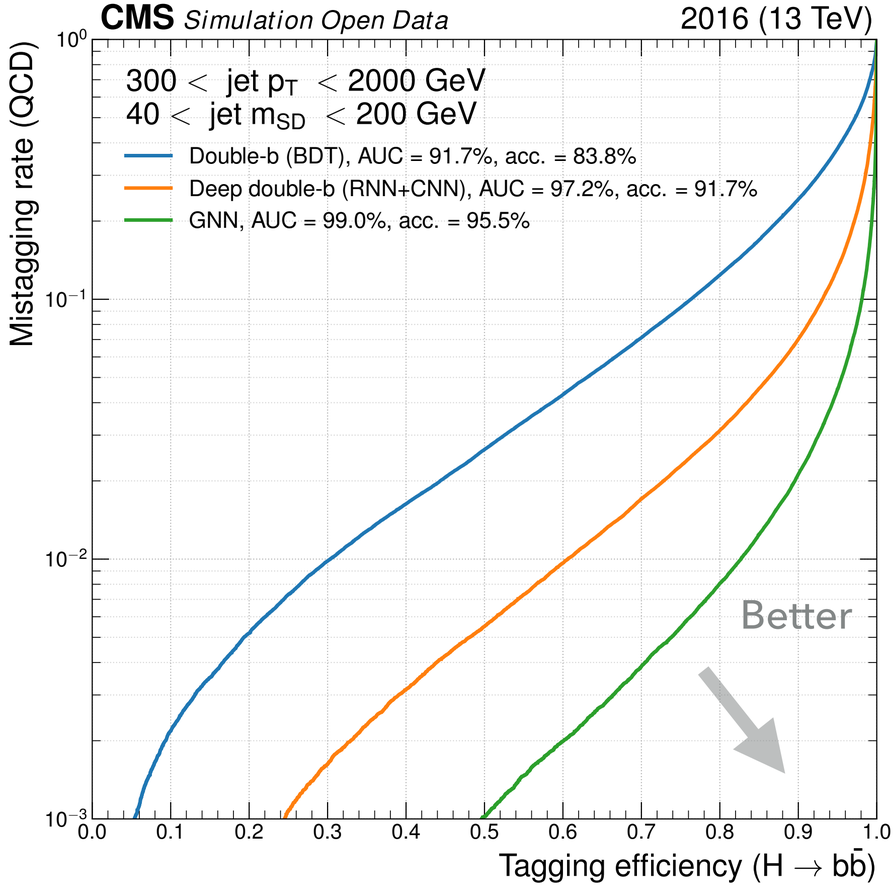
PROJECTS IMPLEMENTING FAIR4HEP PRINCIPLES

UCSD

FAIR AI MODEL BENCHMARK: H->BB INTERACTION NETWORK

- Edge convolutions for particle-particle and particle-vertex connections update particle features; summed particle features used to predict H(bb) or QCD prob.
- FAIR AI model implementation: https://github.com/FAIR4HEP/hbb_interaction_network





BUILT UPON FAIR MODELS: RELEVANT FEATURES & MODEL COMPRESSION

- Can use layerwise relevance propagation and other XAI techniques to determine most relevant features and optimal model size

- Can remove input features and shrink model size dramatically with very minor loss in performance!

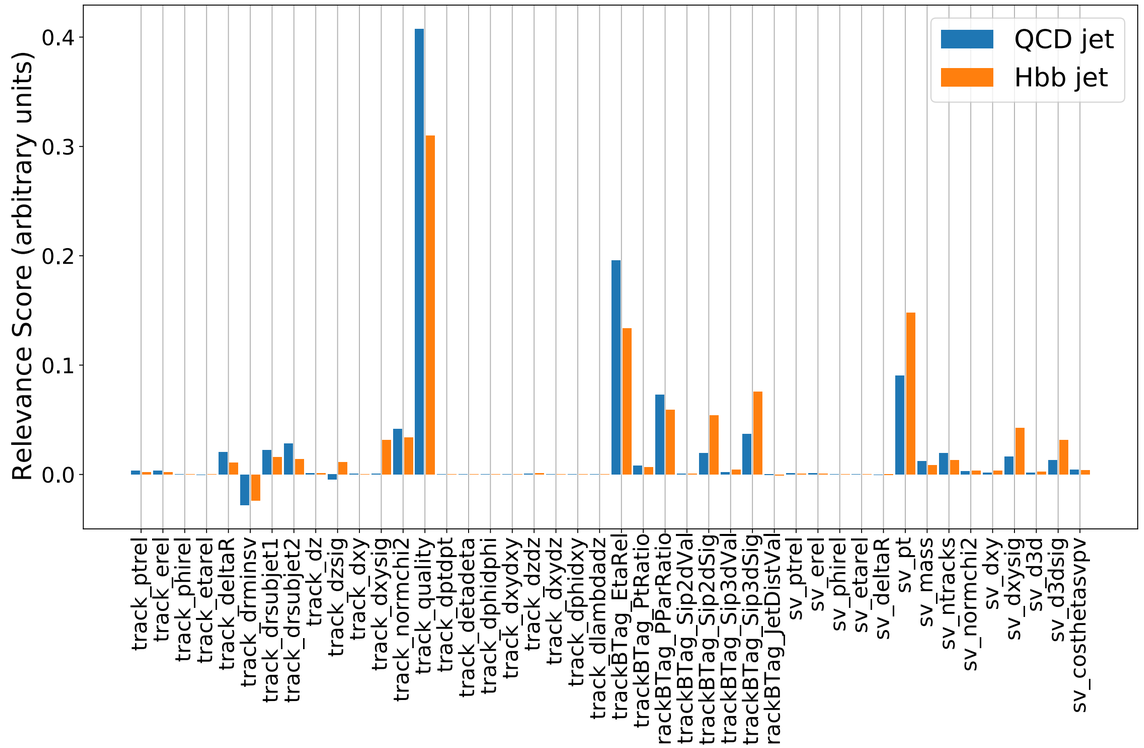


Table 10. The performance of a baseline and ablated models. ΔP represents the number of particle track features that have been dropped and h is the number of nodes in the hidden layers. The fidelity score is measured with respect to a baseline model. Sparsity is measured by the fraction of activation nodes with an RNA score less than 0.2

ΔP, ΔS	h, D_E, D_O	Parameters	AUC score [%]	Fidelity [%]	Sparsity
0, 0 (baseline)	60, 20, 24	25554	99.02	100	0.56
12, 5	32, 16, 16	8498	98.87	96.93	0.52
	32, 8, 8	7178	98.84	96.79	0.48
	16, 8, 8	2842	98.62	96.12	0.40

VISION AND SUMMARY

UCSD

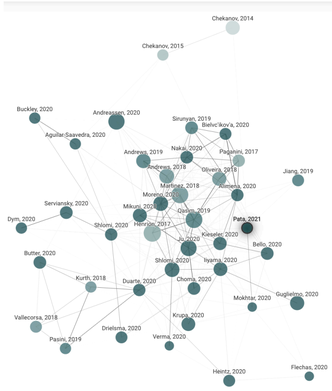
VISION

FAIR data repositories

VISION

FAIR data repositories

Papers / indexing / search / discovery



VISION

FAIR data repositories

opendata

zenodo

AI | MODEL SHARE

GitHub

Storing FAIR AI models

Hugging Face

OpenML

mlflow

DLHub

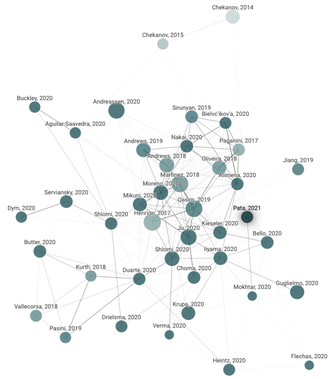
Papers / indexing / search / discovery

Cornell University

arXiv.org

iNSPIRE HEP

CONNECTED PAPERS



VISION

FAIR data repositories

Papers / indexing / search / discovery

Deploying FAIR AI models

Storing FAIR AI models

VISION

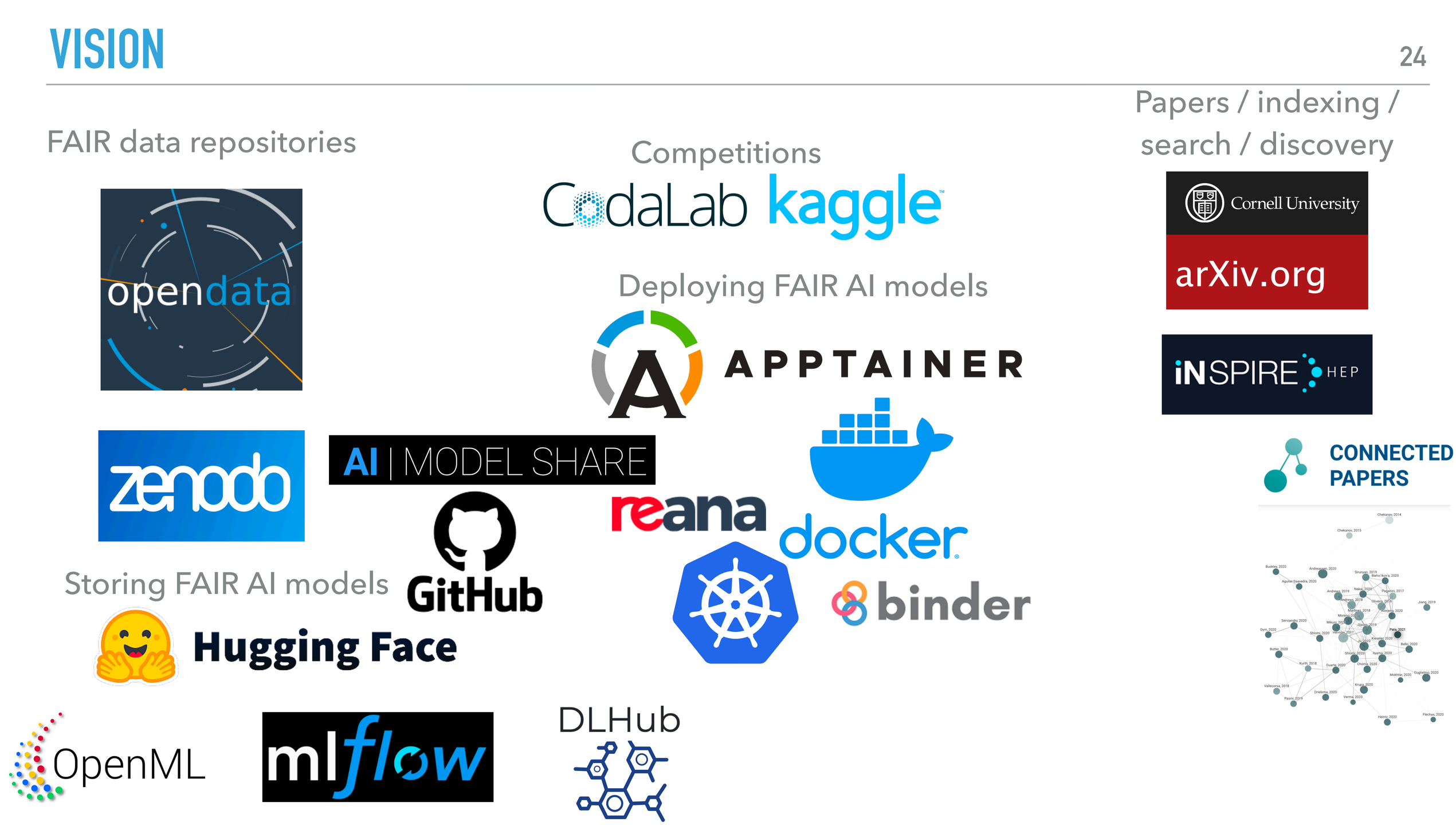
FAIR data repositories

Competitions

Papers / indexing / search / discovery

Deploying FAIR AI models

Storing FAIR AI models



VISION

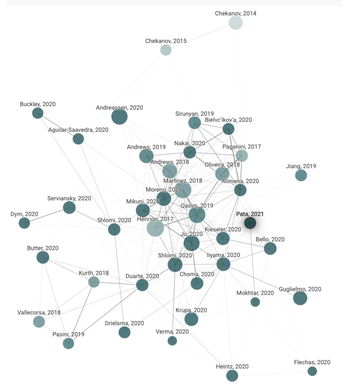
FAIR data repositories

Competitions

Papers / indexing / search / discovery

Deploying FAIR AI models

Storing FAIR AI models



VISION

FAIR data repositories

Competitions

Papers / indexing / search / discovery

Deploying FAIR AI models

Storing FAIR AI models

SUMMARY

- Goal of FAIR4HEP is to interpret and refine what FAIR means for HEP data/models
  - Enable “plug and play” datasets: allow for combinations of different computing resources
- Vision: connected services linking datasets, benchmark models (code), deployment servers, and publications to make everything more FAIR
  - Simpler discovery of new datasets and models
- Projects
  - Evaluate FAIRness of existing public datasets
  - **Standardize FAIR publication of AI models in HEP**
  - **Create example FAIR datasets and AI models**
  - Enhance existing services to make them more FAIR
- Welcome feedback!

FAIR4HEP TEAM (FAIR4HEP.GITHUB.IO)

Eliu Huerta
PI, ANL

Daniel S. Katz
Co-PI, UIUC

Volodymyr Kindratenko
Co-PI, UIUC

Mark Neubauer
Co-PI, UIUC

Zhizhen Zhao
Co-PI, UIUC

Priscilla Cushman
Co-PI, UMN

Andrew Furmanski
Co-PI, UMN

Vuk Mandic
Co-PI, UMN

Roger Rusack
Co-PI, UMN

Ju-Sun
Co-PI, UMN

Phil Harris
Co-PI, MIT

Javier Duarte
Co-PI, UCSD

+ many postdocs, PhD students, MS students, and undergrads!

U.S. DEPARTMENT OF ENERGY | Office of Science

JAVIER DUARTE, BILLY LI

CHEP

MAY 11, 2023

BACKUP

UCSD