

MicroBooNE Open Datasets

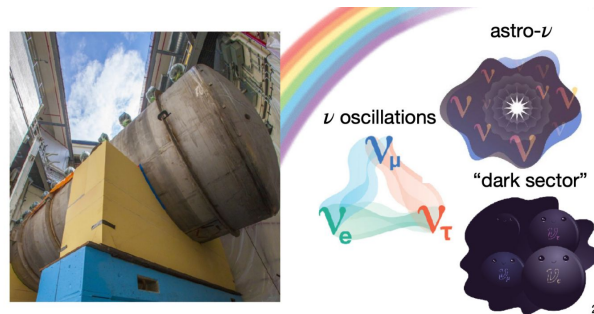
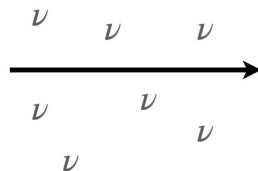
Past Experience and Future Plans

D. Caratelli, G. Cerati, H. Greenlee, K. Terao, M. Toups
On Behalf of the MicroBooNE Collaboration

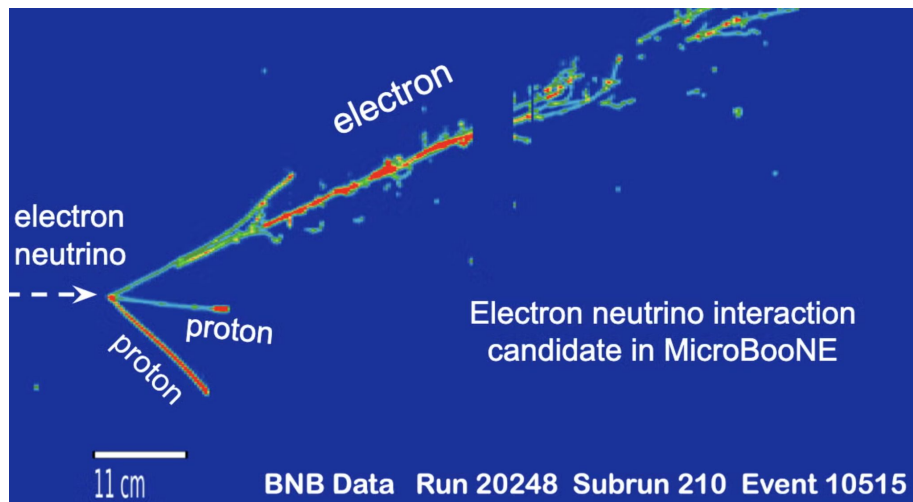
NPML 2026 - Irvine, CA - June 18th 2026

The MicroBooNE Experiment

First large-scale LArTPC to collect neutrino data in US. Six years of beam operations: 2015-2021



- First large-scale LArTPC in US
- Operations milestones:
 - cold electronics,
 - 10s of ms e- lifetime
- Driver for LArTPC reconstruction
- Broad XSEC, BSM, OSC physics program
 - 80+ papers & counting



MicroBooNE Data Characteristics

2.5 x 2.3 x 10.3 meter detector: 85 tonnes of active LAr target.

BNB dataset: 800 MeV mean neutrino energy:

95% ν_μ | 4% anti- ν_μ | 0.5% ν_e .

TPC: ~8,000 wires across 3 wire-planes with 3 mm pitch.

PDS: 32 8" Hamamatsu PMTs.

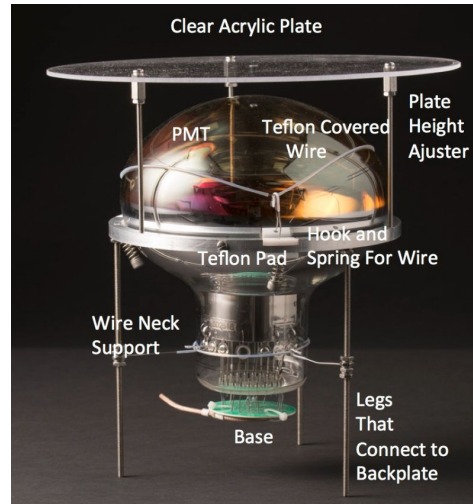
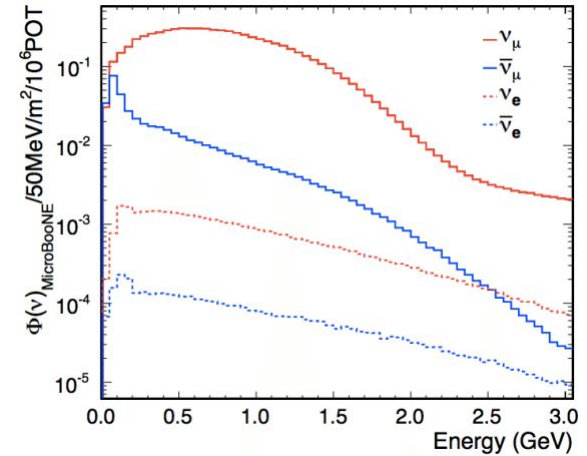
Surface LArTPC:

10 cosmics / 2.3 ms drift window.

Raw data @ 50MB/evt w/ 1 Hz trigger

For 5 years: 9 PB (then add MC, reco, ...)

Curating data-release includes prioritizing
& being efficient in what to store and how.



MicroBooNE Public Data Release (2022)

See [arXiv:2309.15362](https://arxiv.org/abs/2309.15362)

Released:

- 750k BNB and
- 90 ν_e -enhanced “overlay” events.

artROOT & .hdf5 formats

- Targeting LArTPC experts & broader AI/ML communities.

[FAIR principles](#) for data management

Positive feedback from (non-LArTPC) end-users on ease of access and documentation.

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 artroot 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	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

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.

microboone.fnal.gov/documents-publications/public-datasets/

Public Data Documentation & Tutorials

Wire Image Data

Purpose of this notebook

This notebook provides an example for visualizing the recorded data in each event as a set of images, one per TPC plane. The image can be used for visual data processing, such as convolutional neural networks. Recent examples from MicroBooNE are "Semantic Segmentation with a Sparse Convolutional Neural Network for Event Reconstruction in MicroBooNE", arXiv:2012.08513, Phys. Rev. D103, 052012 (2021) and "A Convolutional Neural Network for Multiple Particle Identification in the MicroBooNE Liquid Argon Time Projection Chamber", arXiv:2010.08653, Phys. Rev. D103, 092003 (2021).

In order to use this notebook, samples labeled as "with Wire" need to be used; due to large size of the Wire waveform information, these are samples with reduced number of events. The 2D image shows the time (in TPC clock ticks) vs the wire number, and the color scale is the wire waveform amplitude at each tick in ADC units. Larger ADC values mean larger deposited charge.

Ground truth information with resolution of image pixels is not provided, as it would further increase the size of the sample. However, as shown in the "Hit Labeling.ipynb" notebook, we provide tools to produce ground truth information from hits that have the same 2D structure of these images and thus can be used to define the image ground truth. For instance, one can check if a non-zero pixel lies on the same wire and within $\pm n \times rms$ of the `local_time` of a hit, and in that case assign to the pixel the same label as the hit. An example of how to do this will be demonstrated in this notebook.

Load useful libraries, utilities, and data

```
In [1]: import numpy as np
import matplotlib.pyplot as plt

In [2]: from microboone_utils import *

In [3]: from pynuml.io import File
f = File("bnb_WidthWire_00.h5")

In [4]: tables = ['event_table', 'wire_table', 'hit_table', 'edep_table']
for t in tables: f.add_group(t)
f.read_data(0, 50)
evts = f.build_evt()
```

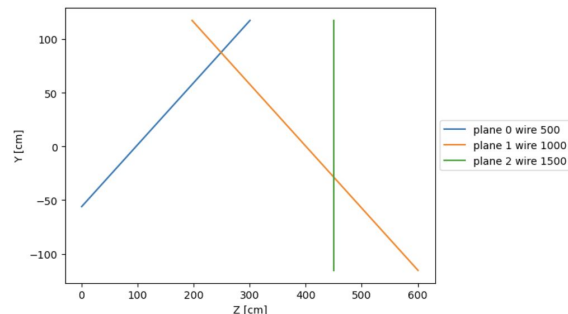
```
y = [start[1],end[1]]
plt.plot(x, y, label = "plane 0 wire %i"%w0, linestyle="--")

start = wireStartPos(1,w1)
end = wireEndPos(1,w1)
x = [start[2],end[2]]
y = [start[1],end[1]]
plt.plot(x, y, label = "plane 1 wire %i"%w1, linestyle="--")

start = wireStartPos(2,w2)
end = wireEndPos(2,w2)
x = [start[2],end[2]]
y = [start[1],end[1]]
plt.plot(x, y, label = "plane 2 wire %i"%w2, linestyle="--")

plt.xlabel('Z [cm]')
plt.ylabel('Y [cm]')
plt.legend(loc='center left', bbox_to_anchor=(1, 0.5))
plt.show()

print('do plane 0 wire %i and plane 1 wire %i cross?'%(w0,w1),doWiresCross(0,w0,1,w1))
if (doWiresCross(0,w0,1,w1)): print('crossing:',wireCrossingYZ(0,w0,1,w1))
print('do plane 1 wire %i and plane 2 wire %i cross?'%(w1,w2),doWiresCross(1,w1,2,w2))
if (doWiresCross(1,w1,2,w2)): print('crossing:',wireCrossingYZ(1,w1,2,w2))
print('do plane 0 wire %i and plane 2 wire %i cross?'%(w0,w2),doWiresCross(0,w0,2,w2))
if (doWiresCross(0,w0,2,w2)): print('crossing:',wireCrossingYZ(0,w0,2,w2))
```

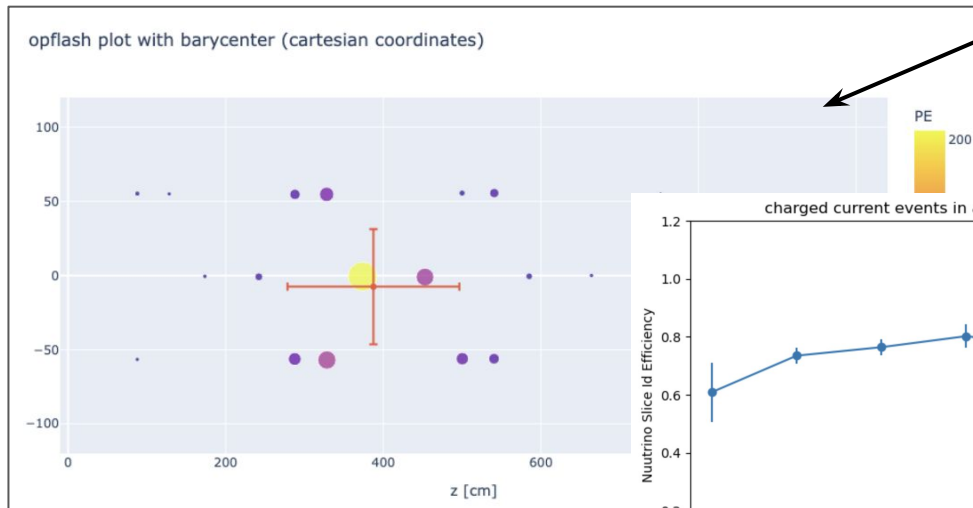


```
do plane 0 wire 500 and plane 1 wire 1000 cross? True
crossing: [87.57235399441043, 248.80003865905684]
do plane 1 wire 1000 and plane 2 wire 1500 cross? True
crossing: [-28.734615075765433, 450.25]
do plane 0 wire 500 and plane 2 wire 1500 cross? False
```

We now check the hit multiplicity in each wire, to find out which wires are not functional or noisy.

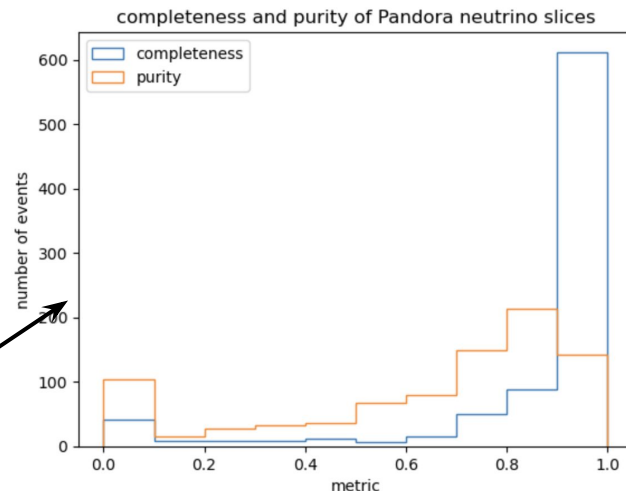
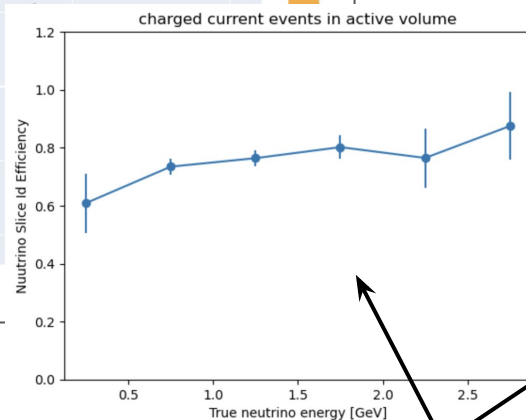
<https://github.com/uboone/OpenSamples>

Public Data Documentation & Tutorials



Interpretation of optical vs. charge “image” comparison with simple “flash-matching” metrics

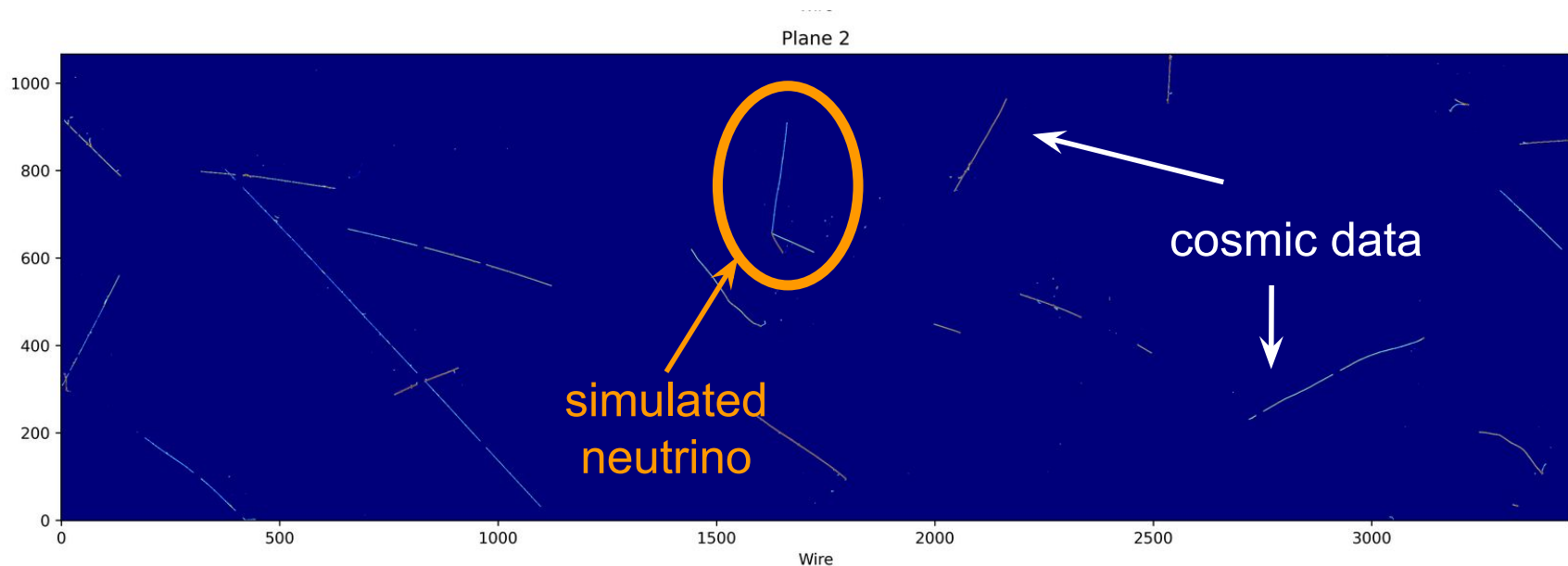
[Eur.Phys.J.C 78 \(2018\) 1, 82](#)



Comparison with Pandora reconstruction in uB samples. Not the latest but still the basis for many current papers and a great benchmark for external users

<https://github.com/uboone/OpenSamples>

MicroBooNE Simulation: Overlays



MicroBooNE simulation uses a data-driven “overlay” strategy:

- Simulate neutrino interaction
- Overlay on top of a data event collected with unbiased trigger.
 - Fully data-driven cosmics, radiological, and electronics noise
 - No truth information for these!

MicroBooNE “Open Data” File Contents

“wire” information (top view) for
10% subset of events:

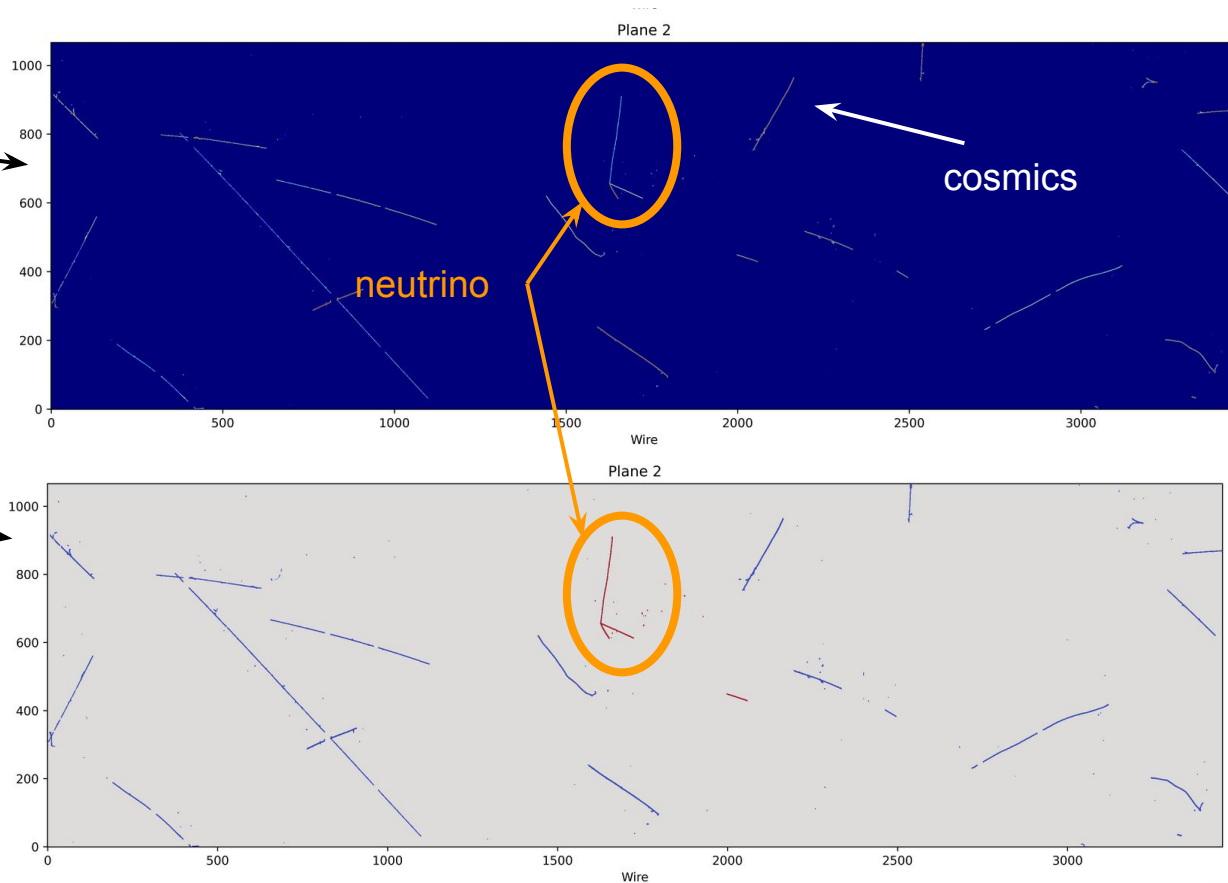
- After noise filtering
- After signal processing

Reconstructed quantities:

- 2D hits
- Pandora reco output

Truth Information / labels:
(Including backtracked
EDEPs)

Access to artROOT files allows
people to re-process and
add/modify information for
further analysis.



Possible Applications

Public data sets allow for a wide range of developments, but primarily can be used for training or testing of AI models.

Some possible applications include:

- Foundation models
- Generative simulation
- Anomaly detection
- Detector surrogate models
- ... [your idea here]

MicroBooNE Public Data Release Impact (so far...)

Nine peer-reviewed publications.

Teams include:

- Mix of uB / LArTPC reconstruction experts
- Mix of ML / LArTPC experts

Data release includes:

- Jupyter notebooks with example tutorials
- Pandora information for performance reference

Availability of well documented .hdf5 formatted samples was seen favourably: lowered barrier for access / usage.

microboone.fnal.gov/documents-publications/public-datasets/

Peer-reviewed Publications

1. A Case Study of Data Management Challenges Presented in Large-Scale Machine Learning Workflows

Lee CS (Northwestern U.), Hewes V (Cincinnati U.), Cerati G (Fermilab), J. Kowalkowski (Fermilab), Aurisano A (Cincinnati U.), Agrawal A (Northwestern U.), Choudhary A (Northwestern U.) and Liao W-K (Northwestern U.)
DOI: 10.1109/CCGrid57682.2023.00017
Published in: 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Bangalore, India, 2023, pp. 71-81

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

MicroBooNE Collaboration · Giuseppe Cerati (Fermilab) for the collaboration.
e-Print: 2309.15362 [hep-ex]
DOI: 10.1051/epjconf/202429508012
Published in: EPJ Web Conf. 295 (2024), 08012

3. Graph neural network for neutrino physics event reconstruction

A. Aurisano (Cincinnati U.), V. Hewes (Cincinnati U.), G. Cerati (Fermilab), J. Kowalkowski (Fermilab), C.S. Lee (Northwestern U.), W. Liao (Northwestern U.), D. Grzenda (Chicago U.), K. Gumpula (Chicago U.), X. Zhang (UCLA and Chicago U.)
e-Print: 2403.11872 [physics.data-an]
DOI: 10.1103/PhysRevD.110.032008 (publication)
Published in: Phys.Rev.D 110 (2024) 3, 3

4. Addressing GPU memory limitations for Graph Neural Networks in High-Energy Physics applications

Lee CS (Northwestern U.), Hewes V (Cincinnati U.), Cerati G (Fermilab), Wang K (Northwestern U.), Aurisano A (Cincinnati U.), Agrawal A (Northwestern U.), Choudhary A (Northwestern U.) and Liao W-K (Northwestern U.)
DOI: 10.3389/fhpcp.2024.1458674 (publication)
Published in: Front. High Perform. Comput. 2:1458674 (2024)

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

Callum Duffy (University Coll. London), Marcin Jastrzebski (University Coll. London), Stefano Vergani (University Coll. London), Leigh H. Whitehead (Cambridge U.), Ryan Cross (Warwick U.), Andrew Blake (Lancaster U.), Sarah Malik (University Coll. London), John Marshall (Warwick U.)
DOI: 10.1103/byy5-zk73 (publication)
Published in: Phys.Rev.D 112 (2025) 9, 092006

6. Optimal Transport for $\$/n^{\circ}05$ Particle Classification in LArTPC Neutrino Experiments

David Caratelli (UCSB), Nathaniel Craig (UCSB, KITP), Chuyue Fang (UCSB), Jessica N. Howard (UCSB, KITP)
DOI: 10.1103/6zws-frnf (publication)
Published in: Phys.Rev.D 113 (2026) 7, 072005

7. NuGraph2 with Context-Aware Inputs: Physics-Inspired Improvements in Semantic Segmentation

Vitor F. Grizzi (Illinois U., Urbana and Argonne), Margaret Voetberg (Fermilab), Giuseppe Cerati (Fermilab), Hadi Meidani (Illinois U., Urbana), V. Hewes (Cincinnati U.)
DOI: 10.1007/JHEP03(2026)149 (publication)
Published in: JHEP 03 (2026) 149

8. NuGraph2 with Explainability: Post-hoc Explanations for Geometric Neural Network Predictions

Margaret Voetberg (Fermilab), Vitor F. Grizzi (Illinois U., Urbana and Argonne), Giuseppe Cerati (Fermilab), Hadi Meidani (Illinois U., Urbana), V. Hewes (Cincinnati U.)
e-Print: 2509.10676 [hep-ex]

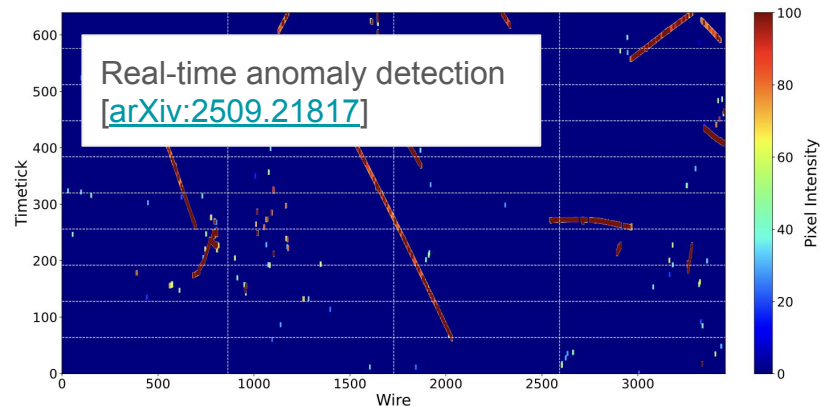
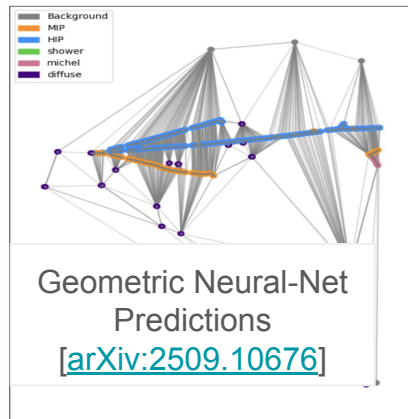
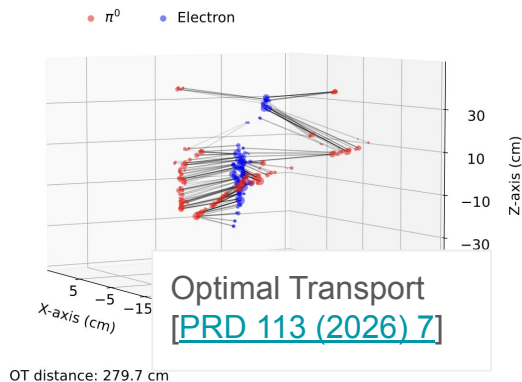
9. Real-time Anomaly Detection for Liquid Argon Time Projection Chambers

Seokju Chung (Columbia U.), Jack Clevee (Columbia U.), Akshay Malige (Columbia U.), Georgia Karagiorgi (Columbia U.), Lino Gerlach (Princeton U.), Adrian A. Pol (Princeton U.), Isobel Ojalvo (Princeton U.)
e-Print: 2509.21817 [hep-ex]

Other Conference Presentations

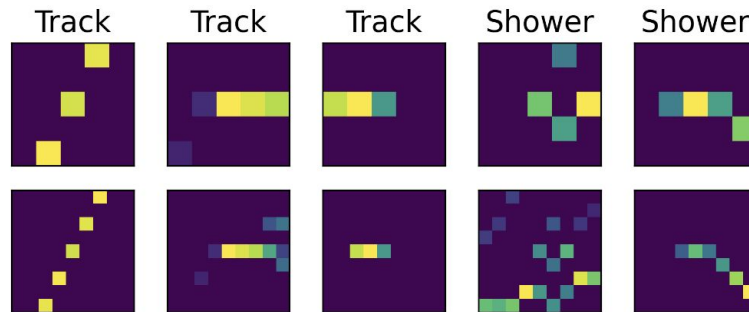
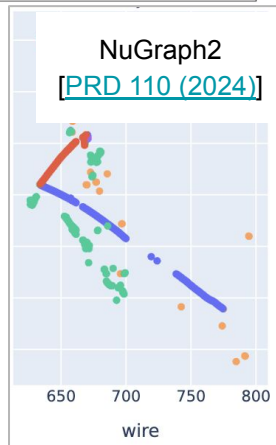
1. Online tagging and triggering with deep learning AI for next generation particle imaging detector, M. Bhattacharya (Fermilab), CHEP 2023.
2. SparsePixels: Efficient Convolution for Sparse Data on FPGAs, H. Tsoi (University of Pennsylvania), Fast Machine Learning for Science Conference 2025.

MicroBooNE Public Data Release Impact (so far...)



Some recent examples.

Common theme: team of 1+2
LArTPC experts plus external
collaborators.



Quantum Machine Learning Hit-Based Classifier
[PRD 112 (2025) 9, 092006]

Impact: re-integration in experimental workflow

These developments are starting to have a real impact

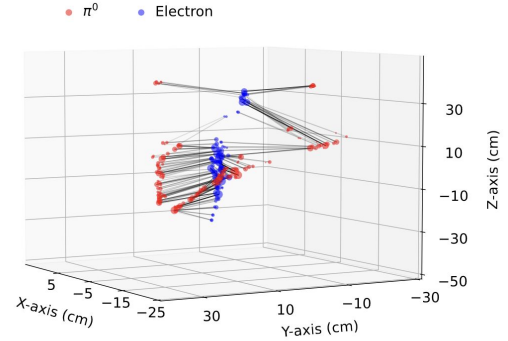
ν Optimal Transport
[PRD 113 (2026) 7]

LArTPC / AI+ML
community

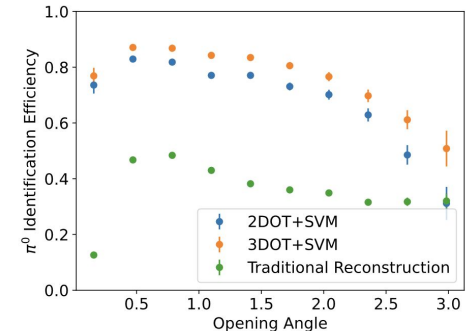


π^0 rejection without shower counting. Integrating in a first MicroBooNE analysis and then expand to DUNE

“neutrino OT”

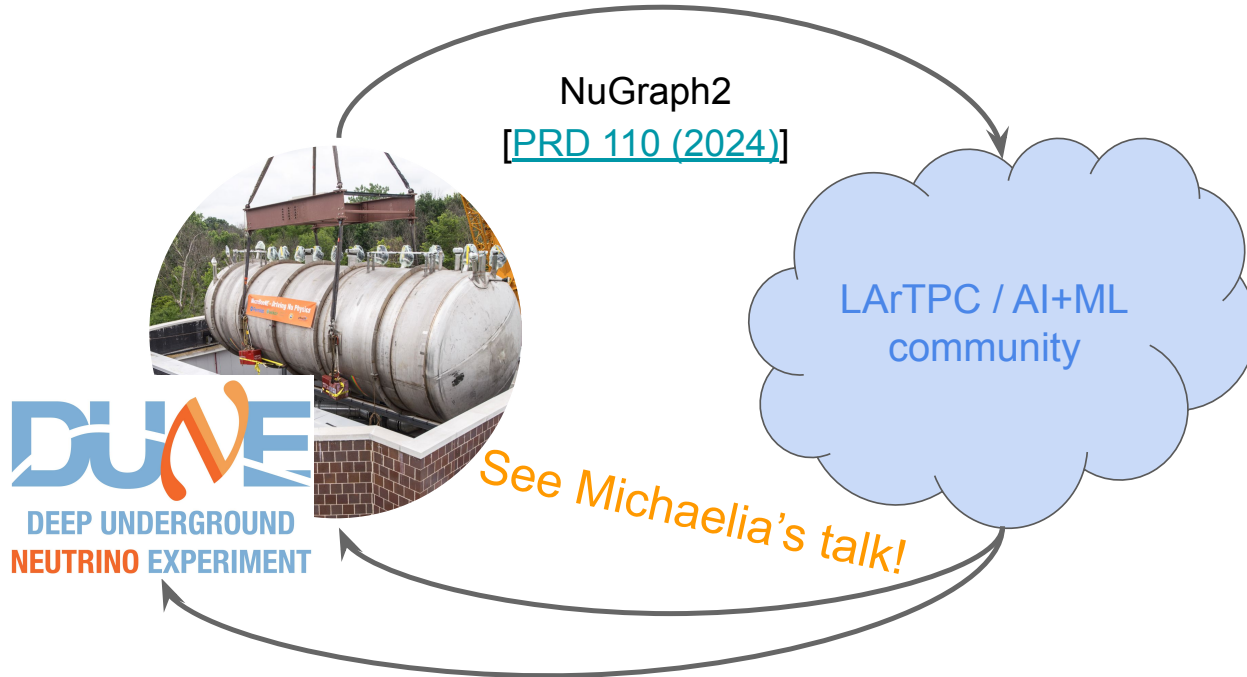


OT distance: 279.7 cm



Impact: re-integration in experimental workflow

These developments are starting to have a real impact

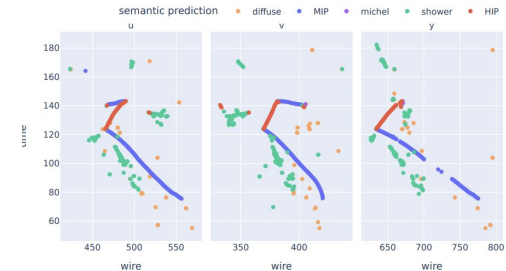


NuGraph2 development bringing major improvements in reconstruction for MicroBooNE and being deployed in other LArTPC experiments

NuGraph2

a)

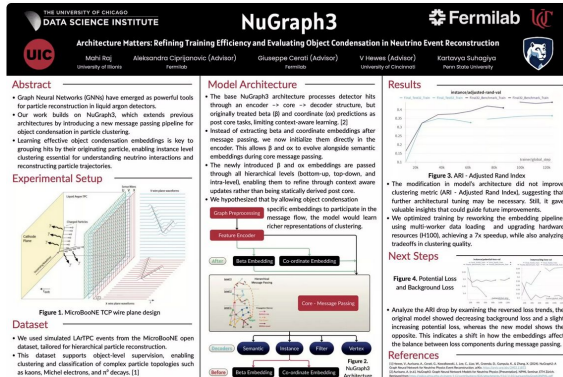
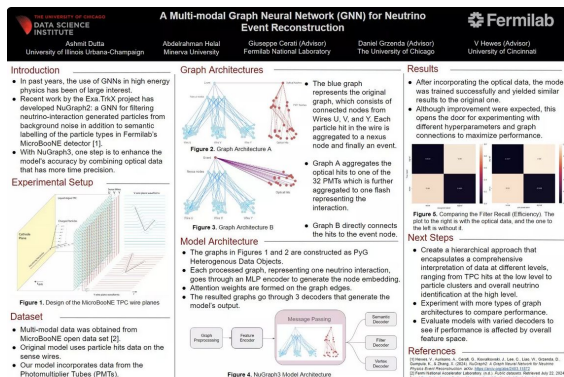
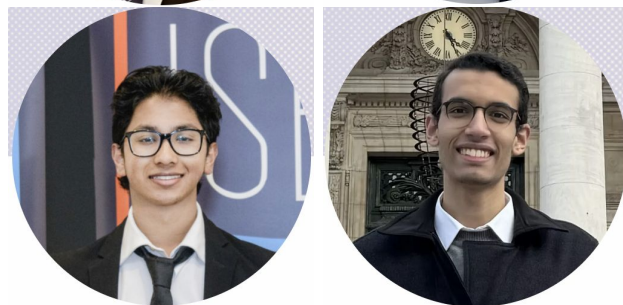
Precision, Baseline					
True label \ Assigned label	MIP	HIP	Shower	Michel	Diffuse
Diffuse	0.002	0.011	0.025	0.048	0.78
Michel	0.001	0.0011	0.013	0.6	0.026
Shower	0.0025	0.0066	0.89	0.12	0.075
HIP	0.0074	0.9	0.015	0.019	0.038
MIP	0.99	0.078	0.059	0.21	0.081



Impact: training & workforce development

Provides opportunity for undergraduate students to work on real science problems and contribute to cutting-edge research.

- Great for projects for interns / summer students. E.g. through UChicago DSI internship w/ G. Cerati & V. Hewes.



Impact: outreach & education

Public Data can also provide the basis for engaging student in coursework or outreach activities.

[arXiv:2509.18859](https://arxiv.org/abs/2509.18859)

- Hands-on labs or independent research projects for class
- Source for visuals / input data for outreach

A sterile-neutrino search using data from the MicroBooNE liquid-argon time projection chamber performed in an undergraduate teaching laboratory

John Waiton¹, Joseph Bateman¹, Justin J. Evans¹, Ole Gunnar Finnerud¹, Elena Gramellini¹, Roxanne Guenette¹, Pawel Guzowski¹, Aleksander Kedziora¹, Stefan Söldner-Rembold²

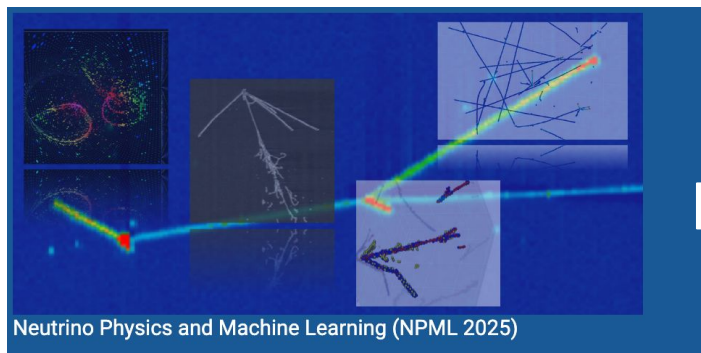
¹University of Manchester, Manchester M13 9PL, UK

²Imperial College of Science, Technology and Medicine, London SW7 2BZ, UK

E-mail: elena.gramellini@manchester.ac.uk

July 2025

New opportunities



Oct 20 – 31, 2025
University of Tokyo
Asia/Tokyo timezone

Enter your search term

Overview
Registration
Timetable
Satellite Workshop
What's unconference?
Contribution List
Childcare

Satellite Workshop

The satellite workshop for Data Olympic at NPML (DO-NPML) preparation will take place at Kavli IPMU in the week before the main conference.

- When: October 20th (Monday) to 24th (Friday).
- Where: Kavli IPMU 1F Lecture Hall (20th to 22nd) and 4F/5F Balcony A/B (23rd and 24th)

The goal of the workshop is to design and develop a large-scale public datasets and associated scientific research challenges for the neutrino community. The workshop dedicates much time in a discussion style to take your inputs rather than talks and presentations.



DOE Genesis Mission

<https://genesis.energy.gov> - A National Mission to Accelerate Science Through Artificial Intelligence



The Genesis Mission unites DOE National Labs, industry, academia, and more to harness AI for breakthroughs in energy dominance, discovery science, and national security.

- In March 2026, DOE announced funding to advance the Genesis Mission's efforts to tackle the most complex science and technology challenges.
 - Challenge: 14 - Unifying Physics from Quarks to the Cosmos, Focus Area B: AI Accelerated DUNE Science (HEP)
- "We are amid a revolution in computing, driven by artificial intelligence and quantum information technologies, that will transform how science is done. Genesis Mission has the goal of **doubling the productivity and impact of U.S. research** and development by pairing scientists with intelligent systems that reason, simulate, and experiment at extraordinary speed."

American Science Cloud

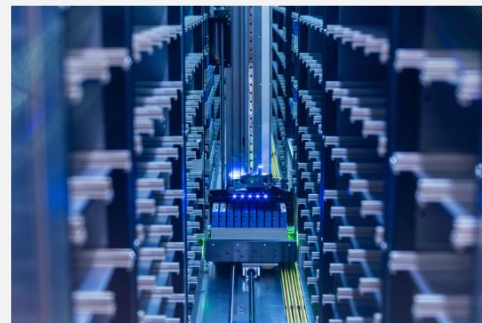
Fermi Data Platform

The Fermi Data Platform, or FDP, provides a high-capacity infrastructure for storing, managing and sharing scientific data within the American Science Cloud.

The platform offers data access through **WebDAV** and **XRootD** protocols, with **Globus** file transfer support coming soon, and uses metadata synchronized with the global AmSC catalog.

Impact: FDP is supporting the data needs of high-energy physics AI applications and is expanding to serve data from across the DOE complex. It is part of the AmSC infrastructure empowering researchers with the tools they need to accelerate discovery.

initial storage: 3 PB



Motivation for New / Expanded Data Releases

Opportunities:

1. GENESIS mission
2. Data Challenges / schools
3. Growing LArTPC community spanning many detectors (DUNE prototypes, SBN)

Studies not possible with existing release:

1. Triggering studies with raw waveform data
2. Detector monitoring over long time-periods
3. Systematic studies of detector response variation
4. Studies on actual data

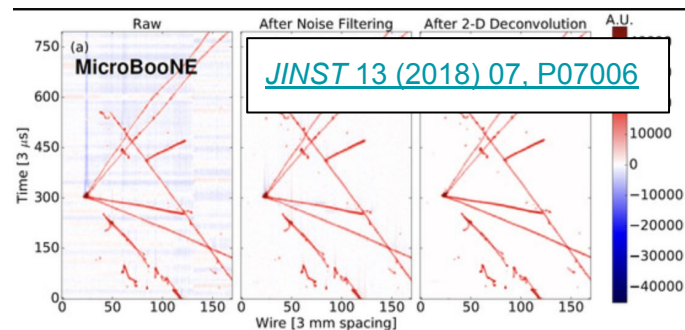
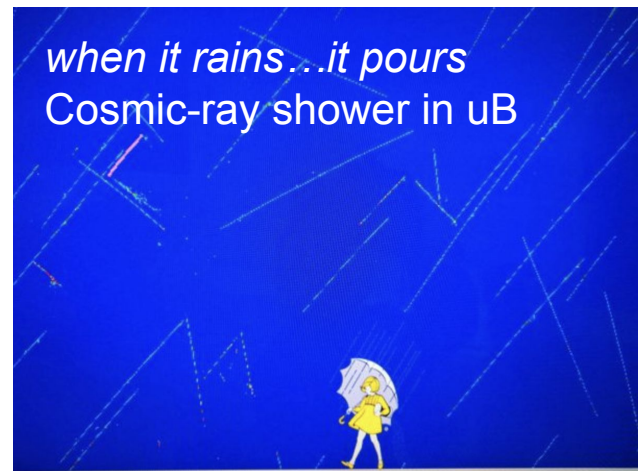
Expanded MicroBooNE Release: Plans

Multi-year cosmic-ray data & simulated neutrino samples:

Multi-year dataset which encompasses time-dependent variations and well-understood calibrations. Can benefit the development of low-level detector status monitoring, new calibration workflows, ...

Raw Waveforms for real data and simulated neutrinos:

Can spur the development of trigger and other online data analysis for LArTPCs.



Expanded MicroBooNE Release: Plans

Detector Variation Samples:

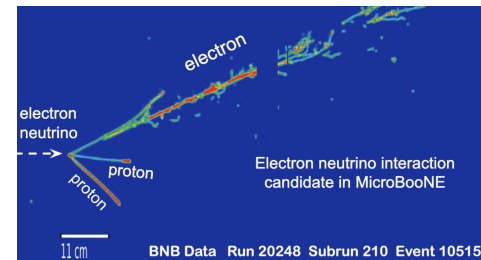
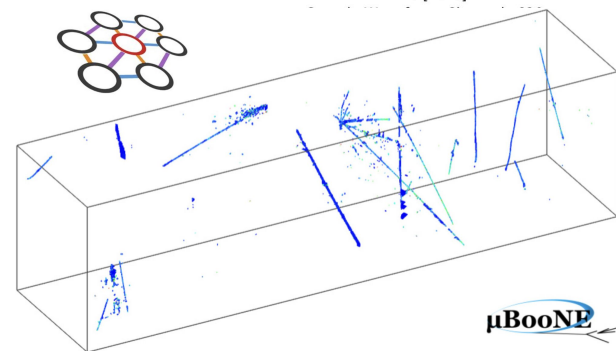
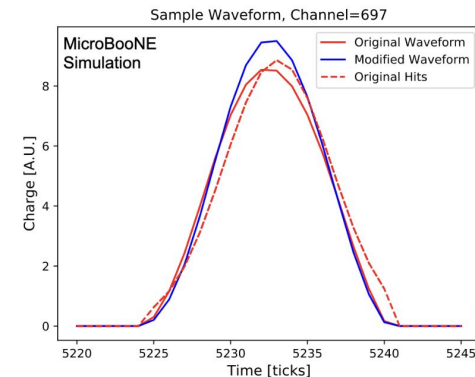
Provides a way to benchmark or aid training to account for detector response uncertainties.

Reconstructed 3D space-points:

Visual output of LArTPC data. Easy to interpret by non-LArTPC experts

Small Neutrino Dataset:

Real neutrinos ideal for data-challenges and engaging development topics



Expanded MicroBooNE Release: Plans

Interested in feedback on:

- Experience with using MicroBooNE Open Data
- Suggestions for future uses / applications of this data
- Requests for datasets you would find useful

In particular

- Are there venues (schools, courses) where you see this data being useful?
- Experience with or suggestions for documentation / code tutorials.
- Applications you are working on that could benefit from this data.

Backup