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

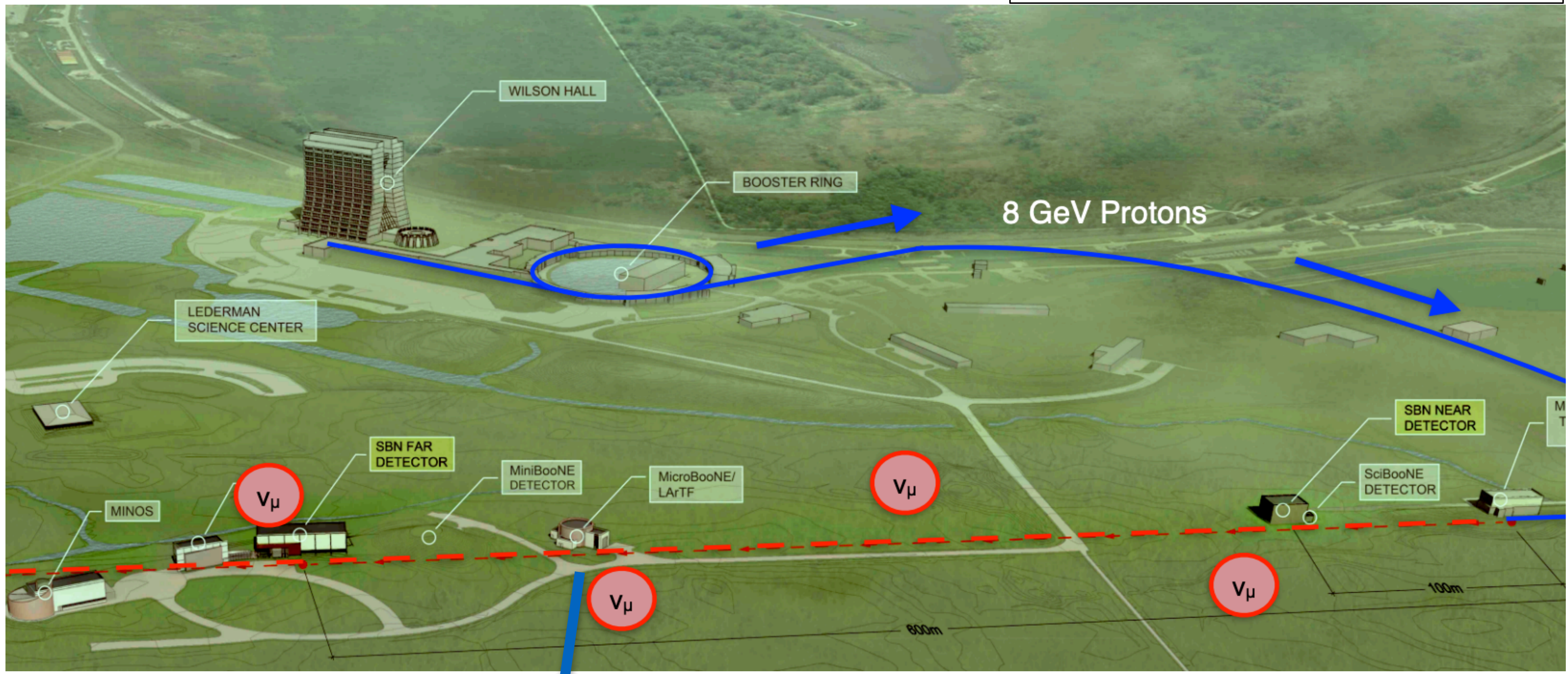
G. Cerati (FNAL)

AD4HEP — Nevis Labs

June 18, 2025

# MicroBooNE

SBN program: arXiv:1503.01520




JINST 12, P02017 (2017)

- 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/



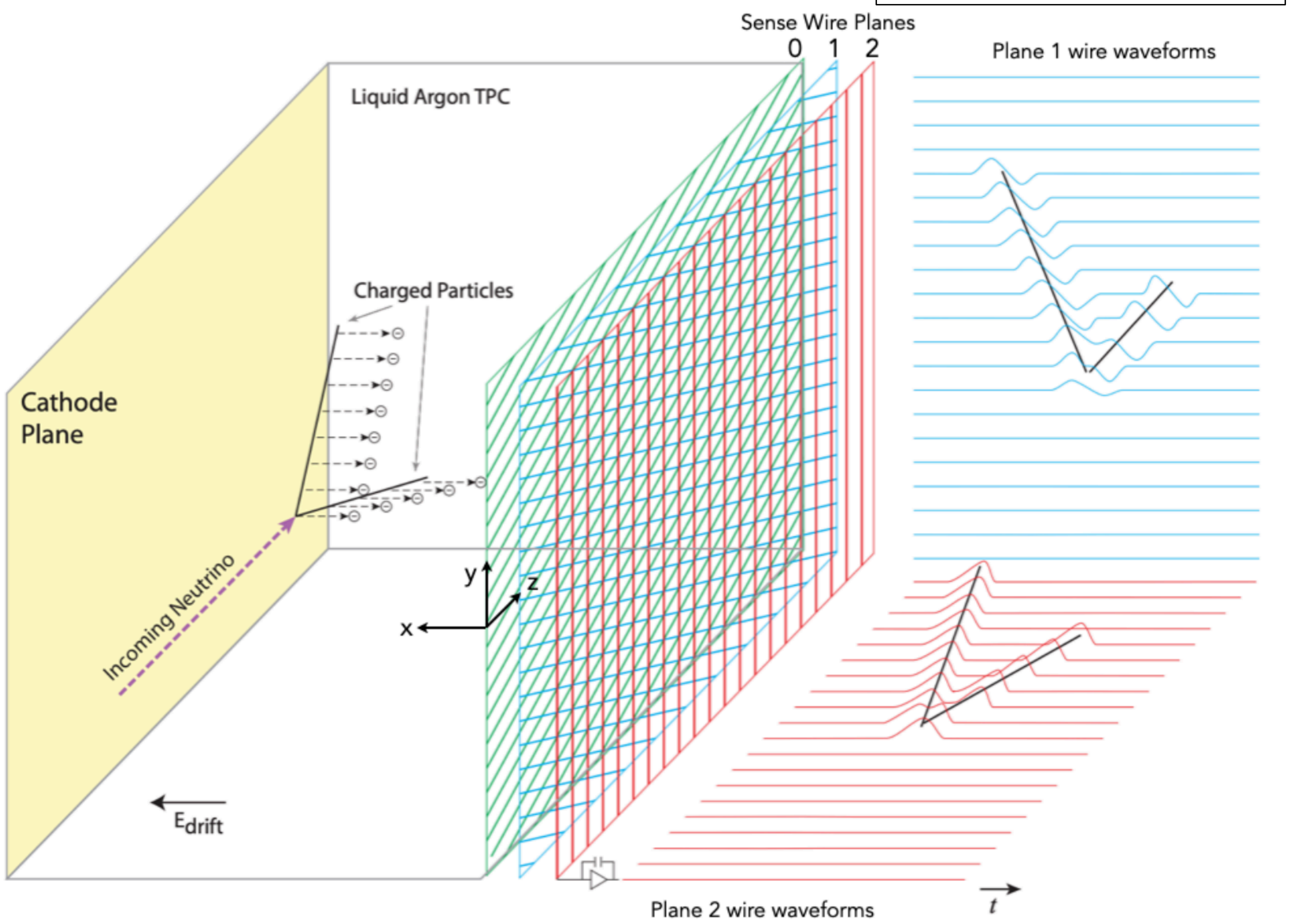
Fermilab

# 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

JINST 12, P02017 (2017)




3 planes allow for 3D reco

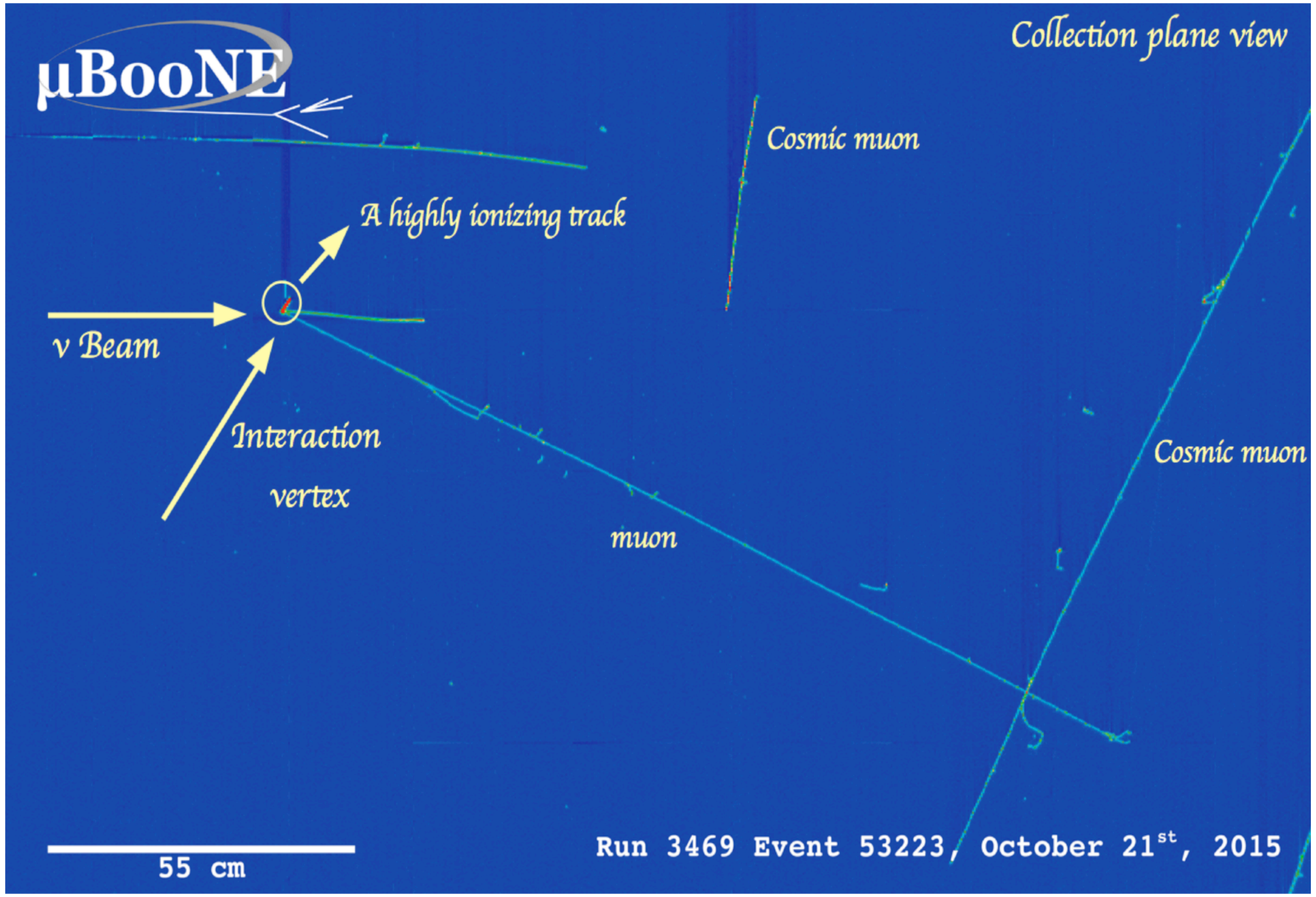


Fermilab

# 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)



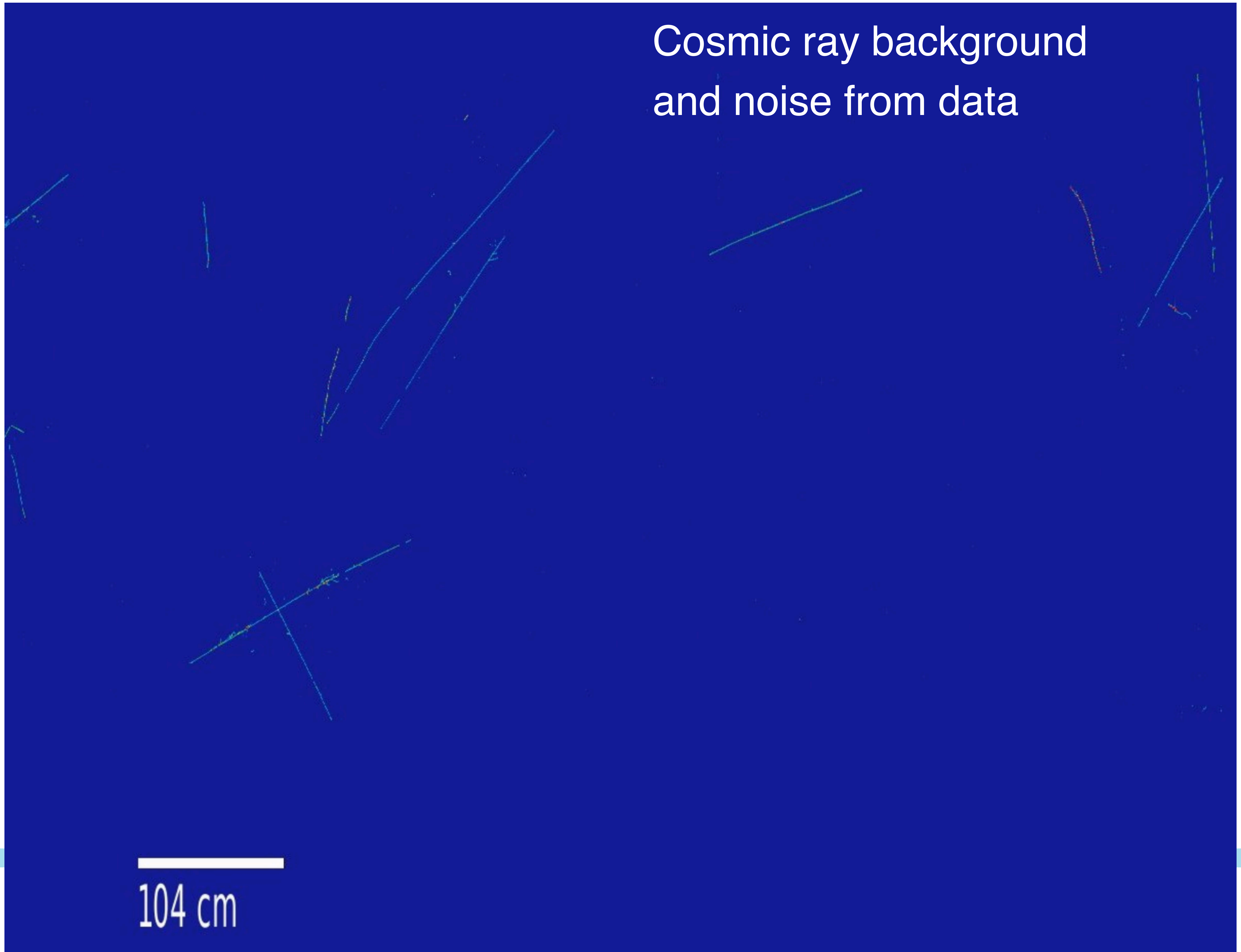

axes: time vs wire - color scale: charge

# 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 $\nu_e$

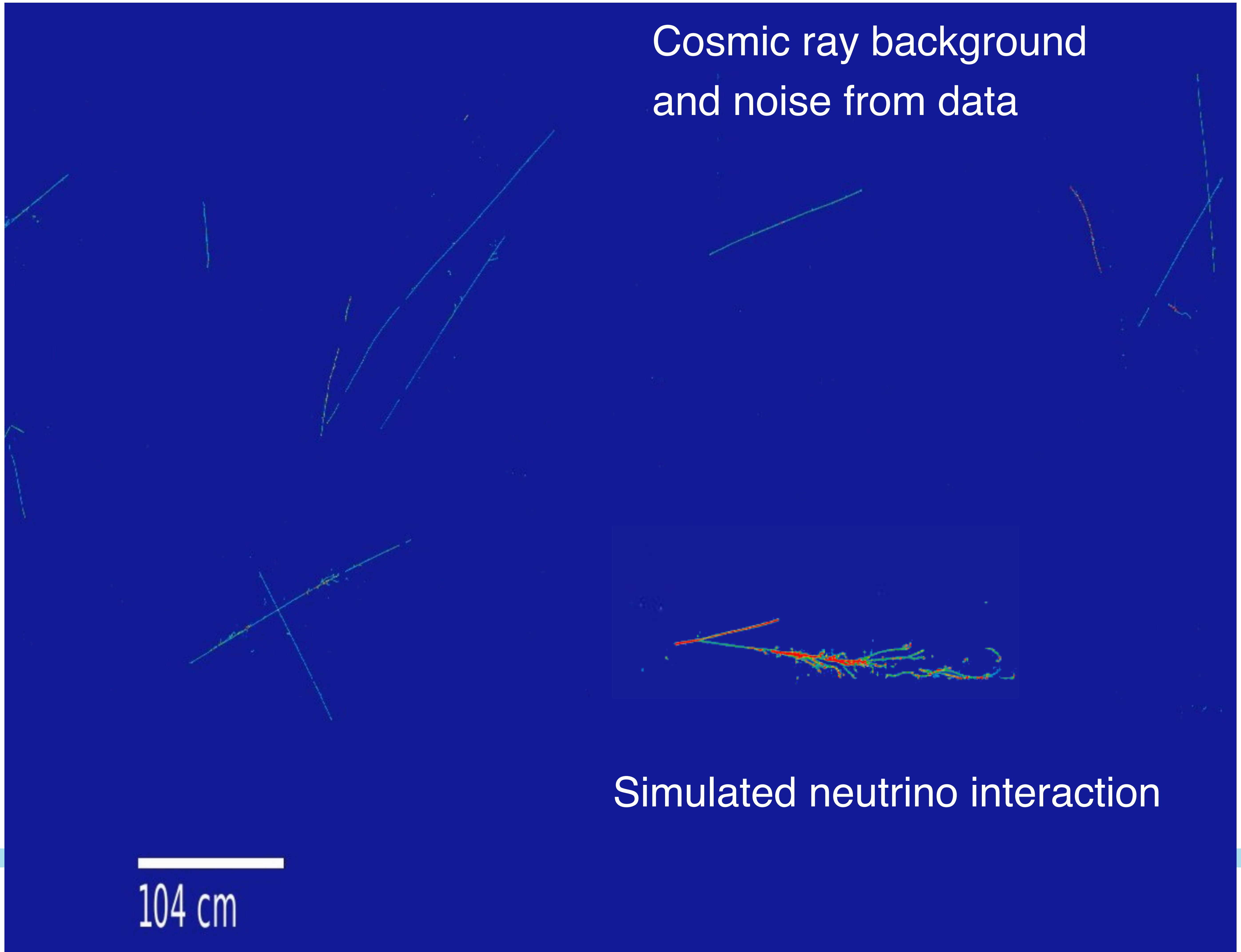





Fermilab

# Implementation of open samples: overview

- Open two "overlay" samples: BNB **inclusive** and BNB intrinsic $\nu_e$






Fermilab

# Implementation of open samples: overview

- Open two “overlay” samples: BNB **inclusive** and BNB intrinsic $\nu_e$

- 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

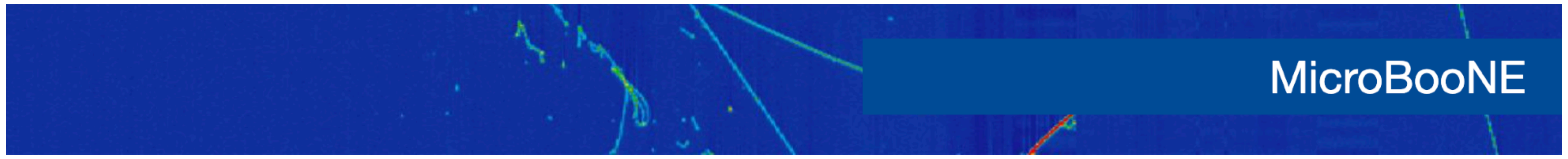
# Dataset definitions

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 | HDF5 | | | artroot | | |
|---|---|---|---|---|---|---|---|
| | | N events | N files | size | N events | N files | size |
| Inclusive, NoWire | 10.5281/zenodo.8370883 | 753,467 | 18 | 195 GB | 1,046,139 | 24436 | 6.4 TB |
| Inclusive, WithWire | 10.5281/zenodo.7262009 | 24,332 | 18 | 44 GB | 24,332 | 720 | 136 GB |
| Electron neutrino, NoWire | 10.5281/zenodo.7261921 | 89,339 | 20 | 31 GB | 89,339 | 2151 | 761 GB |
| Electron neutrino, WithWire | 10.5281/zenodo.7262140 | 19,940 | 20 | 39 GB | 19,940 | 540 | 170 GB |

# Access point

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




**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

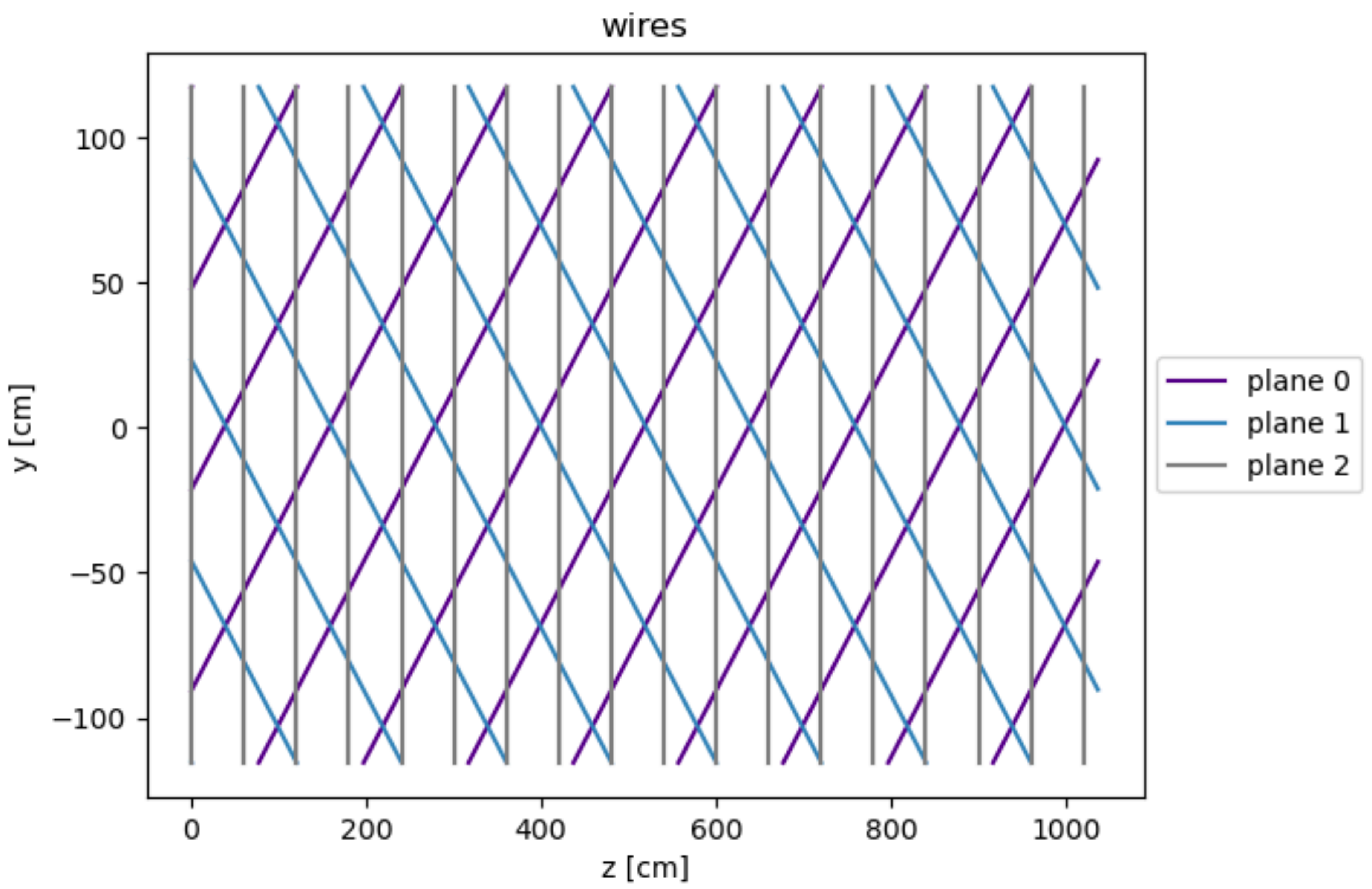
- It also includes a documentation of the file content, in a table with brief description of each element stored in the dataset

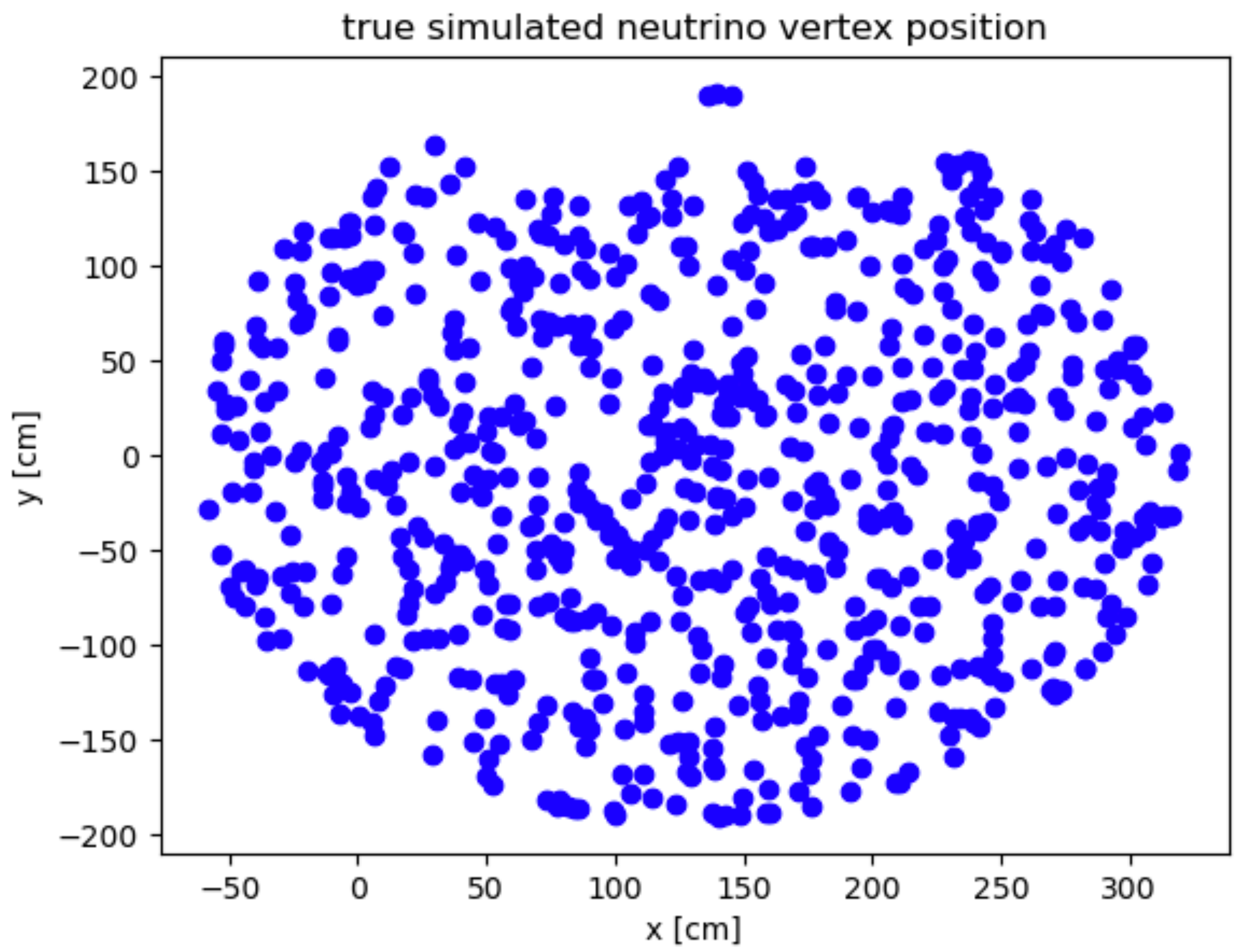
| File Entry | Type | N Elements | Description |
|---|---|---|---|
| / | Group | | Main entry point of the file. |
| /event_table | Group | | Table storing information about a single detector readout and a single simulated neutrino interaction. |
| /event_table/event_id | Dataset | 3 | Run/Subrun/Event 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. |

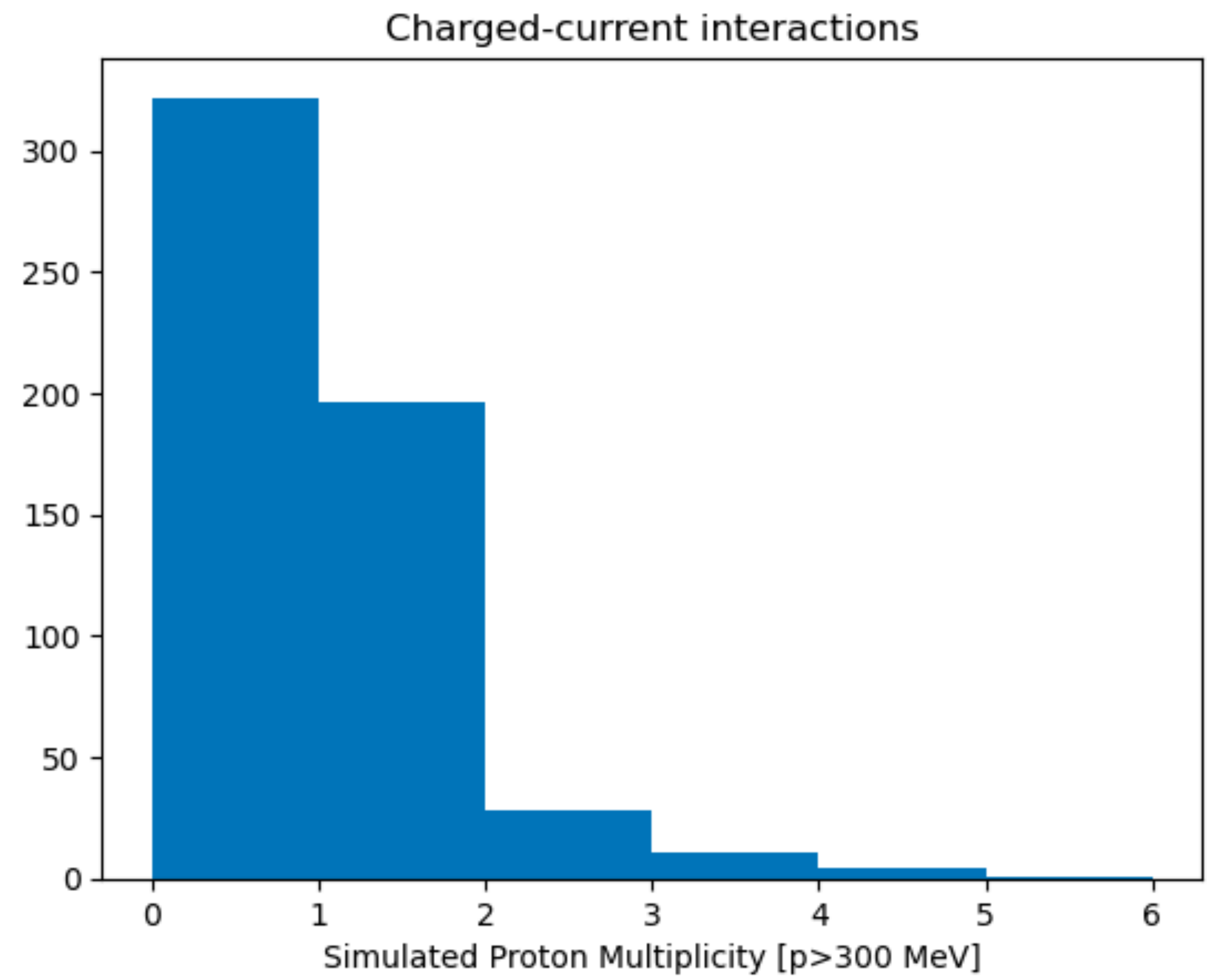
# Highlights from notebooks: Sample Exploration

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.



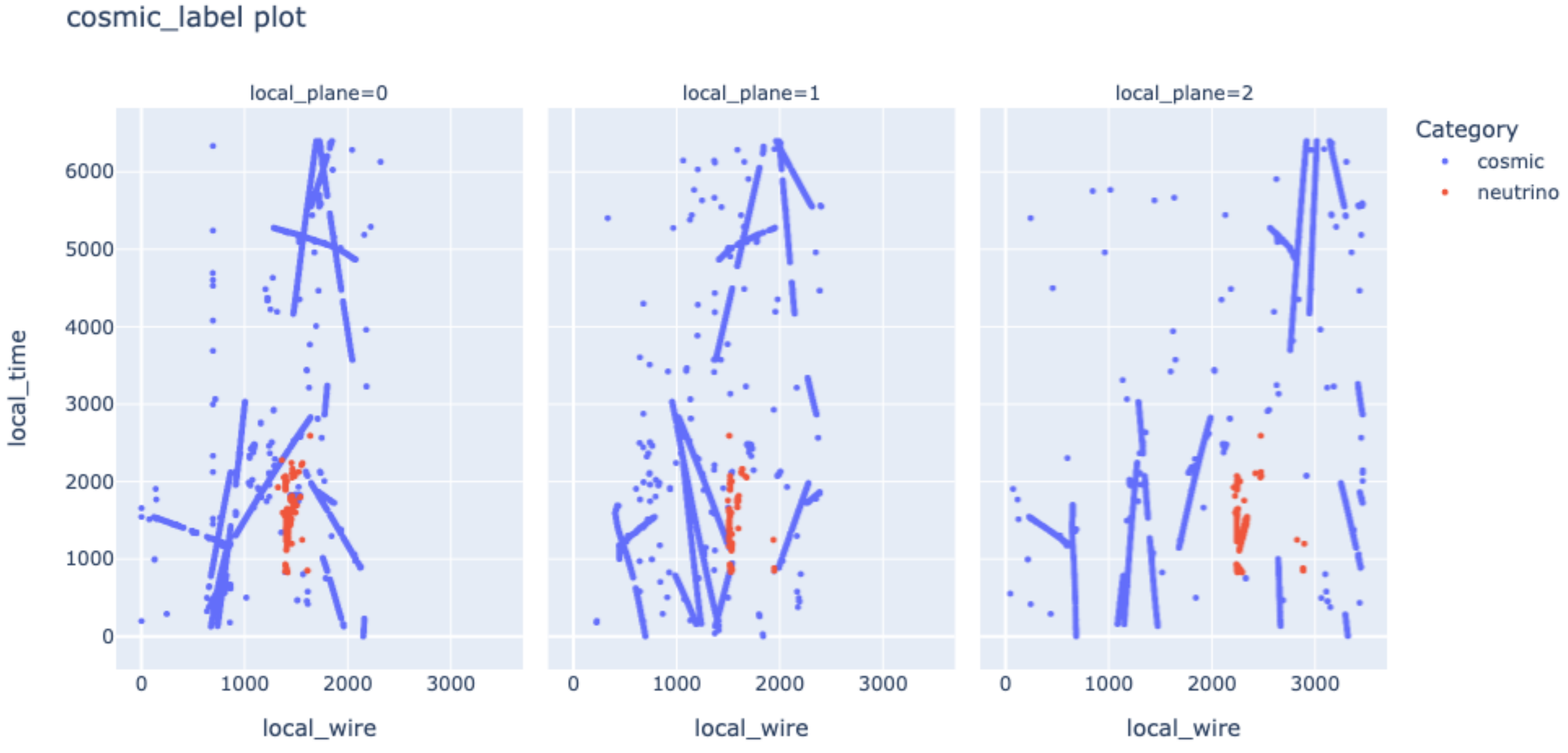



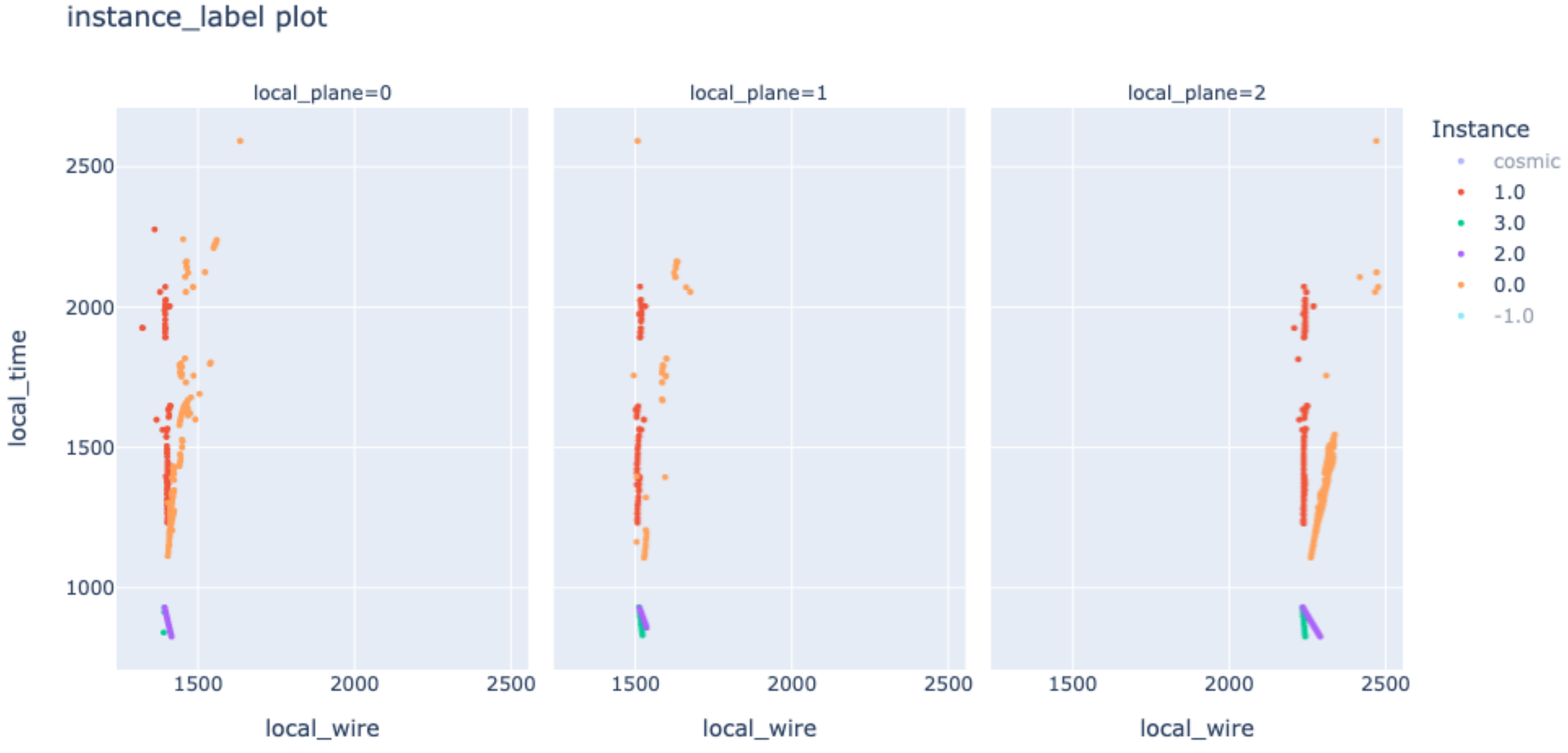



# Highlights from notebooks: Hit Labeling

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

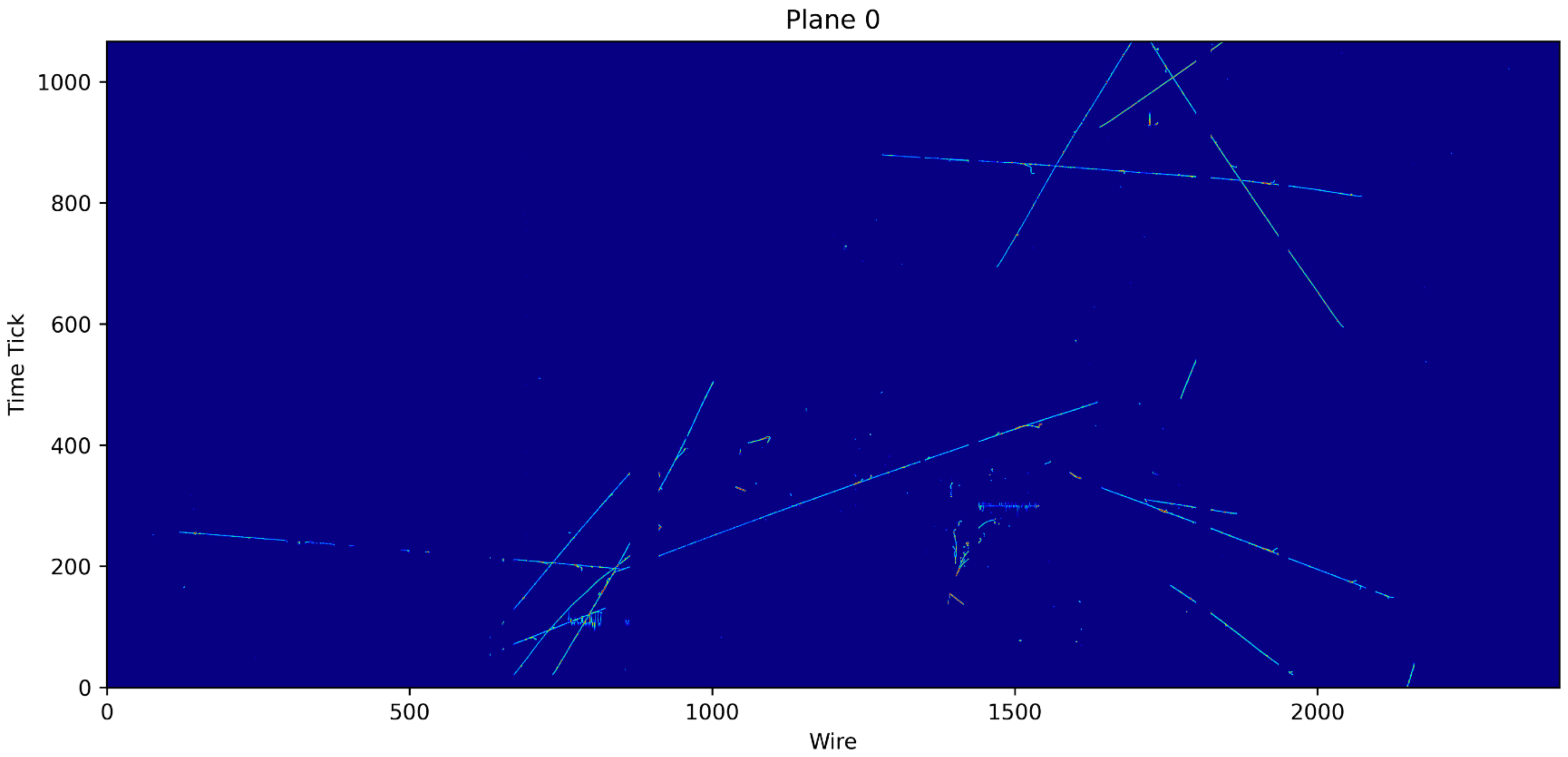
Needs “WithWire” samples containing waveform info

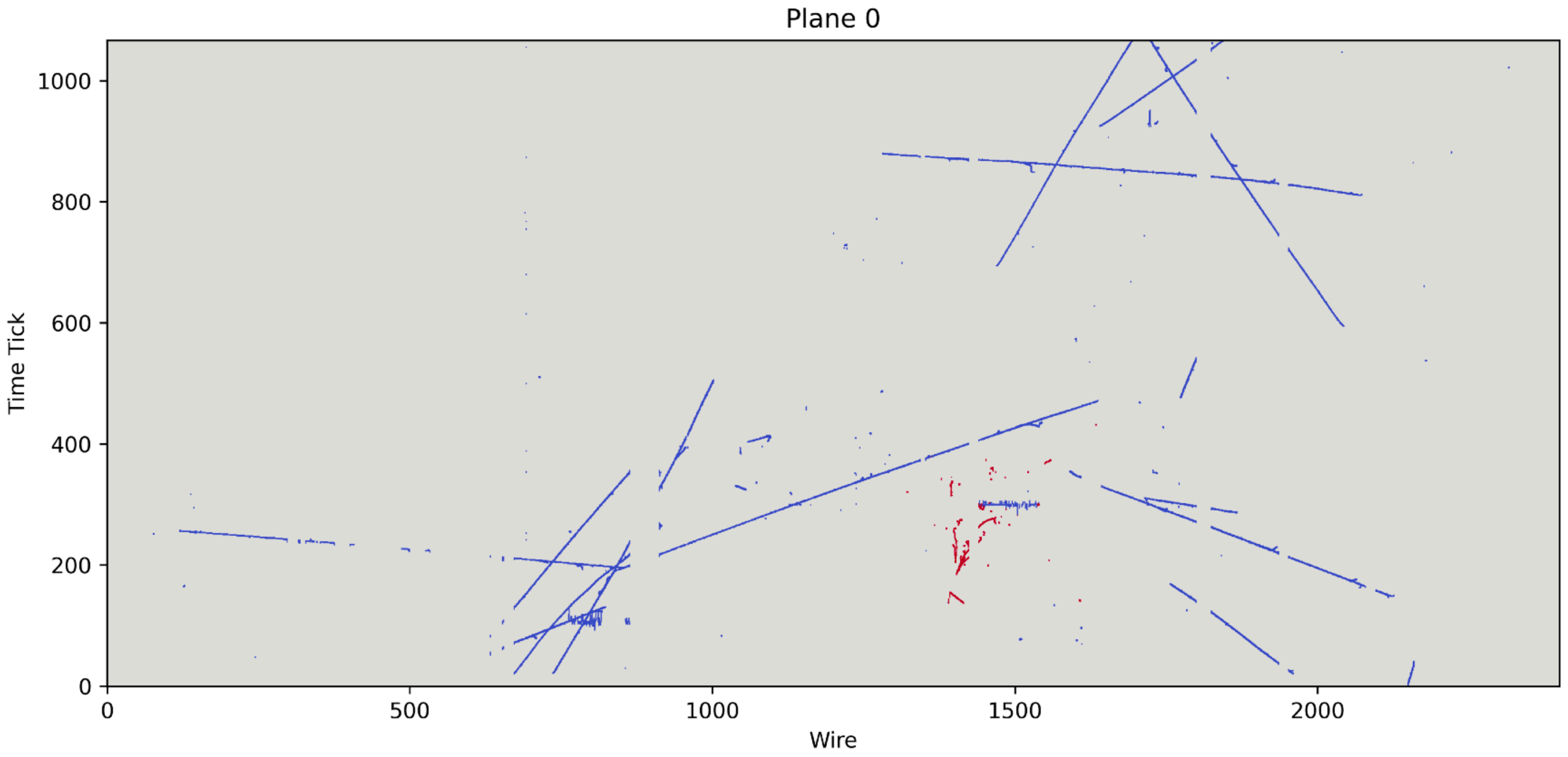
This notebook demonstrates the TPC data visualization in image format.

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

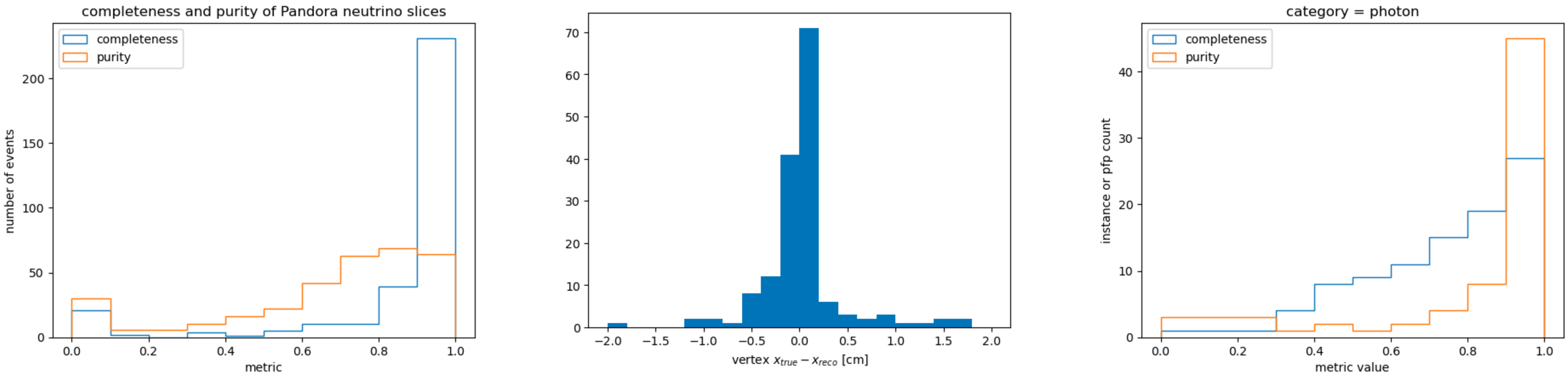
Eur.Phys.J.C 78 (2018) 1, 82

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 = $Nhit_{true,found}$ / $Nhit_{found}$

completeness = $Nhit_{true,found}$ / $Nhit_{true}$






Fermilab

# Highlights from notebooks: Optical Information

JINST 12, P02017 (2017)

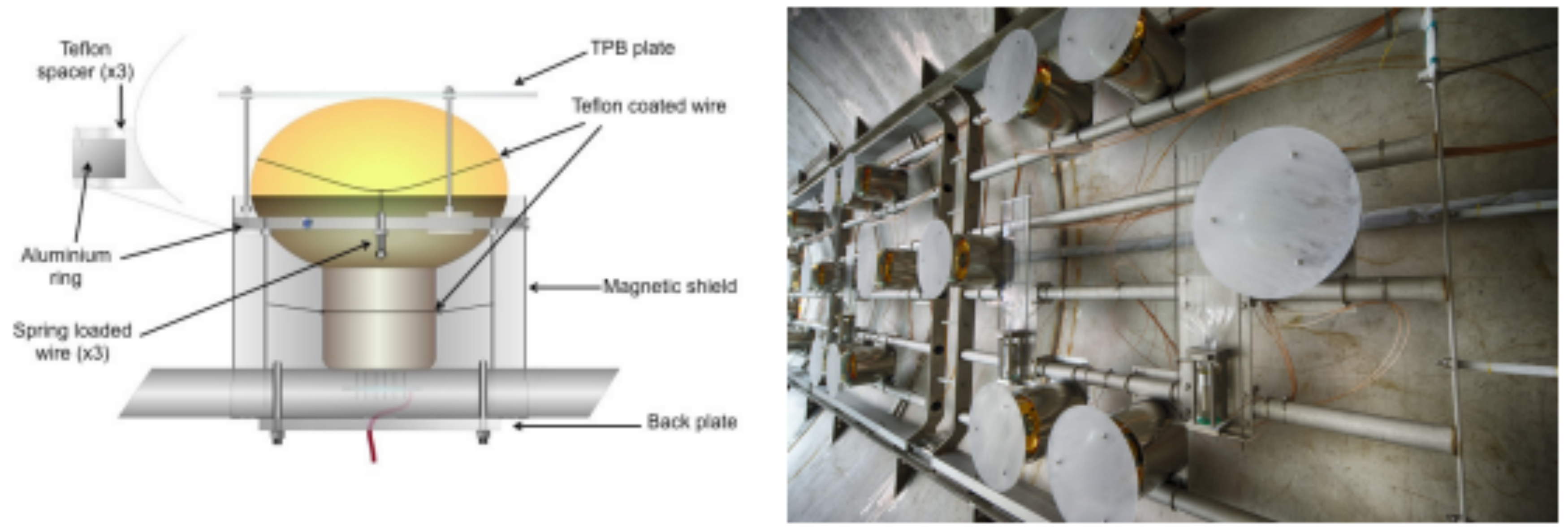
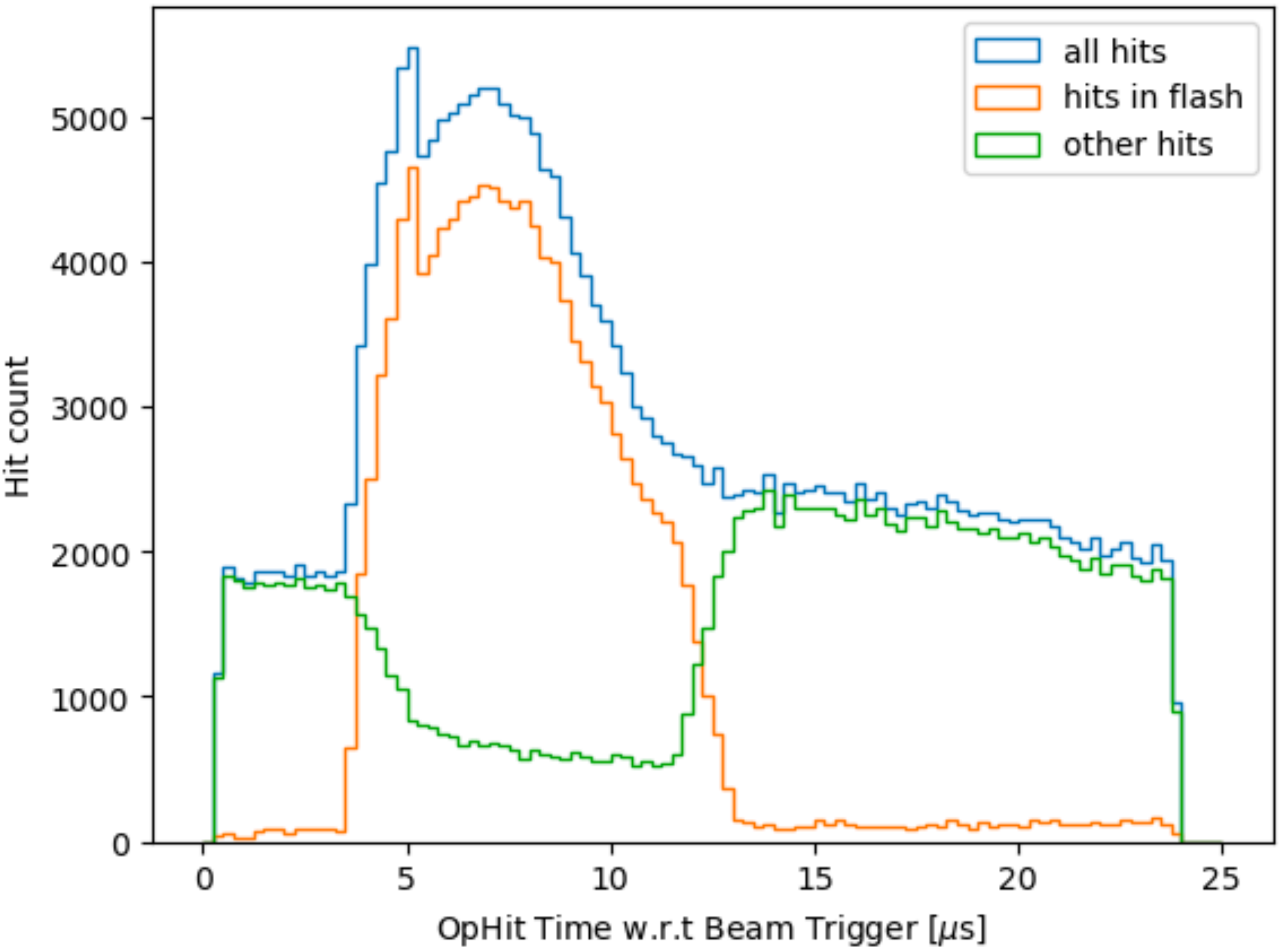
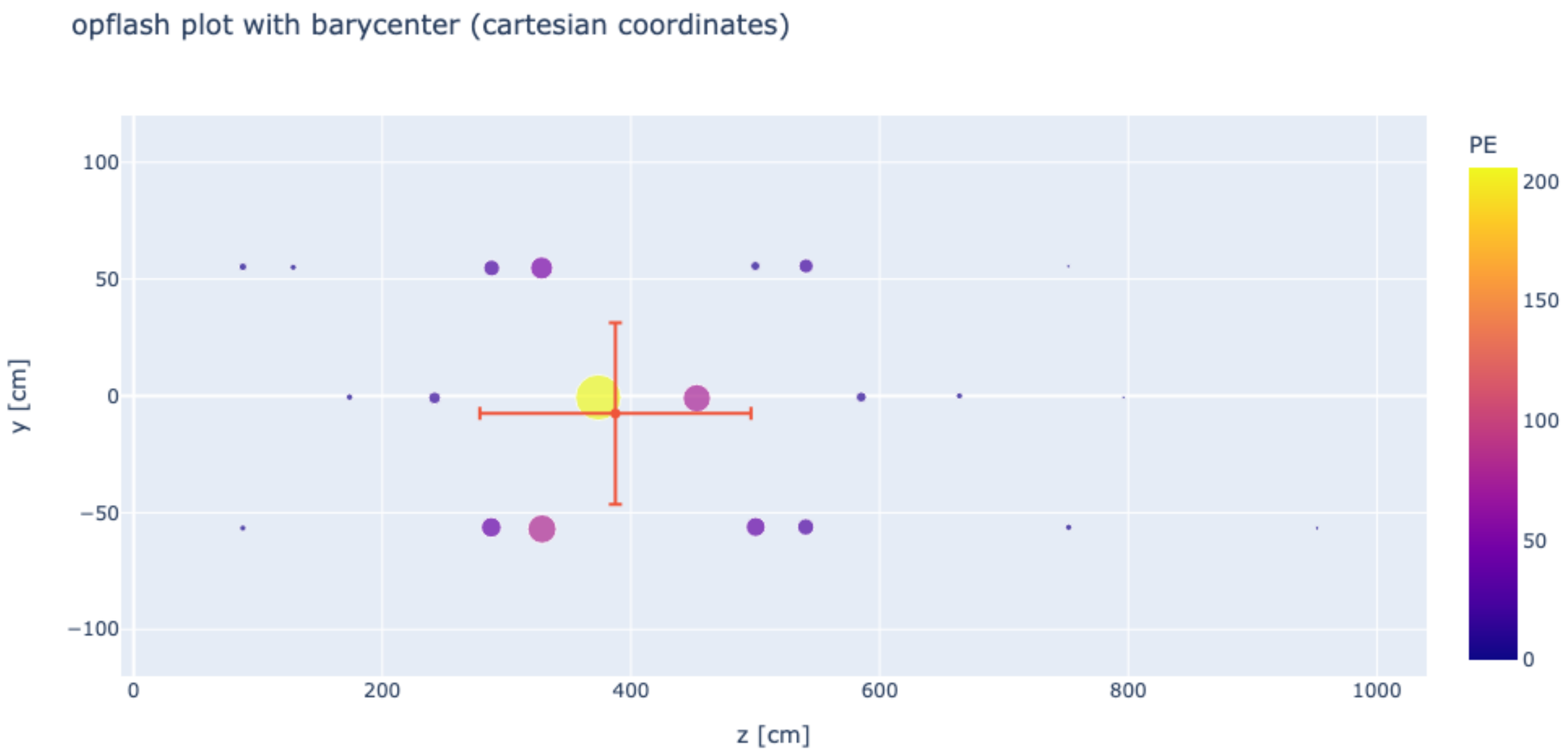


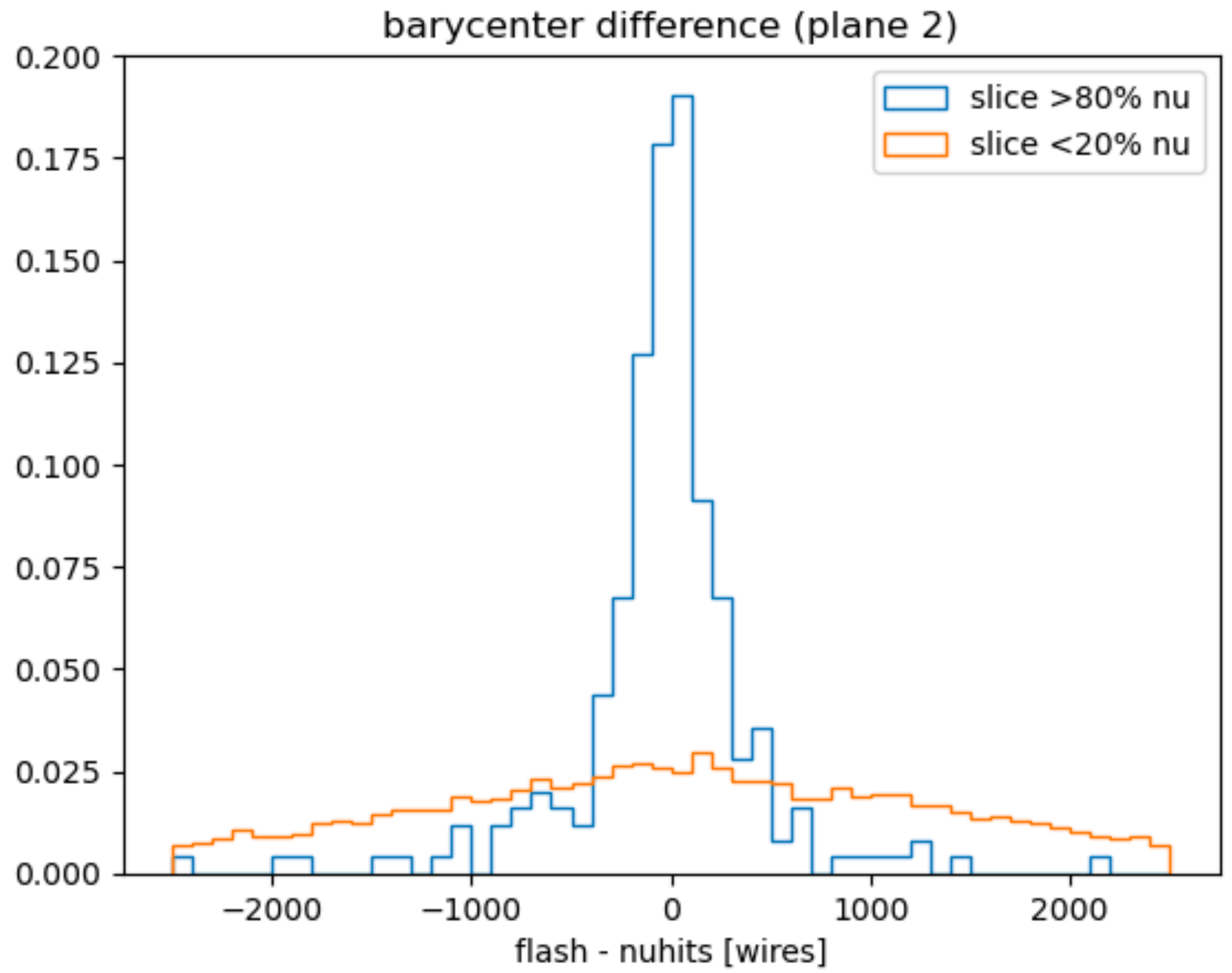

Figure 29. Left: diagram of the optical unit; Right: units mounted in MicroBooNE, immediately prior to LArTPC installation.

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






Fermilab

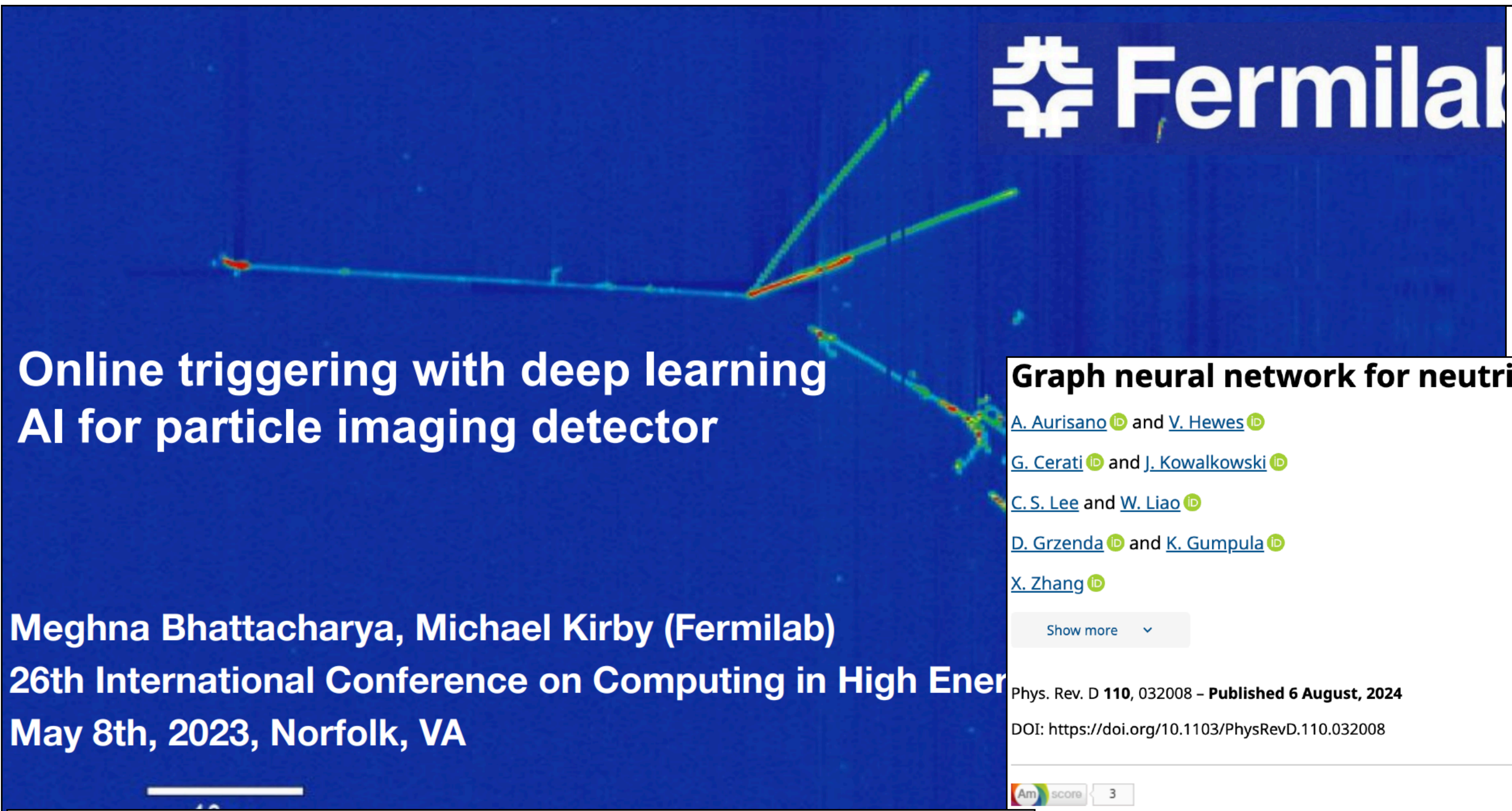
# Conclusions

- 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 several hundreds of downloads!

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




# Optimal Transport for $e/\pi^0$ Particle Classification in LArTPC Neutrino Experiments

David Caratelli,[1,*] Nathaniel Craig,[1,2,†] Chuyue Fang,[1,‡] and Jessica N. Howard[2,§]

[1] *University of California, Santa Barbara*
[2] *Kavli Institute for Theoretical Physics, Santa Barbara, CA USA*
(Dated: June 13, 2025)

The efficient classification of electromagnetic activity from $\pi^0$ and electrons remains an open problem in the reconstruction of neutrino interactions in Liquid Argon Time Projection Chamber
his problem using the mathematical framework of Optimal
ssfully employed for event classification in other HEP contexts
olution calorimetry of LArTPCs. Using a publicly available
oNE collaboration, we show that OT methods achieve state-
e in $e/\pi^0$ classification. The success of this first application
methods for LArTPC-based neutrino experiments.

# Graph neural network for neutrino physics event reconstruction

PDF

A. Aurisano and V. Hewes
G. Cerati and J. Kowalkowski
C. S. Lee and W. Liao
D. Grzenda and K. Gumpula
X. Zhang

Show more

Share

Phys. Rev. D **110**, 032008 – **Published 6 August, 2024**

DOI: https://doi.org/10.1103/PhysRevD.110.032008

Export Citation



**Abstract**

Liquid argon time projection chamber (LArTPC) detector technology offers a wealth of high-resolution information on particle interactions, and leveraging that information to its full potential requires sophisticated automated reconstruction techniques. This article describes nugraph2, a graph neural network for low-level reconstruction of simulated neutrino interactions in a LArTPC detector. Simulated neutrino interactions in the MicroBooNE detector geometry are described as heterogeneous graphs, with energy depositions on each detector plane forming nodes on planar subgraphs. The network utilizes a multihead attention message-passing mechanism to perform background filtering and semantic labeling on these graph nodes, identifying those associated with the primary physics interaction with 98.0% efficiency and labeling them according to particle type with 94.9% efficiency. The network operates directly on detector observables across multiple two-dimensional representations but utilizes a three-dimensional-context-aware mechanism to encourage consistency between these representations. Model inference takes 0.12 s/event on a CPU and 0.005 s/event batched on a GPU. This architecture is designed to be a general-purpose solution for particle reconstruction in neutrino physics, with the potential for deployment across a broad range of detector technologies, and offers a core convolution engine that can be leveraged for a variety of tasks beyond the two described in this paper.

*… and more to come!*

EPJ Web of Conferences **295**, 08012 (2024)
https://doi.org/10.1051/epjconf/202429508012

# MicroBooNE Public Data Sets: A Collaborative Tool for LArTPC Software Development

**Giuseppe Cerati[*] for the MicroBooNE Collaboration**

Fermi National Accelerator Laboratory, Kirk and, Pine St Batavia, IL 60510

[*] e-mail: MICROBOONE_INFO@fnal.gov

**Published online:** 6 May 2024

**Abstract**

Among liquid argon time projection chamber (LArTPC) experiments MicroBooNE is the one that continually took physics data for the longest time (2015-2021), and represents the state of the art for reconstruction and analysis with this detector. Recently published analyses include oscillation physics results, searches for anomalies and other BSM signatures, and cross section measurements. LArTPC detectors are being used in current experiments such as ICARUS and SBND, and being planned for future experiments such as DUNE. MicroBooNE has recently released to the public two of its data sets, with the goal of enabling collaborative software developments with other LArTPC experiments and with AI or computing experts. These data sets simulate neutrino interactions on top of off-beam data, which include cosmic ray background and noise. The data sets are released in two formats: the native art/ROOT format used internally by the collaboration and familiar to other LArTPC experts, and the HDF5 format which contains reduced and simplified content and is suitable for usage by the broader community. This contribution presents the open data sets, discusses their motivation, the technical implementation, and the extensive documentation – all inspired by FAIR principles. Finally, opportunities for collaborations are discussed.

Sea
Help

## High Energy Physics – Experiment

*[Submitted on 16 Mar 2025 (v1), last revised 15 May 2025 (this version, v2)]*

# LArTPC hit-based topology classification with quantum machine learning and symmetry

Callum Duffy, Marcin Jastrzebski, Stefano Vergani, Leigh H. Whitehead, Ryan Cross, Andrew Blake, Sarah Malik, John Marshall

We present a new approach to separate track-like and shower-like topologies in liquid argon time projection chamber (LArTPC) experiments for neutrino physics using quantum machine learning. Effective reconstruction of neutrino events in LArTPCs requires accurate and granular information about the energy deposited in the detector. These energy deposits can be viewed as 2-D images. Simulated data from the MicroBooNE experiment and a simple custom dataset are used to perform pixel-level classification of the underlying particle topology. Images of the events have been studied by creating small patches around each pixel to characterise its topology based on its immediate neighbourhood. This classification is achieved using convolution-based learning models, including quantum-enhanced architectures known as quanvolutional neural networks. The quanvolutional networks are extended to symmetries beyond translation. Rotational symmetry has been incorporated into a subset of the models. Quantum-enhanced models perform better than their classical counterparts with a comparable number of parameters but are outperformed by classical models, which contain an order of magnitude more parameters. The inclusion of rotation symmetry appears to benefit only large models and remains to be explored further.