MicroBooNE Public Data Sets:
a Collaborative Tool for LArTPC Software Development

G. Cerati (FNAL), on behalf of the MicroBooNE Collaboration

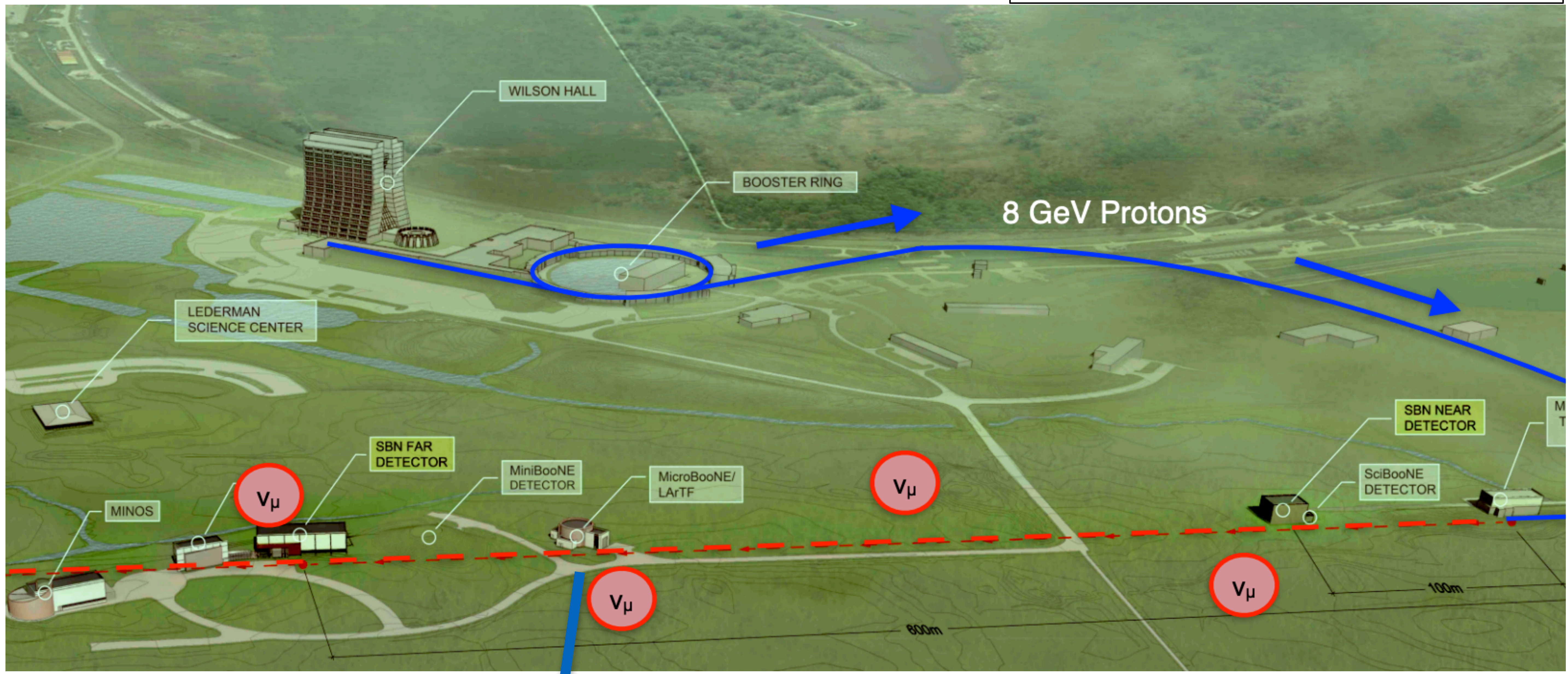
CHEP23 - Norfolk, VA

May 09, 2023

MicroBooNE

- Neutrino experiment at Fermilab, designed to test the MiniBooNE anomaly
  - ~same beam (**BNB**) and distance from source
- Broader experimental program:
  - Test short-baseline oscillations as part of SBN
  - BSM physics searches
  - nu-Ar cross sections
- Physics operations: 2015-2021
- Analyzed about 1/2 data, producing over 50 publications: https://microboone.fnal.gov/documents-publications/

SBN program: arXiv:1503.01520

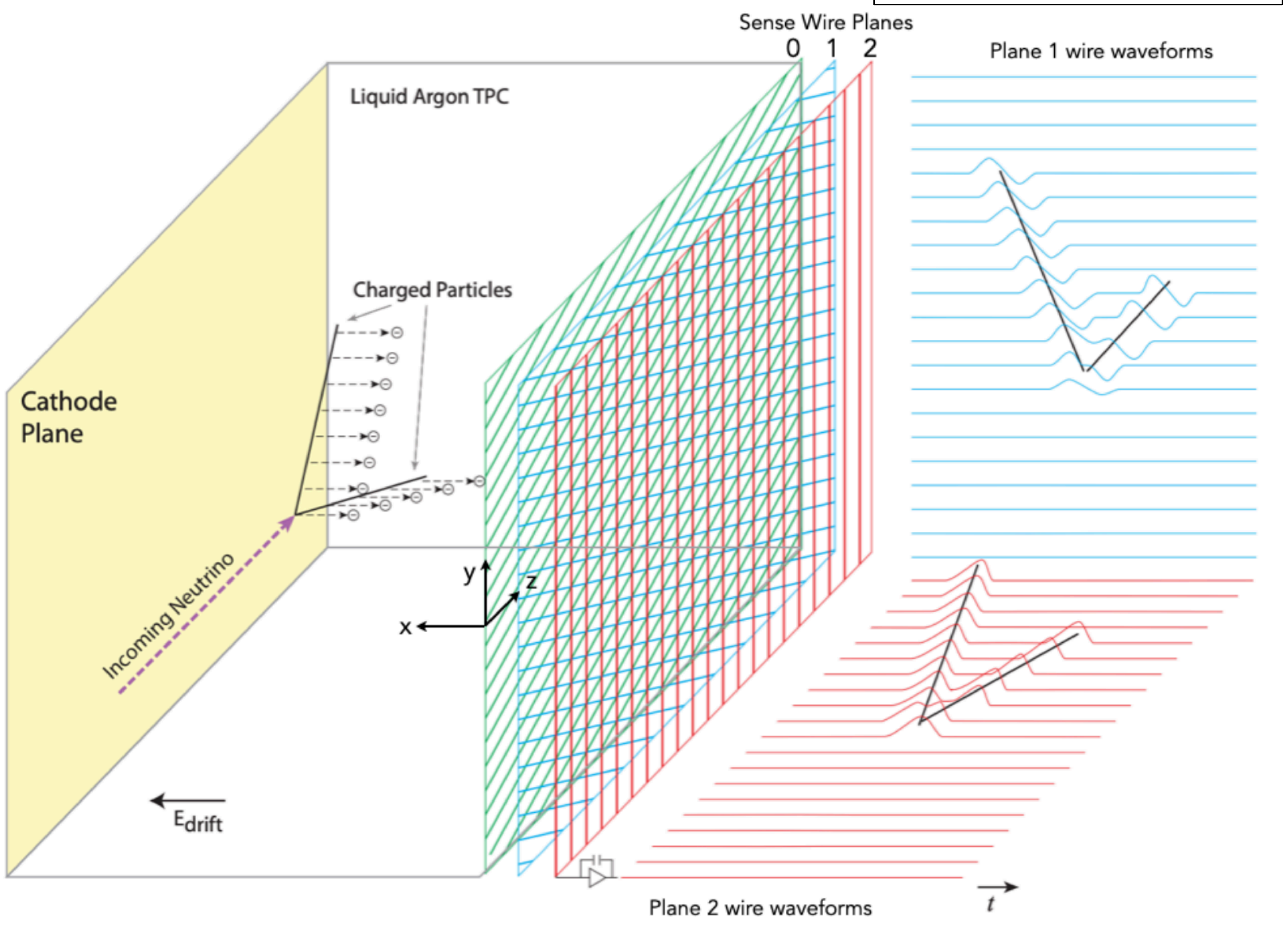


JINST 12, P02017 (2017)

MicroBooNE’s Liquid Argon Time Projection Chamber (LArTPC)

- Charged particles produced in neutrino interactions ionize the argon, ionization electrons drift in electric field towards anode planes
- Sense wires detect the incoming charge, producing beautiful detector data images

JINST 12, P02017 (2017)



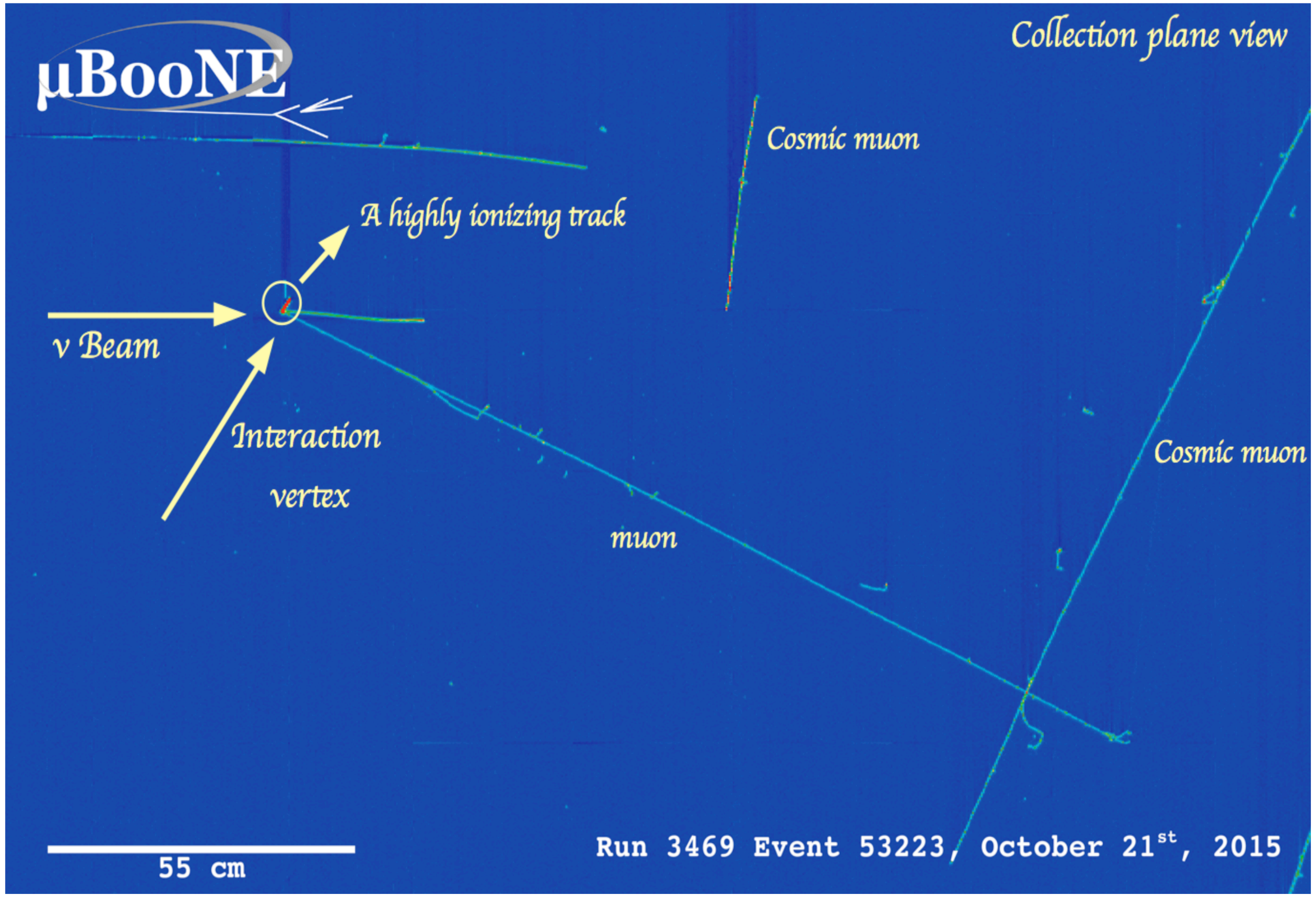
3 planes allow for 3D reco

Fermilab

MicroBooNE's Liquid Argon Time Projection Chamber (LArTPC)

- Charged particles produced in neutrino interactions ionize the argon, ionization electrons drift in electric field towards anode planes
- Sense wires detect the incoming charge, producing beautiful detector data images
- Full detail of neutrino interaction with O(mm) spatial resolution and calorimetric information
- Fast scintillation light detected by Optical system (PMT) for trigger & cosmic rejection

JINST 12, P02017 (2017)



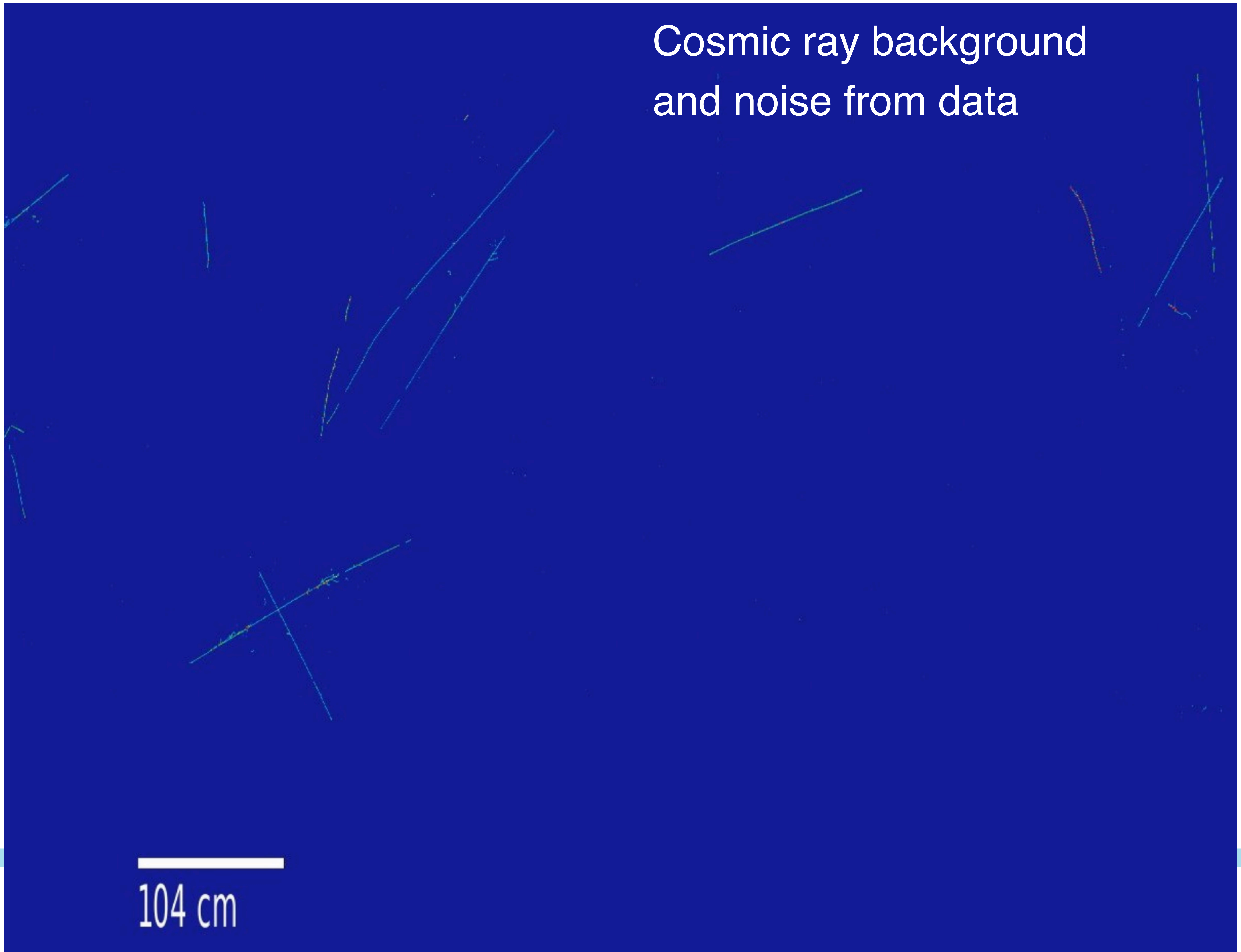
axes: time vs wire - color scale: charge

MicroBoNE open samples: motivation

- **Establish MicroBooNE as state of the art LArTPC technology.**
  - Attested primarily by our publications, but public datasets provide direct reference point.
- **Efficient collaboration with colleagues in LArTPC experiments, as well as computer scientists.**
  - SW development collaborations don't need an MoU and nor external public datasets.
  - Facilitate integration of tools with other LArTPC experiments (SBN and DUNE).
  - The output of external collaborations is directly usable within MicroBooNE.
- **Potentially attract developments from beyond our community.**
  - Data challenges, etc.

Implementation of open samples: overview

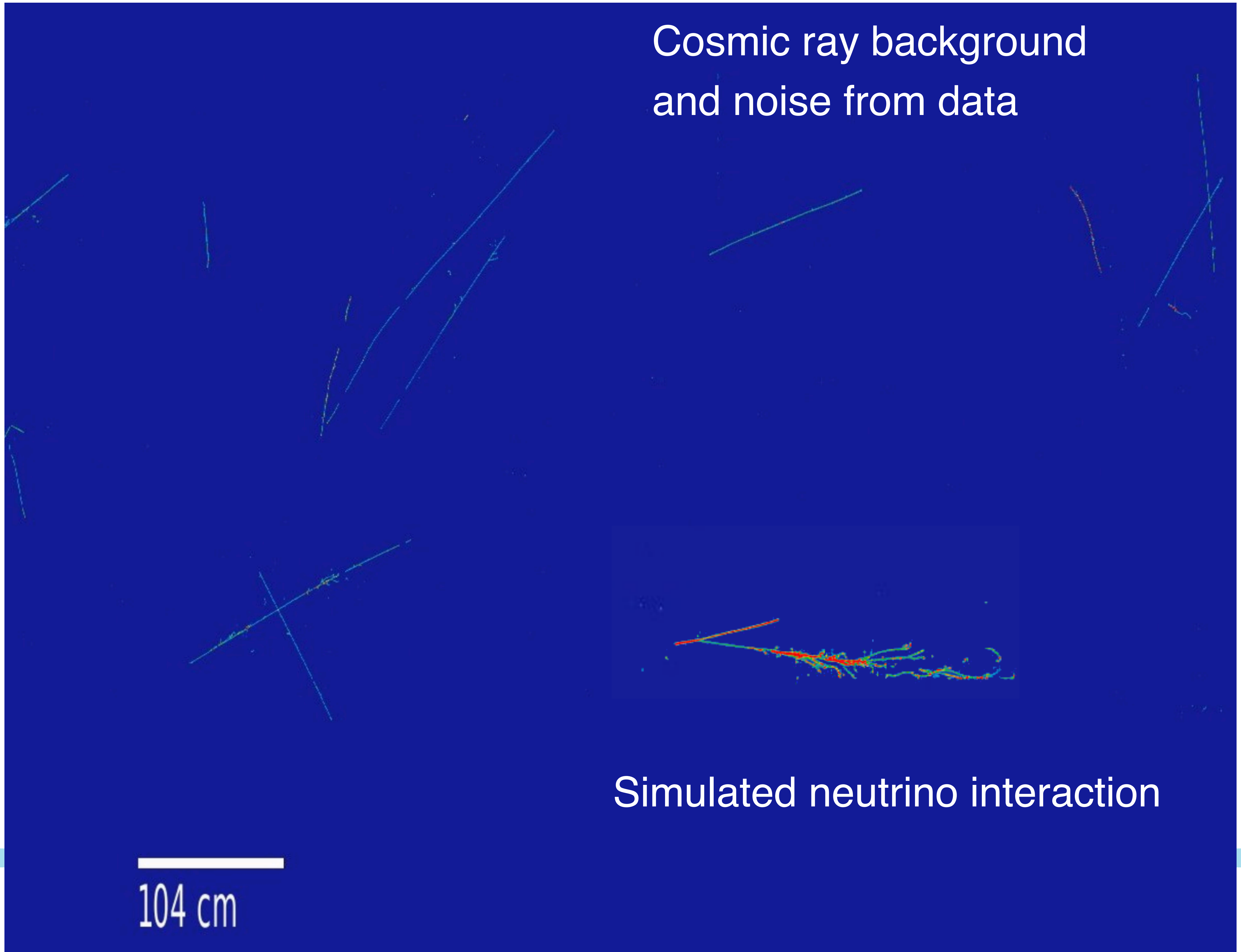
- Open two “overlay” samples: BNB **inclusive** and BNB intrinsic ν_e



Fermilab

Implementation of open samples: overview

- Open two “overlay” samples: BNB **inclusive** and BNB intrinsic ν_e



Fermilab

Implementation of open samples: overview

- Open two “overlay” samples: BNB **inclusive** and BNB intrinsic ν_e

- Inspired by **FAIR** principles (findable, accessible, interoperable, reusable data)

- Two formats: regular reconstructed **art/ROOT** and **HDF5**
  - respectively targeting LArTPC and broader data & computer science communities
- HDF5 files stored on **Zenodo**, providing citable DOI (digital object identifier) & versioning
- Artroot files stored on persistent **dCache** pool area and made accessible with **xrootd**
  - list of xrood urls stored with the corresponding HDF5 files on Zenodo

- Samples available under “cc-by” license. Template text for acknowledgment is provided.
  - requesting resulting software products to be made available

Dataset definitions

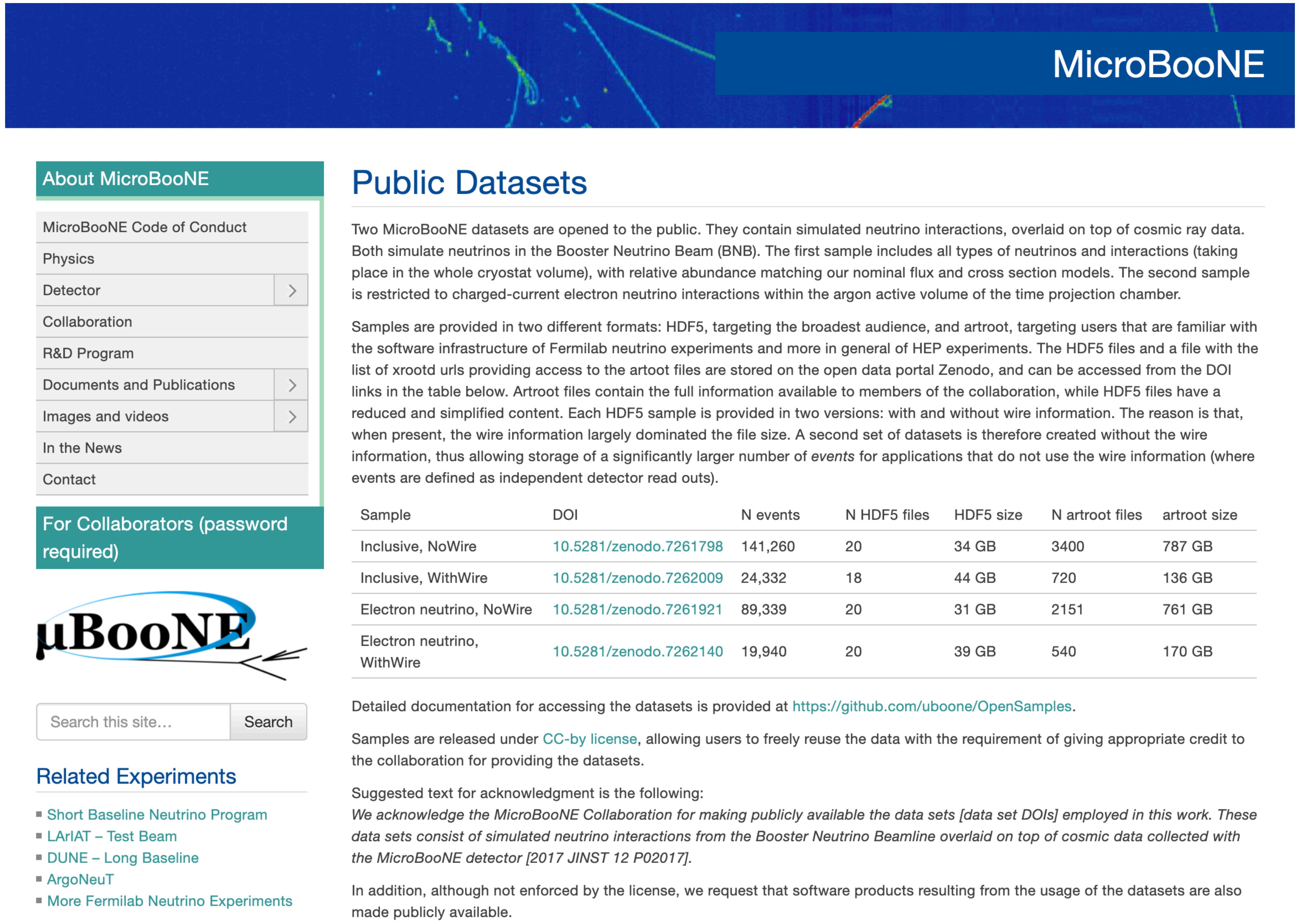
Each HDF5 sample comes in two flavors: with and without wire information (waveform).
Due to size requirements, sample with this information contain less events.

Sample	DOI	N events	N HDF5 files	HDF5 size	N artroot files	artroot size
Inclusive, NoWire	10.5281/zenodo.7261798	141,260	20	34 GB	3400	787 GB
Inclusive, WithWire	10.5281/zenodo.7262009	24,332	18	44 GB	720	136 GB
Electron neutrino, NoWire	10.5281/zenodo.7261921	89,339	20	31 GB	2151	761 GB
Electron neutrino, WithWire	10.5281/zenodo.7262140	19,940	20	39 GB	540	170 GB

Open BNB inclusive sample is a subsample of what internally available. We may open a larger sample upon request and if technically feasible.

Access point

- Entry point is the MicroBooNE website:
  - https://microboone.fnal.gov/documents-publications/public-datasets/



MicroBooNE

About MicroBooNE

- MicroBooNE Code of Conduct
- Physics
- Detector
- Collaboration
- R&D Program
- Documents and Publications
- Images and videos
- In the News
- Contact

For Collaborators (password required)

μBooNE

Search this site… Search

Related Experiments

- Short Baseline Neutrino Program
- LArIAT – Test Beam
- DUNE – Long Baseline
- ArgoNeuT
- More Fermilab Neutrino Experiments

Public Datasets

Two MicroBooNE datasets are opened to the public. They contain simulated neutrino interactions, overlaid on top of cosmic ray data. Both simulate neutrinos in the Booster Neutrino Beam (BNB). The first sample includes all types of neutrinos and interactions (taking place in the whole cryostat volume), with relative abundance matching our nominal flux and cross section models. The second sample is restricted to charged-current electron neutrino interactions within the argon active volume of the time projection chamber.

Samples are provided in two different formats: HDF5, targeting the broadest audience, and artroot, targeting users that are familiar with the software infrastructure of Fermilab neutrino experiments and more in general of HEP experiments. The HDF5 files and a file with the list of xrootd urls providing access to the artoot files are stored on the open data portal Zenodo, and can be accessed from the DOI links in the table below. Artroot files contain the full information available to members of the collaboration, while HDF5 files have a reduced and simplified content. Each HDF5 sample is provided in two versions: with and without wire information. The reason is that, when present, the wire information largely dominated the file size. A second set of datasets is therefore created without the wire information, thus allowing storage of a significantly larger number of *events* for applications that do not use the wire information (where events are defined as independent detector read outs).

Sample	DOI	N events	N HDF5 files	HDF5 size	N artroot files	artroot size
Inclusive, NoWire	10.5281/zenodo.7261798	141,260	20	34 GB	3400	787 GB
Inclusive, WithWire	10.5281/zenodo.7262009	24,332	18	44 GB	720	136 GB
Electron neutrino, NoWire	10.5281/zenodo.7261921	89,339	20	31 GB	2151	761 GB
Electron neutrino, WithWire	10.5281/zenodo.7262140	19,940	20	39 GB	540	170 GB

Detailed documentation for accessing the datasets is provided at https://github.com/uboone/OpenSamples.

Samples are released under CC-by license, allowing users to freely reuse the data with the requirement of giving appropriate credit to the collaboration for providing the datasets.

Suggested text for acknowledgment is the following:
We acknowledge the MicroBooNE Collaboration for making publicly available the data sets [data set DOIs] employed in this work. These data sets consist of simulated neutrino interactions from the Booster Neutrino Beamline overlaid on top of cosmic data collected with the MicroBooNE detector [2017 JINST 12 P02017].

In addition, although not enforced by the license, we request that software products resulting from the usage of the datasets are also made publicly available.

Description

Links to Zenodo

Link to documentation

Info about license and citation

art/ROOT format: definition and documentation

- Target users of this format is the **LArTPC community**, i.e. physicists already familiar with the LArSoft software environment

- art/ROOT files include the **full information** available to the Collaboration members, both at simulation and reconstruction level

- Documentation assumes **prior knowledge** of these tools and consists of:
  - description of the samples and list of data products stored
    - https://github.com/uboone/OpenSamples/blob/v01/file-content-artroot.md
  - links to documentation websites (LArSoft, xrootd, etc…)
  - instructions to setup the software release (uboonecode and LArSoft) from CVMFS
  - link to module for creating HDF5 files as example of how to access the artroot content

HDF5 format: scope and file content

- HDF5 include a **reduced subset** of the art/ROOT information
  - In a simplified format for **usage by non-experts**. Still, designed to allow a wide range applications.

- The following information is stored in the HDF5 files:
  - Noise-filtered and deconvolved wire waveforms in regions of interest
  - TPC Hit information
  - Optical Hit and Flash information
  - MC Truth information
    - incoming neutrino properties, energy deposits as associated to hits, Geant4 particles

- In addition we provide information for **benchmarking** purposes:
  - Based on the Pandora reconstruction package [Eur. Phys. J. C78, 1, 82 (2018)]
  - E.g.: neutrino identification, track-shower classification, interaction and cluster hit mapping,…

Documentation - HDF5

- Documentation mainly consists of **notebooks** for demonstration of usage:
  - https://github.com/uboone/OpenSamples/tree/v01
  - Recipe for installing required packages in a **conda environment** with minimal dependencies
  - Use **pynuml** for handling file I/O
  - Notebooks are also briefly introduced to clarify their purpose
  - **Auxiliary tools**: functions for basic detector navigation and minimal plotting utils

MicroBooNE open samples

Two MicroBooNE datasets are opened to the public. They contain simulated neutrino interactions, overlaid on top of cosmic ray data. Both simulate neutrinos in the Booster Neutrino Beam (BNB). The first sample includes all types of neutrinos and interactions (taking place in the whole cryostat volume), with relative abundance matching our nominal flux and cross section models. The second sample is restricted to charged-current electron neutrino interactions within the argon active volume of the time projection chamber.

Samples are provided in two different formats: HDF5, targeting the broadest audience, and artroot, targeting users that are familiar with the software infrastructure of Fermilab neutrino experiments and more in general of HEP experiments. The HDF5 files and a file with the list of xrootd urls providing access to the artoot files are stored on the open data portal Zenodo, and can be accessed from the DOI links in the table below. Artroot files contain the full information available to members of the collaboration, while HDF5 files have a reduced and simplified content. Each HDF5 sample is provided in two versions: with and without wire information. The reason is that, when present, the wire information largely dominated the file size. A second set of datasets is therefore created without the wire information, thus allowing storage of a significantly larger number of *events* for applications that do not use the wire information (where events are defined as independent detector read outs).

Sample	DOI	N events	N HDF5 files	HDF5 size	N artroot files	artroot size
Inclusive, NoWire	10.5281/zenodo.7261798	141,260	20	34 GB	3400	787 GB
Inclusive, WithWire	10.5281/zenodo.7262009	24,332	18	44 GB	720	136 GB
Electron neutrino, NoWire	10.5281/zenodo.7261921	89,339	20	31 GB	2151	761 GB
Electron neutrino, WithWire	10.5281/zenodo.7262140	19,940	20	39 GB	540	170 GB

HDF5 format

This section provides documentation on how to access the information included in the HDF5 files. Examples demonstrating how to use the data is provided in the form of jupyter notebooks. The full description of the file content is also provided.

The HDF5 format is a product of the HDF5 group. In the notebookes we open the files using the `File` class from pynuml, which internally relies on h5py. We also use p5concat to merge files and to add auxiliary data for faster lookup of related information across different tables.

Jupyter notebooks

Local Setup

Fermilab

Documentation - HDF5

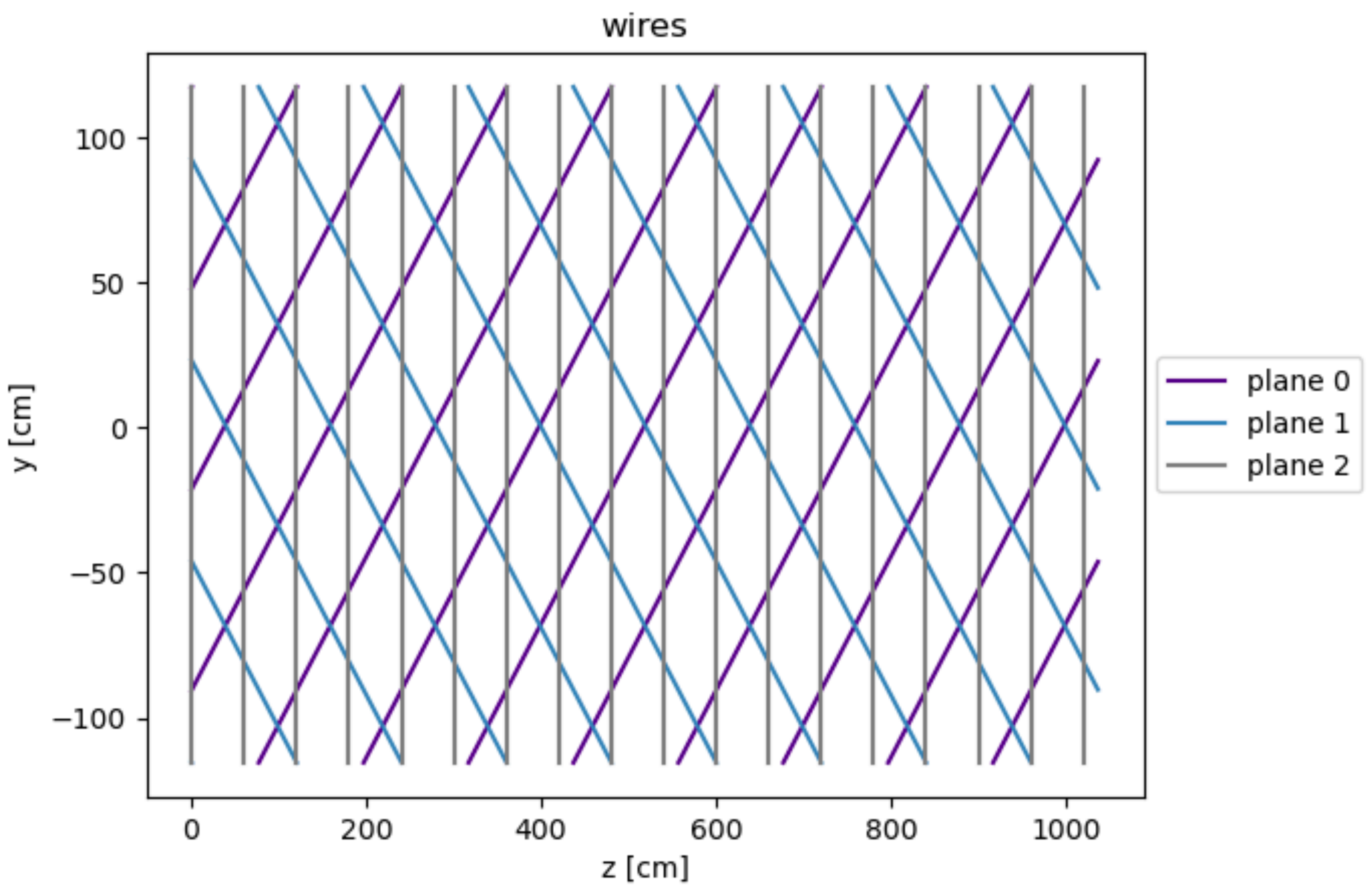
- It also includes a documentation of the file content, in a table with brief description of each element stored in the dataset

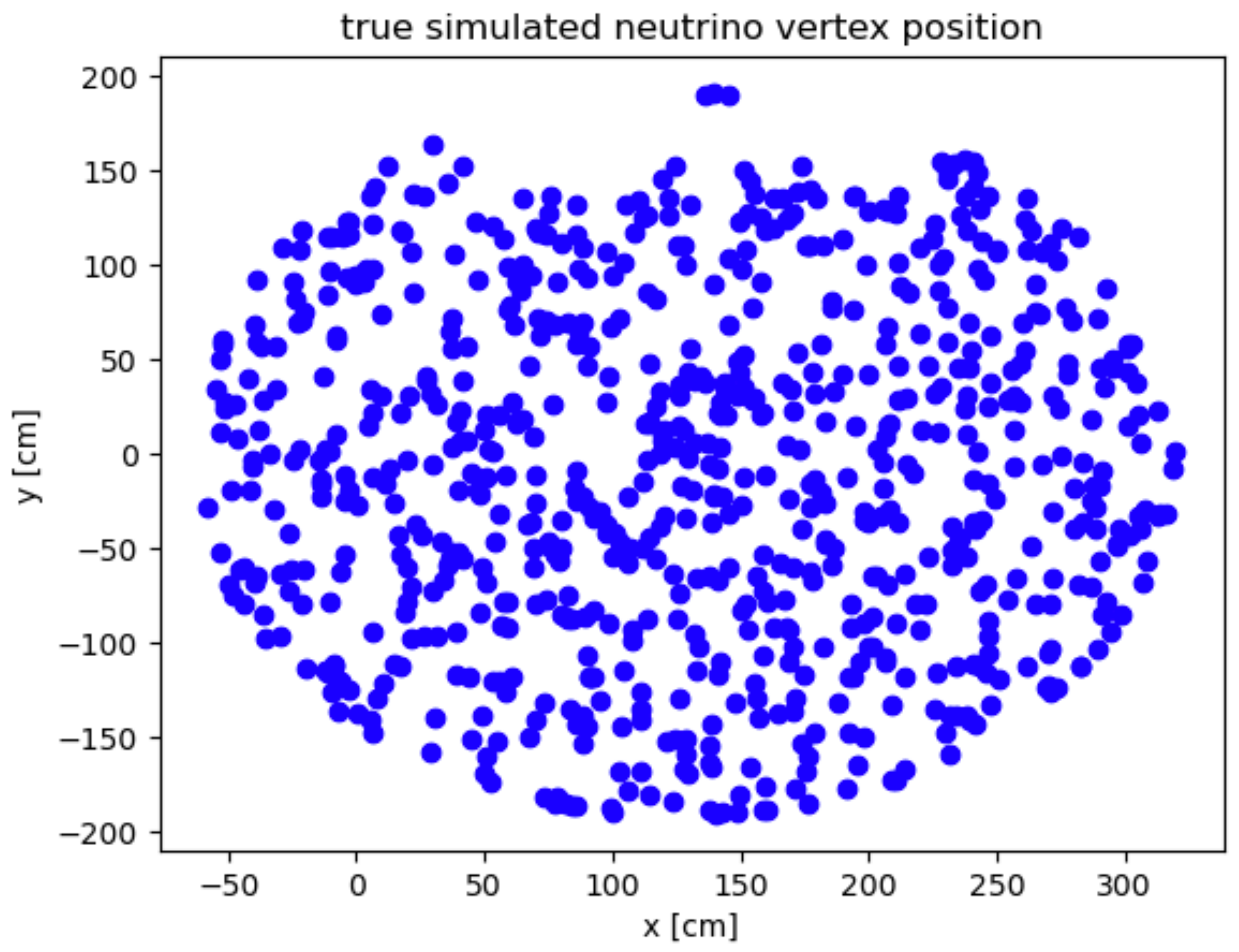
File Entry	Type	N Elements	Description
/	Group		Main entry point of the file.
/event_table	Group		Table storing information about a single detector readout and a single simulated neutrino interaction.
/event_table/event_id	Dataset	3	Run/Subrun/Event number for a detector readout.
/event_table/event_id.seq_cnt	Dataset	2	Auxiliary information added in post-processing step for simple grouping and fast access of table entries separated by event.
/event_table/is_cc	Dataset	1	If 1 the simulated neutrino interaction is charged-current, if 0 it is neutral-current.
/event_table/lep_energy	Dataset	1	Simulated energy of the lepton outgoing from the neutrino interaction (in GeV).
/event_table/nu_dir	Dataset	3	Initial direction of the simulated neutrino interacting in the detector (3D cartesian coordinates).
/event_table/nu_energy	Dataset	1	Simulated energy of the interacting neutrino (in GeV).
/event_table/nu_pdg	Dataset	1	Particle Data Group (PDG) particle code for the interacting neutrino. See https://pdg.lbl.gov/2022/reviews/rpp2022-rev-monte-carlo-numbering.pdf.
/event_table/nu_vtx	Dataset	3	Simulated position of neutrino interaction (3D cartesian coordinates, in cm). This quantity is to be used to compare with e.g. the detector boundaries.

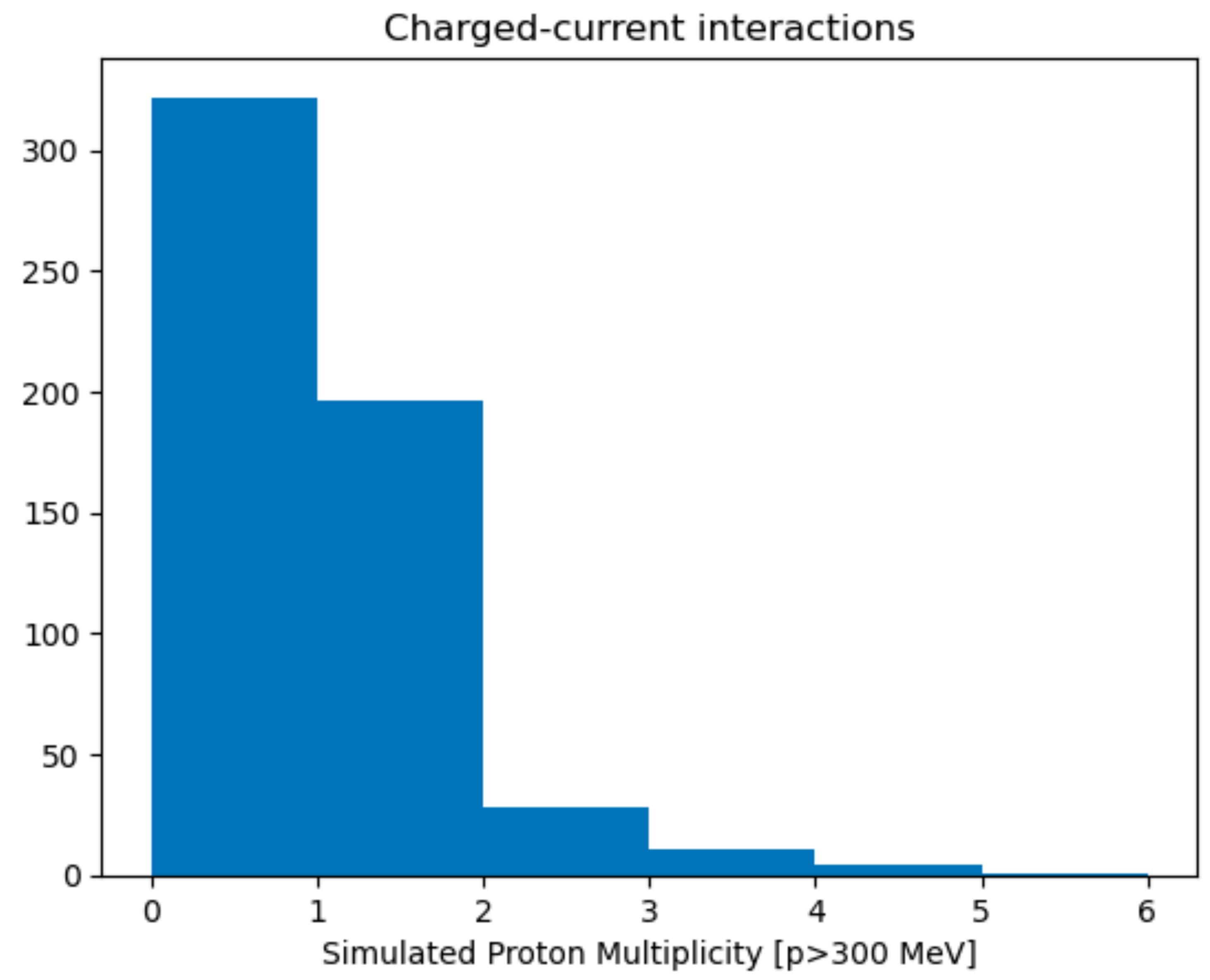
Highlights from notebooks: Sample Exploration

Goal of this notebook is to familiarize with the sample content and with tools provided to understand the LArTPC detector properties.

E.g.: wire positions and intersections, neutrino interaction position in the cryostat, simulated particle multiplicities.





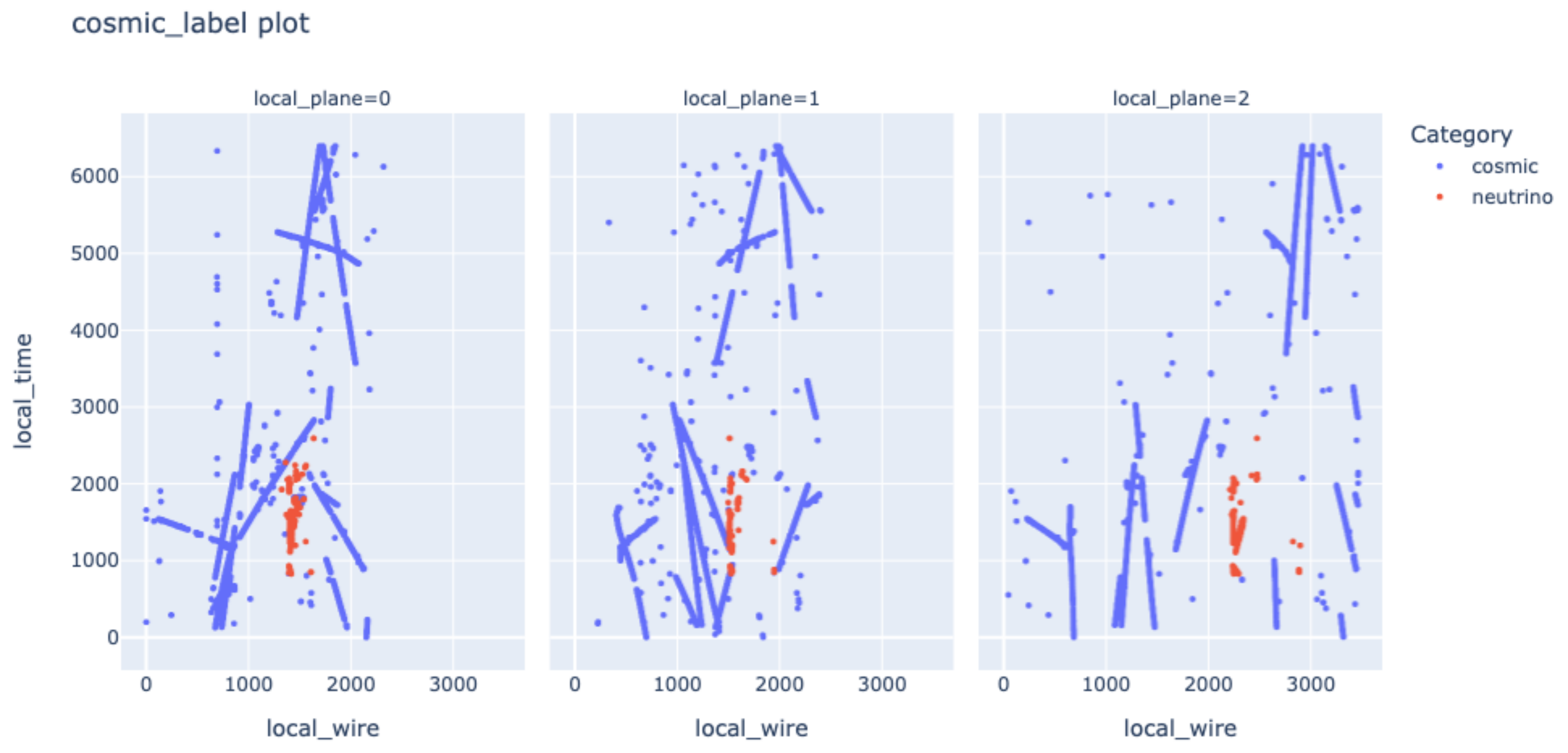


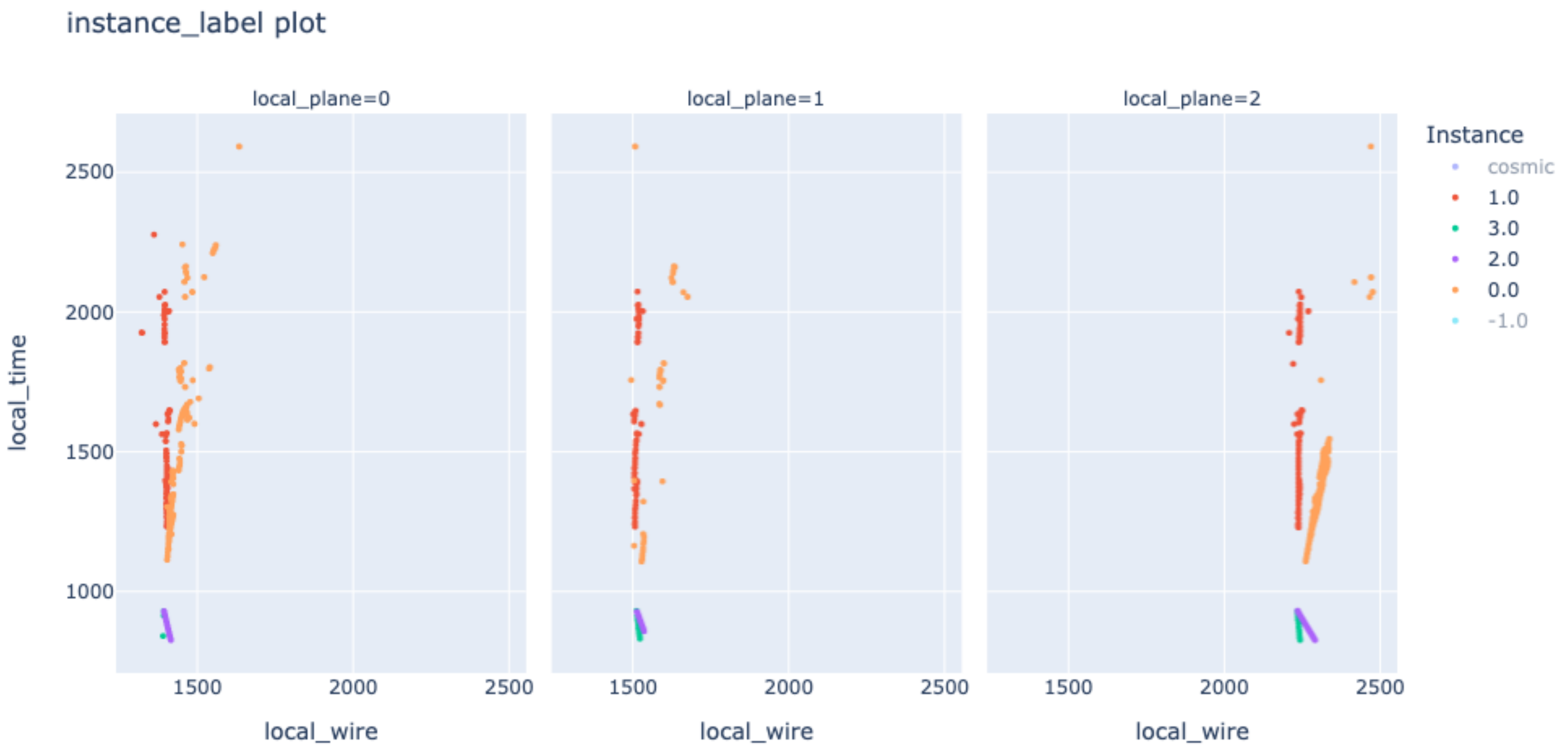
Fermilab

Highlights from notebooks: Hit Labeling

Goal of this notebook is to demonstrate ground-truth labeling of TPC hits according to different categorizations. Each categorization can be the target of specific algorithms / network training.

E.g.: neutrino identification, semantic segmentation, instance segmentation.





Highlights from notebooks: WireImage

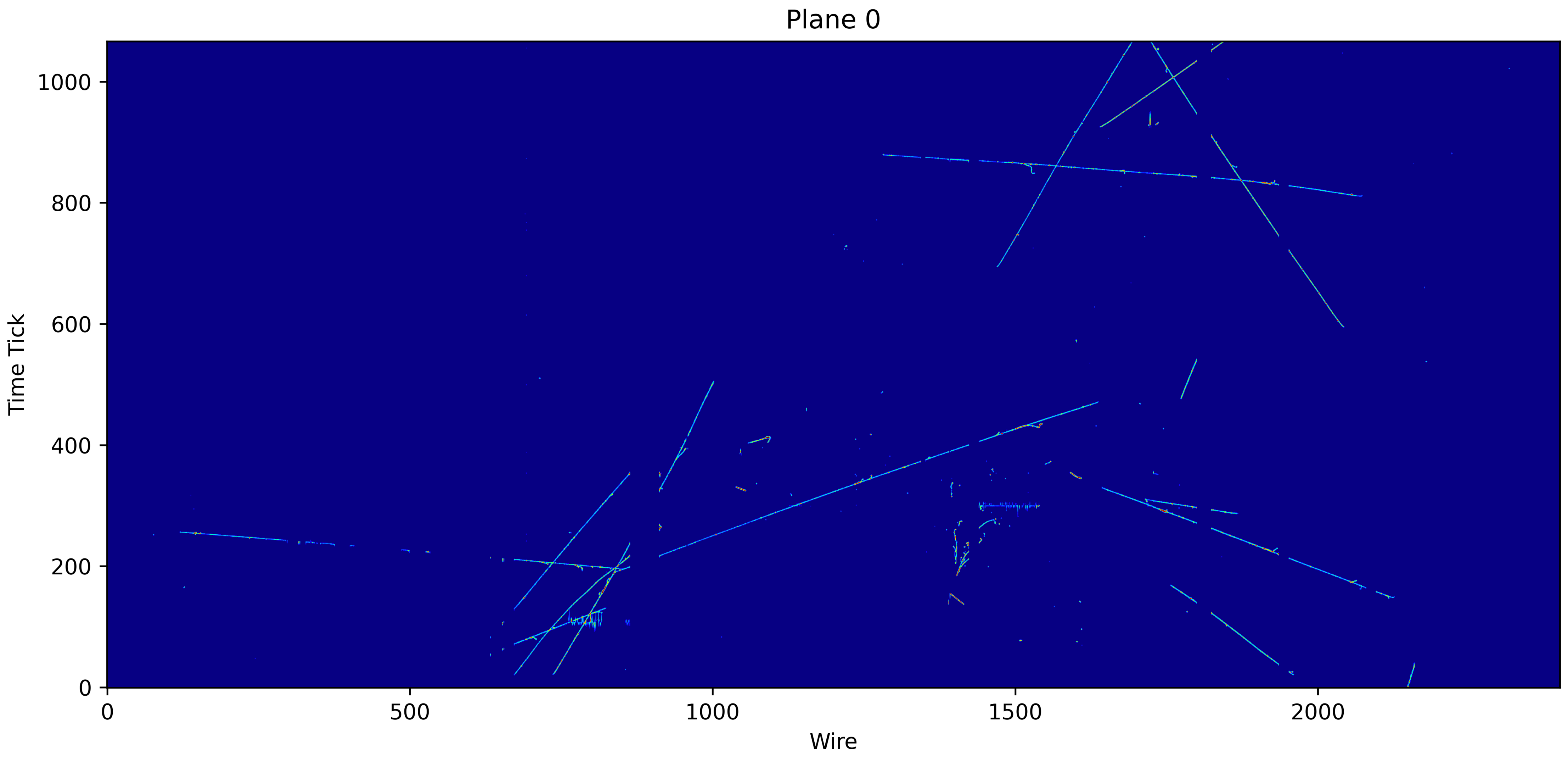
Needs “WithWire” samples containing waveform info

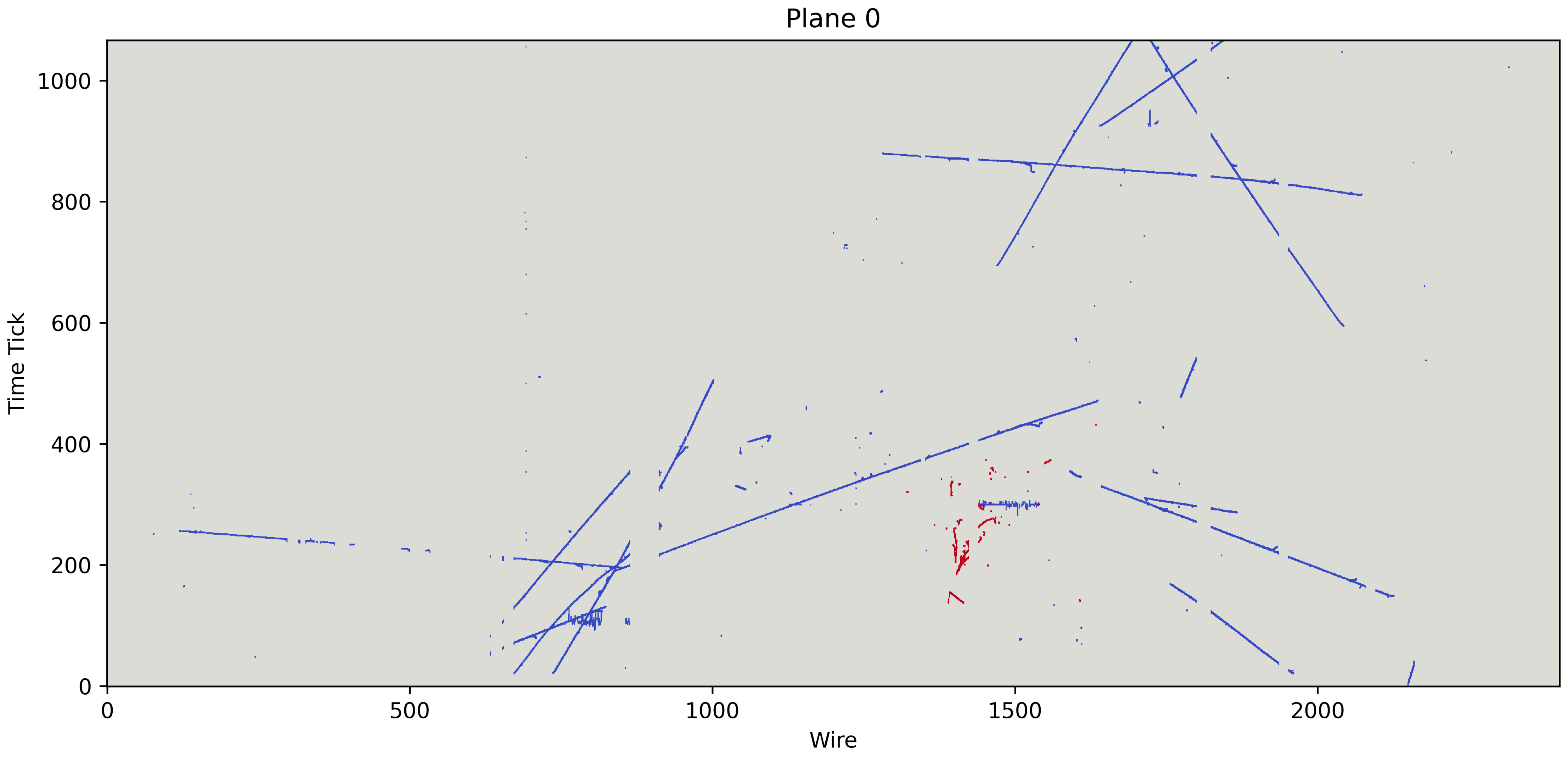
This notebook demonstrates the TPC data visualization in image format.

It can be used for visual data processing, e.g. Convolutional Neural Networks.

Ground truth at wire level not provided, but can be extracted matching the waveform and hit information.

Phys. Rev. D103, 052012 (2021)
Phys. Rev. D103, 092003 (2021)
Phys. Rev. D99, 092001 (2019)
JINST 12, P03011 (2017)





Highlights from notebooks: Pandora metrics

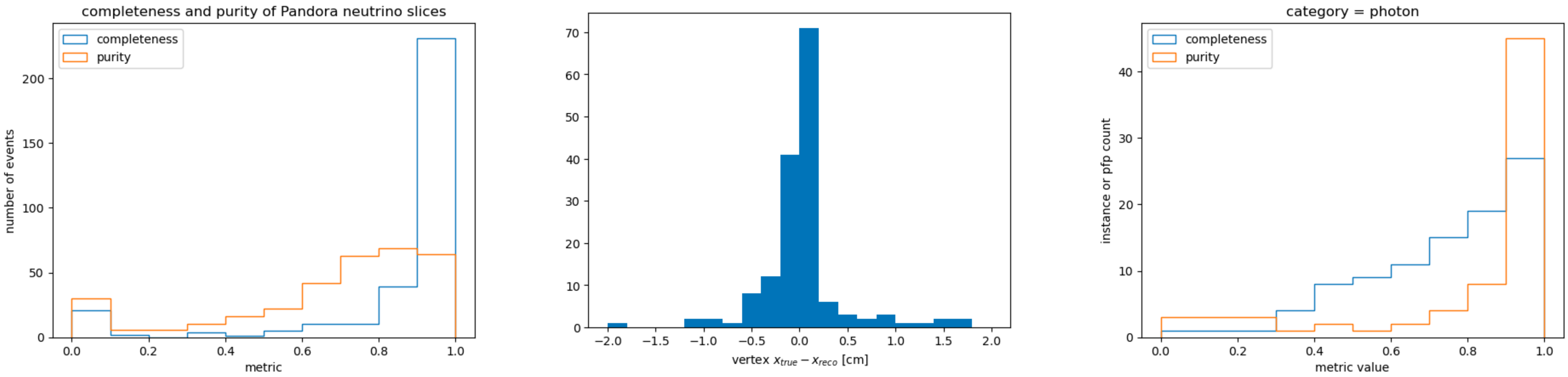
Eur.Phys.J.C 78 (2018) 1, 82

Purpose of this notebook is to introduce the definition of important metrics, and produce performance results obtained using Pandora.

E.g.: Purity and completeness at neutrino interaction or particle level, vertex resolution.

purity = $\mathrm{Nhit}_{\mathrm{true,found}}$ / $\mathrm{Nhit}_{\mathrm{found}}$

completeness = $\mathrm{Nhit}_{\mathrm{true,found}}$ / $\mathrm{Nhit}_{\mathrm{true}}$



Highlights from notebooks: Optical Information

JINST 12, P02017 (2017)

Purpose of this notebook is to demonstrate the usage of the optical detector information.

E.g.: Optical Hit properties, their clustering in time into "flash" objects, comparison of flash and neutrino TPC hit barycenters

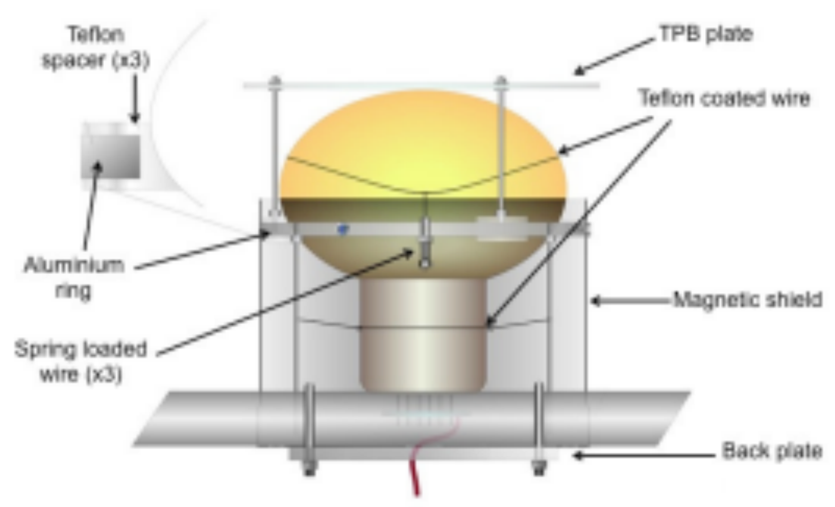
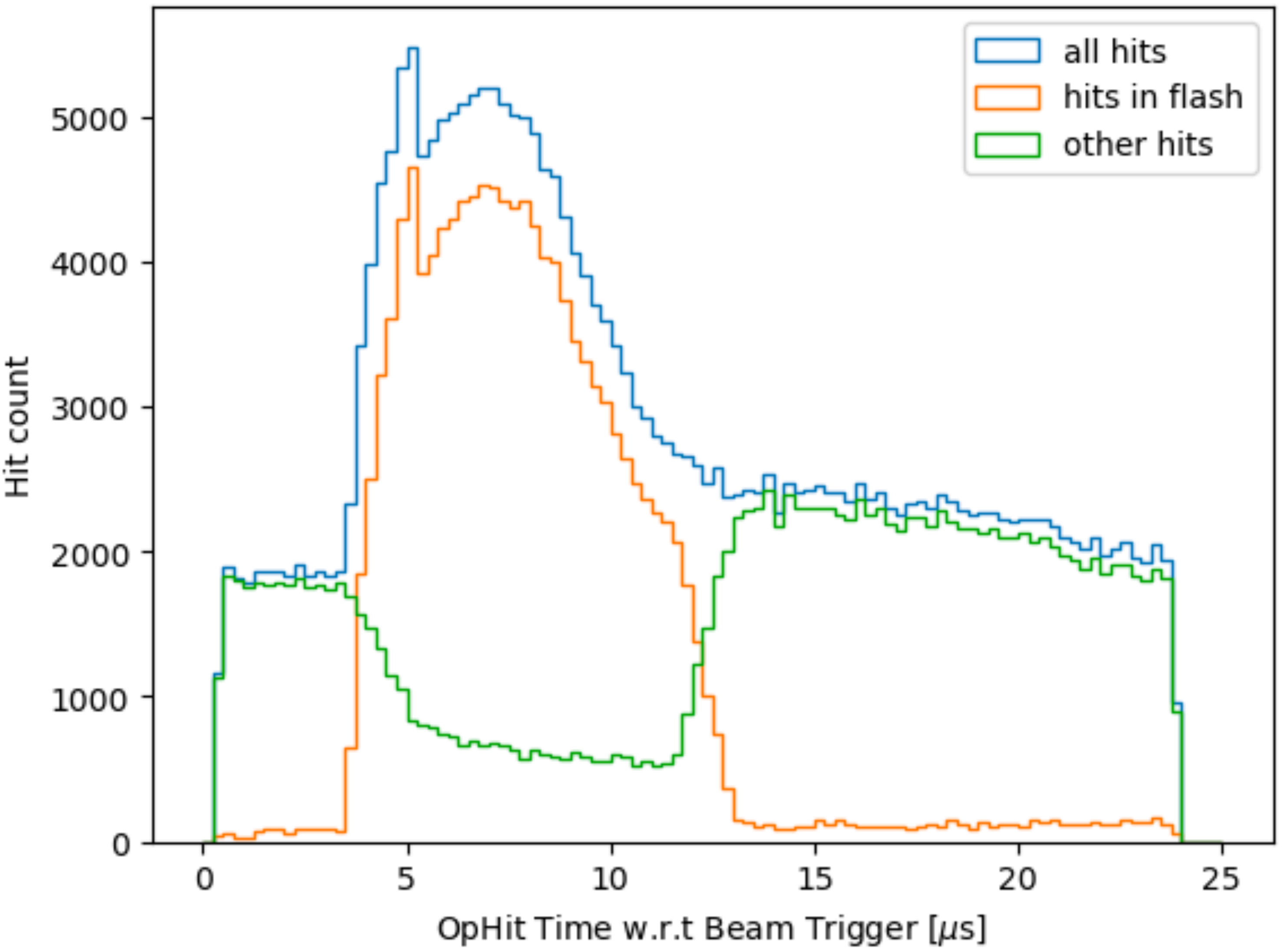
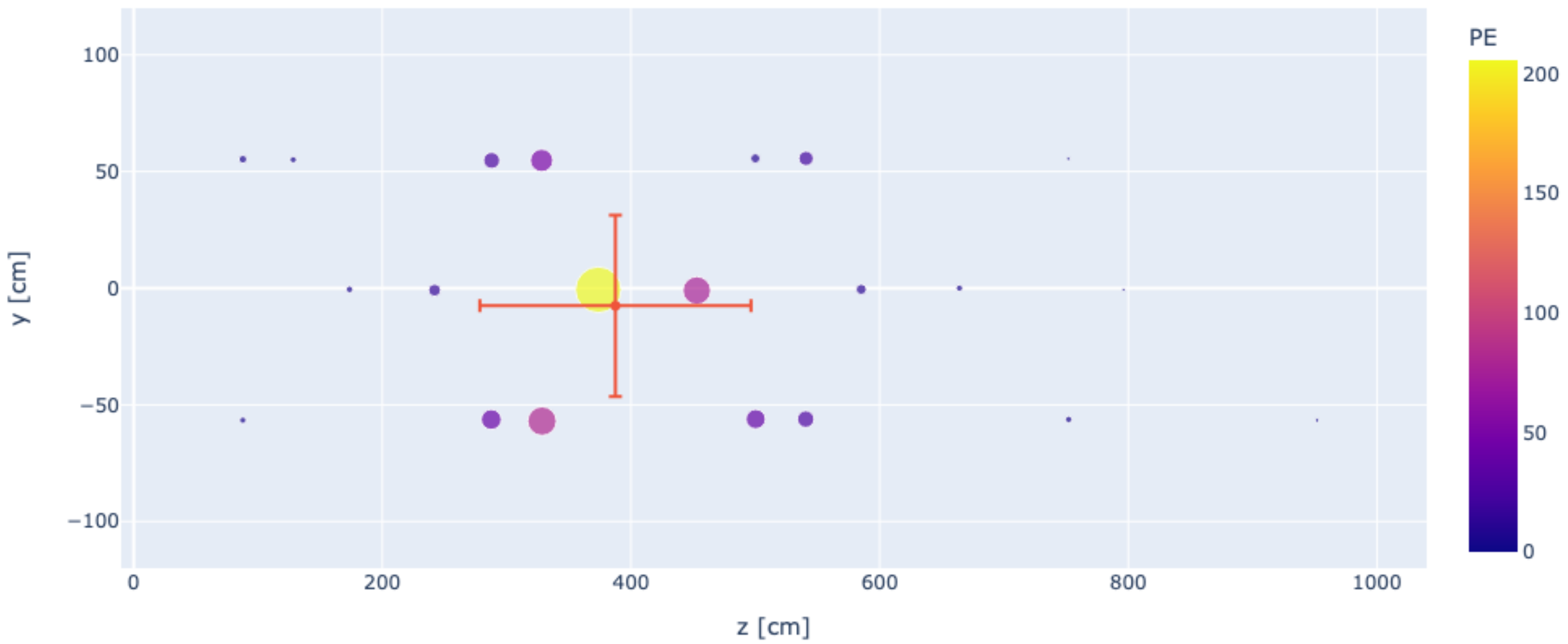
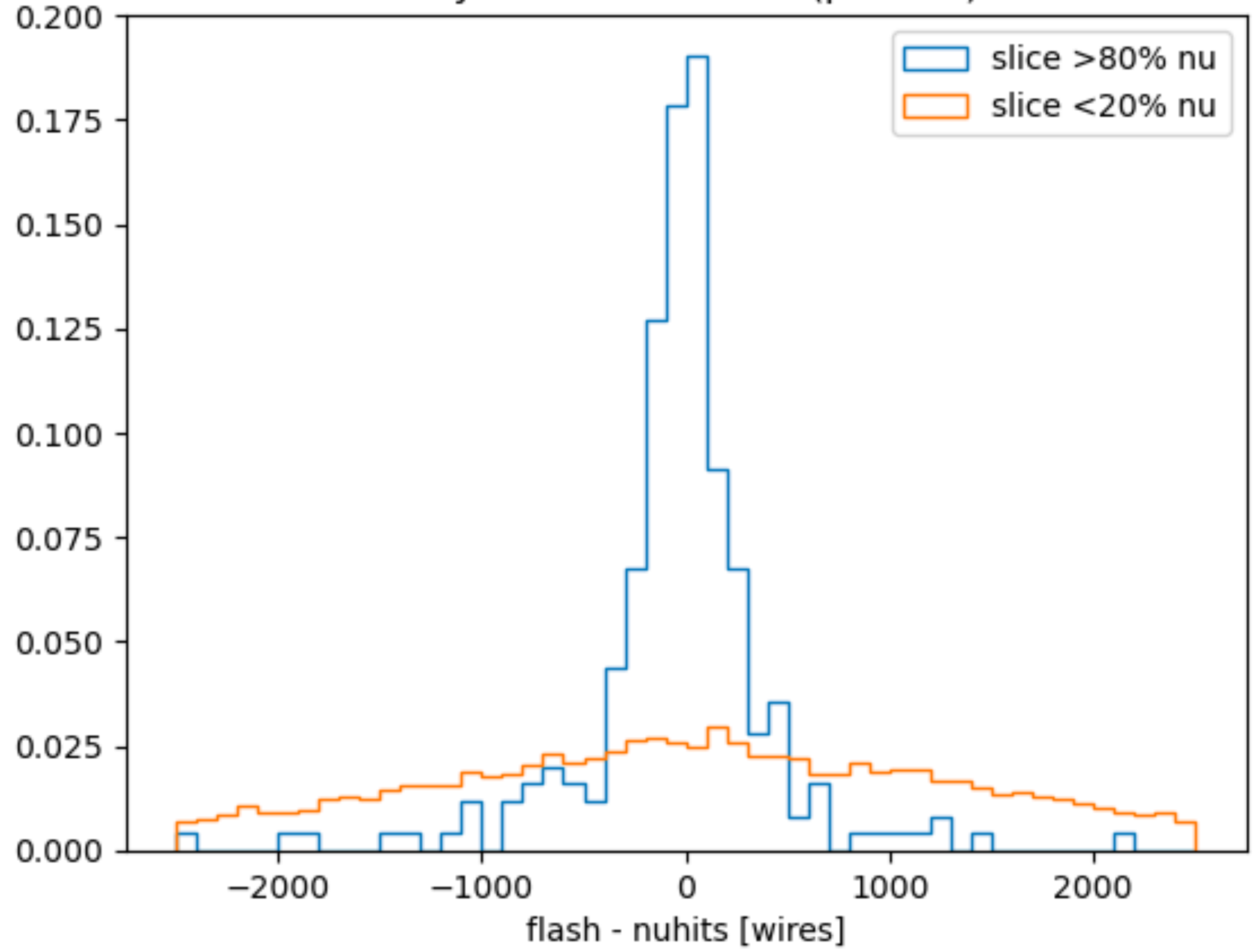


Figure 29. Left: diagram of the optical unit; Right: units mounted in MicroBooNE, immediately prior to LArTPC installation.







Fermilab

Conclusions

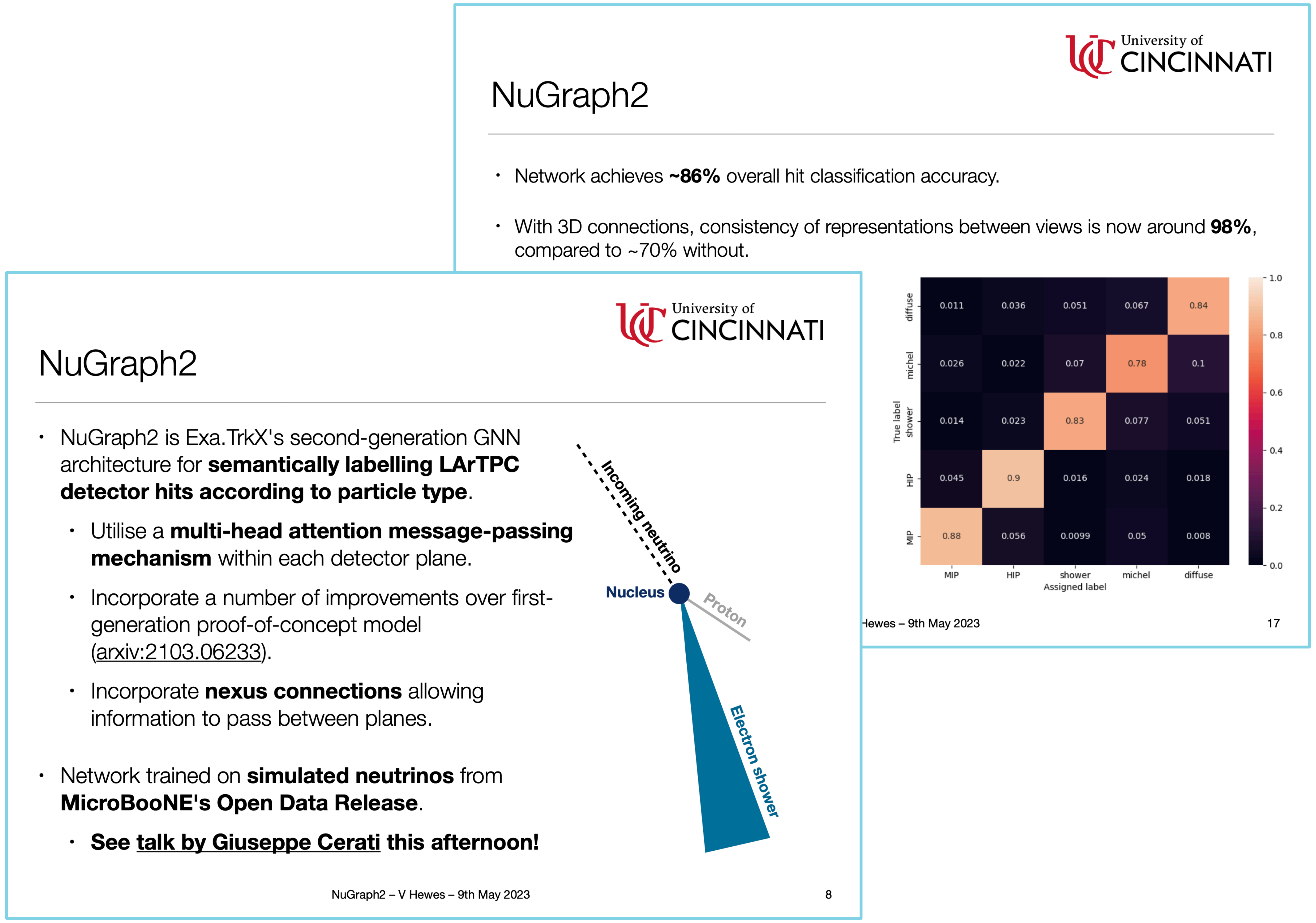
- MicroBooNE has opened samples for collaborative software development, available on Zenodo and via xrootd

- Software development and AI applications for LArTPC can benefit from them:
  - format can target images/CNN or other applications (e.g. GNN based on hits)
  - rich documentation for usage of these data sets
  - size of sample is enough for training
  - labeling examples can represent targets of ML applications
  - reference metrics from Pandora
  - enable porting of application to/from MicroBooNE

- Stats on Zenodo indicate hundreds of downloads already!

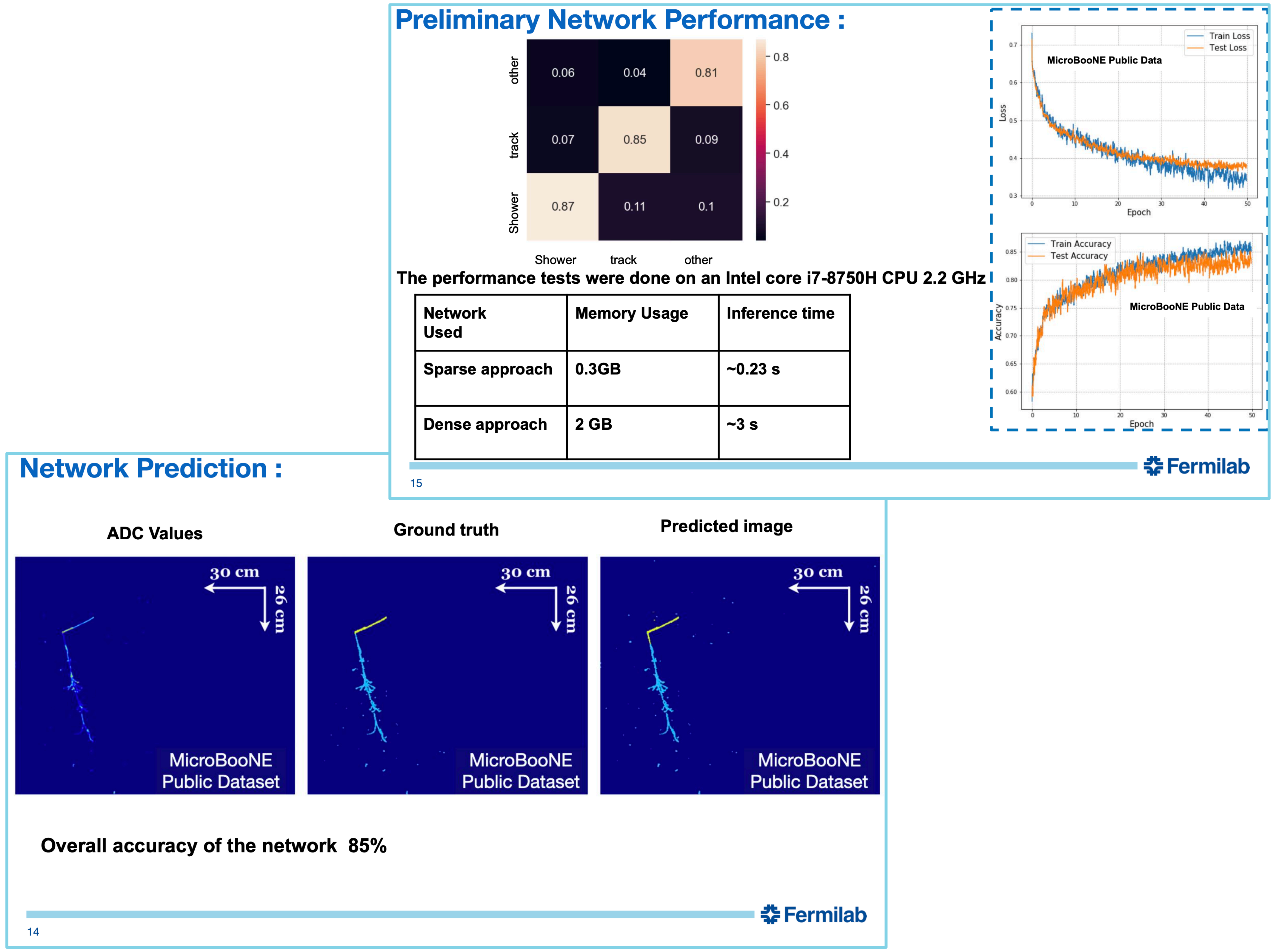
- Please reach out if you have questions or requests for more data/information!

MicroBooNE open data @ CHEP23



Graph Neural Network for 3D Reconstruction in Liquid Argon Time Projection Chambers

V Hewes - Track 9



Online tagging and triggering with deep learning AI for next generation particle imaging detector

Meghna Bhattacharya - Track 2

… to be continued!?

Backup

FAIR Principles

- How to release a dataset that can be usable by the widest possible set of users?

- FAIR Principles provide guidelines for this purpose
  - **F**indable
  - **A**ccessible
  - **I**nteroperable
  - **R**eusable
  - Info at https://www.go-fair.org/fair-principles/

FAIR Principles

- **Findable**
  - The first step in (re)using data is to find them. **Metadata and data should be easy to find** for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- **Accessible**
  - Once the user finds the required data, she/he/they need to **know how they can be accessed**, possibly including authentication and authorisation.

FAIR Principles

- **Interoperable**
  - The data usually need to be **integrated with other data**. In addition, the data need to interoperate with **applications or workflows** for analysis, storage, and processing.
- **Reusable**
  - The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, **metadata and data should be well-described so that they can be replicated and/or combined** in different settings.

List of principles

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata (defined by R1 below)
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
- A1.1 The protocol is open, free, and universally implementable
- A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

- I1. (Meta)data use a formal, accessible, 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 are richly described with a plurality of accurate and relevant attributes
- R1.1. (Meta)data are released with a clear and accessible data usage license
- R1.2. (Meta)data are associated with detailed provenance
- R1.3. (Meta)data meet domain-relevant community standards

FAIR Principles in HEP

https://fair4hep.github.io/

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

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[27]) 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[28] 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 2. Interoperable and **Reusable** principle assessment checks for 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.[29,30], using public software[31] 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.

arXiv:2108.02214