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

Giuseppe Cerati[1], for the MicroBooNE Collaboration*

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

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.

1 Introduction

MicroBooNE [1] is a neutrino experiment at Fermilab designed to test the MiniBooNE anomaly [2] as it is located along the same Booster Neutrino Beam (BNB) and at a similar distance from source. MicroBooNE's goals are not limited to probing this anomaly, and span a broader experimental program including tests of short-baseline oscillations as part of the Short Baseline Neutrino program (SBN) [3], searches for beyond-Standard Model particles, and the measurement of neutrino-argon interaction cross sections. MicroBooNE's physics operations took place between 2015 and 2021. To this date, MicroBooNE has analyzed about half of the collected beam data and produced over 50 publications.

MicroBooNE's detector [1] is a liquid argon time projection chamber (LArTPC). The working principle of the MicroBooNE LArTPC is the following: charged particles produced in neutrino-argon interactions ionize the argon as they travel in the detector active volume ($2.56 \times 2.32 \times 10.36$ m). Ionization electrons drift in an electric field towards anode planes. Here, sense wires detect the incoming charge. About 8200 wires are arranged in three planes oriented in different directions (0, ±60 degrees), allowing for 3D reconstruction of the particle trajectories with O(mm) spatial resolution. The amplitude of the signal detected on the

---

*e-mail: MICROBOONE_INFO@fnal.gov

wires provides calorimetric information for energy measurements. The fast scintillation light emitted by the argon is detected by the optical system, made of 32 photo-multiplier tubes (PMTs), and is used for triggering and cosmic rejection.

In this document we describe the release of MicroBoNE data sets for the purpose of collaborative software development. This release is motivated by the following arguments:

- Establish MicroBooNE as state of the art LArTPC technology. This is already attested by our publication record, but public data sets provide a direct reference point for any LArTPC software development.

- Efficient collaboration of members of MicroBooNE with colleagues in other LArTPC experiments, as well as with computer scientists. Until this data release, software development collaborations were required to have an approved memorandum of understanding to share data sets outside the Collaboration or to use other public data sets. Being able to use MicroBooNE data sets implies that the output of external collaborations is directly usable within MicroBooNE without further tuning.

- Potentially attract developments from beyond our community, through public initiatives such as data challenges.

The next sections describe the technical implementation of the data release, as well as the relative documentation. Finally, conclusions and future prospects are discussed.

2 Implementation of open samples

With this data release, MicroBooNE aims at reaching out to the largest possible set of developers and at enabling the widest range of applications. In the process, we therefore followed as much as possible the FAIR principles for scientific data management (findable, accessible, interoperable, reusable data). For more information on these principles, see e.g. ref. [4]. The MicroBooNE open samples are advertized from the MicroBooNE website[1], and are made available on the Zenodo open data repository. The MicroBooNE website contains a brief description of the data set, links to Zenodo and to documentation, and information about license and citation. Zenodo, provides citable DOI (digital object identifier) and versioning. Samples are made available under the "cc-by" license, allowing users to utilize the data in any way, including modifying and redistributing, as long as credit is given to the original authors. A suggested text for acknowledging the Collaboration is provided. The Collaboration requests that, whenever possible, resulting software products are also made publicly available, although this is not required by the license.

The data sets release consists of "overlay" samples, where with overlay we refer to events from off-beam data taking with an overlaid simulated neutrino interaction. These events provide data-driven cosmic ray background and noise, as well as Monte Carlo truth information for the neutrino interaction. Neutrino interactions in the open samples are either the inclusive set of neutrino interactions as expected from the BNB nominal flux or the subset of charged-current electron neutrino interactions in the BNB. We will refer to these as "inclusive BNB" and "intrinsic ν_e", respectively. Inclusive BNB interactions are simulated in the full cryostat volume, while intrinsic ν_e interactions are simulated in the LArTPC active volume.

The data is released in two formats. The first one is the art/ROOT data format [5, 6]. These are the same files used within the collaboration, thus making available to the public all reconstructed and simulated data products. This data format targets HEP physicists, and in particular the LArTPC community, that are likely already familiar with this data format. art/ROOT files are stored on a dedicated persistent dCache pool area that is accessible with

---

[1]https://microboone.fnal.gov/documents-publications/public-datasets/

Table 1: Summary of MicroBooNE Open Data Sets. The column "Wire" refers to whether the wire waveform information is stored in the HDF5 files or not.

Interaction	DOI	HDF5				art/ROOT		
		Wire	Events	Files	Size	Events	Files	Size
Inclusive BNB	[11]	No	753,467	18	195 GB	1,046,139	24,436	6.4 TB
Inclusive BNB	[12]	Yes	24,332	18	44 GB	24,332	720	136 GB
Intrinsic ν_e	[13]	No	89,339	20	31 GB	89,339	2,151	761 GB
Intrinsic ν_e	[14]	Yes	19,940	20	39 GB	19,940	540	170 GB

xrootd [7] without requiring any virtual organization credentials. The list of xrootd urls is stored on Zenodo. The second format is HDF5 [8], targeting usage by the broader data and computer science communities. HDF5 files include a subset of the art/ROOT information, with a simplified layout for ease of use. Nevertheless, this data format contains the most useful information and is designed to allow a wide range applications. The following information is stored in the HDF5 files:

1. Noise-filtered and deconvolved wire waveforms in regions of interest.

2. LArTPC hit information.

3. Optical hit and flash information.

4. Monte Carlo truth information (incoming neutrino properties, energy deposits as associated to hits, Geant4 [9] particles).

In addition we provide information for the purpose of benchmarking new developments against the state-of-the-art reconstruction performance provided by the Pandora multi-algorithm pattern recognition toolkit [10]. These include the interaction and cluster hit mapping, a multivariate track-shower classification, the neutrino flavor identification. HDF5 files stored on Zenodo, at the same DOI as the xrootd urls of corresponding art/ROOT data set. Each HDF5 sample comes in two flavors: with and without wire waveform information listed as 1. in the list above. Due to size requirements, samples with this information contain less events. A summary of the samples can be found in table 1.

3 Documentation

Documentation is of utmost importance for an open data release, as it allows external users to become familiar with the content of the data and to understand which applications it can be used for. The art/ROOT format targets users from the LArTPC community, i.e. physicists already familiar with the LArSoft [15] software environment. Therefore, the documentation for this format assumes prior knowledge of LArSoft-related tools, and consists of:

- A description of the samples and list of data products stored[2].

- Links to websites with documentation about the related software tools (LArSoft, xrootd, etc.).

- A recipe to setup the software release (uboonecode and LArSoft) from CVMFS.

- A link to the module for creating HDF5 files, providing an example of how to access the art/ROOT content.

---

[2]https://github.com/uboone/OpenSamples/blob/v01/file-content-artroot.md

Documentation for the HDF5 data sets instead needs to be more detailed, as these format targets users not necessarily familiar with LArTPC experiments and the software they use. We chose to document the usage of this format mainly through a demonstration of usage with a few Jupyter notebooks[3]. These notebooks are described in the subsection below. A recipe for installing the required packages in a Conda environment to run the notebooks is provided. This environment has minimal dependencies so that users can easily install it and add any other package they may need for their applications on top. Dependencies include pynuml[4] for file I/O handling. All notebooks contain a brief introduction to clarify their purpose. We also provide a set of auxiliary tools, such as functions for basic detector navigation and minimal plotting utilities. Documentation also includes a description of the file content, formatted as a table with a brief explanation of each element stored in the data set.

3.1 Notebooks Demonstrating Usage of HDF5 Files

The first notebook, "Sample Exploration", is meant to help the user become familiar with the sample content and with tools provided to understand the detector properties. As shown in figure 1, the notebook demonstrates how to determine the wire positions and their intersections, as well properties of the neutrino interaction like the interaction position (vertex) in the cryostat and multiplicities of simulated particles.
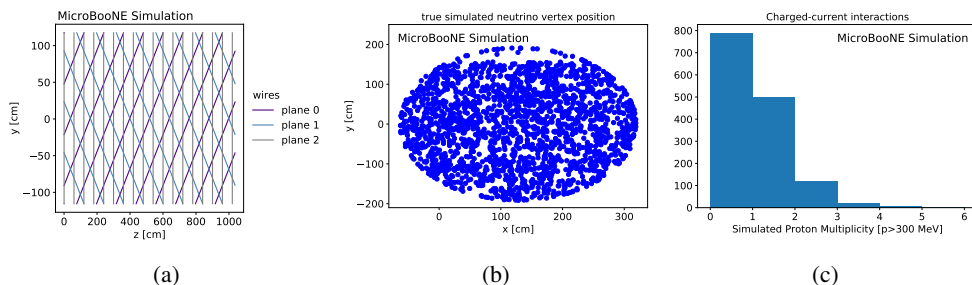

Figure 1: Example of plots from the "Sample Exploration" notebook: the detector positions of wires in each plane, displaying one wire every 200 (a), position of the simulated neutrino interaction vertex in the plane transverse to the beam (b), number of protons with momentum above 300 MeV produced in BNB neutrino interactions (c).

The second notebook, "Hit Labeling", provides examples of ground-truth labels of TPC hits according to different categorizations. Each categorization can be the target of specific algorithms or network training. These categorizations are the neutrino identification (hits from neutrino interaction vs from noise or cosmic ray background), semantic segmentation (categorization of hits based on the type of particle that produced them), and instance segmentation (labeling of hits according to the particle instance that produced them). Examples are shown in figure 2.

The "WireImage" notebook is the only one that requires HDF5 files with the wire waveform information. This notebook demonstrates the LArTPC data visualization in image format. It can be used for visual data processing methods, such as Convolutional Neural Networks (CNN). Examples of CNNs developed by the MicroBooNE Collaboration can be found

---

[3]https://github.com/uboone/OpenSamples/tree/v01
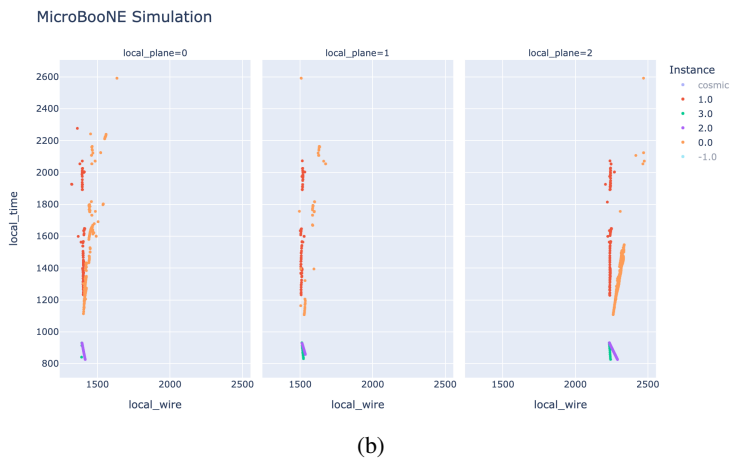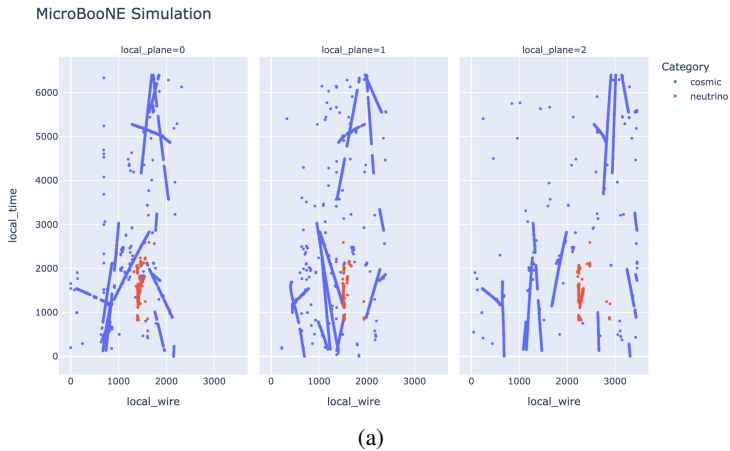[4]https://github.com/vhewes/pynuml

Figure 2: Example of plots from the "Hit Labeling" notebook: categorization of hits in the 3 planes based on whether they originate from a neutrino or cosmic ray interaction (a), categorization of neutrino hits in different particle instances (b).

at refs [16–19]. Ground truth at waveform level is not directly provided, but the notebook demonstrates how it can be extracted by matching the waveform with the hit-level ground truth information. Example images of the wire waveform and of the extracted ground truth for a specific event are shown in figure 3.

Purpose of the "Pandora metrics" notebook is to introduce the definition of important metrics used to benchmark the reconstruction software performance, and produce results for these metrics using Pandora. As shown in figure 4, these metrics include the neutrino vertex position resolution, as well as purity and completeness at the level of the neutrino interaction or at the level of particle instances. Purity and completeness are respectively defined as the fraction of reconstructed or true hits correctly identified.

Finally, as the other notebooks focused on the LArTPC information, the "Optical Information" notebook demonstrates the usage of the optical detector information. This notebook shows how to access optical hit properties, and how the MicroBooNE optical reconstruction
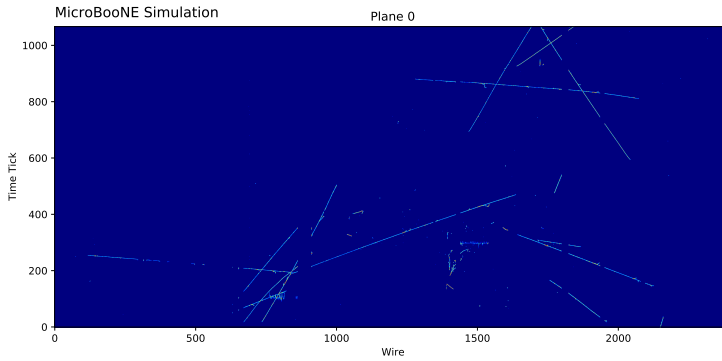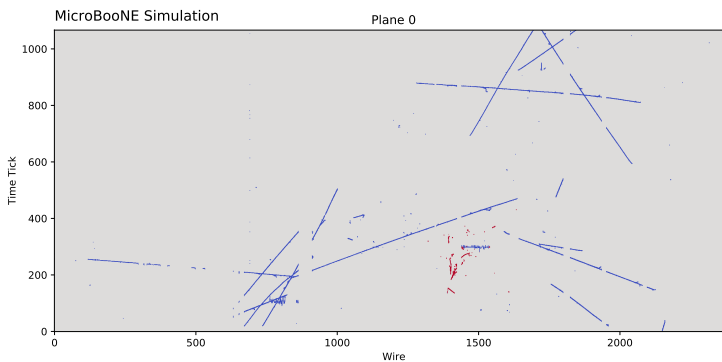
Figure 3: Example of plots from the "WireImage" notebook: for the same event as in figure 2, the wire waveform image for plane 0 (a) and the corresponding ground truth in terms of neutrino identification (b) are shown.

clusters them in time into "flash" objects. It also shows how to use the light information to help identifying the neutrino hits in the TPC, e.g. by comparing the flash barycenter and with the one from the neutrino TPC hits. Examples are shown in figure 5.

4 Conclusions and Outlook

MicroBooNE has released data sets for collaborative software development, and made them available on Zenodo and via xrootd. A wide range of applications can be developed based on these samples, with data formats for both image processing (e.g. CNN) and hit-based algorithms (GNN or other traditional algorithms). The size of the sample is enough for neural network training, and examples for extracting the ground truth for supervised learning are provided. The rich documentation for the usage of these data sets includes a set of performance metrics with reference results from the Pandora algorithms.

The statistics on the Zenodo website indicate that the data sets have been dowloaded hundreds of times already. The first developments based on these samples were presented at the conference.
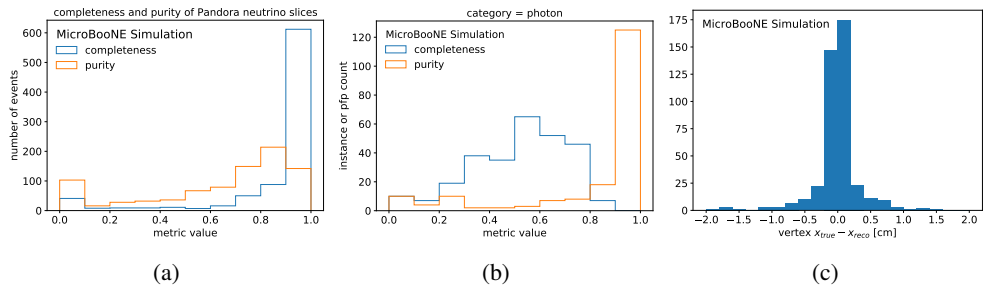
(a) (b) (c)

Figure 4: Example of plots from the "Pandora metrics" notebook: the purity and completeness for the neutrino interaction identification (a) and for the reconstruction and clustering of hits from photons (b), resolution for the neutrino vertex x coordinate (c).
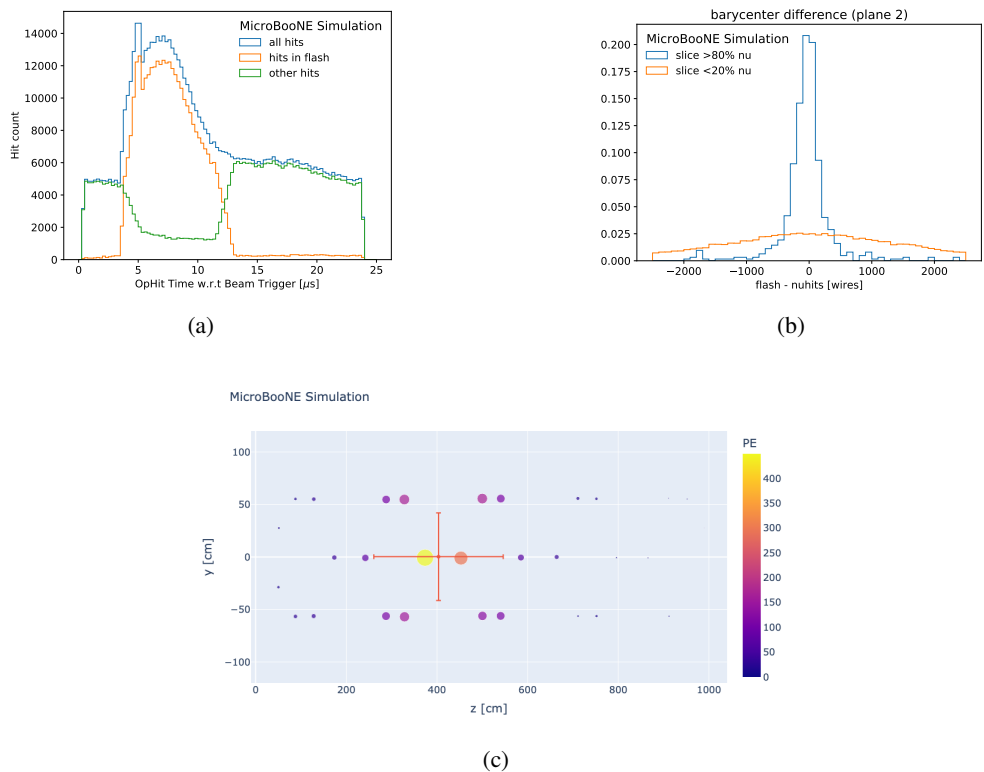

(a) (b)

(c)

Figure 5: Example of plots from the "Optical Information" notebook: the time of the optical hits with respect to the beam trigger (a), the difference in wire units between the barycenter of the optical flash and the one of the neutrino hits as identified by Pandora (b), display of the 32 PMT detectors with size and color weighted by the number of photoelectrons in a specific event, as well the flash barycenter and its width (c).

References

[1] R. Acciarri et al. (MicroBooNE), JINST **12**, P02017 (2017), `1612.05824`

[2] A.A. Aguilar-Arevalo et al. (MiniBooNE), Phys. Rev. D **103**, 052002 (2021), `2006.16883`

[3] R. Acciarri et al. (MicroBooNE, LAr1-ND, ICARUS-WA104) (2015), `1503.01520`

[4] Y. Chen et al. (2021), `2108.02214`

[5] C. Green, J. Kowalkowski, M. Paterno, M. Fischler, L. Garren, Q. Lu, Journal of Physics: Conference Series **396**, 022020 (2012)

[6] R. Brun, F. Rademakers, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **389**, 81 (1997), new Computing Techniques in Physics Research V

[7] A. Dorigo, P. Elmer, F. Furano, A. Hanushevsky, WSEAS Transactions on Computers **4**, 348 (2005)

[8] S. Lee et al., Parallel Comput. **110**, 102877 (2022), `2205.01168`

[9] S. Agostinelli et al. (GEANT4), Nucl. Instrum. Meth. A **506**, 250 (2003)

[10] R. Acciarri et al. (MicroBooNE), Eur. Phys. J. C **78**, 82 (2018), `1708.03135`

[11] P. Abratenko et al., *MicroBooNE BNB Inclusive Overlay Sample (No Wire Info)* (2023), `https://doi.org/10.5281/zenodo.8370883`

[12] P. Abratenko et al., *MicroBooNE BNB Inclusive Overlay Sample (With Wire Info)* (2022), `https://doi.org/10.5281/zenodo.7262009`

[13] P. Abratenko et al., *MicroBooNE BNB Electron Neutrino Overlay Sample (No Wire Info)* (2022), `https://doi.org/10.5281/zenodo.7261921`

[14] P. Abratenko et al., *MicroBooNE BNB Electron Neutrino Overlay Sample (With Wire Info)* (2022), `https://doi.org/10.5281/zenodo.7262140`

[15] E.L. Snider, G. Petrillo, J. Phys. Conf. Ser. **898**, 042057 (2017)

[16] P. Abratenko et al. (MicroBooNE), Phys. Rev. D **103**, 052012 (2021), `2012.08513`

[17] P. Abratenko et al. (MicroBooNE), Phys. Rev. D **103**, 092003 (2021), `2010.08653`

[18] C. Adams et al. (MicroBooNE), Phys. Rev. D **99**, 092001 (2019), `1808.07269`

[19] R. Acciarri et al. (MicroBooNE), JINST **12**, P03011 (2017), `1611.05531`