FERMILAB-SLIDES-24-0134-CSAID

NuGraph2:
A Graph Neural Network for Neutrino Event Reconstruction

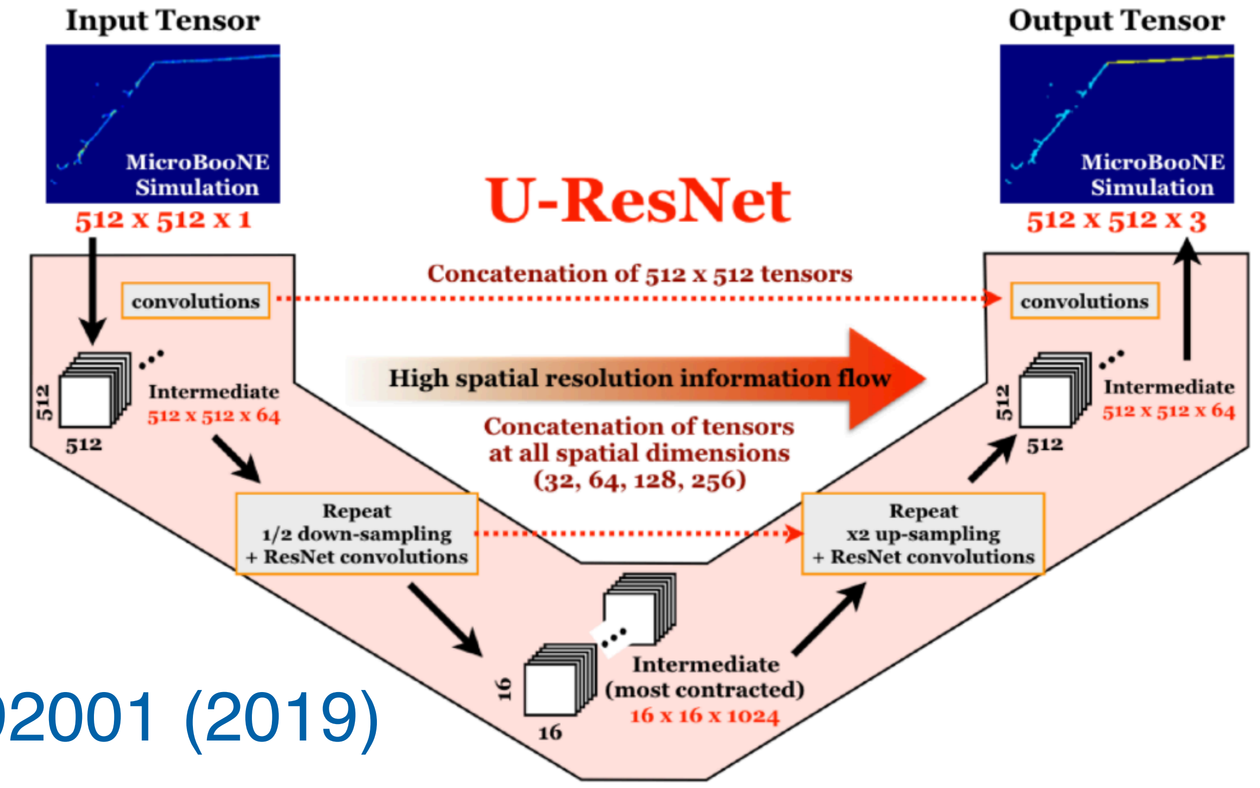
G. Cerati (FNAL)

Neutrino Physics and Machine Learning 2024

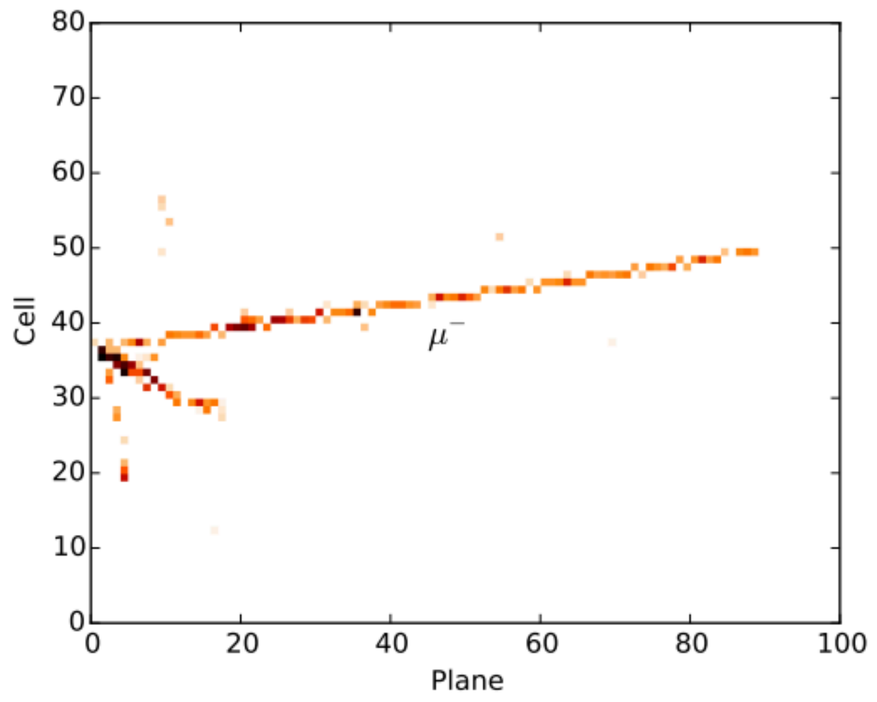
Zurich — June 27, 2024

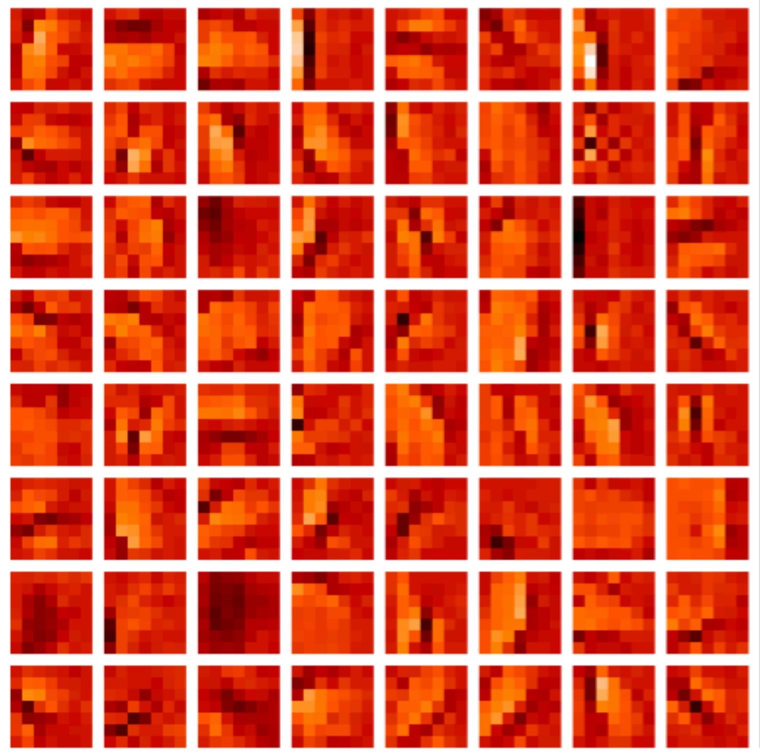
Introduction

- CNN-based networks reshaped event reconstruction for neutrino physics



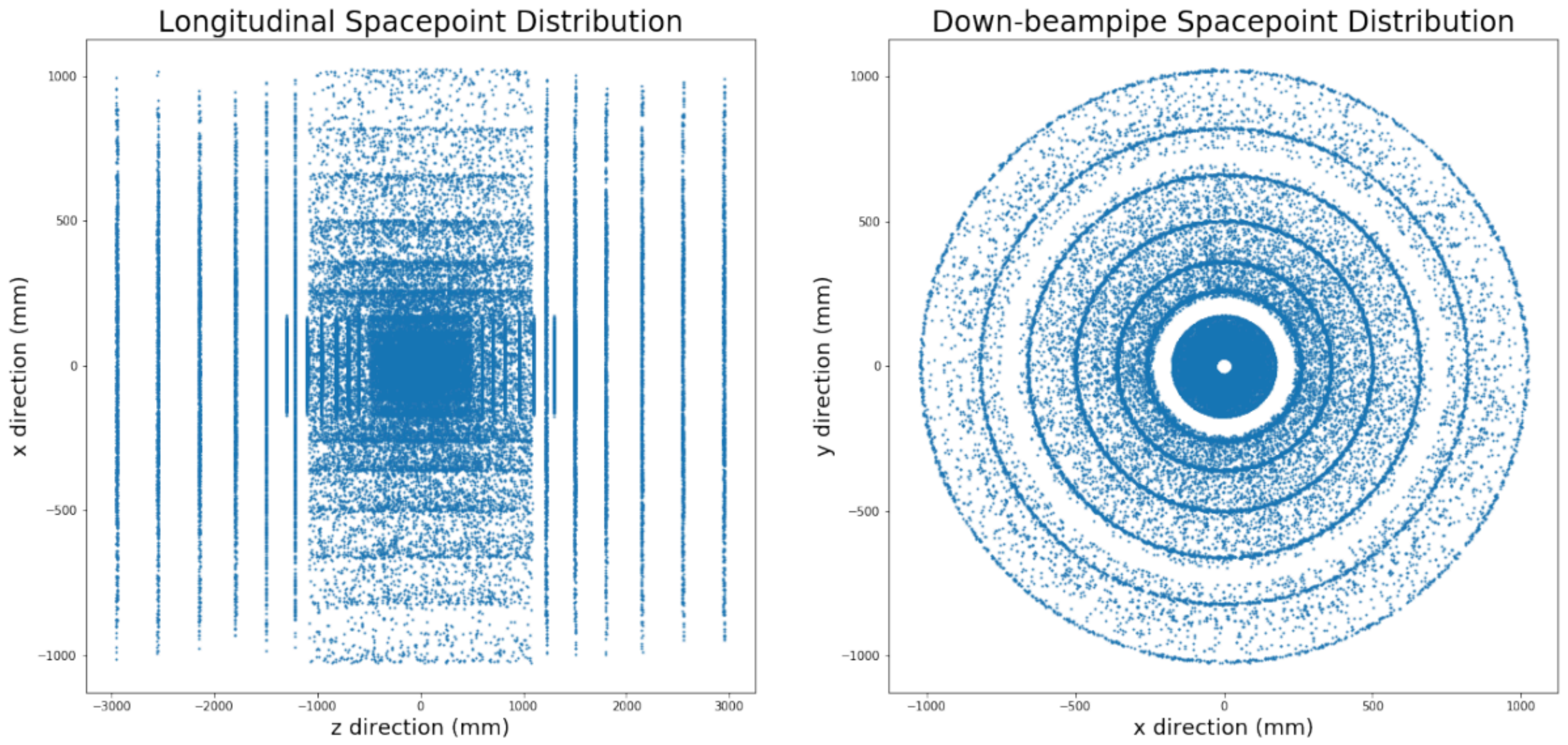
Phys. Rev. D99, 092001 (2019)

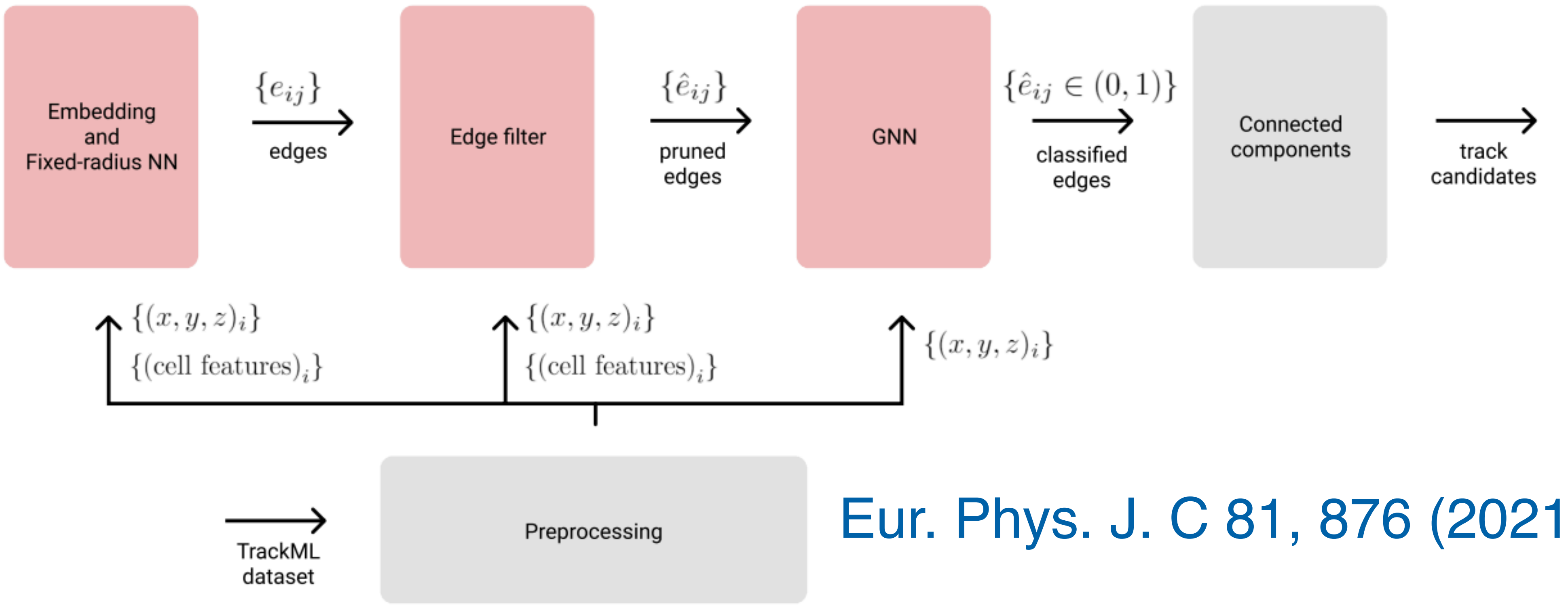




2016 JINST 11 P09001

- GNN proved to be promising for track reconstruction at the LHC
  - naturally sparse
  - no image pre-processing
  - flexible structure





Eur. Phys. J. C 81, 876 (2021)

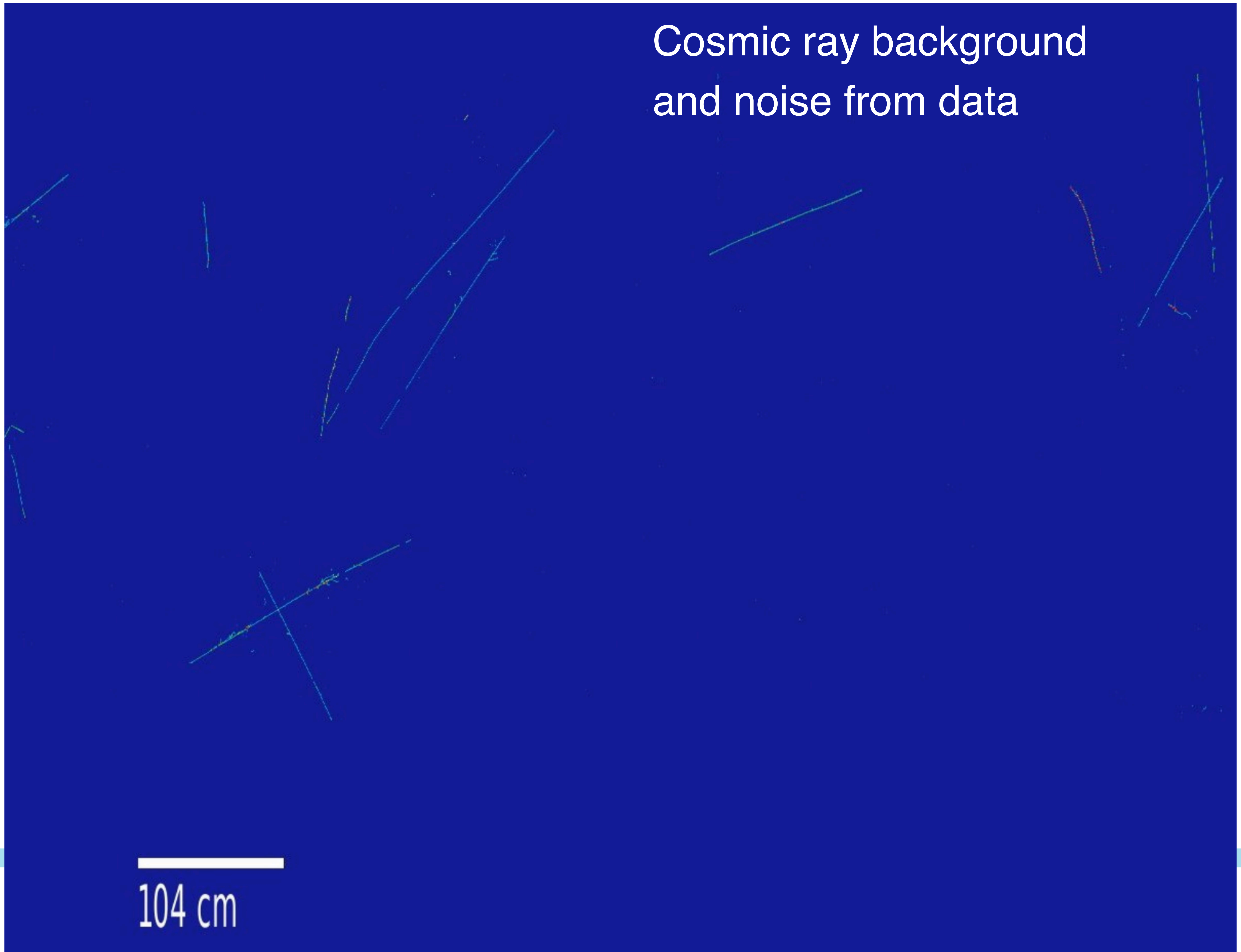
Fermilab

Chapter 1: The Data

MicroBooNE Open Samples

arXiv:2309.15362
https://microboone.fnal.gov/documents-publications/public-datasets/

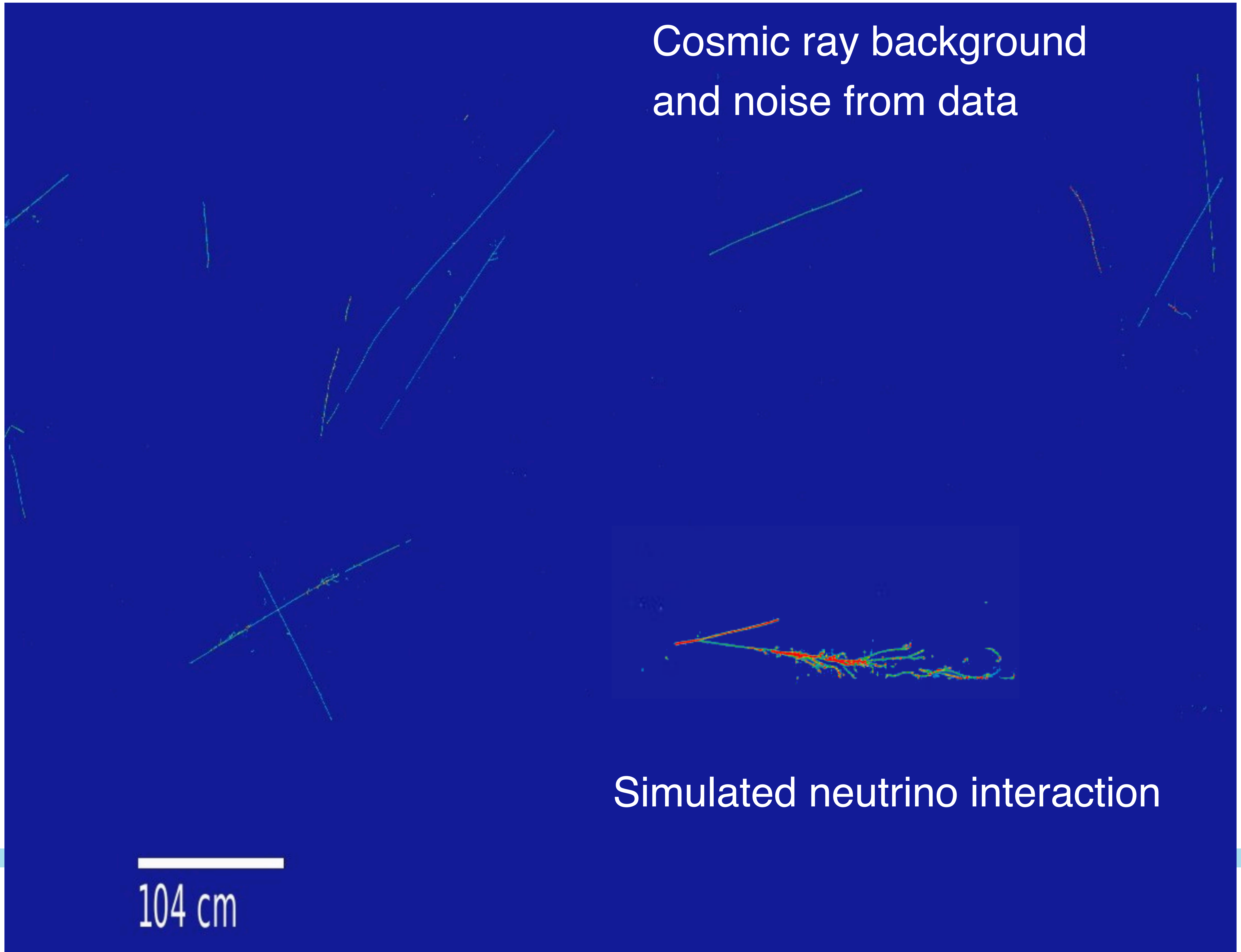
- Two “overlay” samples: BNB **inclusive** and BNB intrinsic ν_e



MicroBooNE Open Samples

arXiv:2309.15362
https://microboone.fnal.gov/documents-publications/public-datasets/

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



Fermilab

MicroBooNE Open Samples — Overview

- Inspired by **FAIR** principles (findable, accessible, interoperable, reusable data)
- Samples available under “cc-by” license. Template text for acknowledgment is provided.
  - requesting resulting software products to be made available
- Two formats: targeting LArTPC and broader data & computer science communities
  - **art/ROOT** is the same format as used by the collaboration.
    - Files are stored on persistent **dCache** pool area and made accessible with **xrootd**
  - **HDF5** include a reduced subset of the art/ROOT information in a simplified format for usage by non-experts.
    - Files stored on **Zenodo**, providing citable DOI (digital object identifier) & versioning.
- Extensive documentation and tutorials are also made public.
  - Notebooks show how to access the data, demonstrate useful applications, define reference performance metrics

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

Chapter 2: The Model

Breaking news!

NuGraph2: A Graph Neural Network for Neutrino Physics Event Reconstruction

A. Aurisano and V Hewes
University of Cincinnati, Cincinnati, OH 45221, USA

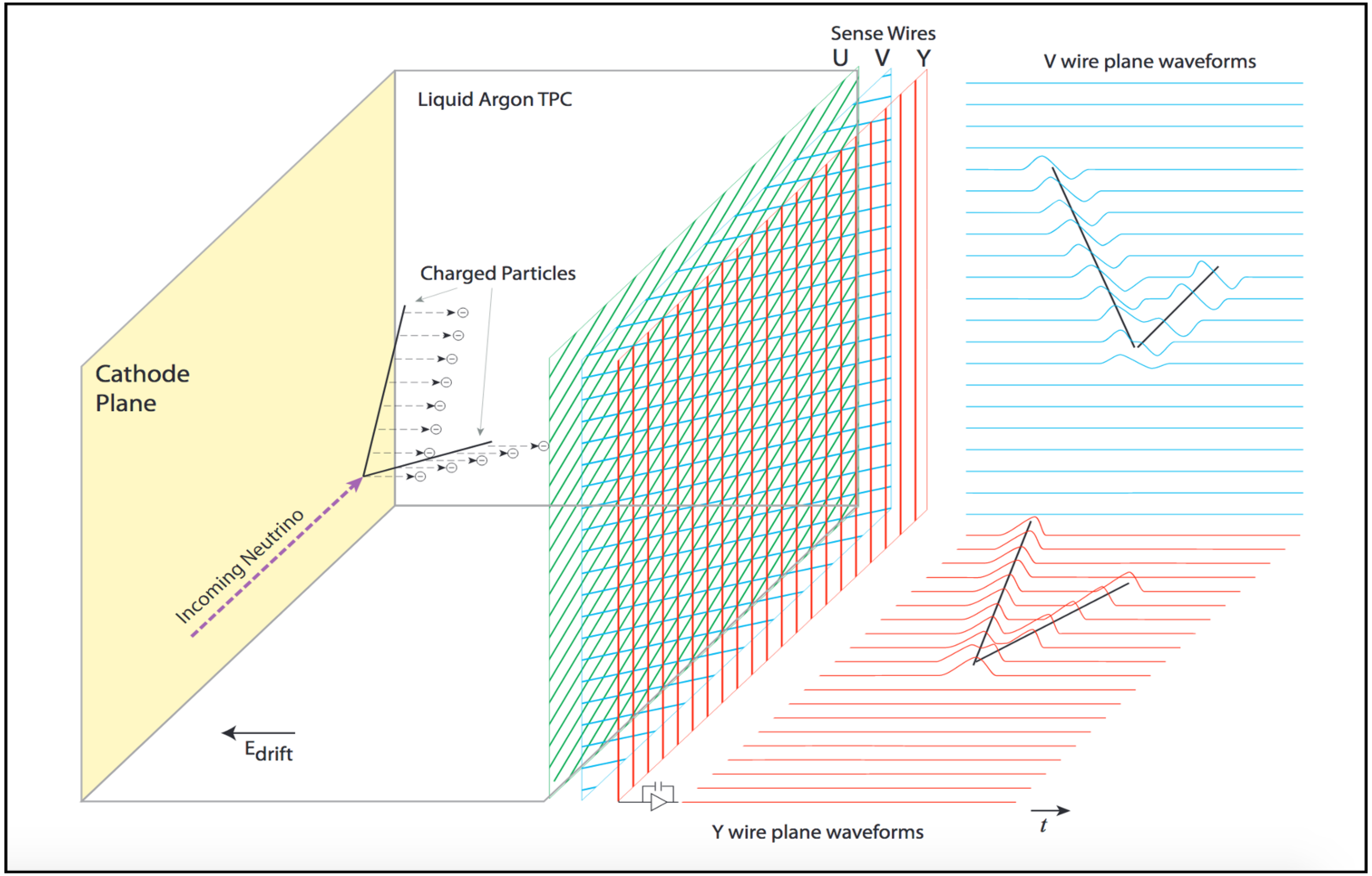
G. Cerati and J. Kowalkowski
Fermi National Accelerator Laboratory, Batavia, IL 60510, USA

C. S. Lee and W. Liao
Northwestern University, Evanston, Il 60208, USA

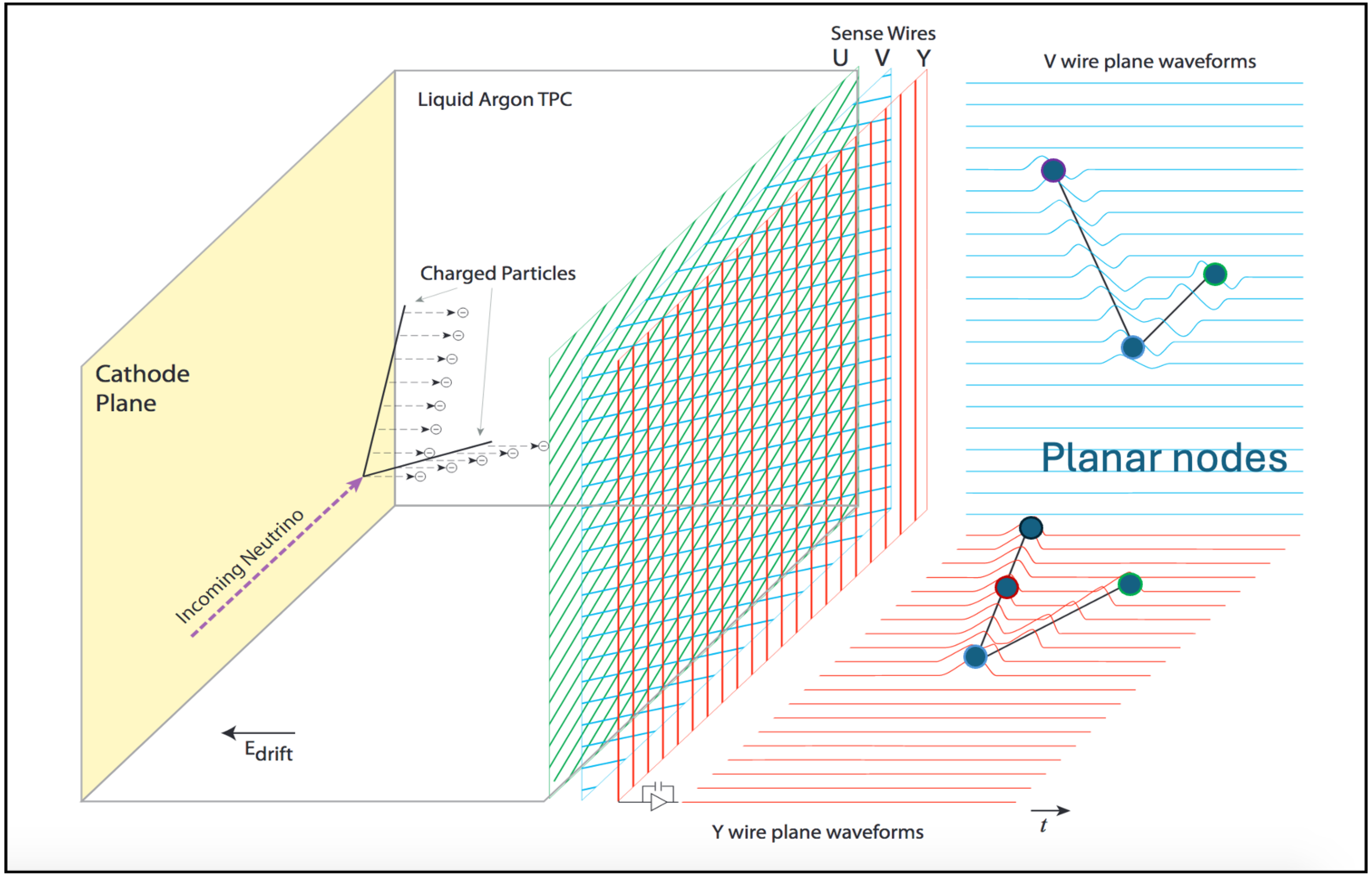
D. Grzenda, K. Gumpula, and X. Zhang[a]
Data Science Institute, University of Chicago, Chicago, IL 60637, USA

- Paper accepted by Phys. Rev. D [arXiv:2403.11872]
- Preprocessed training data set and trained model available on Zenodo
  - https://zenodo.org/records/12169756

Graph Construction

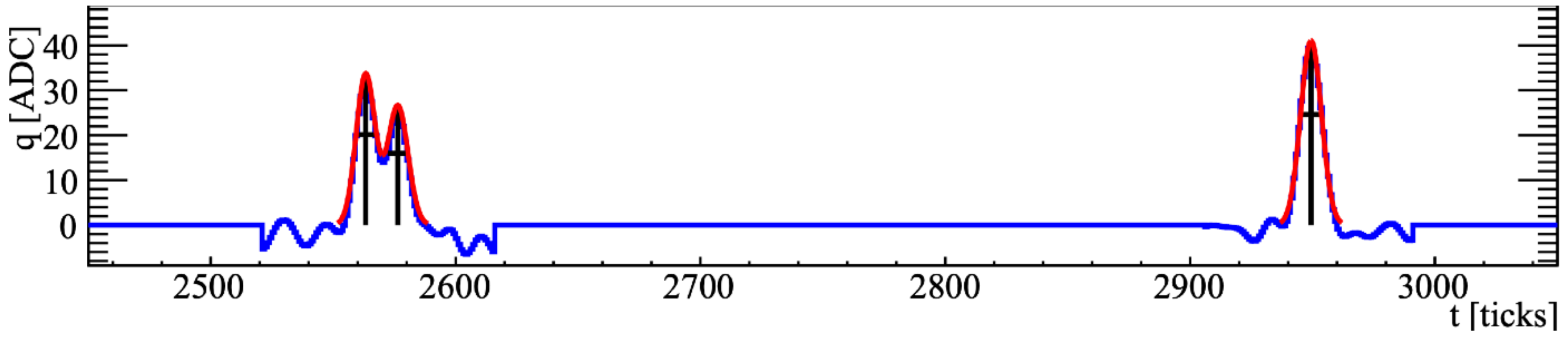


Graph Construction

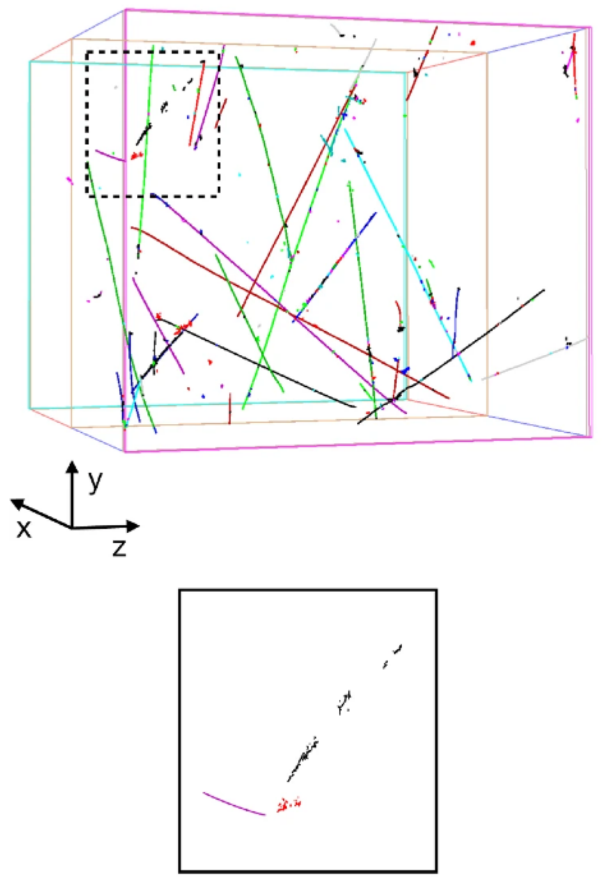


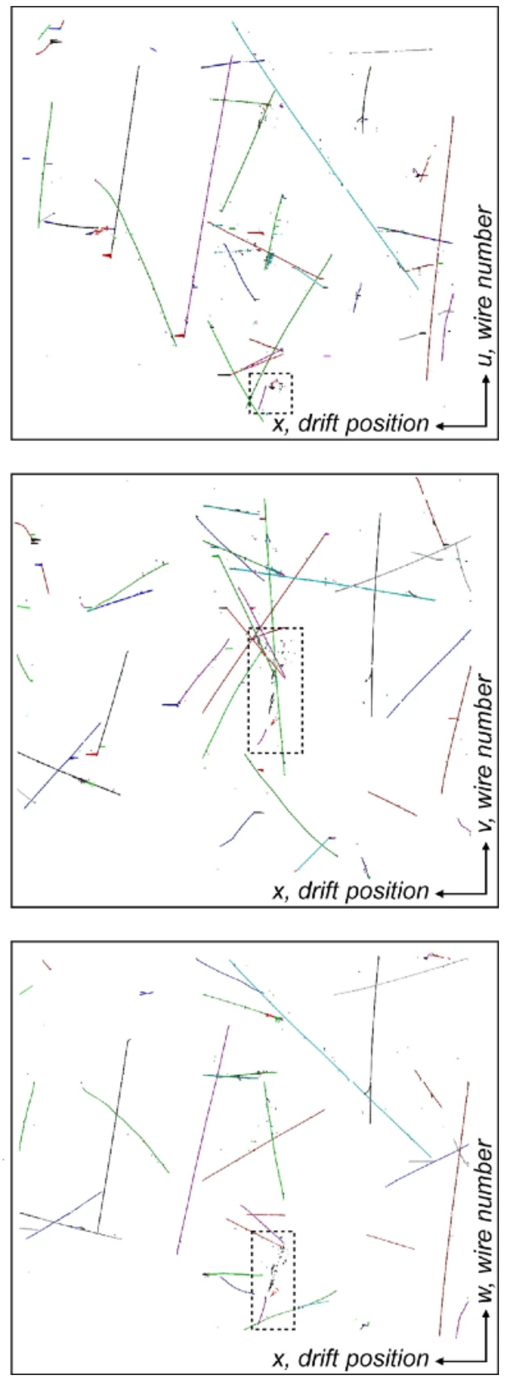
- Main inputs to the GNN are the Hits
  - hits are Gaussian fits to waveforms
  - features: wire, peak time, integral, RMS
  - currently using Hits associated to the neutrino interaction by Pandora

JINST 17 (2022) 01, P01026

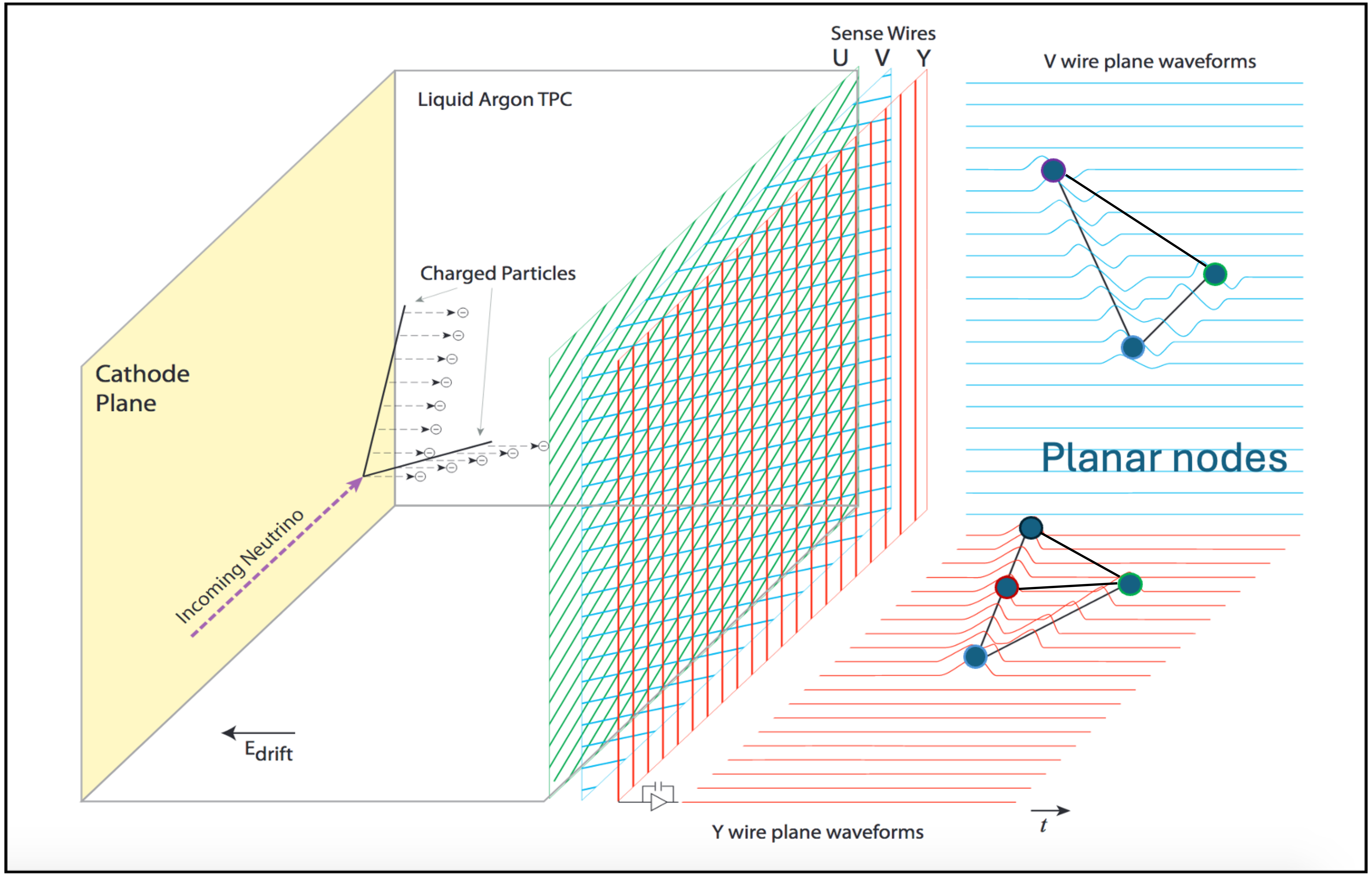


Eur. Phys. J. C 83, 618 (2023)



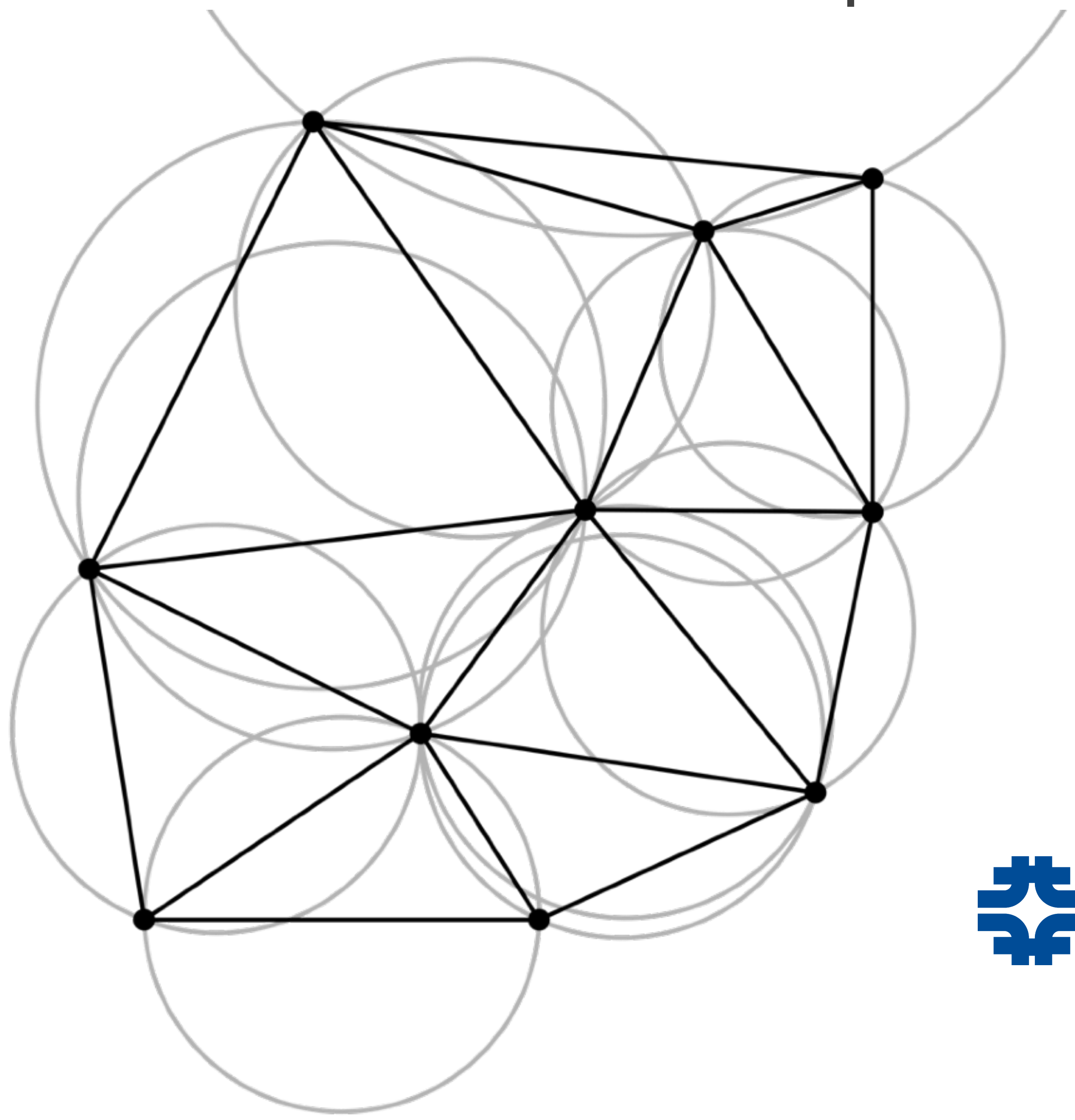


Graph Construction



- Main inputs to the GNN are the Hits
  - hits are Gaussian fits to waveforms
  - features: wire, peak time, integral, RMS
  - currently using Hits associated to the neutrino interaction by Pandora

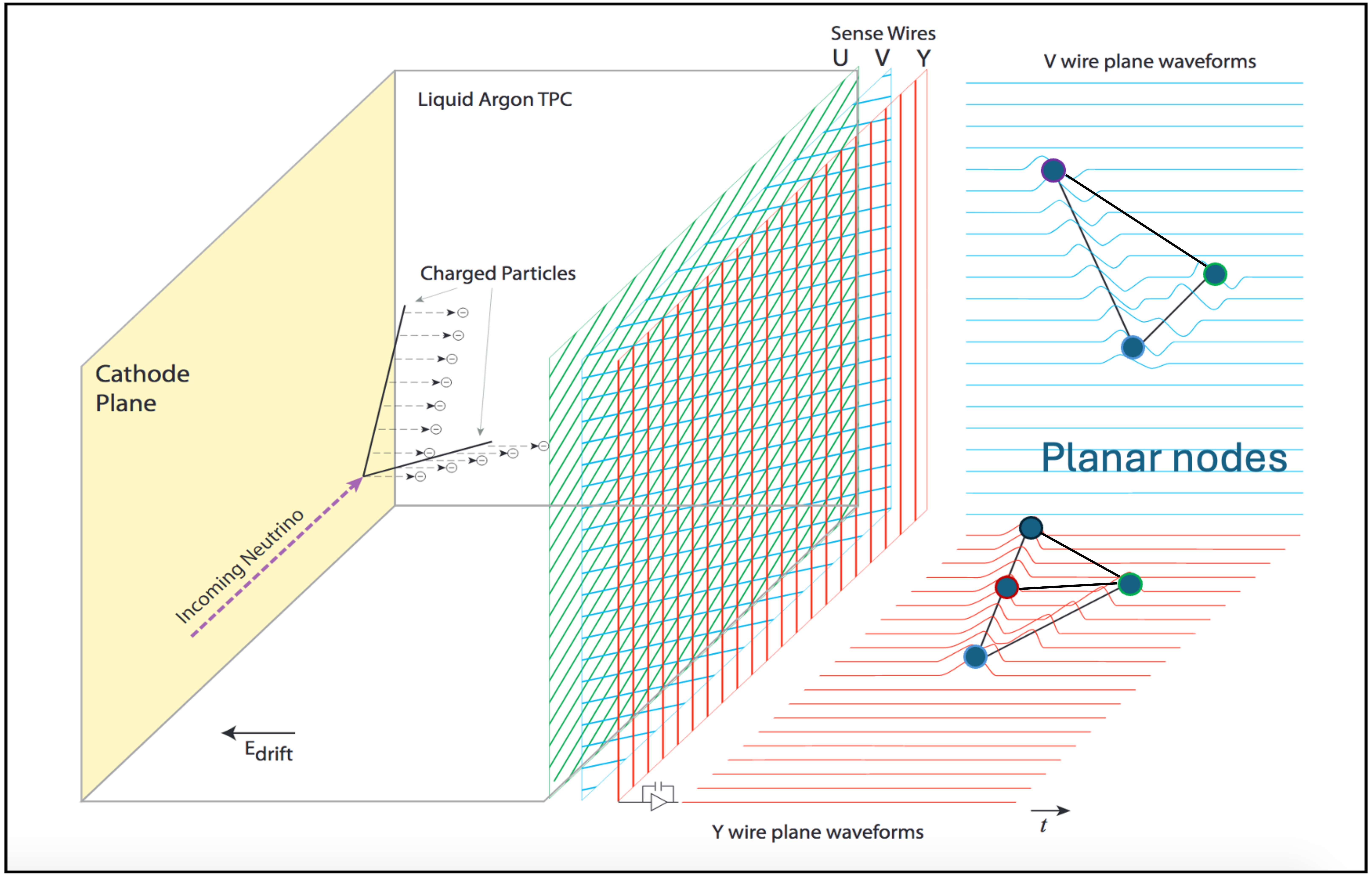
- Within each plane hits are connected in a graph using Delaunay triangulation
  - fully connected graph
  - both long and short distance edges
  - connect across unresponsive wire regions



wikipedia

Fermilab

Graph Construction

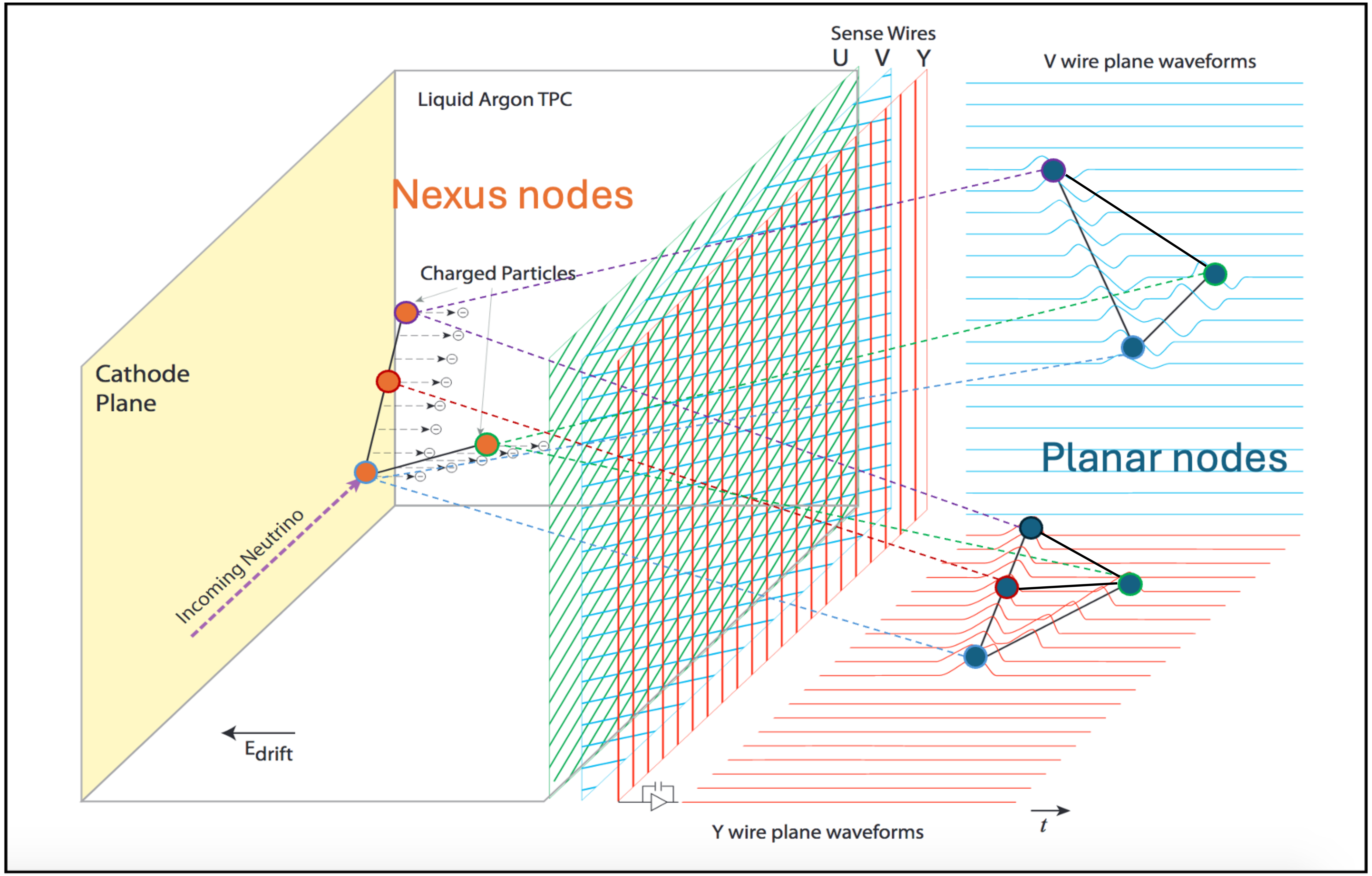


- Main inputs to the GNN are the Hits
  - hits are Gaussian fits to waveforms
  - features: wire, peak time, integral, RMS
  - currently using Hits associated to the neutrino interaction by Pandora
- Within each plane hits are connected in a graph using Delaunay triangulation
  - fully connected graph
  - both long and short distance edges
  - connect across unresponsive wire regions

2023 J. Phys.: Conf. Ser. 2438 012091

Edge Type	Data Type	Labeling Scheme	Accuracy
Delaunay	**2D**	**Simple**	**86.24%**
Window	2D	Simple	76.9%
kNN	2D	Simple	81.14%
Radius	2D	Simple	78.58%

Graph Construction

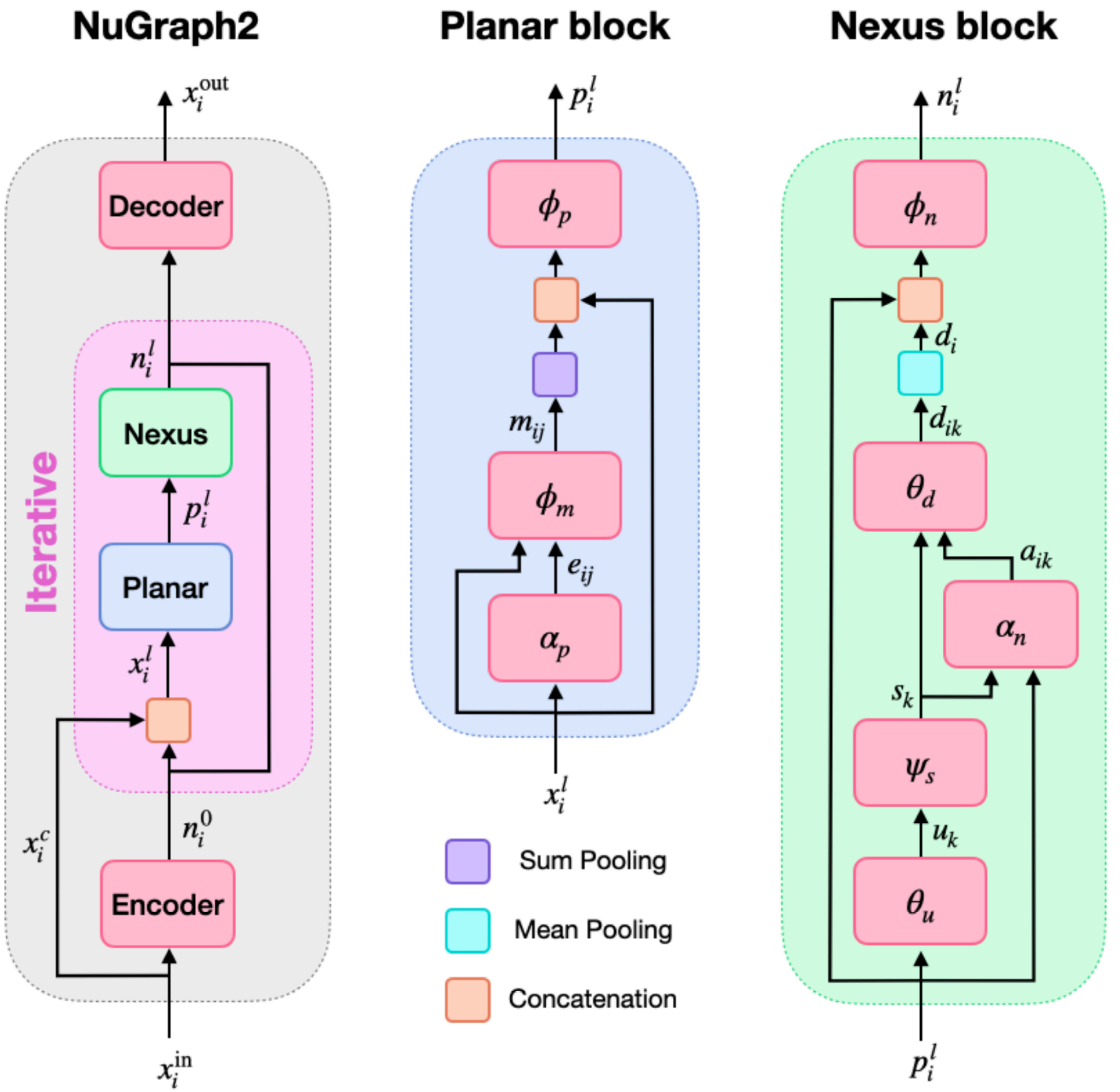


- Main inputs to the GNN are the Hits
  - hits are Gaussian fits to waveforms
  - features: wire, peak time, integral, RMS
  - currently using Hits associated to the neutrino interaction by Pandora
- Within each plane hits are connected in a graph using Delaunay triangulation
  - fully connected graph
  - both long and short distance edges
  - connect across unresponsive wire regions
- Hit associations to 3D SpacePoints create "nexus" connections across graphs in each plane
  - Currently defined by "Space Point Solver"
  - SPs are not connected among themselves
  - No input features for SPs

Fermilab

Network architecture

- NuGraph2's architecture is an iterative message-passing network.
- Each message-passing iteration consists of two phases:
  - Planar block: pass messages internally in each plane.
  - Nexus block: pass messages up to 3D nexus nodes to share context information.
- Messages are based on a categorical embedding:
  - Each semantic category is provided with a separate set of embedded features, which are convolved independently.
  - Context information is exchanged between different particle types via a categorical cross-attention mechanism.



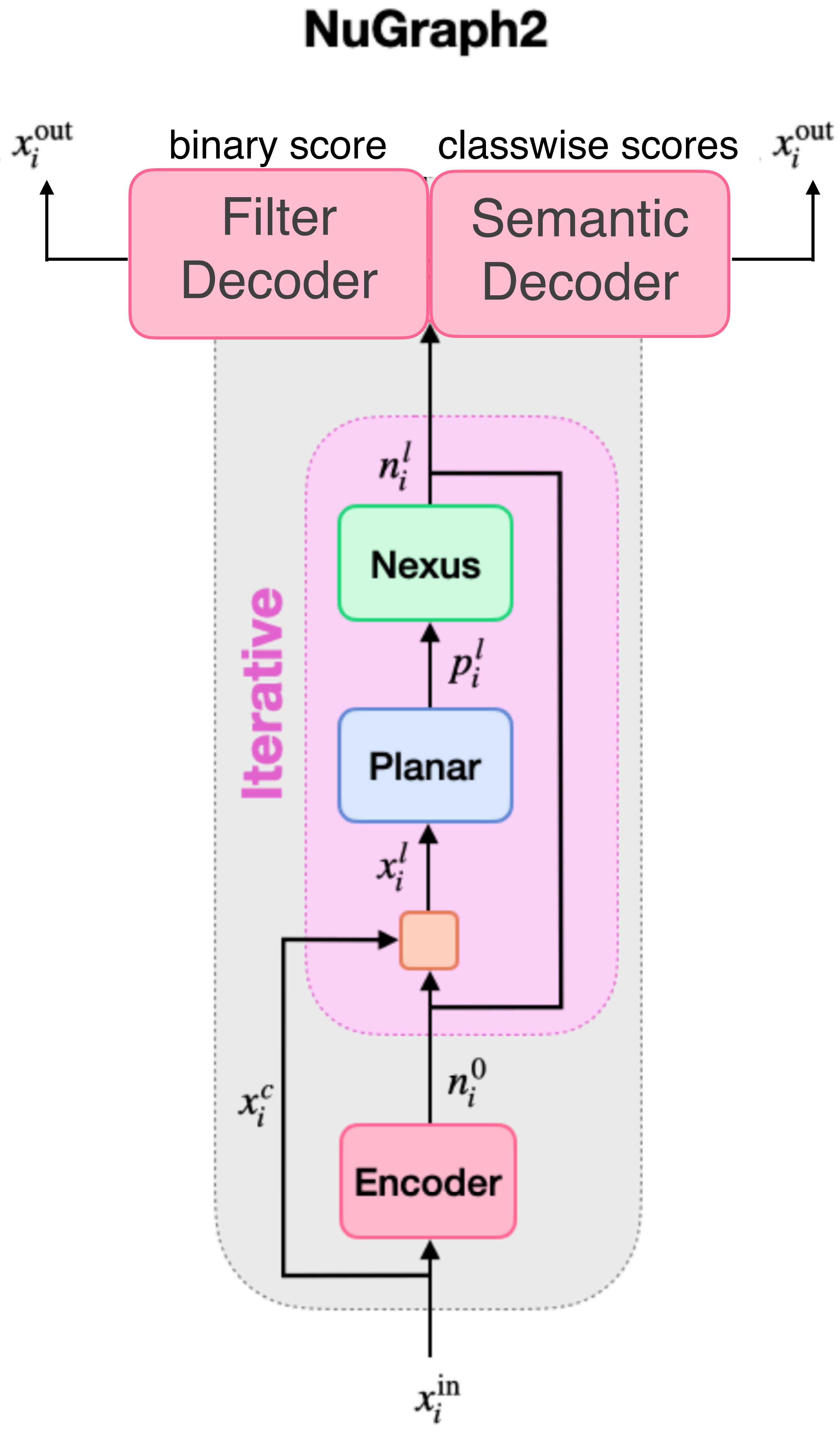
Fermilab

Decoders

- The last step at the end of the message passing network are the decoder steps

- Paper describes two node classifications decoders:
  - Semantic: classify each hit by particle type
  - Filter: separate hits from neutrino interaction from background
  - Output both class-wise scores from the semantic decoder and a binary score from the filter decoder
  - Same learned features are used as input to all decoders
  - Different loss functions weighted based on per-task variance (arXiv:1705.07115)

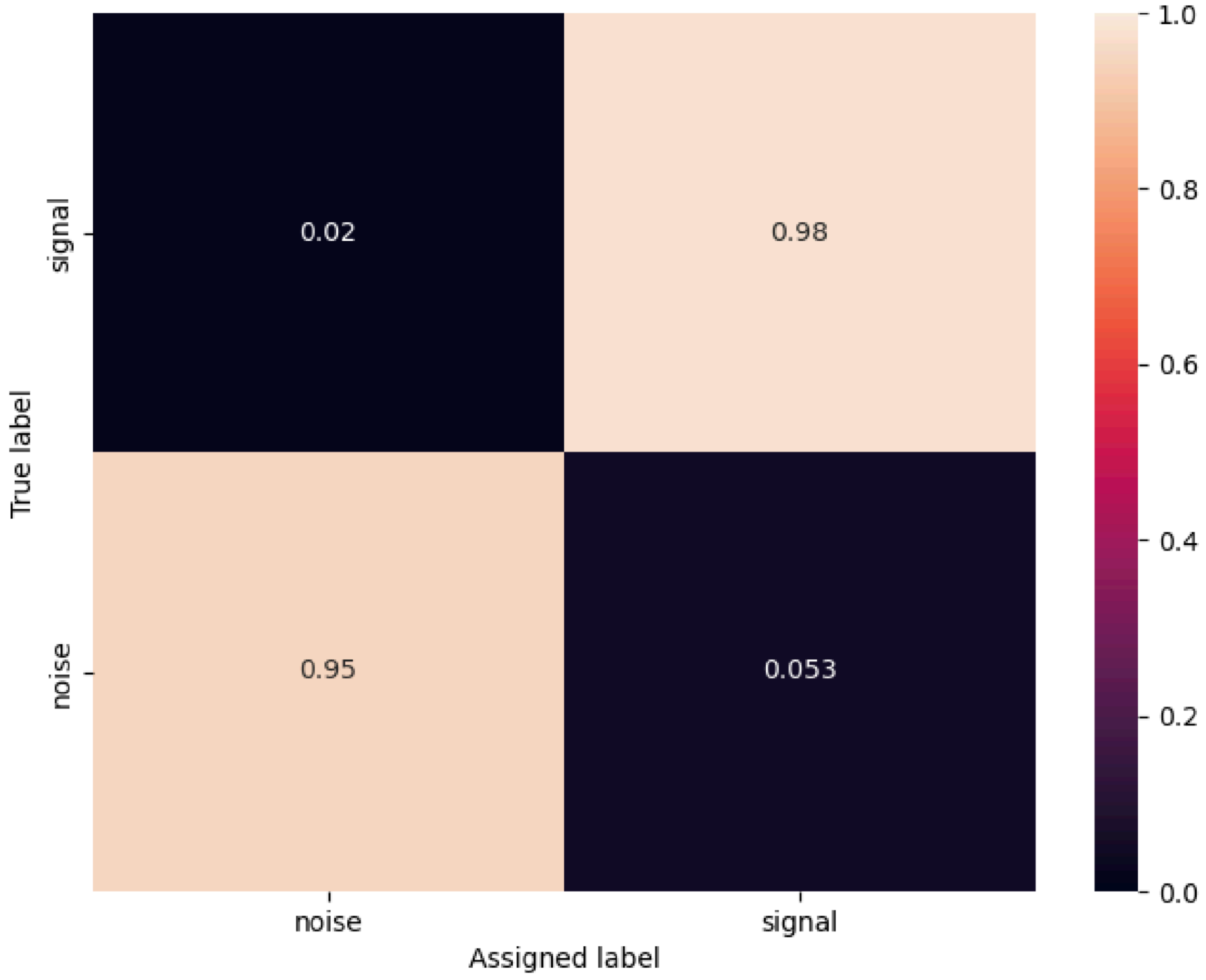
- Work in progress on more decoders: neutrino flavor, vertex regression, object condensation
  - see Adam's talk!



Performance on Simulation: Filter

- Decoder trained to separate neutrino-induced hits from background (noise or cosmic-induced hits)
  - Pandora slicing tends to prioritize completeness over purity

- Performance metrics:
  - recall and precision: ~0.98

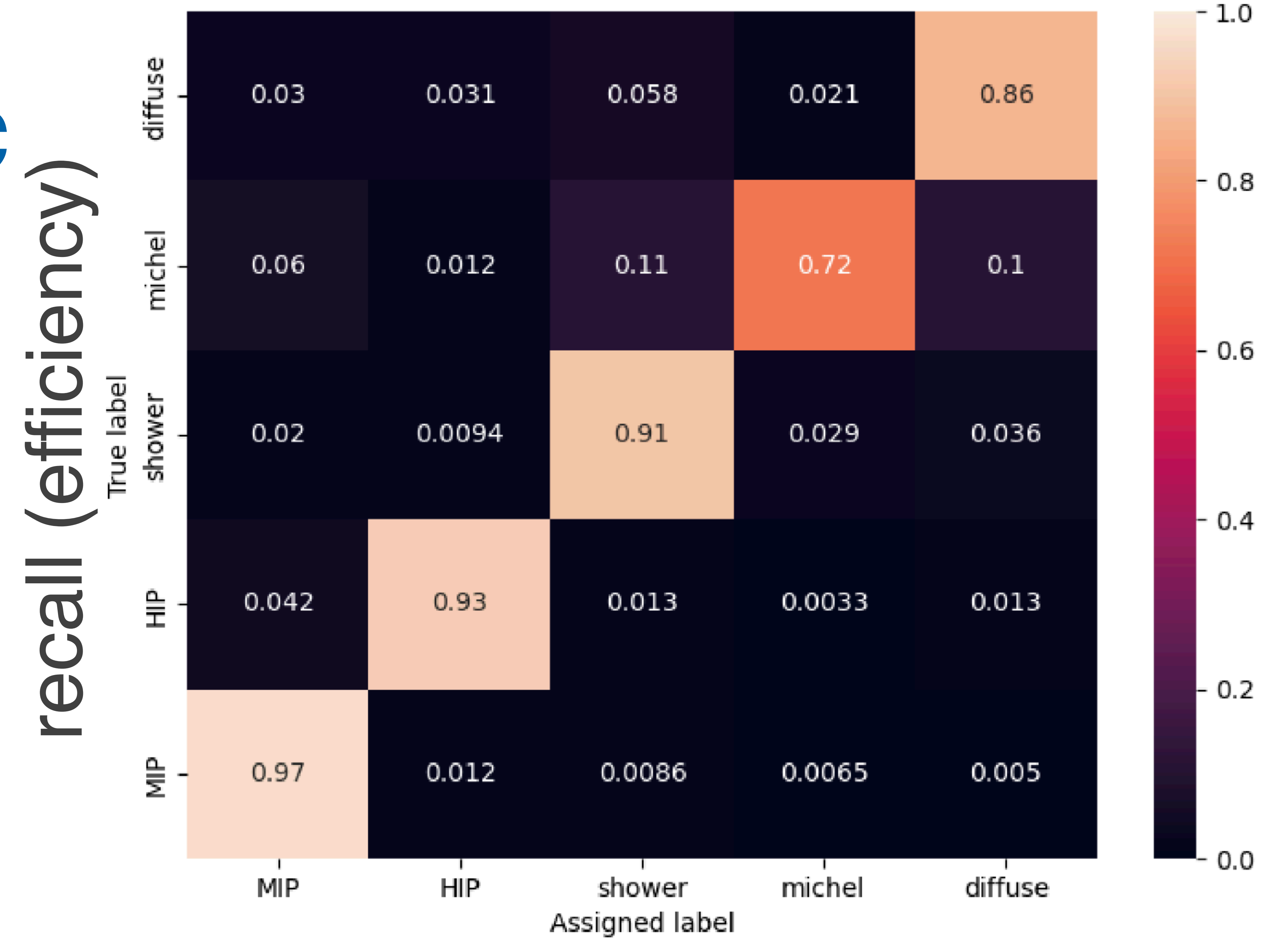


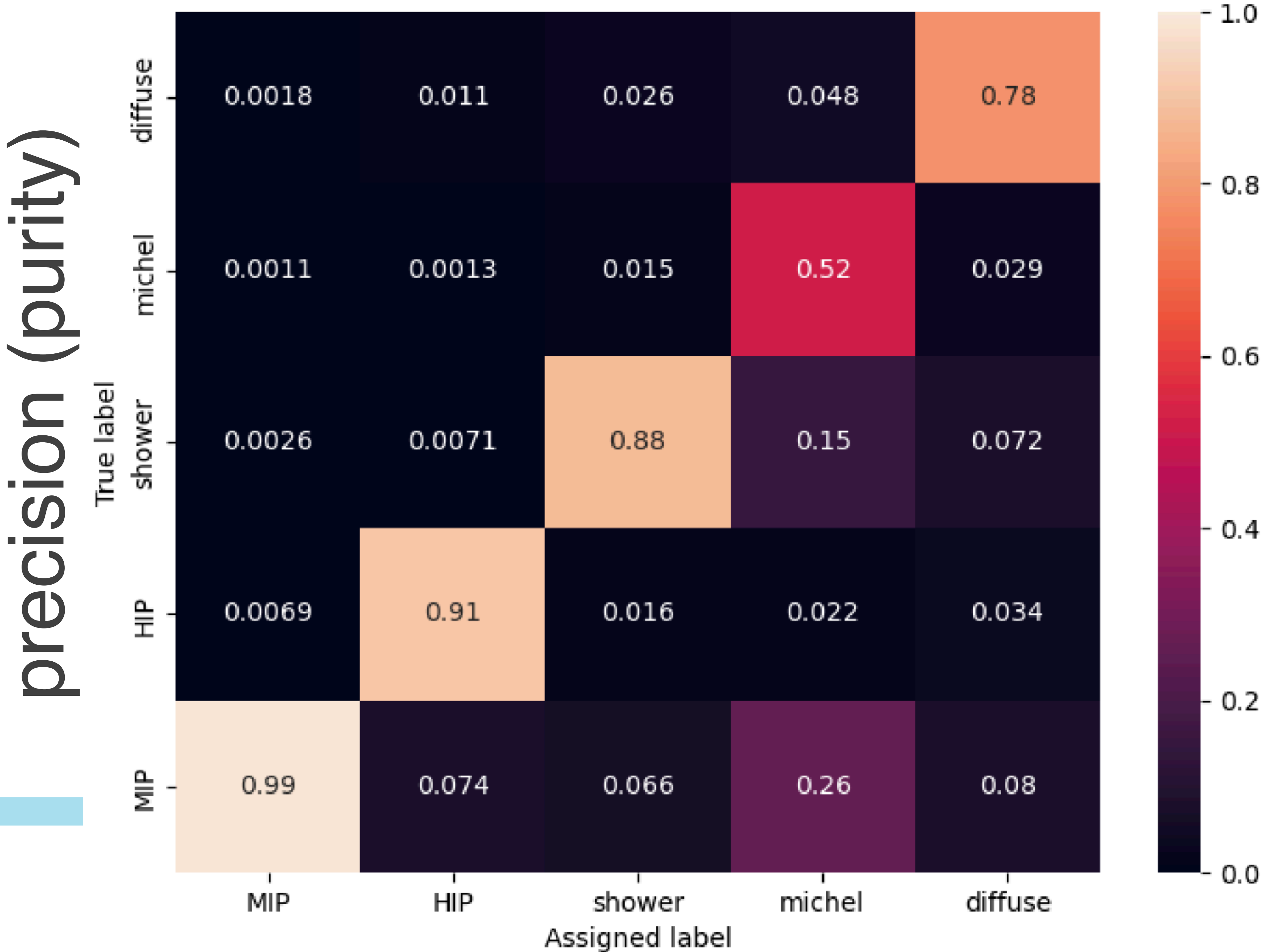
Performance on Simulation: Semantic

- Decoder trained to classify each neutrino-induced hit according to particle type

- Use five semantic categories:
  - MIP: Minimum ionizing particles (muons, charged pions)
  - HIP: Highly ionizing particles (protons)
  - EM showers (primary electrons, photons)
  - Michel electrons
  - Diffuse activity (Compton scatters, neutrons)

- Performance metrics:
  - recall and precision: ~0.95
  - consistency between planes around 98%
    - compared to ~70% without 3D nexus edges

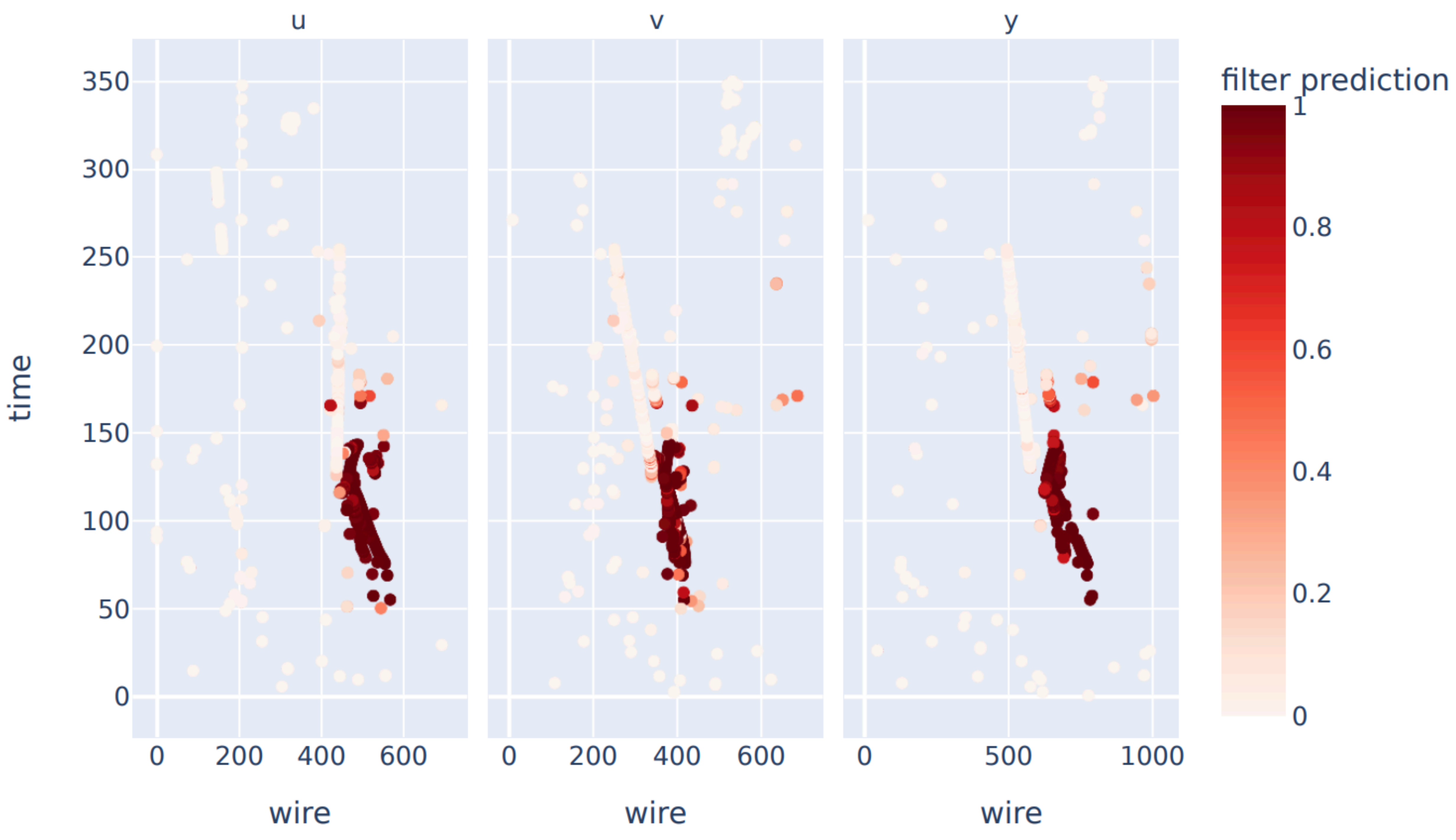




Performance on Simulation: Event Display

- Filter successfully rejects hits that are not from the neutrino interaction, including cosmic tracks that are close to it

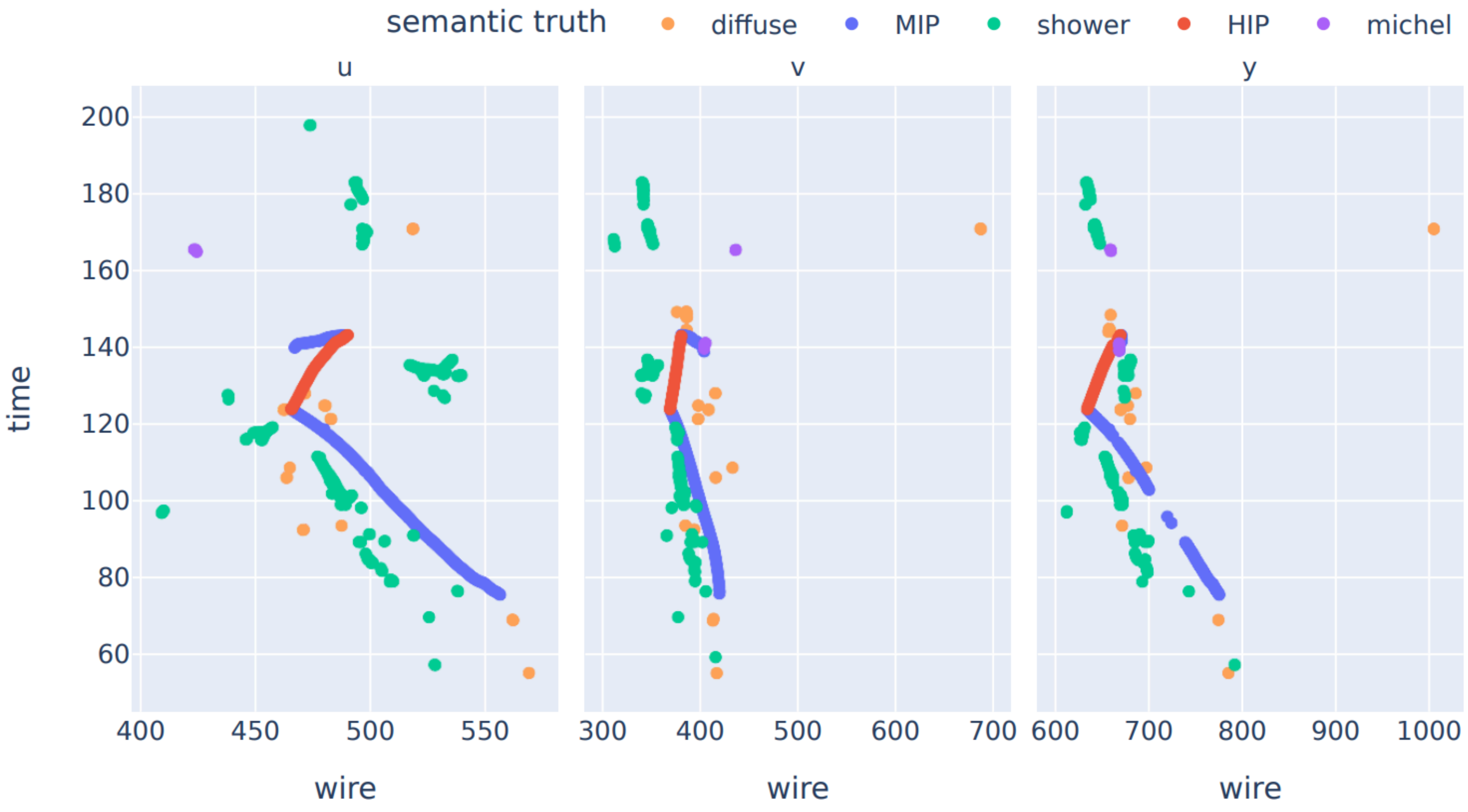
(a) Filter truth



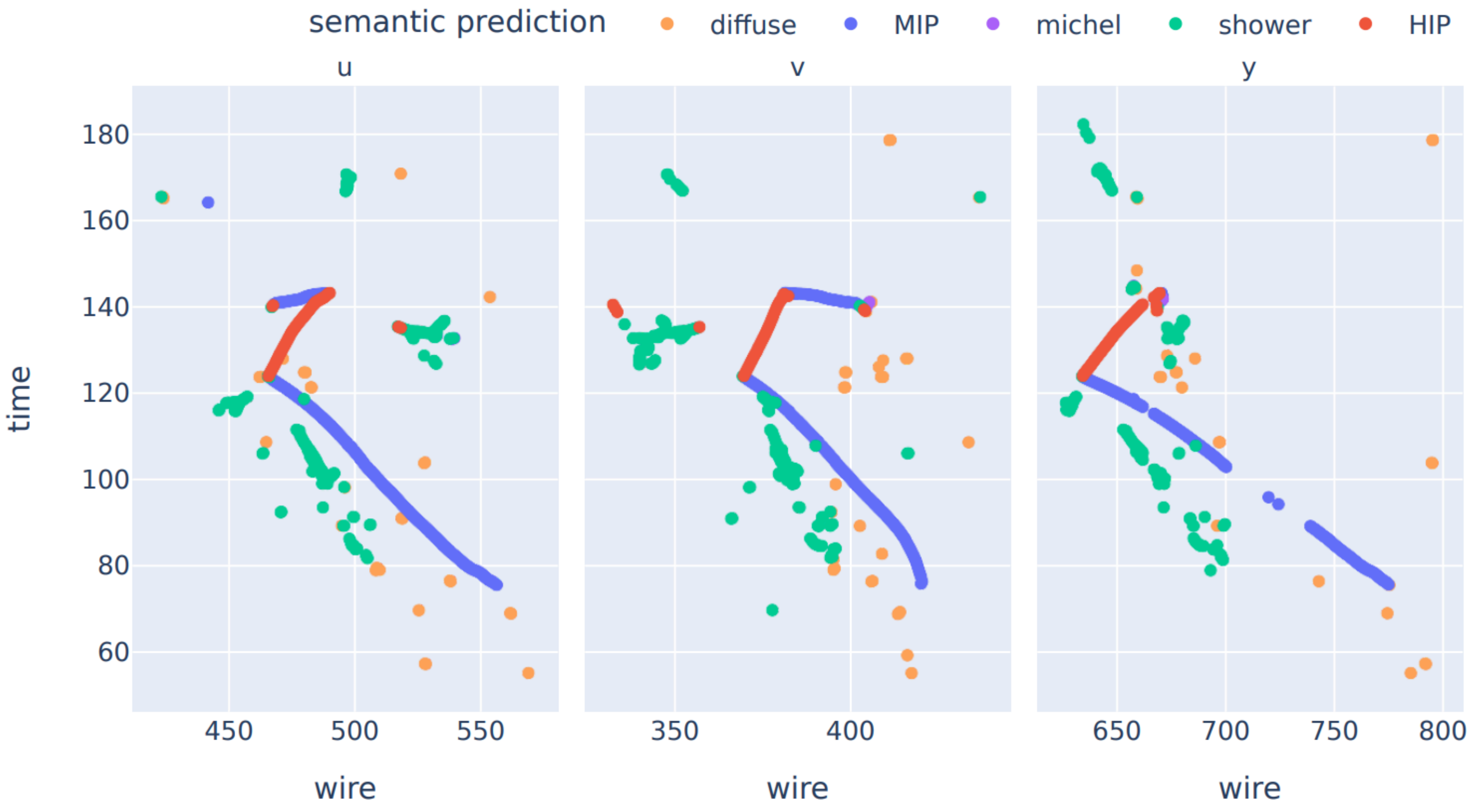
(b) Filter prediction

Performance on Simulation: Event Display

- Semantic classification correctly classifies hits classes both in events with a simple topology and also in higher multiplicity events.



(c) Semantic truth, filtered by truth



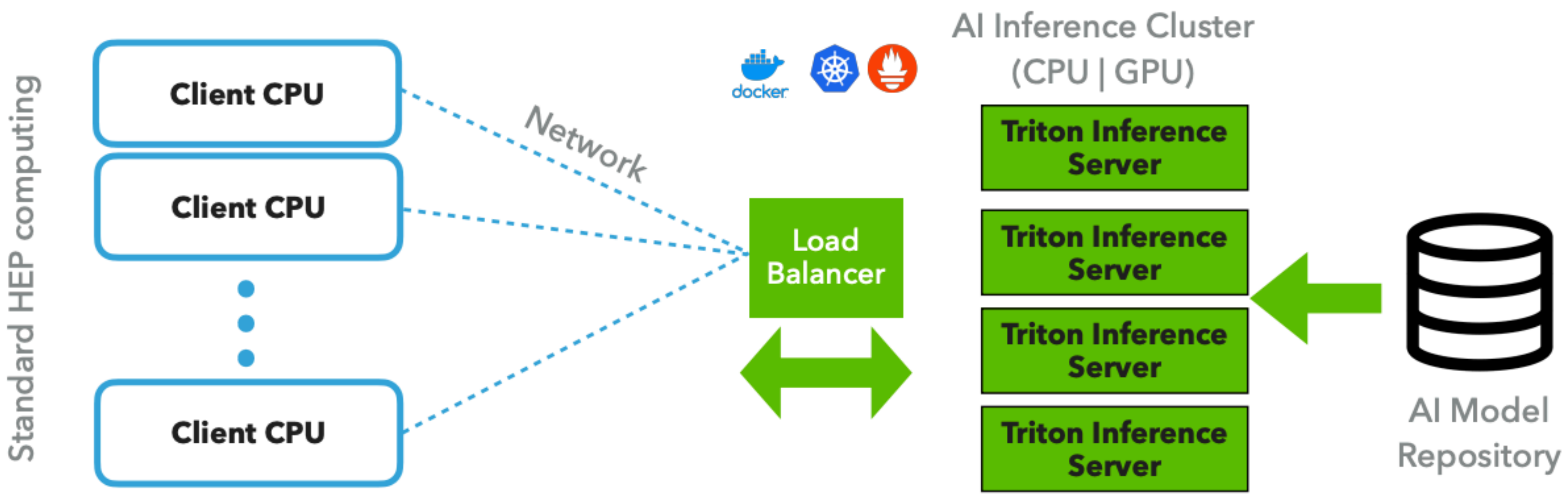
(d) Semantic prediction, filtered by prediction

Chapter 3: Deploying NuGraph2

Integration in LArSoft

- NuGraph2 is integrated in the software framework for LArTPC experiments, LArSoft
- Model compiled with JIT and run using the libtorch C++ library.
  - Integrated a package for Delaunay triangulation as well.
  - Inference results are stored in the Event record for usage in downstream reconstruction and analysis.
- Inference module takes 0.75 s per event event on CPU, including graph construction
- Enables running in production workflows for LArTPC experiments!
- Currently exploring more flexible integration methods based on NVIDIA Triton inference server (NuSonic: arXiv:2009.04509)

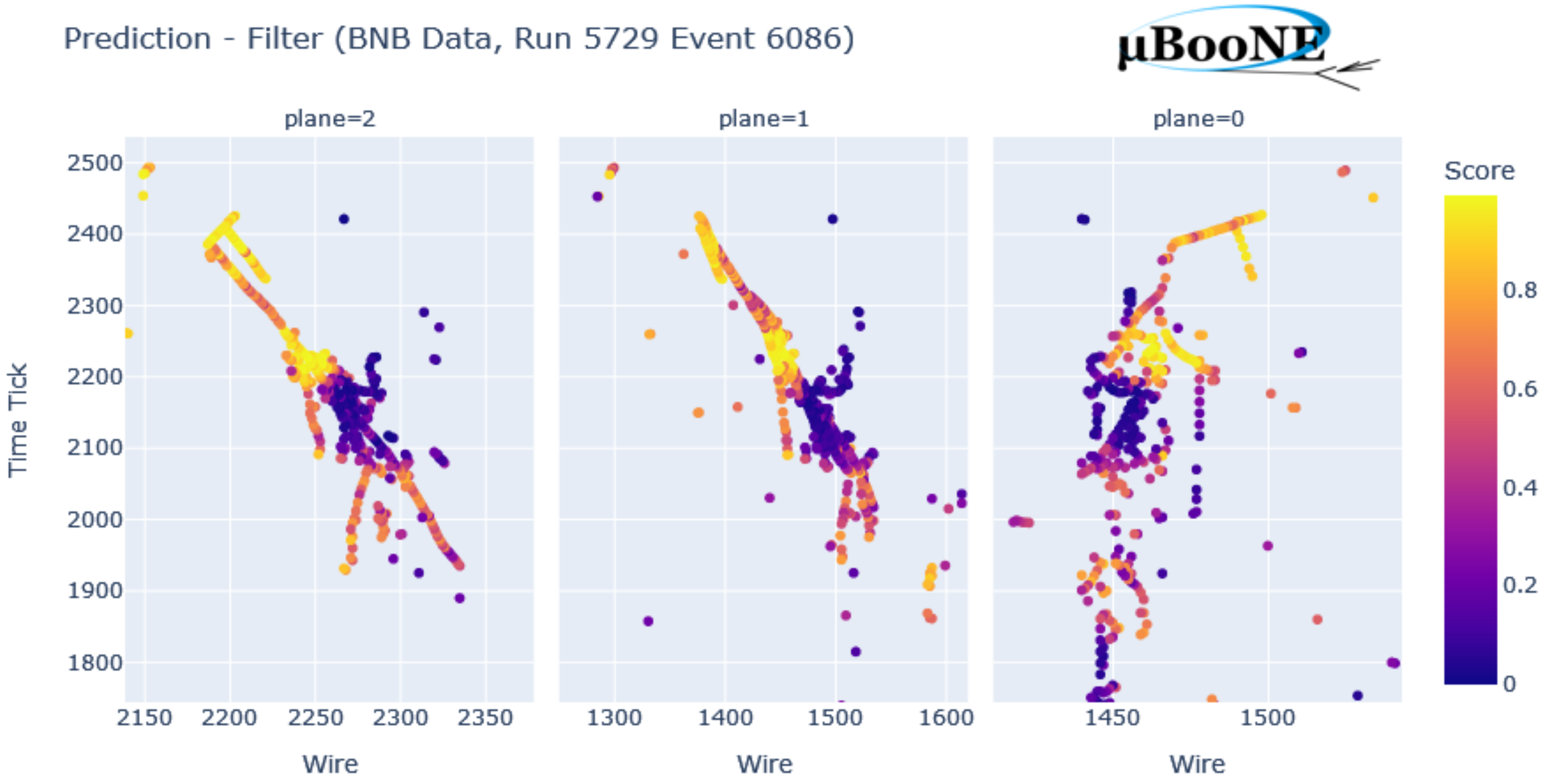
TimeTracker printout (sec)	Min	Avg	Max	Median	RMS	nEvts
Full event	0.0450458	3.36097	87.7468	0.237533	12.1151	74
source:RootInput(read)	0.000725606	0.00255304	0.019421	0.00131291	0.00392539	74
reco:nuslhits:NuSliceHitsProducer	0.0411265	0.116099	0.55599	0.0900547	0.0817036	74
reco:sps:SpacePointSolver	0.000110578	2.48479	85.3879	0.000217748	11.6239	74
reco:NuGraph:NuGraphInference	4.7356e-05	0.74844	5.22709	8.83935e-05	1.14997	74
[art]:TriggerResults:TriggerResultInserter	1.4952e-05	2.38511e-05	6.7179e-05	2.1032e-05	9.54137e-06	74
end_path:rootOutput:RootOutput	2.915e-06	4.5257e-06	1.9485e-05	3.9445e-06	2.18303e-06	74
end_path:rootOutput:RootOutput(write)	0.000867838	0.008697	0.0783238	0.00176224	0.0132931	74

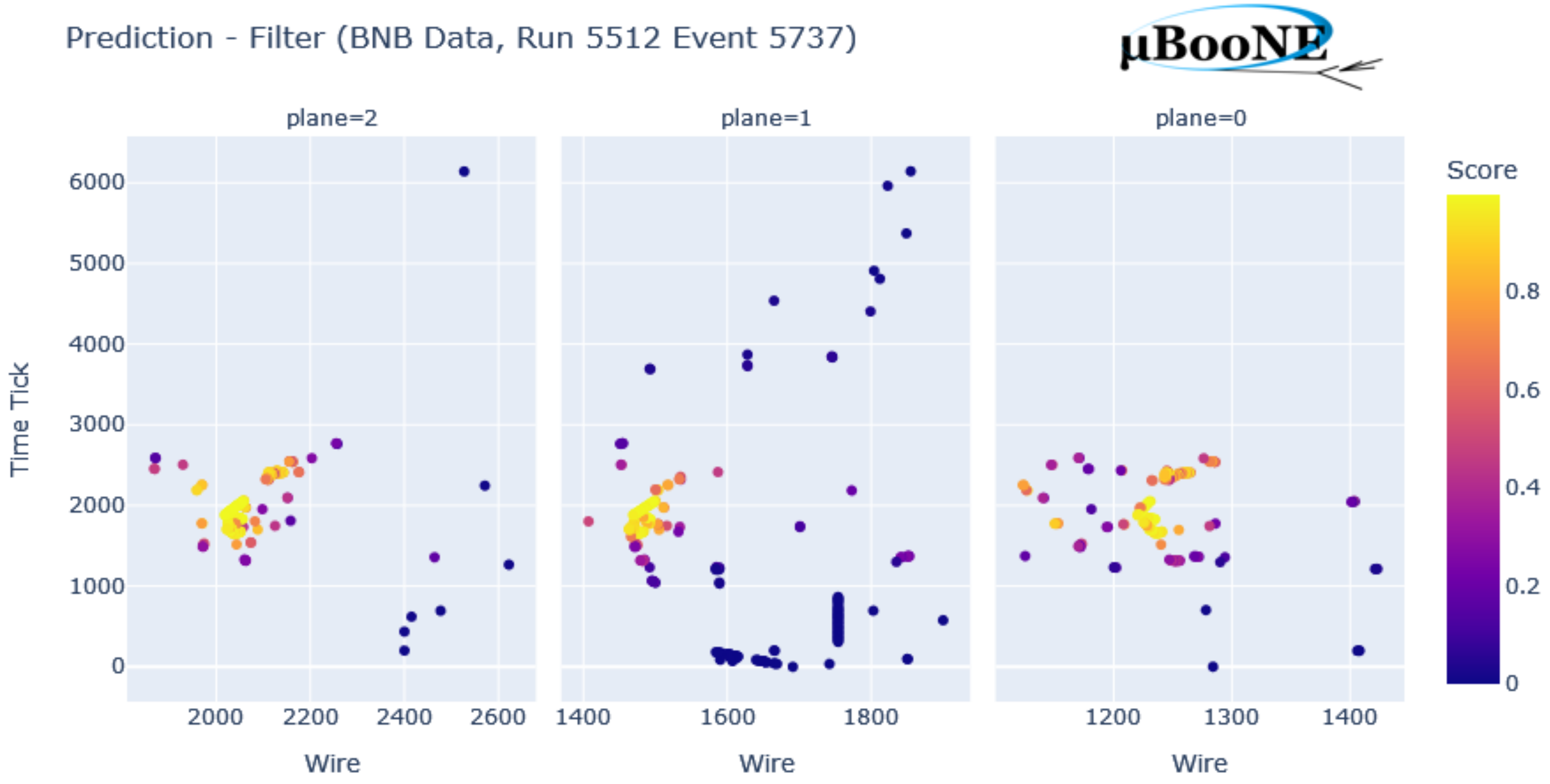


Fermilab

First Look at Performance in MicroBooNE Data

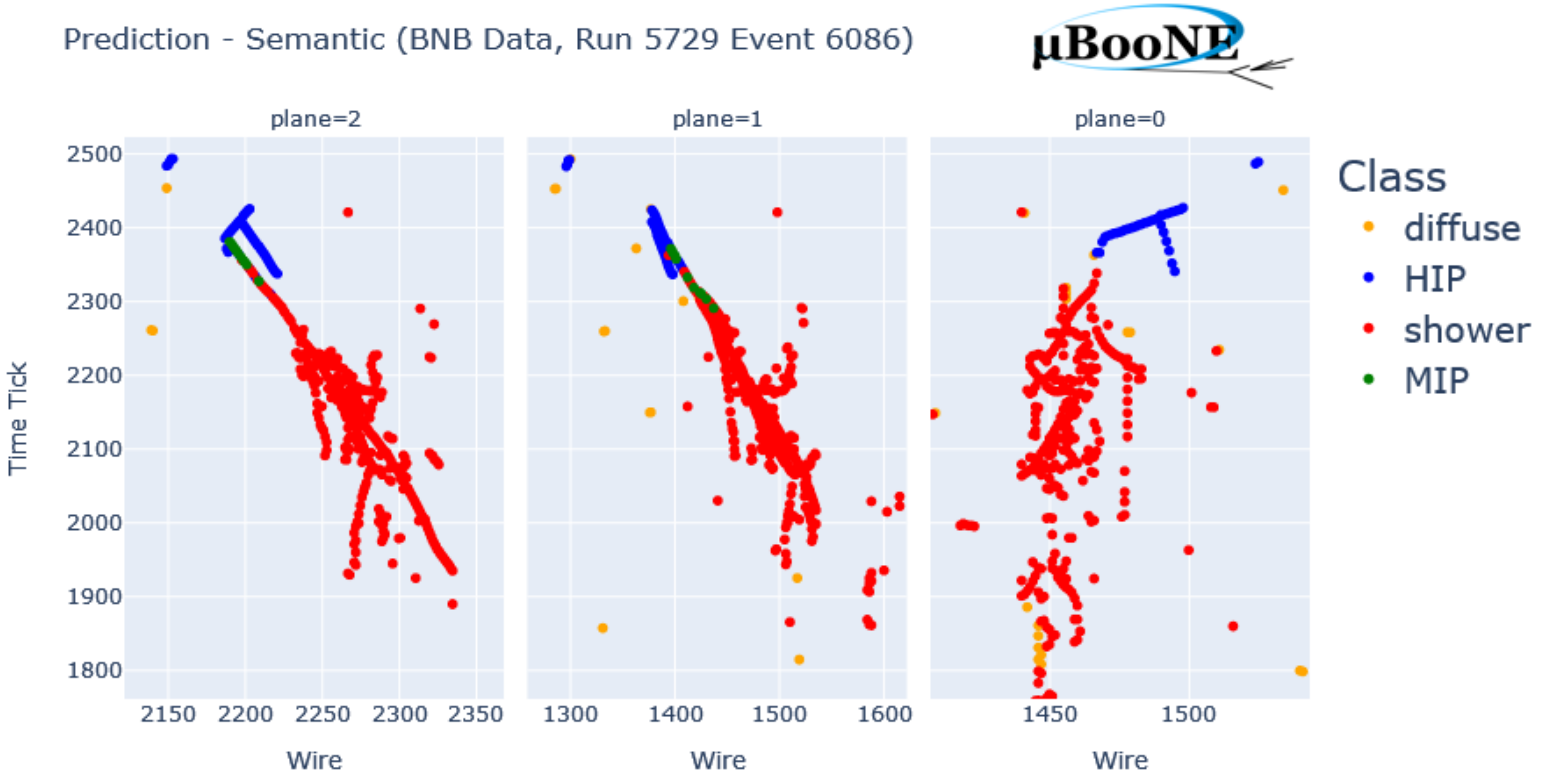
- First tests on MicroBooNE data events passing a loose ν_e CC preselection
- The filter decoder seems to overly reject shower hits from the neutrino interaction, so domain shifts between the training data set and the application data are being investigated.

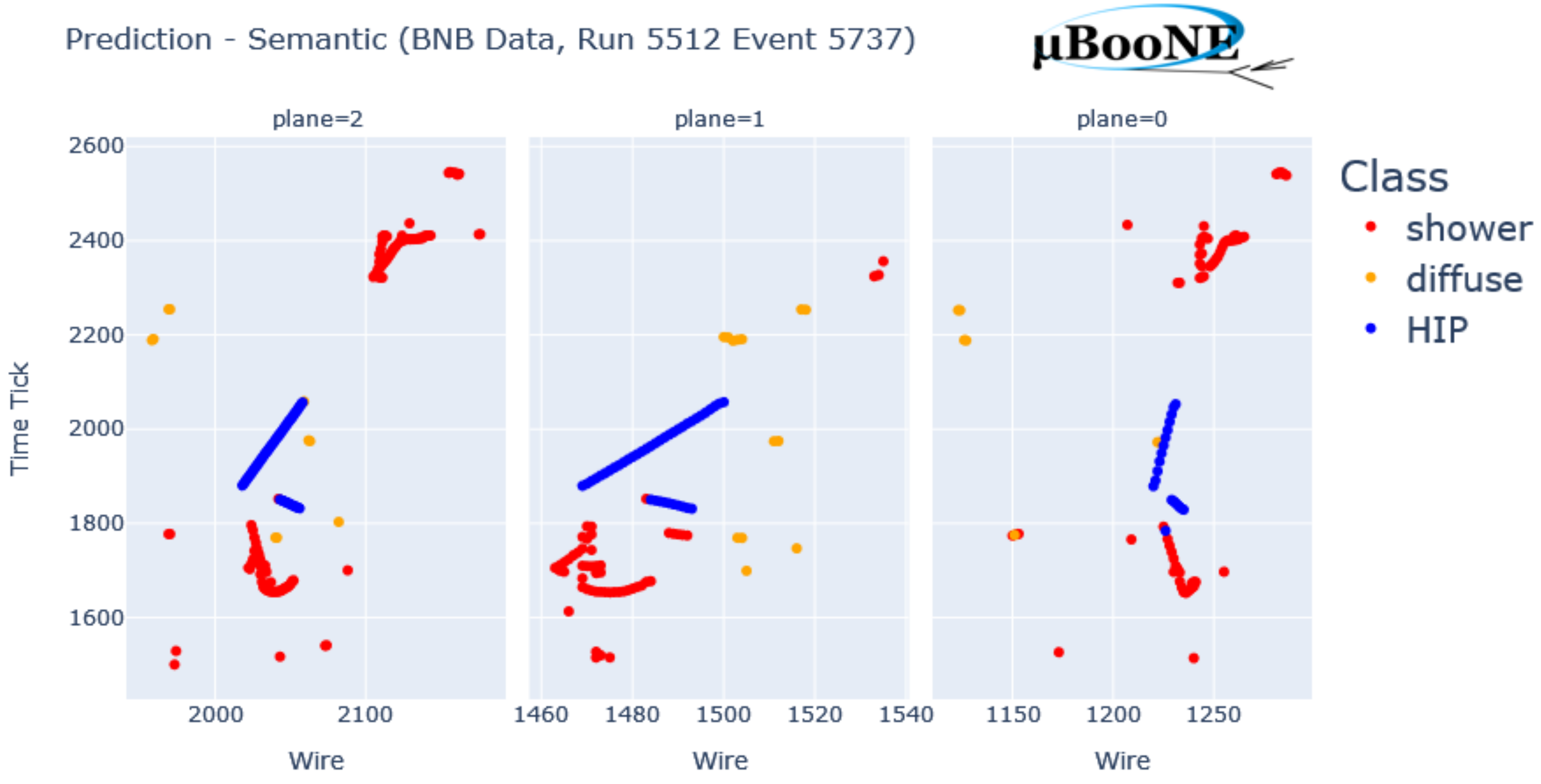




First Look at Performance in MicroBooNE Data

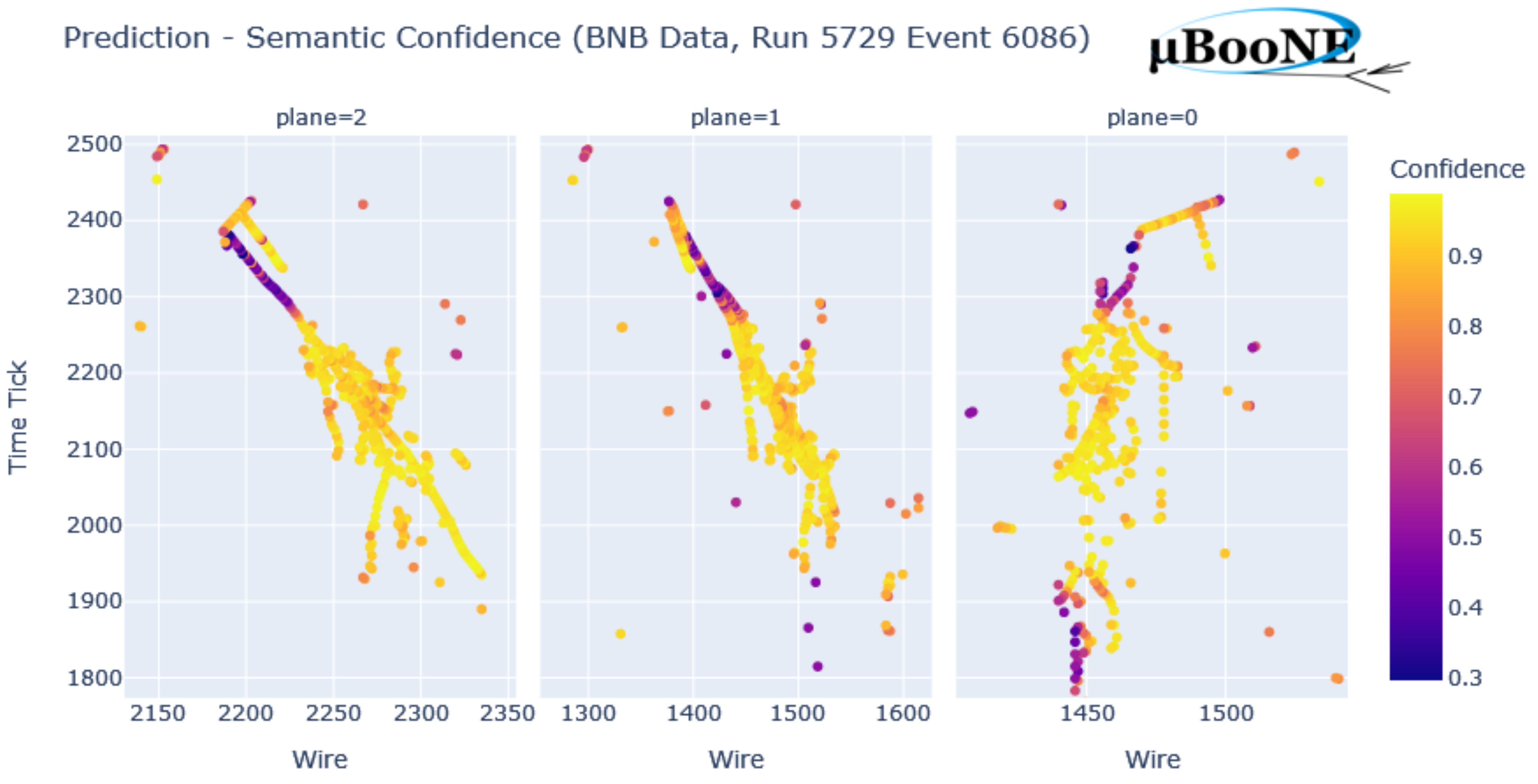
- First tests on MicroBooNE data events passing a loose ν_e CC preselection
- Encouraging performance for the semantic decoder: NuGraph2 correctly tags shower hits both from primary electrons (left plot) and photons (right) from π0 background.

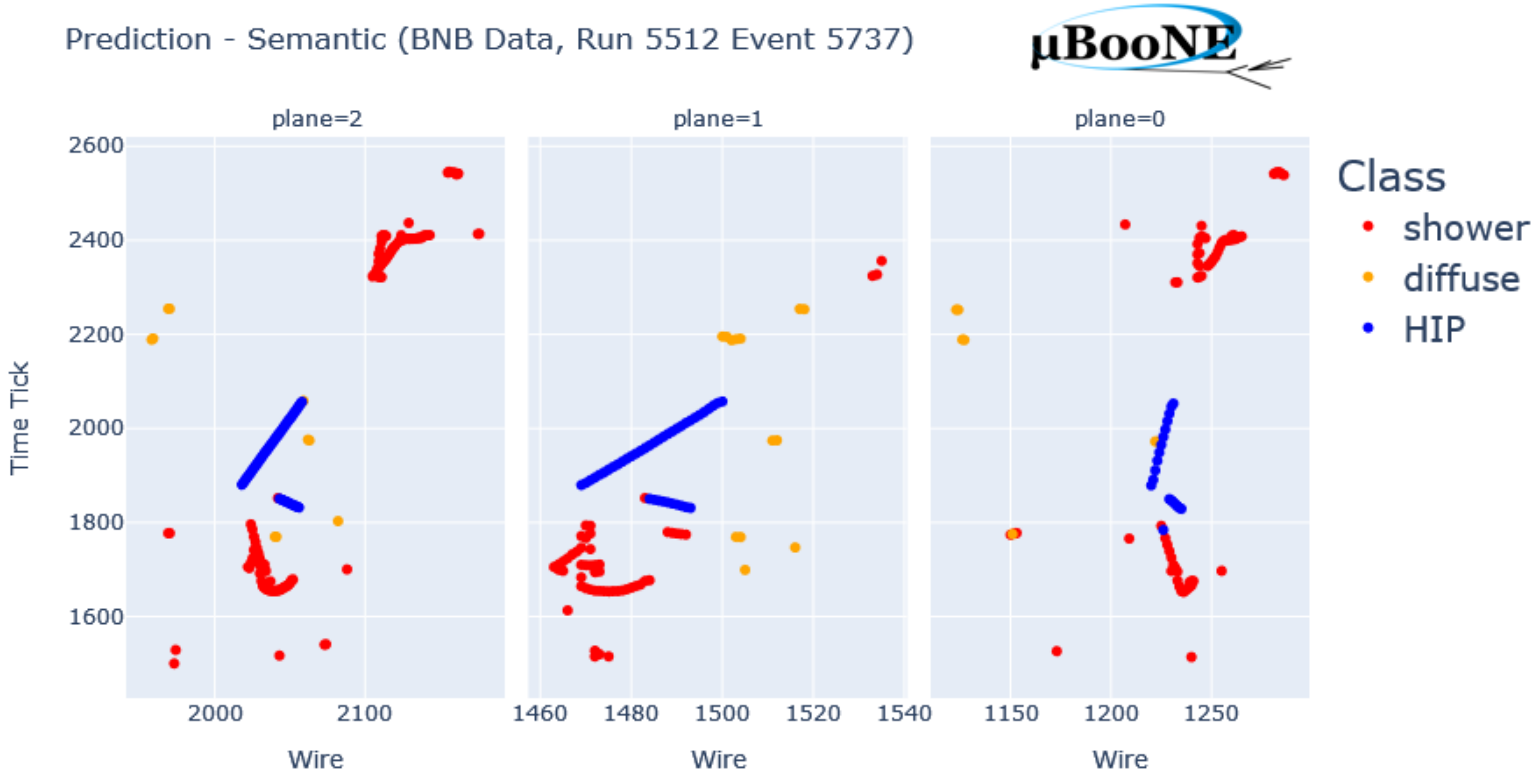




First Look at Performance in MicroBooNE Data

- First tests on MicroBooNE data events passing a loose ν_e CC preselection
- Encouraging performance for the semantic decoder: NuGraph2 correctly tags shower hits both from primary electrons (left plot) and photons (right) from π0 background.



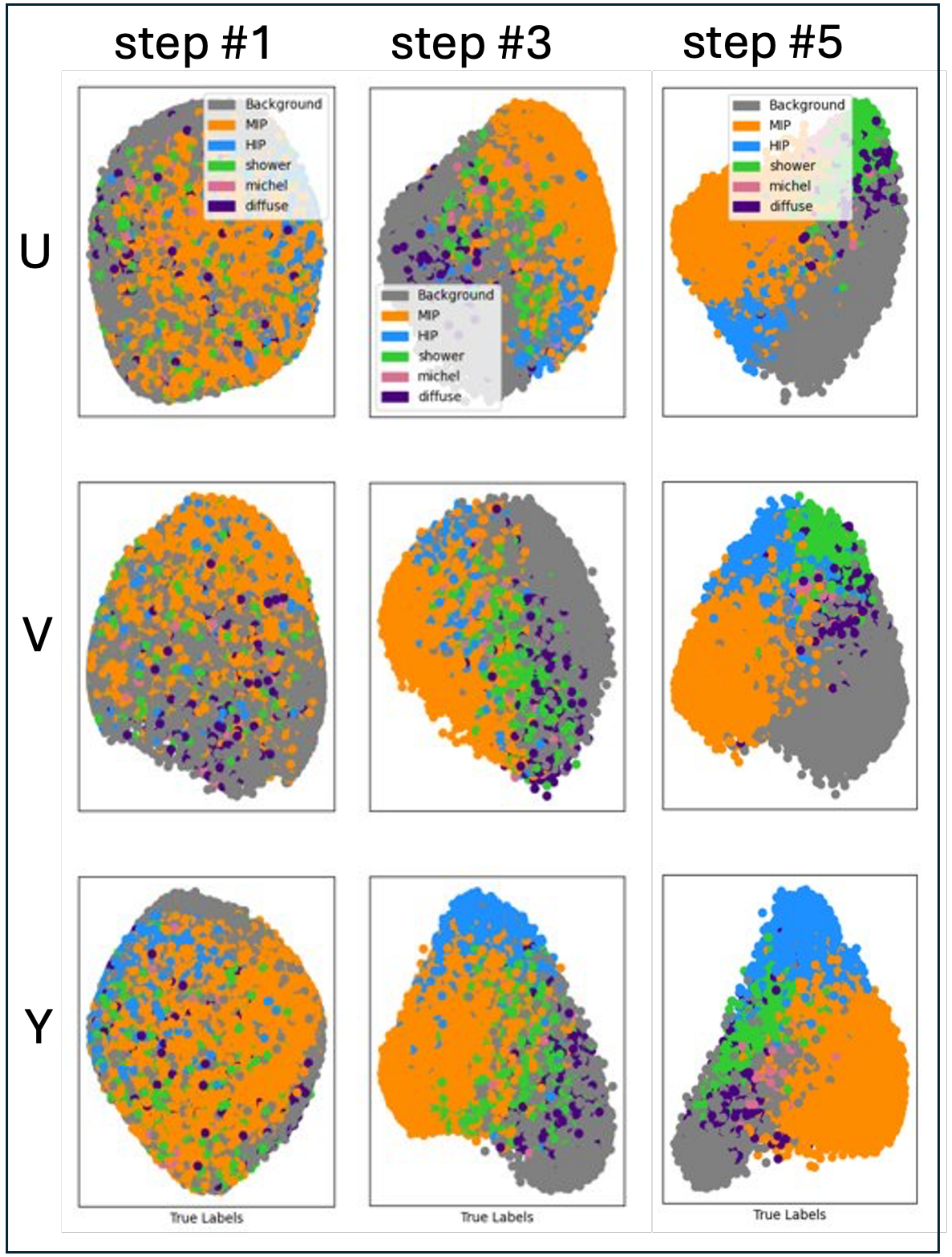


Fermilab

Chapter 4: Interacting with NuGraph2

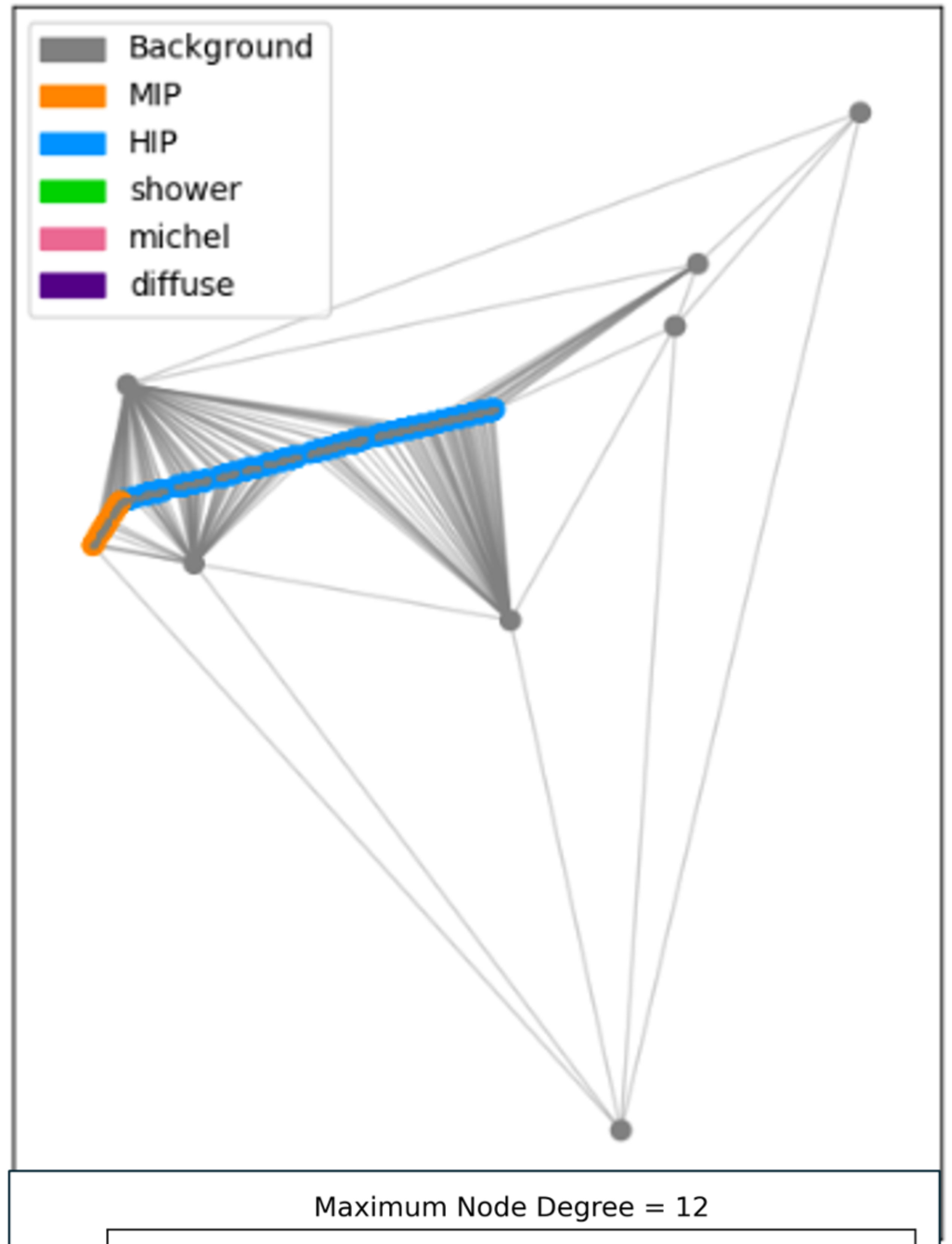
Network Explainability: Latent Space

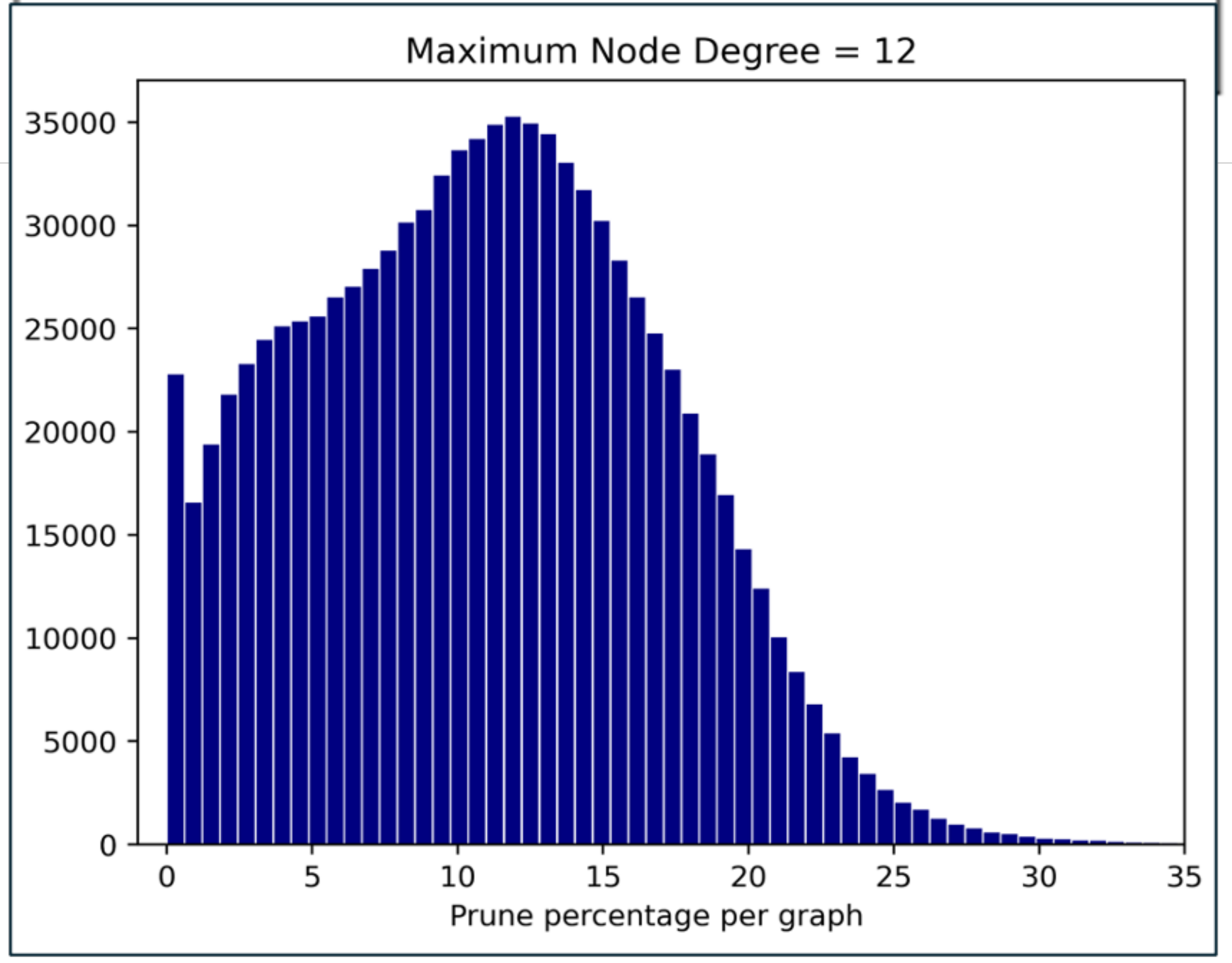
- Explainability: Goal is to “open the black box” to build confidence and drive developments.
  - Find that some “standard” tools for GNN interpretability (e.g. GNNExplainer) struggle with our network
- First study is the visualization of latent space:
  - Cluster latent node features (320D space) and project in 2D for visualization
  - Clear separation between different categories is achieved by the last (5th) network message-passing iteration



Network Explainability: “Hub Nodes”

- Understanding the role of “hub” nodes
  - Feature of Delaunay triangulation: detached nodes have large edge multiplicity and connect nodes within and across objects
  - These nodes introduce a large degree of redundancy, but also create a bridge for nodes at beginning/end of an object
- Performed a pruning test:
  - Aiming at understanding how many edges are essential and what are their properties
  - Find that 12 is the lowest upper bound in multiplicity without affecting performance, when pruning edges uniformly in terms of length
- Demonstrates that there is a degree of redundancy up to a few 10% of the edges, and that both short and long edges matter.
  - Can also lead to network speedups, both for training and inference

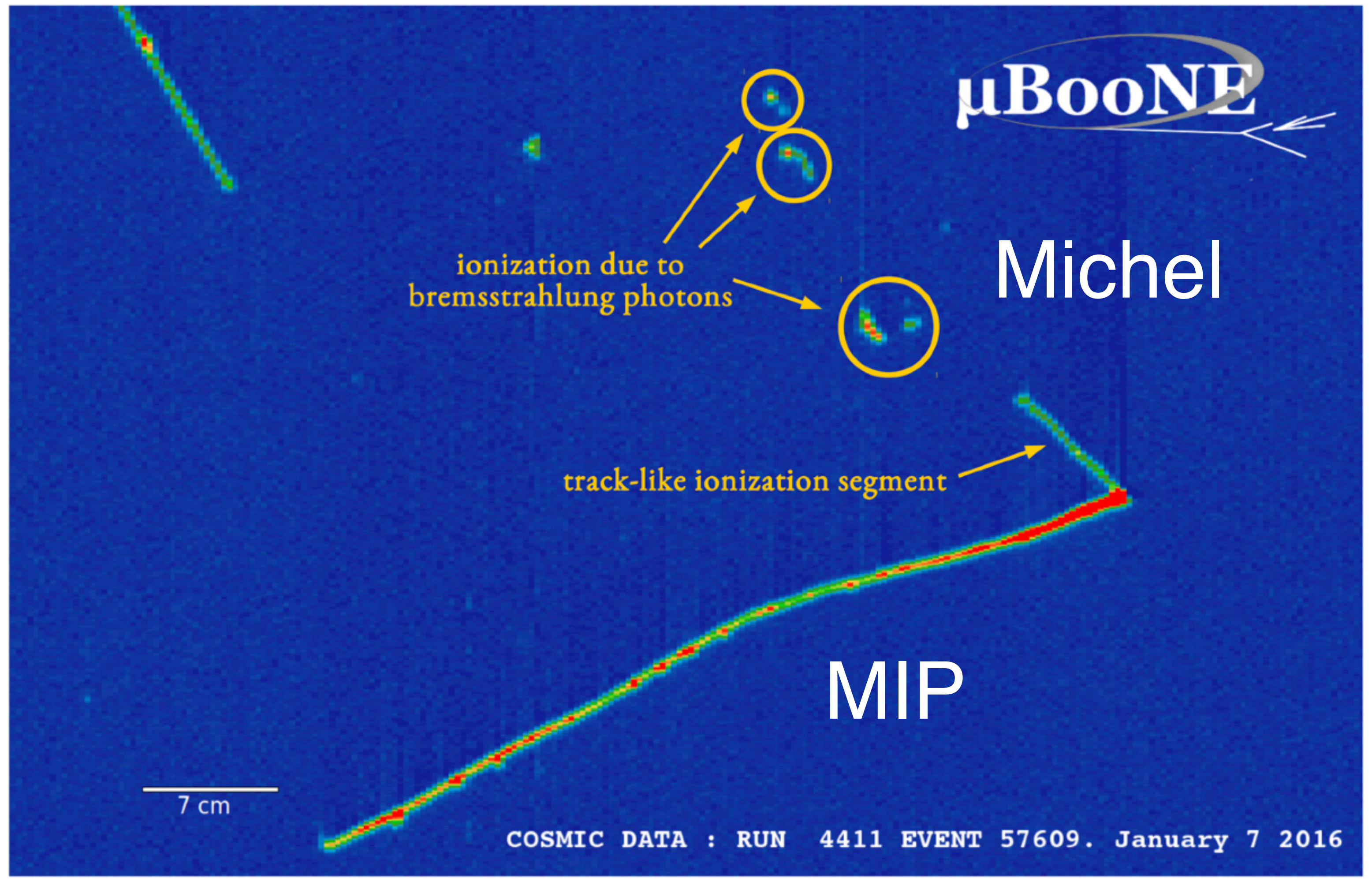




Fermilab

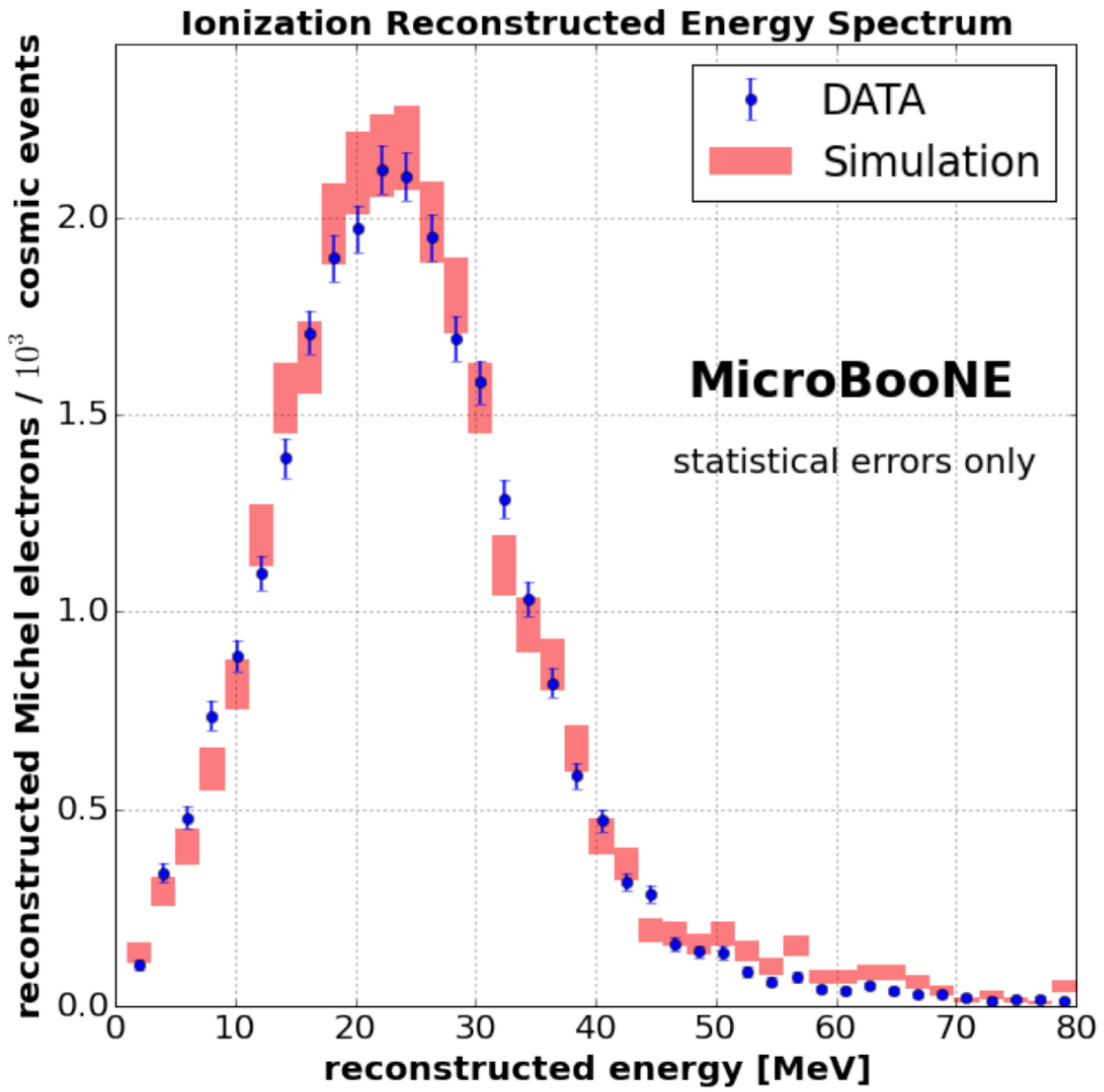
Injecting Physics Domain Knowledge: Michel electrons

- Michel electrons are the class least represented in our training dataset and the one with worst performance in terms of semantic classification
- Can we find ways to supplement the limited training data set and drive the network to learn better this category?



JINST 12, P09014 (2017)

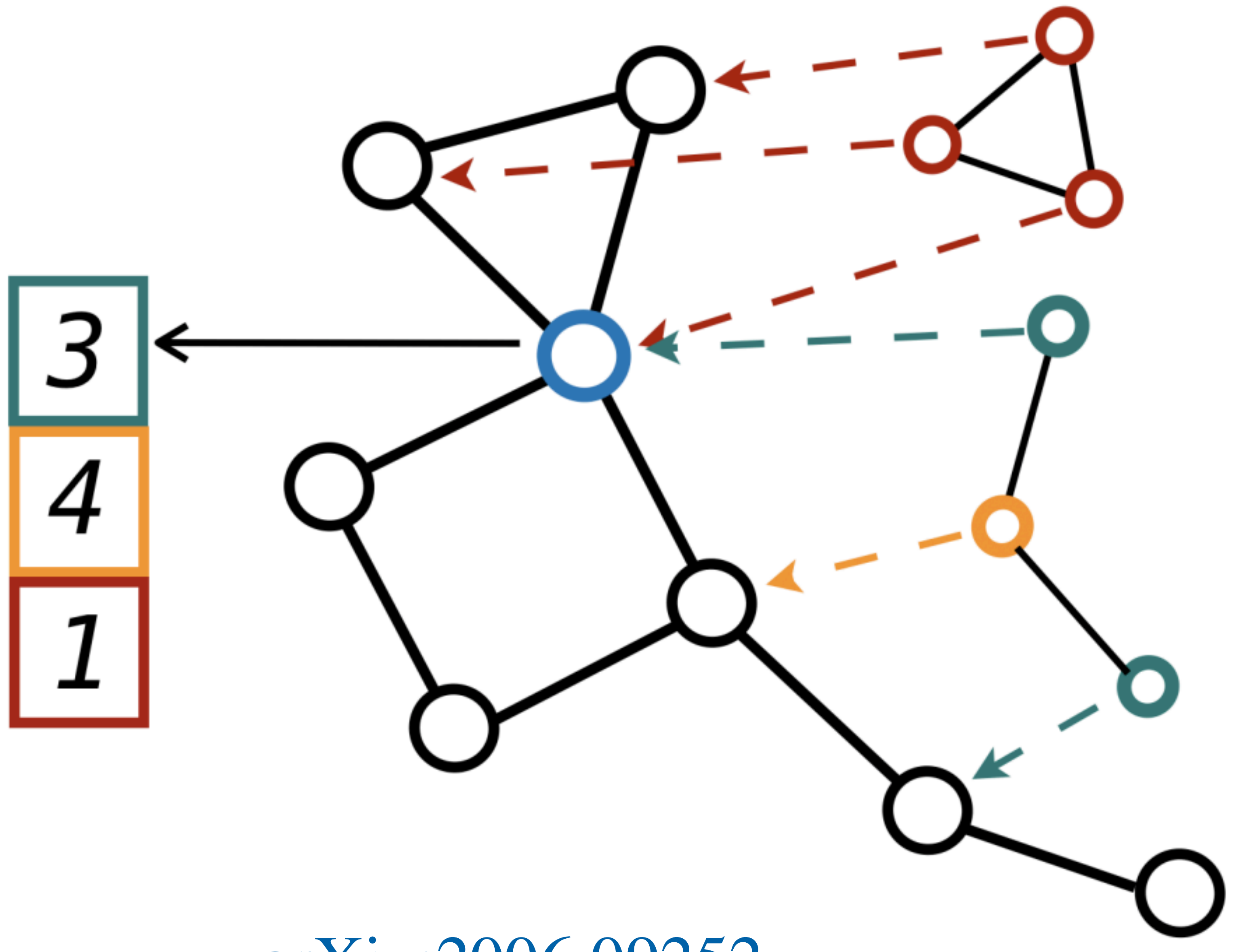
- A few ideas are being explored:
  - Michel electrons are the product of the decay at rest of a muon (MIP). Teach this correlation to the network by adding a decoder that predicts the fraction of hits in each class in the event
  - The Michel energy spectrum is a well defined function. Teach this property to the network by penalizing events where the sum of predicted Michel hit integrals is not compatible with the expected p.d.f.



- Results for these tests are coming soon!

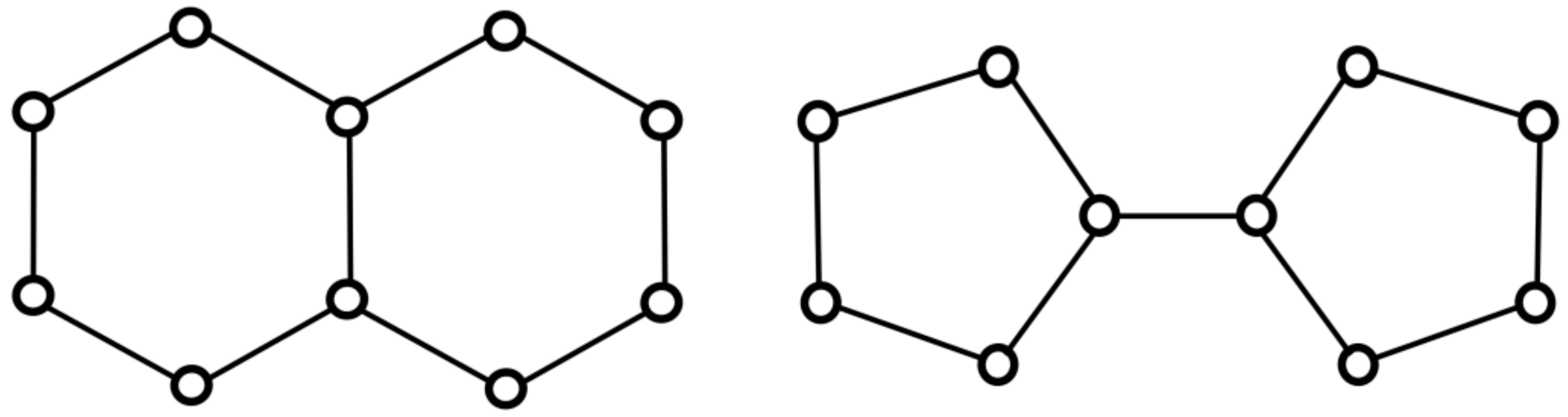
Injecting Physics Domain Knowledge: Augmented Features

- It turns out that GNNs are not aware of the structural role of nodes
  - They do not learn the graph structure
  - GNNs do not distinguish graphs that are isomorphic according to the Wesfeiler-Lehman test



arXiv:2006.09252

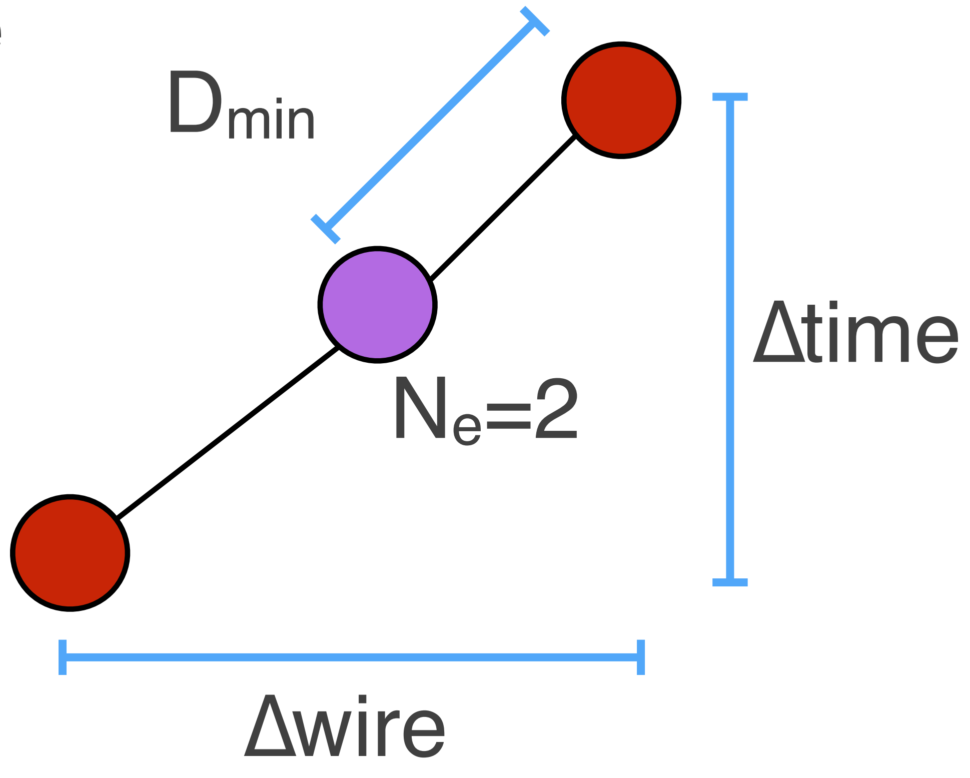
- Adding the graph structural information (e.g. triangles, circles) may help with classification
  - This can be implemented by a structure-aware message passing which contains structural information about the nodes



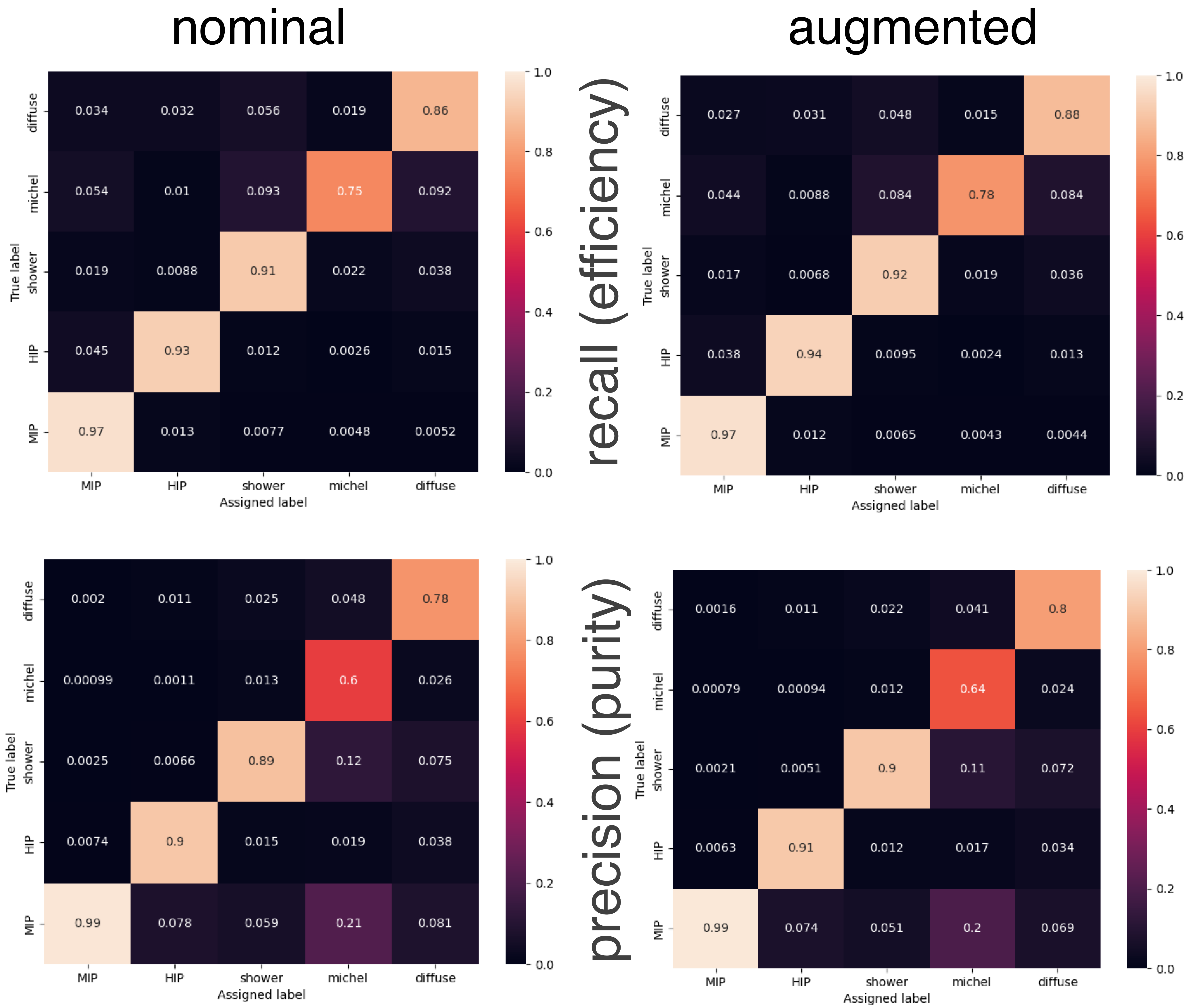
Indistinguishable molecules by the WL test and thus message passing NN

Injecting Physics Domain Knowledge: Augmented Features

- Add structural and non-local features to nodes:
  - Δtime, Δwire between 2 closest nodes
  - distance to closest node D_{min}
  - edge multiplicity N_e



- Improves the network performance across the board
  - ~5% (relative) improvement for the Michel category



Conclusions

- NuGraph2 is a GNN for reconstruction in LArTPC detectors
  - competitive performance for filter and semantic classification tasks
- Deployment in experiments' workflows in ongoing
  - integrated in LArSoft, promising results in data
- Work ongoing to interpret and further improve the network
  - stay tuned for Adam's talk for NuGraph3 developments!

Acknowledgements

U.S. DEPARTMENT OF ENERGY
Office of Science

DISCOVERY PARTNERS INSTITUTE
PART OF THE UNIVERSITY OF ILLINOIS SYSTEM

M. Voetberg, J. Kowalkowski (FNAL), H. Meidani, V. Grizzi (UIUC),
D. Caratelli, M. Fang (UCSB), A. Aurisano, V Hewes (UCincinnati),
D. Grzenda, K. Gumpula, X. Zhang (UChicago),
C. S. Lee and W. Liao (NorthwesternU)

This document was prepared by the Exa.TrkX/NuGraph Collaboration using the resources of the Fermi National Accelerator Laboratory (Fermilab), a U.S. Department of Energy, Office of Science, Office of High Energy Physics HEP User Facility. Fermilab is managed by Fermi Research Alliance, LLC (FRA), acting under Contract No. DE-AC02-07CH11359.

Fermilab