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/
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

3 planes allow for 3D reco
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
MicroBooNE 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 $\nu_e$

Cosmic ray background and noise from data
Implementation of open samples: overview

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

![Image of cosmic ray background and noise from data]

- 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) and versioning
- two flavors: with and without wire info, due to size restrictions
- files stored on persistent dCache pool area and made accessible with
- list of urls stored with the corresponding HDF5 files on Zenodo
- Samples available under “cc-by” license. Template text for acknowledgment is provided.

![Image of simulated neutrino interaction]

Cosmic ray background and noise from data

Simulated neutrino interaction
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.

<table>
<thead>
<tr>
<th>Sample</th>
<th>DOI</th>
<th>N events</th>
<th>N HDF5 files</th>
<th>HDF5 size</th>
<th>N artroot files</th>
<th>artroot size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inclusive, NoWire</td>
<td>10.5281/zenodo.7261798</td>
<td>141,260</td>
<td>20</td>
<td>34 GB</td>
<td>3400</td>
<td>787 GB</td>
</tr>
<tr>
<td>Inclusive, WithWire</td>
<td>10.5281/zenodo.7262009</td>
<td>24,332</td>
<td>18</td>
<td>44 GB</td>
<td>720</td>
<td>136 GB</td>
</tr>
<tr>
<td>Electron neutrino, NoWire</td>
<td>10.5281/zenodo.7261921</td>
<td>89,339</td>
<td>20</td>
<td>31 GB</td>
<td>2151</td>
<td>761 GB</td>
</tr>
</tbody>
</table>

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:
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](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 crystal volume), with relative abundances 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 atoolkit, 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 root tills providing access to the atoolkit files are stored on the open data portal Zenodo, and can be accessed from the DOI links in the table below. Atoolkit 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).

<table>
<thead>
<tr>
<th>Sample</th>
<th>DOI</th>
<th>N events</th>
<th>N HDF5 files</th>
<th>HDF5 size</th>
<th>N atoolkit files</th>
<th>atoolkit size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inclusive, NoWire</td>
<td>10.5281/zenodo.7261796</td>
<td>141,290</td>
<td>20</td>
<td>34 GB</td>
<td>3400</td>
<td>787 GB</td>
</tr>
<tr>
<td>Inclusive, Wire</td>
<td>10.5281/zenodo.7262009</td>
<td>24,332</td>
<td>18</td>
<td>44 GB</td>
<td>720</td>
<td>136 GB</td>
</tr>
<tr>
<td>Electron neutrino, NoWire</td>
<td>10.5281/zenodo.7261921</td>
<td>89,339</td>
<td>20</td>
<td>31 GB</td>
<td>2151</td>
<td>761 GB</td>
</tr>
</tbody>
</table>

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 HDF Group. In the notebooks we open the files using the F tell class from pynh5, which internally relies on HDF5. We also use pynuml to merge files and to add auxiliary data for faster lookup of related information across different tables.

Jupyter notebooks

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

<table>
<thead>
<tr>
<th>File Entry</th>
<th>Type</th>
<th>N Elements</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>/</td>
<td>Group</td>
<td></td>
<td>Main entry point of the file.</td>
</tr>
<tr>
<td>/event_table</td>
<td>Group</td>
<td></td>
<td>Table storing information about a single detector readout and a single simulated neutrino interaction.</td>
</tr>
<tr>
<td>/event_table/event_id</td>
<td>Dataset</td>
<td>3</td>
<td>Run/Subrun/Event number for a detector readout.</td>
</tr>
<tr>
<td>/event_table/event_id_seq_cnt</td>
<td>Dataset</td>
<td>2</td>
<td>Auxiliary information added in post-processing step for simple grouping and fast access of table entries separated by event.</td>
</tr>
<tr>
<td>/event_table/is_cc</td>
<td>Dataset</td>
<td>1</td>
<td>If 1 the simulated neutrino interaction is charged-current, if 0 it is neutral-current.</td>
</tr>
<tr>
<td>/event_table/lep_energy</td>
<td>Dataset</td>
<td>1</td>
<td>Simulated energy of the lepton outgoing from the neutrino interaction (in GeV).</td>
</tr>
<tr>
<td>/event_table/nu_dir</td>
<td>Dataset</td>
<td>3</td>
<td>Initial direction of the simulated neutrino interacting in the detector (3D cartesian coordinates).</td>
</tr>
<tr>
<td>/event_table/nu_energy</td>
<td>Dataset</td>
<td>1</td>
<td>Simulated energy of the interacting neutrino (in GeV).</td>
</tr>
<tr>
<td>/event_table/nu_vtx</td>
<td>Dataset</td>
<td>3</td>
<td>Simulated position of neutrino interaction (3D cartesian coordinates, in cm). This quantity is to be used to compare with e.g. the detector boundaries.</td>
</tr>
</tbody>
</table>
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

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.

Needs “WithWire” samples containing waveform info

Phys. Rev. D103, 092003 (2021)
JINST 12, P03011 (2017)
Highlights from notebooks: Pandora metrics

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.

\[
purity = \frac{\text{Nhit}_\text{true,found}}{\text{Nhit}_\text{found}}
\]

\[
\text{completeness} = \frac{\text{Nhit}_\text{true,found}}{\text{Nhit}_\text{true}}
\]
Highlights from notebooks: Optical Information

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.
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!
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!?
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
  - Findable
  - Accessible
  - Interoperable
  - Reusable
  - 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
### Table 1. Findable and Accessible principle assessment checks for the CMS (Hep) Open Dataset.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>F.1.</td>
<td>(Metadata) are assigned globally unique and persistent identifiers.</td>
</tr>
<tr>
<td>F.2.</td>
<td>Data are described with rich metadata.</td>
</tr>
<tr>
<td>F.3.</td>
<td>Metadata clearly and explicitly include the identifier of the data they describe.</td>
</tr>
<tr>
<td>F.4.</td>
<td>Metadata are registered or indexed in a searchable resource.</td>
</tr>
<tr>
<td>A.1.</td>
<td>The protocol is open, free and universally implementable</td>
</tr>
<tr>
<td>A.2.</td>
<td>Metadata should be accessible even when the data is no longer available</td>
</tr>
</tbody>
</table>

### Table 2. Interoperable and Reusable principle assessment checks for the CMS (Hep) Open Dataset.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.1.</td>
<td>(Metadata) use a formal, accessible, shared, and broadly applicable language for knowledge representation.</td>
</tr>
<tr>
<td>I.2.</td>
<td>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.</td>
</tr>
<tr>
<td>I.3.</td>
<td>(Metadata) include qualified references to other metadata.</td>
</tr>
<tr>
<td>R.1.1.</td>
<td>(Metadata) are released with a clear and accessible data usage license.</td>
</tr>
<tr>
<td>R.1.2.</td>
<td>(Metadata) are associated with detailed provenance.</td>
</tr>
<tr>
<td>R.1.3.</td>
<td>Meet Domain-relevant Community Standards:</td>
</tr>
</tbody>
</table>

---

https://fair4hep.github.io/