AI/ML at DUNE

Jianming Bian

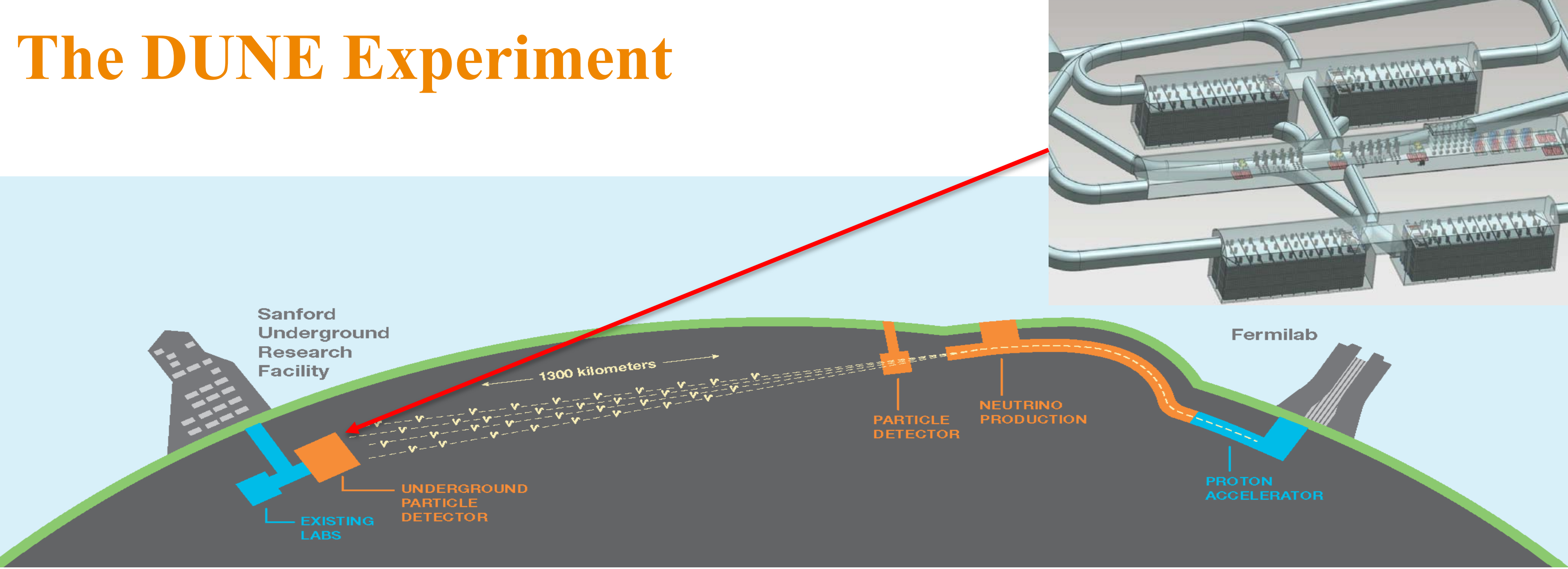
for the DUNE Collaboration

University of California, Irvine

10-29-2025

NPML2025, Tokyo

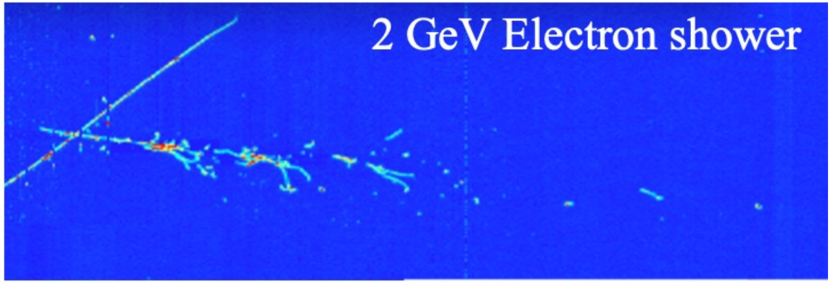
The DUNE Experiment



- New neutrino beam at Fermilab (1.2 MW, upgradeable to 2.4 MW), 1300 km baseline
- Utilize 17 kton Liquid Argon Time Projection Chamber (LArTPC) Far Detector modules at SURF: Horizontal-drift (HD), Vertical-drift (VD)
- Multiple technologies for the Near Detector (ND)
- Far Detector Prototypes at CERN: ProtoDUNEs
- Neutrino oscillation and diverse non-beam neutrino and BSM projects

Introduction

- AI/ML has been implemented in major steps in DUNE reconstruction, enhancing the use of its high-resolution, large-statistics data.
  - Signal Processing and Hit Finding
  - Clustering
  - Kinematics
  - Event-Classification and Particle Identification
  - Non-beam Neutrino reconstruction
- AI/ML has been widely applied to various other aspects of DUNE
  - Simulation
  - Trigger & DAQ
  - Beam design and monitoring
  - Computing workflow
  - Documentation search pipeline
  - Quality Assurance and Quality Control (QA/QC)
- DUNE has applied state-of-the-art machine learning models
- DUNE has utilized and enhanced the AI infrastructure at national laboratories and collaborating institutions



Event display at ProtoDUNE-SP

Reconstruction: Signal Processing and Hit Finding

CNN (U-Net) for regions of interest (ROI) detection in DUNE signal processing algorithm

WireCell

H.W. Yu, M. Bishai, W.Q. Gu, M.F. Lin, X. Qian, Y.H. Ren, A. Scarpelli, B. Viren, H.Y. Wei, H.Z. Yu 2021 JINST 16 P01036

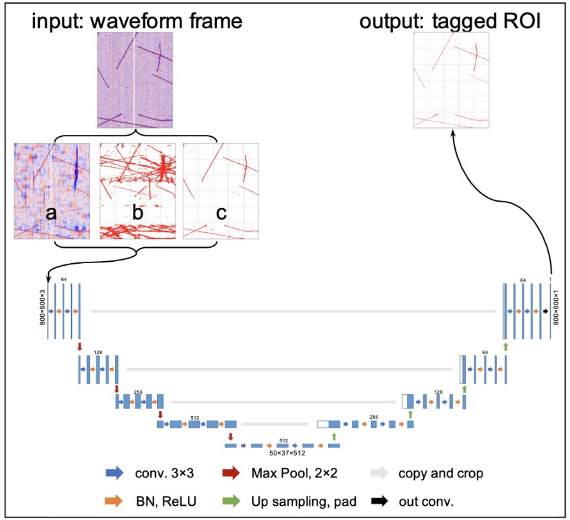


Figure 4: U-Net architecture used in the ROI detection for the DNN LArTPC signal processing. Intermediate 2D images are generated from original images containing the raw waveform. Several of these intermediate images are stacked to a multi-channel 2D image serving as the U-Net input. Numbers of intermediate images can vary. In this example, three are used: a) deconvolved signals from a loose low-frequency filter, b) MP2, and c) MP3. Output of the U-Net is a single channel 2D image labeling each pixel as signal (i.e., inside ROI) or not.

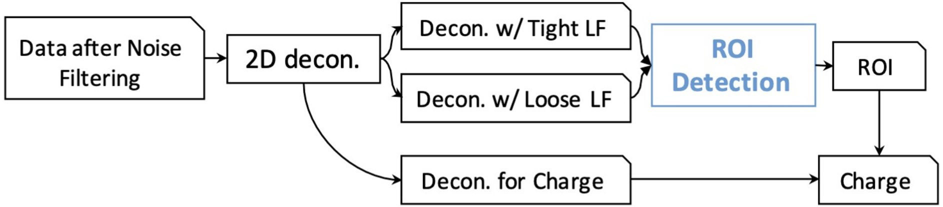
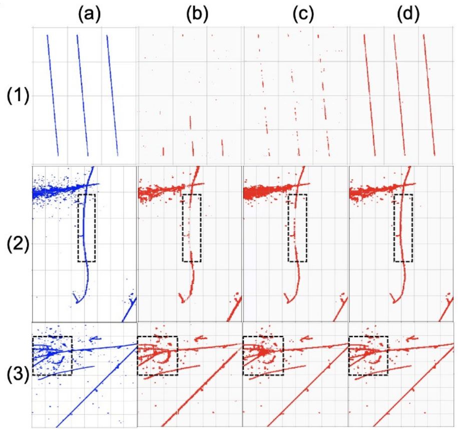


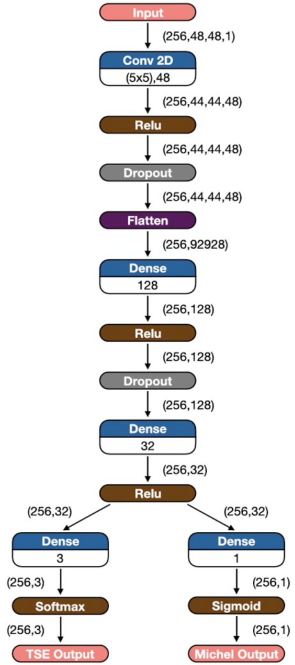
Figure 3: Flowchart of the LArTPC signal processing algorithm in ref. [10]. "LF" stands for low-frequency software filter, while "decon." denotes deconvolution.

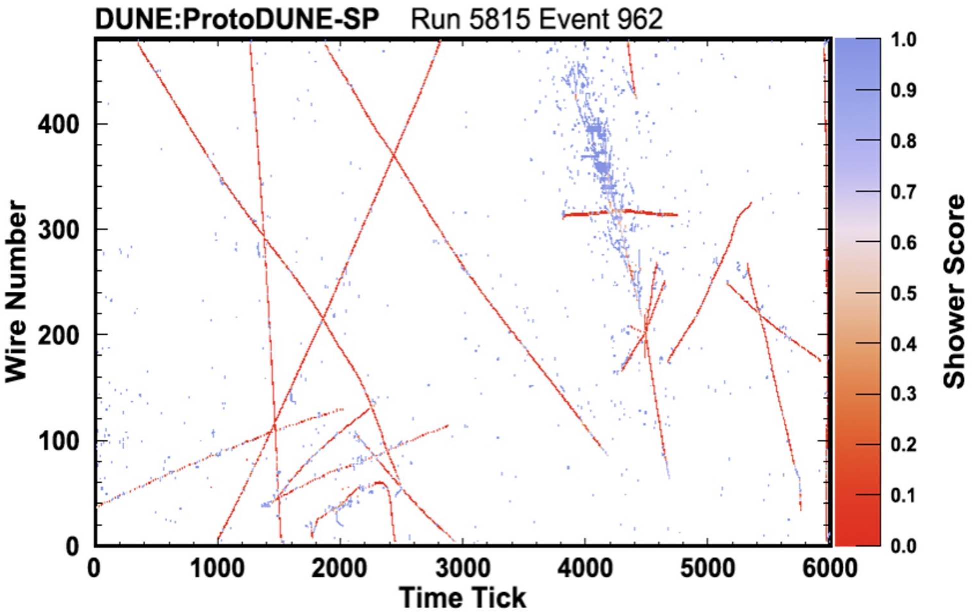


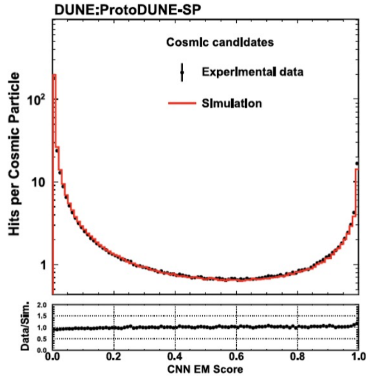
a) Truth
b) Standard signal processing
c) U-Net one plane
d) U-Net using all planes

Reconstruction: Signal Processing and Hit Finding

- CNN developed in ProtoDUNE-SP to separate track-like and shower-like energy deposits; tested on data, Eur.Phys.J.C 82 (2022) 10, 903
- Data-driven training under development with ProtoDUNE data for robust scientific machine learning
- Public Access: https://github.com/DUNE/dune-cvn/tree/master







Reconstruction: Clustering

- A number of software frameworks use deep learning for clustering
- SPINE: end-to-end ML-based reconstruction for the DUNE Near Detector (NDLAr), uses various techniques such as sparse CNNs and GNNs

F.Drielsma, K.Terao, L.Domin and D.H.Koh, NeurIPS proceedings, arXiv:2102.01033

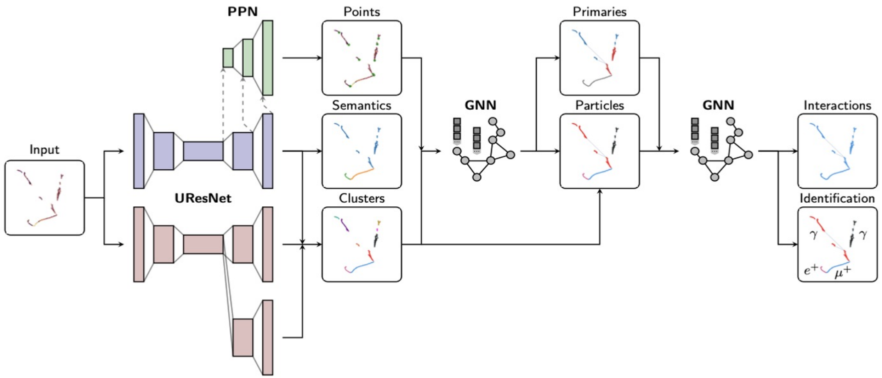
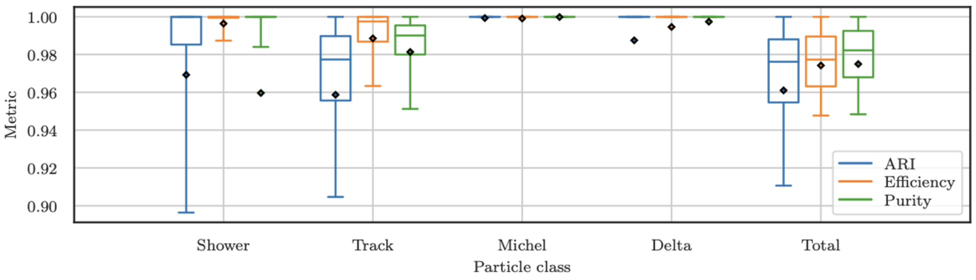


Figure 1: Schematic architecture of the end-to-end, ML-based reconstruction chain for LArTPCs.

SPINE particle clustering performance



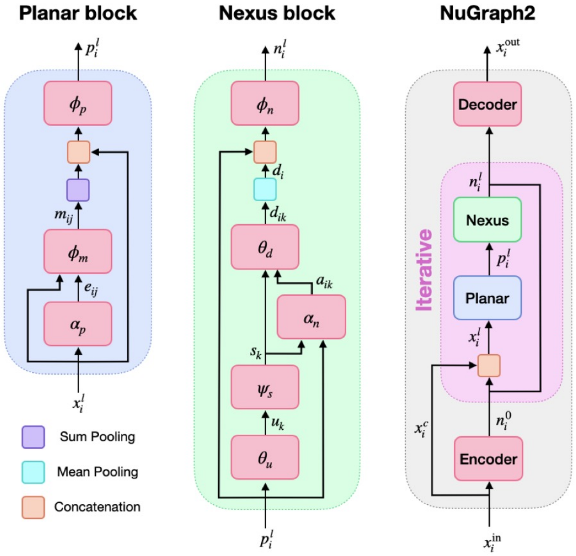
New development of SPINE see Francois Drielsma's talk

GNN in 2x2 ND prototype, see Jessie Micallef's talk

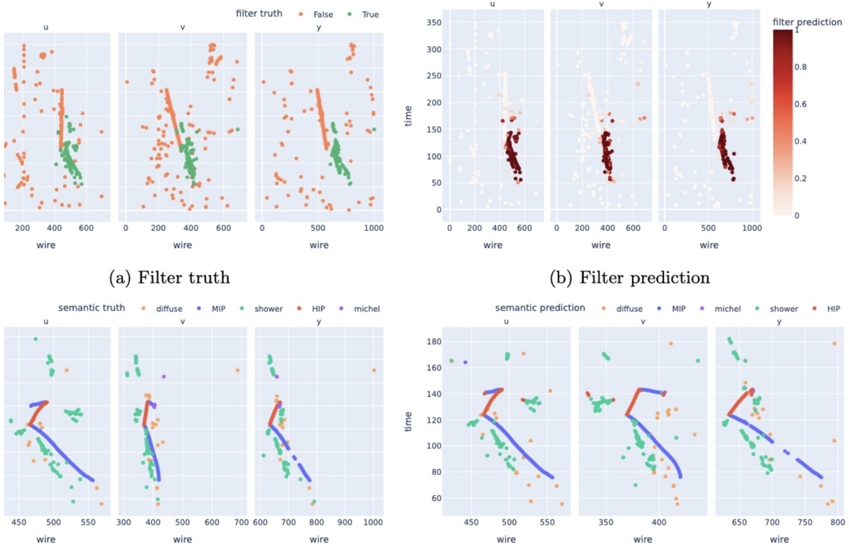
Reconstruction: Clustering

- NuGraph: GNN-based approach to LArTPC event reconstruction; Developing and implementing on DUNE

[A.Aurisano, V.Hewes, G.Cerati, J.Kowalkowski, C.S.Lee, W.Liao, D.Grzenda, K.Gumpula and X.Zhang, Phys. Rev. D 110, no.3, 3 (2024)]



Nu2Graph particle clustering



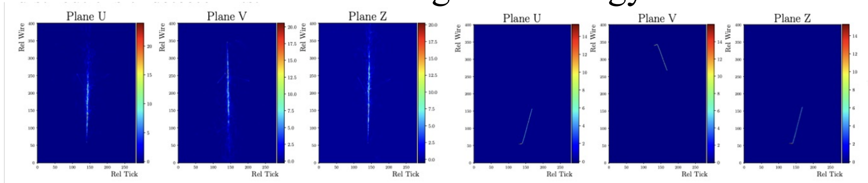
(a) Filter truth
(b) Filter prediction
(c) Semantic truth, filtered by truth
(d) Semantic prediction, filtered by prediction

Reconstruction: Kinematics

- 2D and 3D Regression CNNs for Neutrino event energy and primary particle energy and direction
- Incorporating energy and momentum conservation into network training for domain-aware scientific machine learning

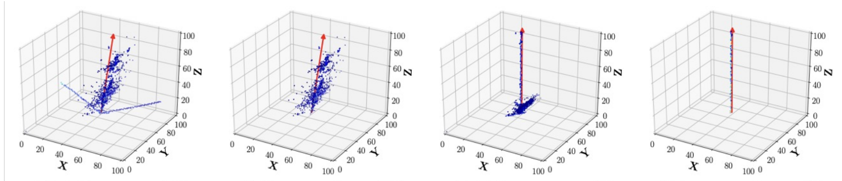
J.Liu, J.Ott, J.Collado, B.Jargowsky, W. Wu,J.Bian, P.Baldi, NeurIPS proceedings, arXiv 2012.06181

2-D images for energy reconstruction

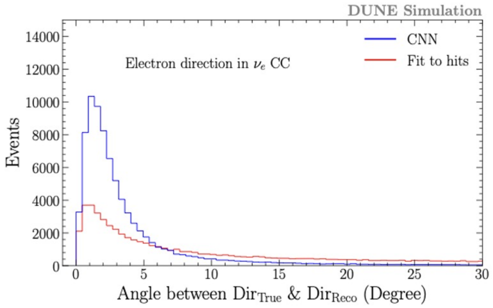


(a) Full-event ν_e CC (b) Full-event ν_μ CC

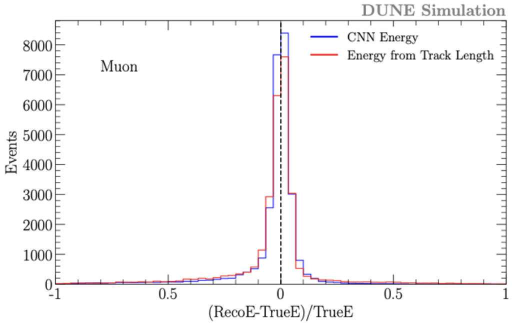
3-D images for angle reconstruction



(a) Full-event ν_e CC (b) Prong-only ν_e CC (c) Full-event ν_μ CC (d) Prong-only ν_μ CC



(a) Angular resolution for ν_e CC



(e) Muon prong energy (contained).

- SPINE and NuGraph also calculate kinematic variables in their workflows

Reconstruction: Kinematics

- Pandora uses a U-Net method for interaction vertex finding, Published in [*Eur. Phys. J. C* **85**, 697 (2025)]

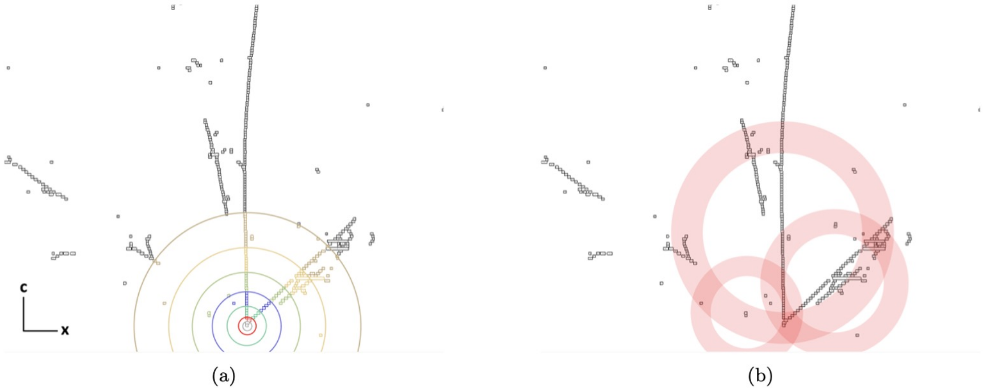
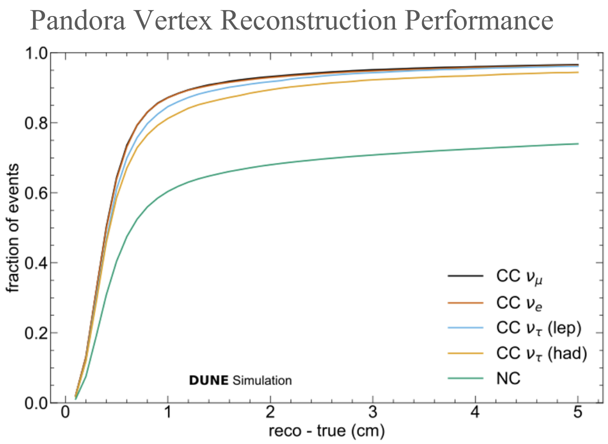
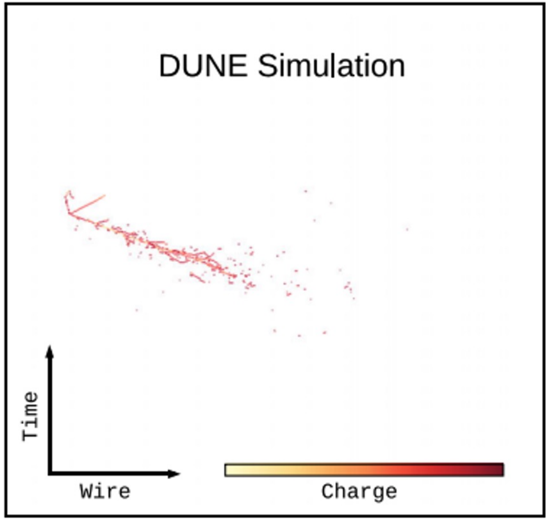


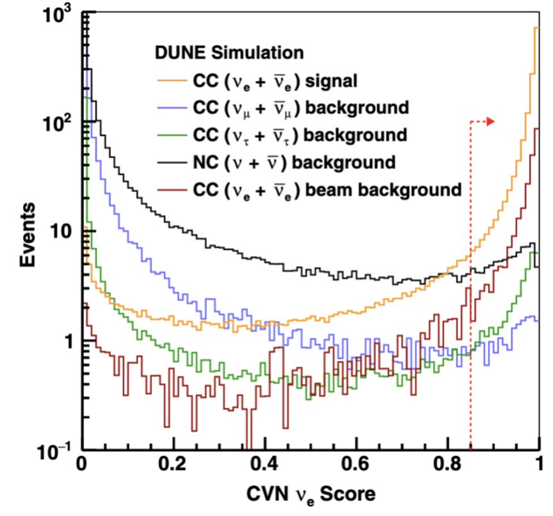
Fig. 3: An example of (a) the input hits and assignment of the first seven of the nineteen true distance classes for those hits and (b) a schematic of the heat map produced by three arbitrary hits during inference, for one view (W) of an event.

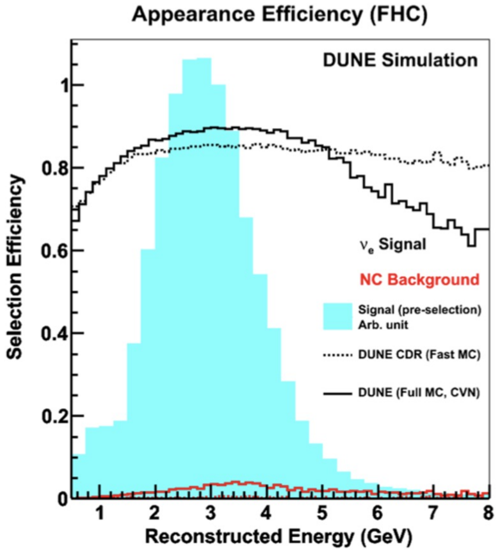


Reconstruction: Event-Classification and Particle Identification

- CNN predicts the flavour of the neutrinos using images of the interactions (CVN), First demonstration that DUNE can achieve the CP violation sensitivity from the CDR, [Phys. Rev. D 102, 092003]
- Data-intensive scientific machine learning for automated scientific inference and data analysis



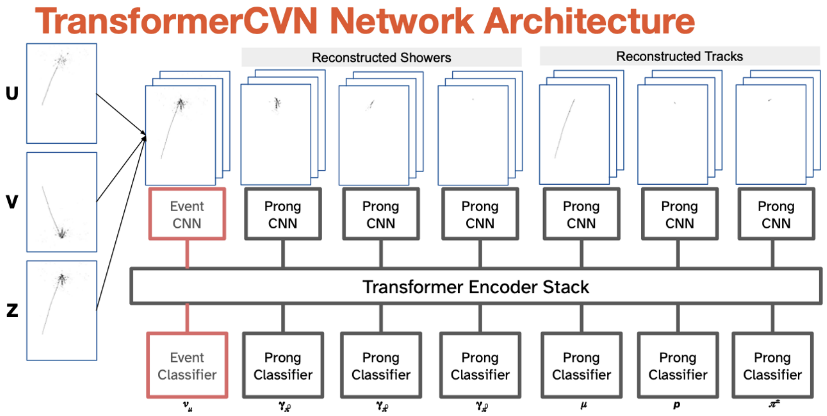


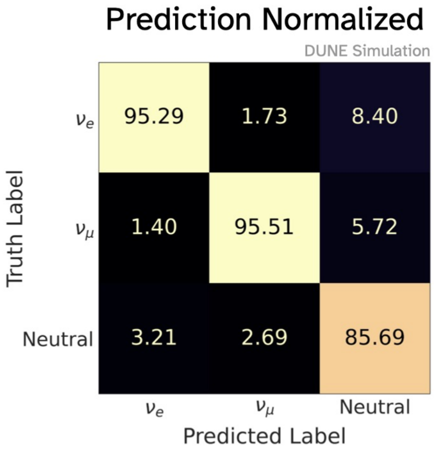


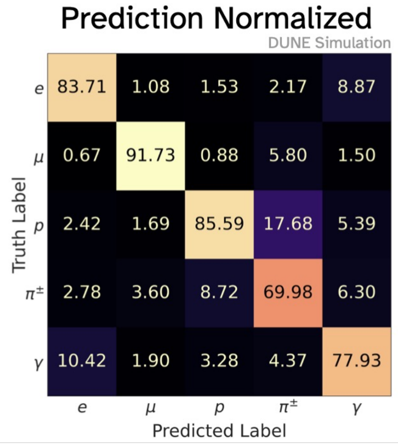
Reconstruction: Event-Classification and Particle Identification

- Transformer and Sparse CNN for simultaneous neutrino flavor classification and final particle identification; attention mechanisms enhance performance and provide interpretability
- Studying interpretability of network decisions using pixel gradients (saliency) and attention scores

Alejandro Yankelevich and Alexander Shmakov
https://indico.bnl.gov/event/21492/contributions/88967/attachments/53875/92140/TransformerCVN_Wire-Cell_Summit_2024.pdf;







- SPINE and NuGraph also perform particle id as part of their respective workflows

ML in particle transport simulation

- Generative machine learning model for fast photon simulation in DUNE FD
- 1D generative MLP as a fast surrogate for Geant4 photon simulation in ProtoDUNE

W Mu, A. Himmel and Bryan Ramson, 2022 Mach. Learn.: Sci. Technol. **3** 015033

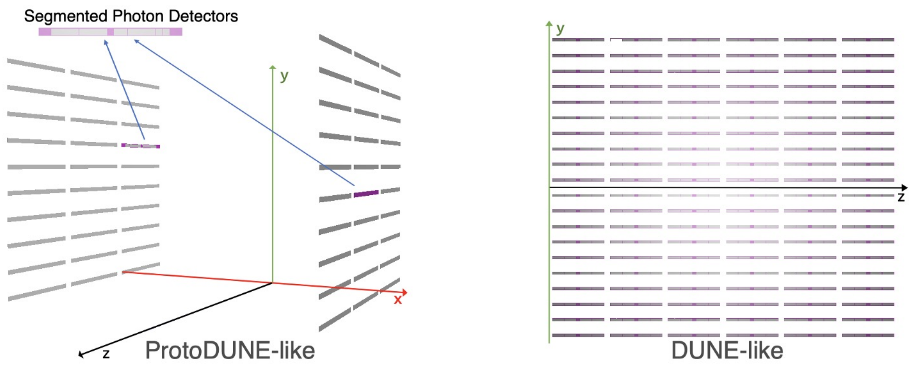


Figure 2. Photon detection systems in ProtoDUNE-like (left) and DUNE-like geometries (right). The grey rectangles are the photon detectors.

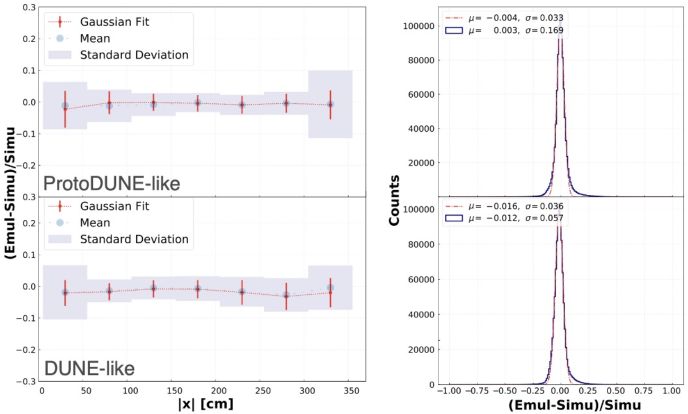


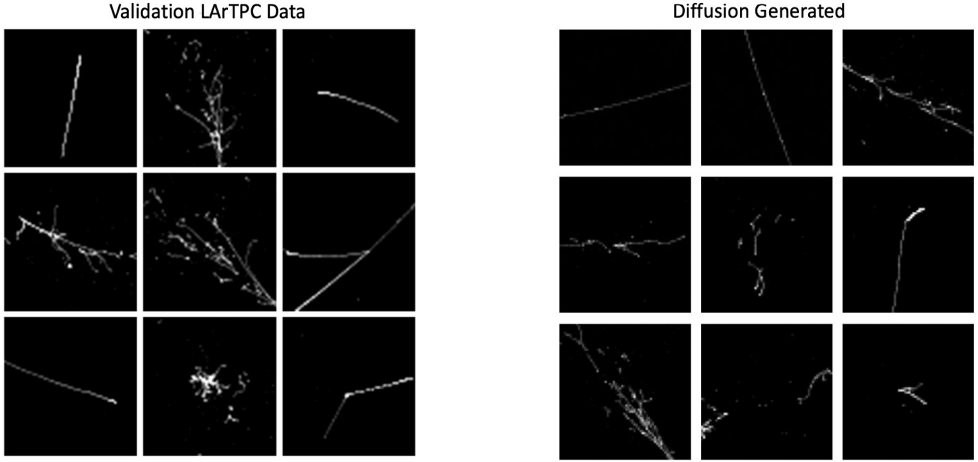
Figure 4. Discrepancy of total detected photons per scintillation vertex. Left: discrepancies variation along the x-axis. Right: overall discrepancies, where the red dashed line is a Gaussian fit. Comparison for ProtoDUNE-like geometry is on the top and DUNE-like geometry on the bottom.

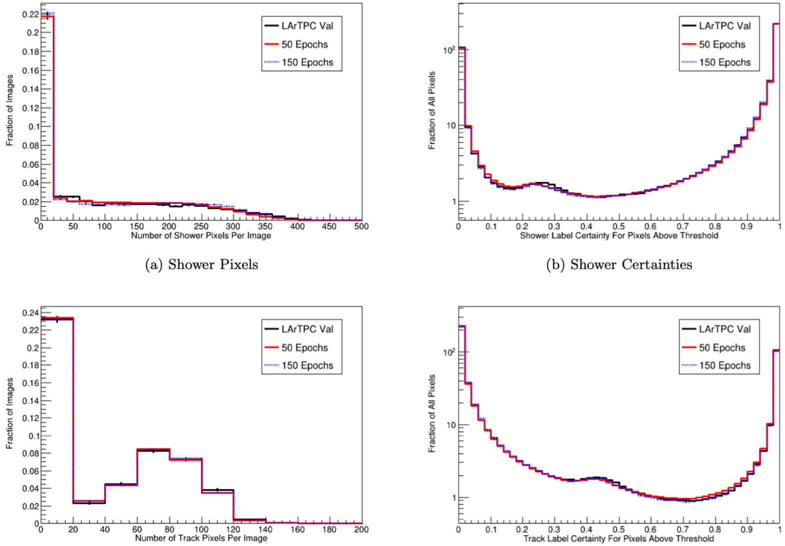
Simulation

ML in particle transport simulation

Score-based Diffusion Models for Generating LArTPC Images

[Zeviel Imani, Taritree Wongjirad, Shuchin Aeron, Phys.Rev.D 109 (2024) 7, 072011]





(a) Shower Pixels (b) Shower Certainties

(c) Track Pixels (d) Track Certainties

Good agreement between Simulation and generation

Simulation

ML in detector response simulation (Digitization)

- Differentiable simulation of DUNE ND-LAr detector response
- Simultaneously adjust multiple detector model parameters (drift field, electron lifetime, recombination, diffusion coefficients, etc.) with data input

S.Gasiorowski, Y.Chen, Y.Nashed, P.Granger, C.Mironov, K.V. Tsang, D.Ratner and K.Terao, Mach. Learn. Sci. Tech. 5, no.2, 025012 (2024)

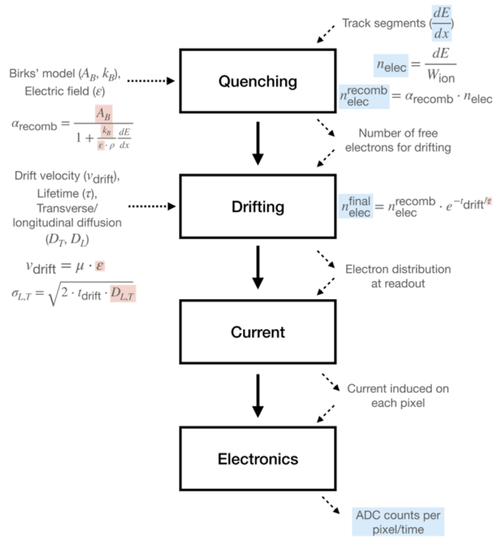


Figure 1. Flow diagram of the simulator, highlighting inputs and outputs of each stage (blue) as well as commonly calibrated model parameters (red).

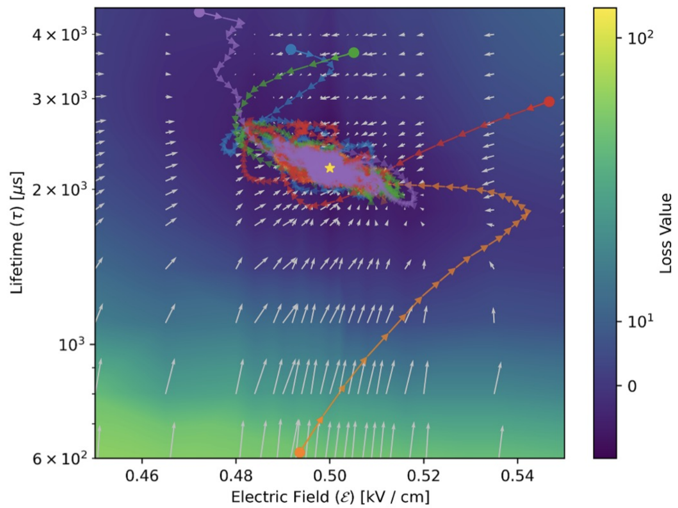
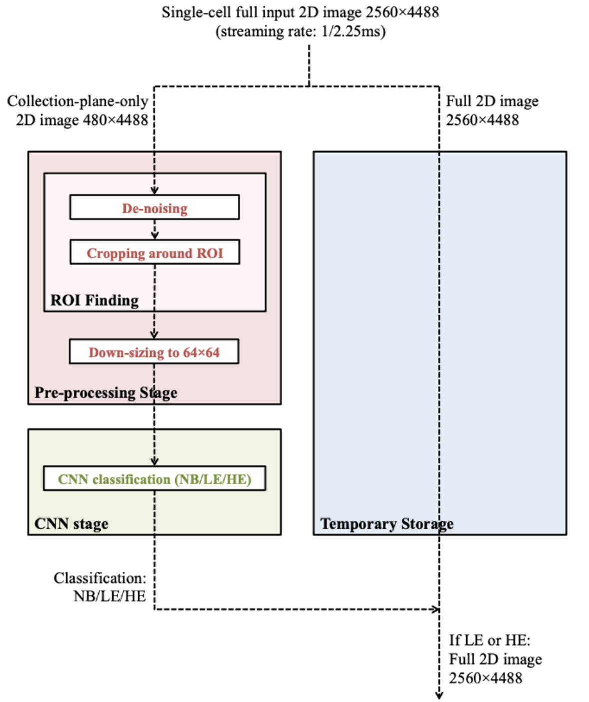


Figure 9. Loss landscape in a 2D parameter space of electric field $\mathcal{E}$ and lifetime τ, averaged across batches. The gold star labels the target parameter values. The negative gradient, shown in the white arrows, points towards the loss landscape minimum at the target. Five example fit trajectories, starting from a variety of different initial points (filled circles) are shown respectively in different colored lines. All fits converge to the target parameter values (the gold star).

Trigger & DAQ

- Real-time Inference with 2D Convolutional Neural Networks on FPGA for Trigger
- Use "High Level Synthesis for Machine Learning” (hls4ml) to test CNN deployment on FPGA



Y.j.Jwa, G.Di Guglielmo, L.Arnold, L.Carloni and G.Karagiorgi, Front. Artif. Intell. 5, 855184 (2022), arXiv 2201.05638

Figure 4. The data processing and data selection scheme under study for potential implementation in the upstream DAQ readout units of the future DUNE FD. The streaming 2D input images contain, > 99.9% of the time, NB data. This overall scheme should select true HE and LE images with > 90% accuracy, and true NB images with > 99.99% accuracy, in order to meet the DUNE FD physics requirements. Additionally, the pre-processing and CNN inference algorithms should meet the computational resources of the DUNE FD upstream DAQ readout units, and the algorithm execution latency should meet the data throughput requirements of the experiment.

Table 10. Estimated resource utilization from Vivado HLS for CNN inference on a Xilinx UltraScale+ (XCKU115) FPGA.

	Block RAM	**DSP Units**	**Flip Flops**	**Look-up Tables**
Available	4320	5520	1326720	663360
CNN02-DS-OP (PQT)	331 (7%)	4309 (78%)	226982 (17%)	163460 (24%)
Q-CNN02-DS-OP (QAT)	187 (4%)	2106 (38%)	142128 (10%)	138715 (20%)

Trigger & DAQ

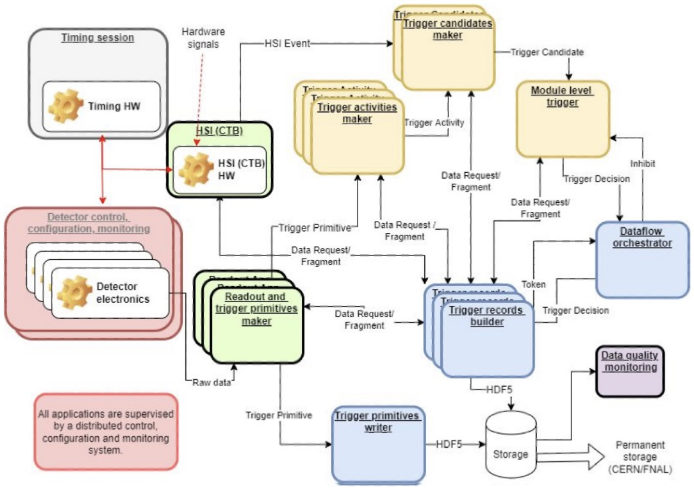
NVIDIA Triton based Online inference framework:

- Deploy AI/ML models on the DAQ cluster to perform software triggering

Andrew J. Mogan, https://indico.cern.ch/event/1387540/contributions/6153596/attachments/2948444/5182096/fast_ml_dunedaq_10_15_24.pdf

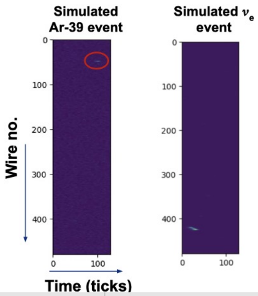
DUNE DAQ Data Flow

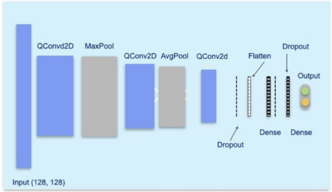
Image source



Initial Event Level Trigger Model

- 2D CNN for rejecting frames containing only radiological background on raw data
- Previously simulated in LArSoft for raw data readout, now developing datasets and training techniques to perform inference on TP based images, much sparser!





Model architecture for 2DCNN
Total Number of Parameters 19,750

Adapted from model by Jwa et al.

CNN Type	Corresponding Step	Purpose
2D CNN	Readout and Trigger Primitives Maker	Filter radiological background (e.g., Ar-39)
1D CNN	Trigger Activities Maker	Identify Regions of Interest (ROI) in TPs
1D Autoencoder + CNN	Trigger Candidates Maker	Remove electronic noise, optimize trigger candidates

ML Trigger Application in ProtoDUNE see Ciaran Hasnip'talk

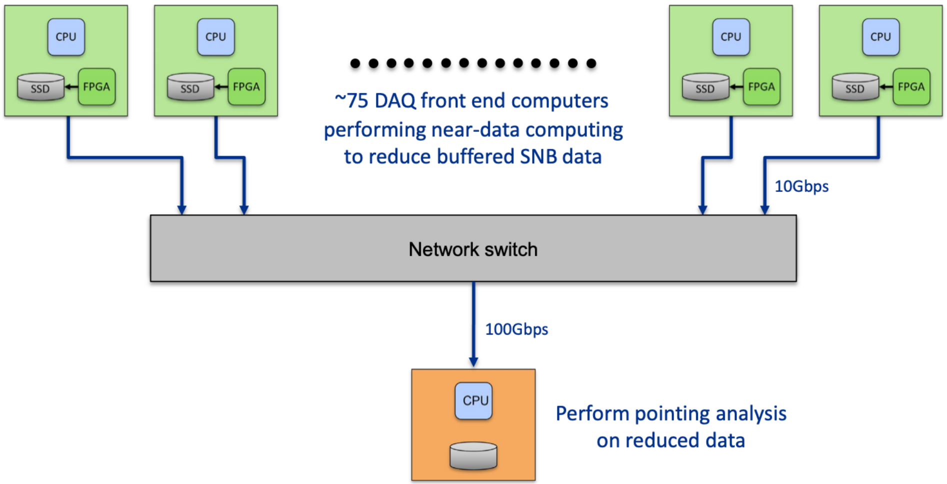
Trigger & DAQ

Intermediate Level Supernova Pointing Trigger for DUNE Using In-storage AI

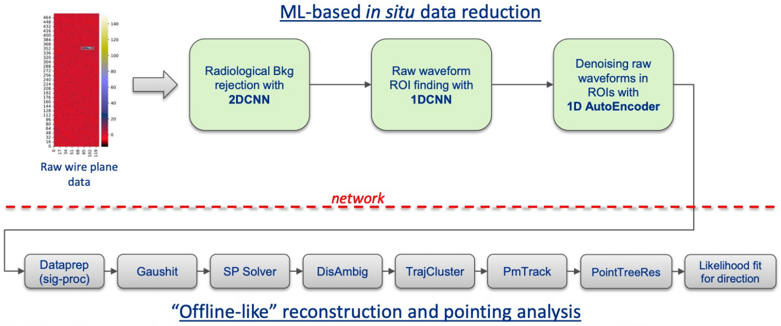
- In-storage AI deployed at the data buffering/storage layer (SSD + FPGA/AI accelerator)
- Performs early ML-based data reduction to enable early supernova pointing information

Michael Wang,
https://lss.fnal.gov/archive/2024/slides/fermilab-slides-24-0090-csaid.pdf

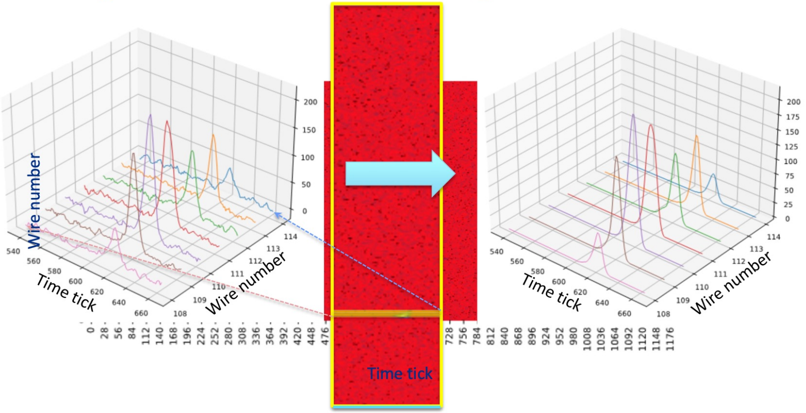
Strategy for fast pointing determination: hardware



Strategy for fast pointing determination: algorithms



ROI finding with 1D-CNN and denoising with 1D-AE



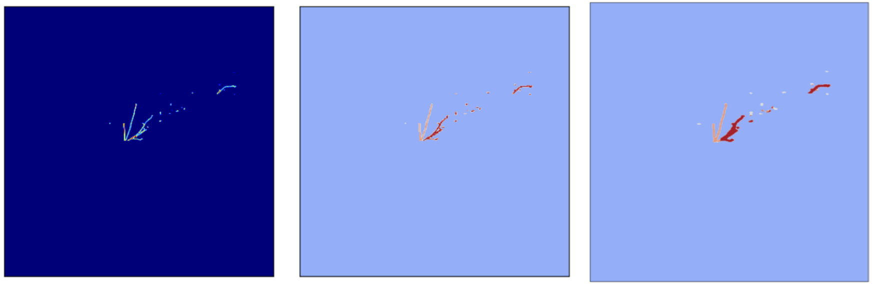
--

Trigger & DAQ

Software Trigger ML Algorithms on DAQ Cluster

- Sparse Convolutional Neural Network for DUNE Supernova Trigger

S. Damish1 and M. Bhattacharya,
https://lss.fnal.gov/archive/2024/pub/fermilab-pub-24-0524-student.pdf



(a)ADC. (b)Ground truth. (c)Model prediction.

FIG. 3. ADC image, its pixel classification, and model's classification prediction.

- Anomaly Detection for Supernova Trigger (CNN based auto-encoder)

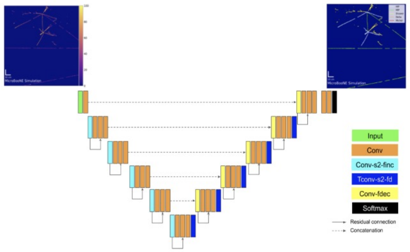
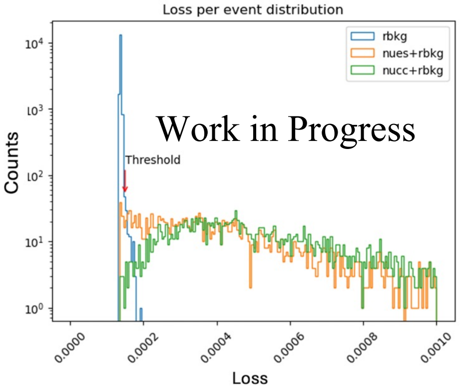


FIG. 1. Representation of the 2D convolution layers that make up the CNN used to classify pixels as belonging to shower or track categories for MicroBooNE.[6]

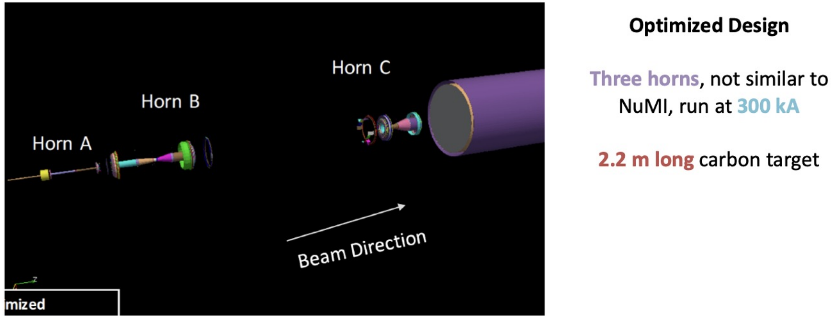


Sunny Seo, FNAL

Beam design and monitoring

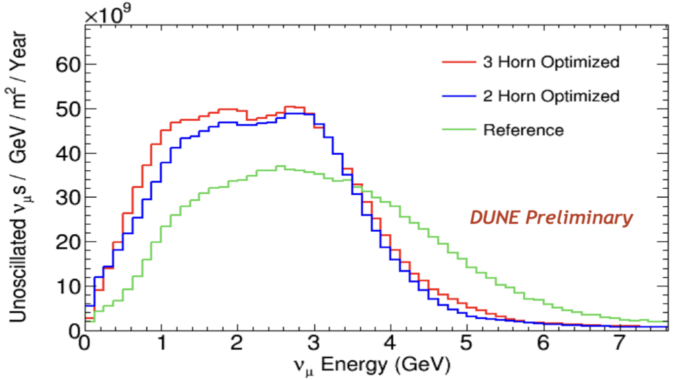
- Genetic algorithm for LBNF neutrino beam optimization

Laura Fields,
https://indico.fnal.gov/event/15204/contributions/30218/attachments/18904/23699/Fields-LBNFDUNEBeamOptimization-June2018.pdf



Genetic algorithm used to optimize the magnetic horn design

- Evolutionary-based method iteratively improves solutions by generating random variations
- Input: Optimizable beam parameters, Output: set of designs most sensitive to CP-violation
- Shapes and positions of numerous parts of the horns optimised
- Significantly increased the CP-violation sensitivity



Beam design and monitoring

- Machine Learning for LBNF/NuMI beam monitoring (ANN)
- Provide key tools for intelligent automation and decision-support for the management and control of complex systems

D.Wickremasinghe, S.Ganguly, K.Yonehara, R.M.Zwaska, Y.Yu and P.Snopok, arXiv:2309.08029

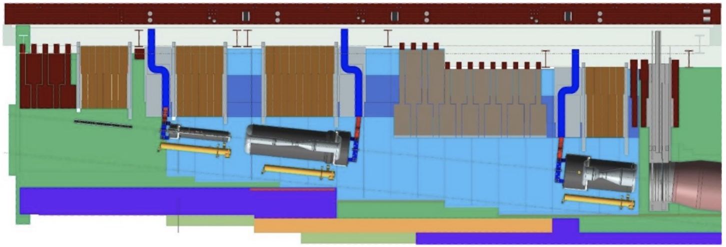


FIG. 2. An illustration of the upstream portion of the LBNF neutrino beamline. A horn-protection baffle, three focusing horns, and the decay pipe are shown inside the target chase respectively from left to right (the beam direction).

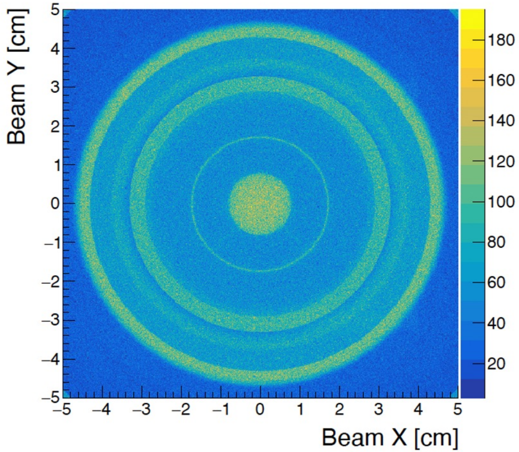


FIG. 6. Two-dimensional horizontal vs. vertical proton beam positions with uniform distribution sample for recorded beam interactions on the LBNF target that has a neutrino candidate at the downstream neutrino detector.

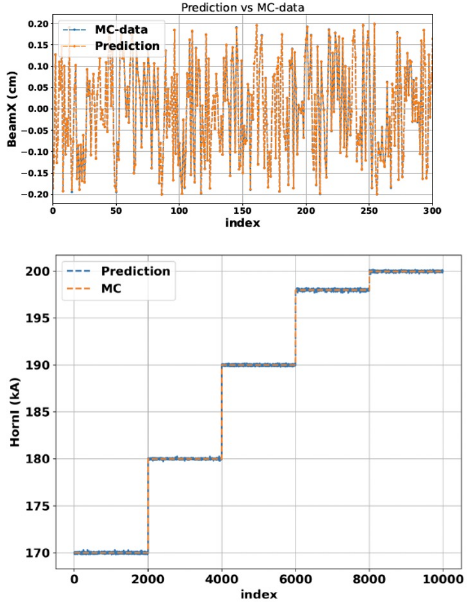
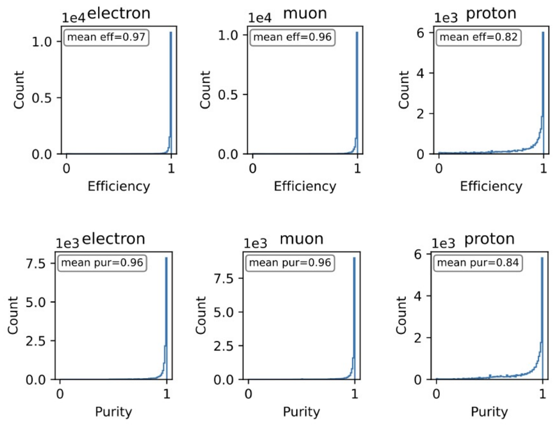


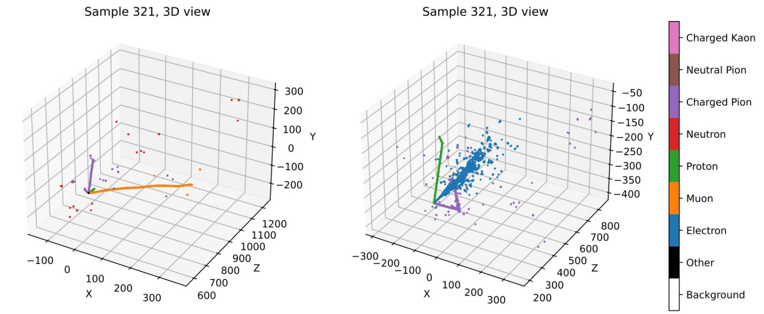
FIG. 29. Top: Horizontal beam position prediction, bottom: horn current prediction using a linear regression model.

Transformer and LLM

Particle Hit Clustering and Identification Using Point Set Transformers in LArTPC

[E.Robles, A.Yankelevich, W.Wu, J.Bian and P.Baldi, JINST 20 (2025) 07, P07030]

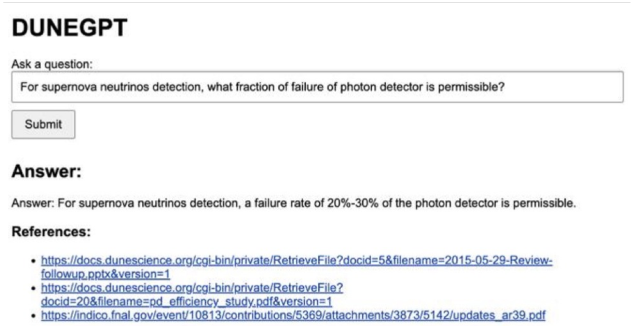




Model	Memory usage (MiB)	Time per sample (s)	Classification OVR AUC	Instance segmentation accuracy
2D R-CNN	440.5 ± 51.04	1.5752 ± 0.091	0.526	0.518
2D GAT	**88.6 ± 7.56**	**0.2300 ± 0.025**	0.833	0.659
2D HPST (ours)	99.1 ± 7.39	0.3542 ± 0.019	0.936	0.779
2D PST (ours)	138.1 ± 11.29	0.2539 ± 0.021	**0.949**	**0.827**
3D GAT	506.1 ± 30.13	0.7216 ± 0.060	0.859	0.727
3D PST (ours)	**170.2 ± 9.65**	**0.1401 ± 0.012**	**0.982**	**0.889**

LLM for DUNE (AI agent): unifies databases and documents, providing referenced answers and source code

[https://indico.cern.ch/event/1493037/contributions/6536623]



DUNEGPT

Ask a question:

For supernova neutrinos detection, what fraction of failure of photon detector is permissible?

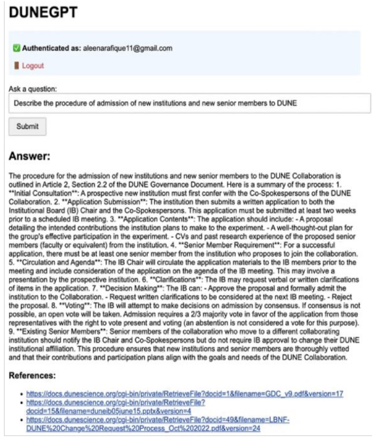
Submit

Answer:

Answer: For supernova neutrinos detection, a failure rate of 20%-30% of the photon detector is permissible.

References:

- https://docs.dunescience.org/cgi-bin/private/RetrieveFile?docid=5&filename=2015-05-29-Review-followup.pptx&version=1
- https://docs.dunescience.org/cgi-bin/private/RetrieveFile?docid=20&filename=pd_efficiency_study.pdf&version=1
- https://indico.fnal.gov/event/10813/contributions/5369/attachments/3873/5142/updates_ar39.pdf



DUNEGPT

Authenticated as: aleenarafique11@gmail.com

Logout

Ask a question:

Describe the procedure of admission of new institutions and new senior members to DUNE

Submit

Answer:

The procedure for the admission of new institutions and new senior members to the DUNE Collaboration is outlined in Article 2, Section 2.2 of the DUNE Governance Document. Here is a summary of the process: 1. **Initial Consultation**: A prospective new institution must first confer with the Co-Spokespersons of the DUNE Collaboration. 2. **Application Submission**: The institution then submits a written application to both the Institutional Board (IB) Chair and the Co-Spokespersons. This application must be submitted at least two weeks prior to a scheduled IB meeting. 3. **Application Contents**: The application should include: - A proposal detailing the intended contributions the institution plans to make to the experiment. - A well-thought-out plan for the group's effective participation in the experiment. - CVs and past research experience of the proposed senior members (faculty or equivalent) from the institution. 4. **Senior Member Requirement**: For a successful application, there must be at least one senior member from the institution who proposes to join the collaboration. 5. **Circulation and Agenda**: The IB Chair will circulate the application materials to the IB members prior to the meeting and include consideration of the application on the agenda of the IB meeting. This may involve a presentation by the prospective institution. 6. **Clarifications**: The IB may request verbal or written clarifications of items in the application. 7. **Decision Making**: The IB can: - Approve the proposal and formally admit the institution to the Collaboration. - Request written clarifications to be considered at the next IB meeting. - Reject the proposal. 8. **Voting**: The IB will attempt to make decisions on admission by consensus. If consensus is not possible, an open vote will be taken. Admission requires a 2/3 majority vote in favor of the application from those representatives with the right to vote present and voting (an abstention is not considered a vote for this purpose). 9. **Existing Senior Members**: Senior members of the collaboration who move to a different collaborating institution should notify the IB Chair and Co-Spokespersons but do not require IB approval to change their DUNE institutional affiliation. This procedure ensures that new institutions and senior members are thoroughly vetted and that their contributions and participation plans align with the goals and needs of the DUNE Collaboration.

References:

- https://docs.dunescience.org/cgi-bin/private/RetrieveFile?docid=1&filename=GDC_v9.pdf&version=17
- https://docs.dunescience.org/cgi-bin/private/RetrieveFile?docid=15&filename=duneib05june15.pptx&version=4
- https://docs.dunescience.org/cgi-bin/private/RetrieveFile?docid=49&filename=LBNF-DUNE%20Change%20Request%20Process_Oct%202022.pdf&version=24

Aleena Rafique, ANL

Transformer and LLM

Vision-Language Model (VLM) for LArTPC Particle Identification:

[D.Sagar, K.Yu, A.Yankelevich, J.Bian and P.Baldi, arXiv:2509.08461]()

Prompt

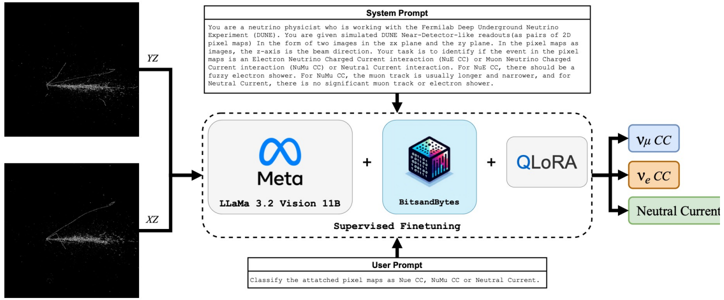


Figure 1: LLaMa 3.2 Vision finetuning pipeline.

Performance

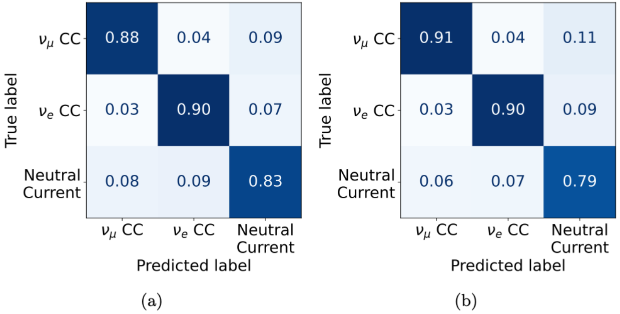


Figure 5: Finetuned LLaMa 3.2 Vision's (a) recall matrix (truth normalized) and (b) precision matrix (prediction normalized).

Interpretability

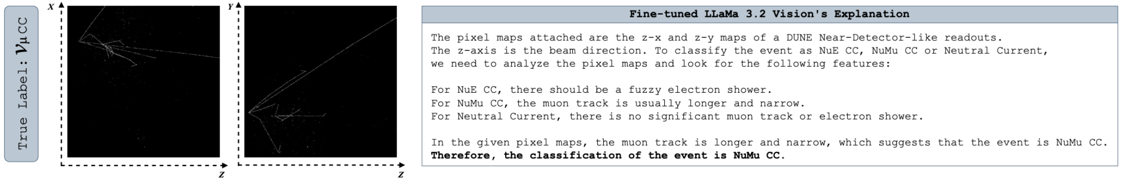


Figure 3: LLaMa 3.2 Vision Prediction Explanation.

Generalization

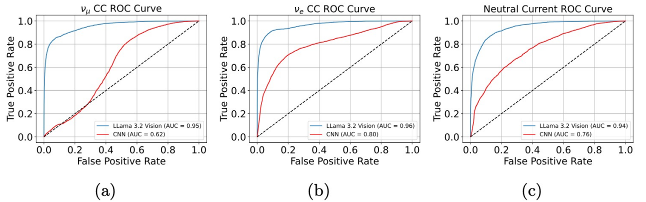
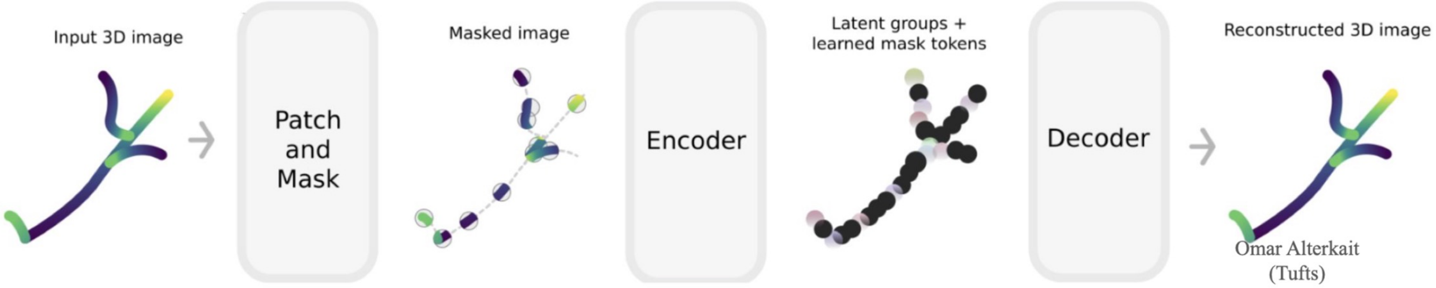


Figure 10: ROC curves for each class (a) ν_μ CC, (b) ν_e CC, and (c) NC comparing performance between the finetuned LLaMa 3.2 Vision and the CNN for generalization testing.

Foundation Models

Toward Foundation Models with HEP Sensory Dataset

First Attempt: 3D Point-based LAr MAE (PoLAr-MAE)

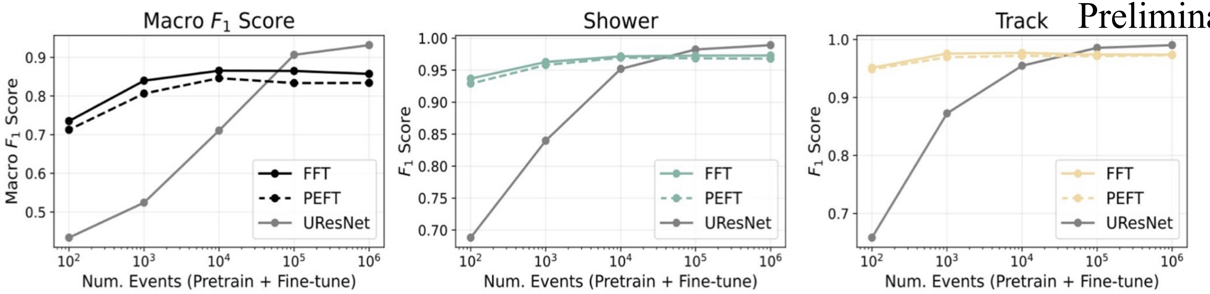


Omar Alterkait (Tufts)

S.Young, Y.j.Jwa and K.Terao, arXiv:2502.02558

Fine Tuning: Scalability + Benchmark v.s. Supervised UResNet

Preliminary



Notes on fine-tuning:

- FFT = full model
- PEFT = encoder frozen

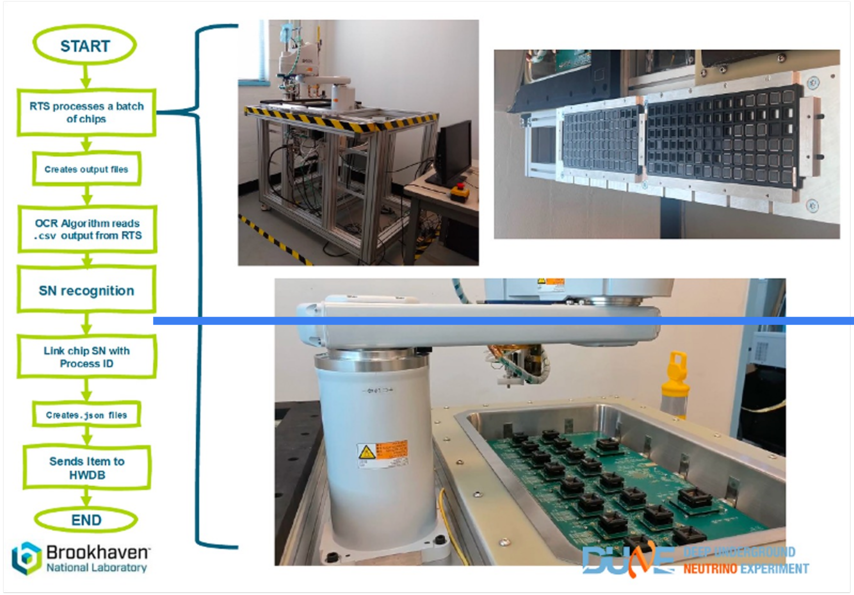
Findings/Confirmations:

- MAE is very data efficient
  - Quality even at 100!
  - Saturates ~1E4
  - Larger model for >1E4

Quality Assurance and Quality Control (QA/QC)

- AI for DUNE cold electronics QA/QC

AI has been utilized for quality assurance and quality control (QA/QC) in DUNE detector construction. This example demonstrates the application of AI-enhanced Optical Character Recognition (OCR) for serial number identification in DUNE cold electronics chips QA/QC within the Robotic Test Stand (RTS).



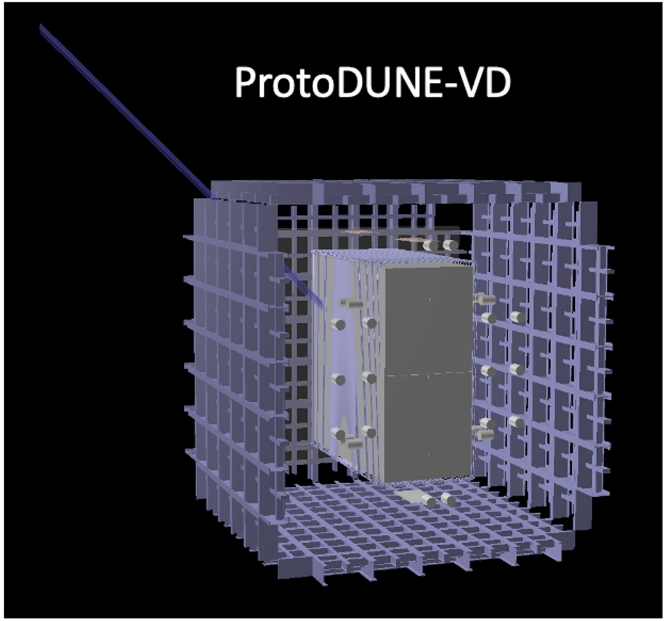
DUNE Tech. Prelim.

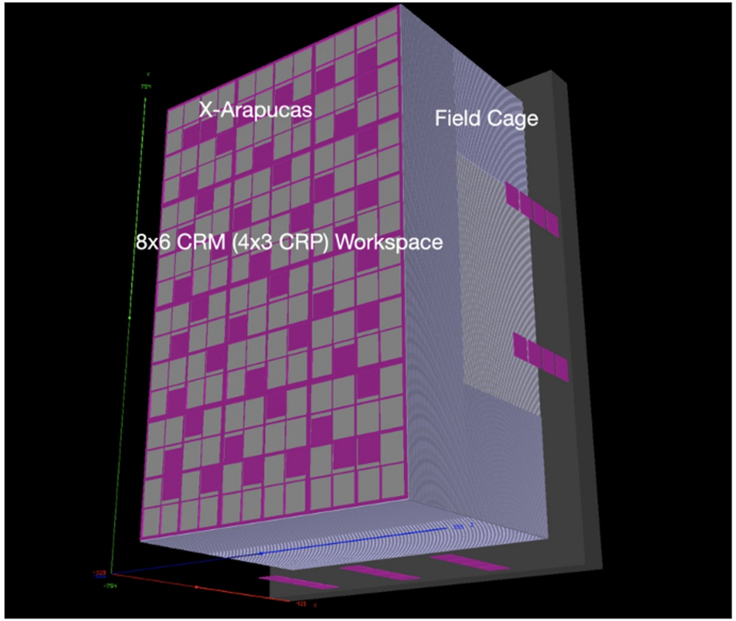
Kieran Wall, BNL

AI-assisted code generation

- Human-AI workflow for DUNE GEANT4 Geometry Construction
- AI-assisted code generation

DUNE Tech. Prelim.





Nitish Nayak, BNL

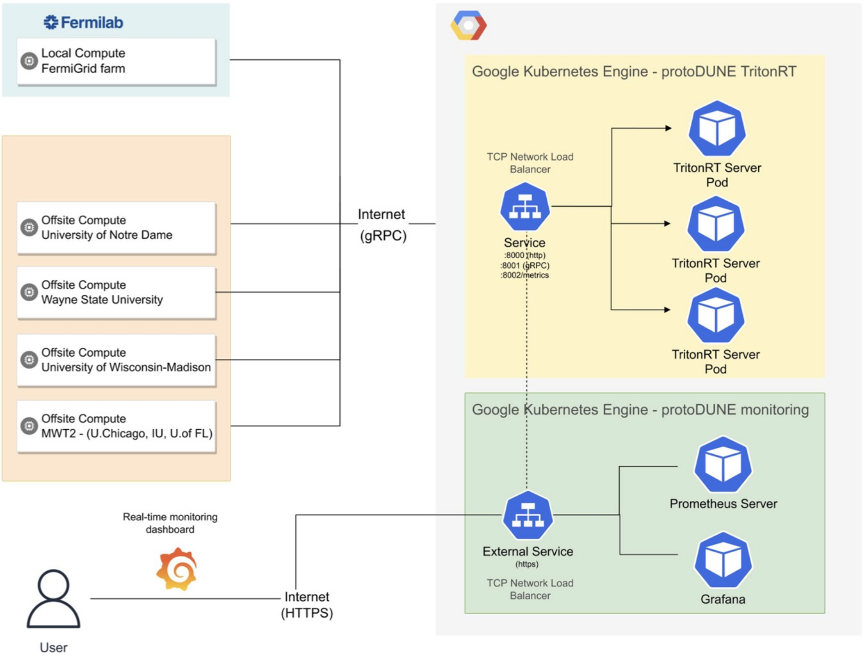
AI Infrastructure

- National Lab HPC Facilities:
  - Fermilab EAF: Experimental Analysis Facility (EAF)
  - SLAC Shared Scientific Data Facility (S3DF)
  - ANL Polaris
- Distributed HTC GPU Clusters: FermiGrid, QMUL, OSG etc
- National HPC accessed via HEPCloud: NERSC/Perlmutter
- Other potential HPC Centers: ORNL, Commercial cloud etc

- All-GPU full-scale Near Detector simulation on NERSC https://www.nersc.gov/news-publications/nersc-news/science-news/2023/nersc-supports-first-all-gpu-full-scale-physics-simulation/
- DUNE utilized NERSC's GPU supercomputers to complete the first full-scale physics detector simulation entirely on GPUs
- Simulate neutrino interactions in the DUNE Near Detector with 12M-pixel readout at unprecedented speed
- Demonstrate GPU-powered NERSC enables large-scale DUNE simulations, paving the way for full detector modeling

- Accelerating Machine Learning Inference with GPUs in ProtoDUNE Data Processing

[T.Cai, K.Herner, T.Yang, M.Wang, M.A.Flechas, P.Harris, B.Holzman, K.Pedro and N.Tran, Comput. Softw. Big Sci. 7, no.1, 11 (2023)]



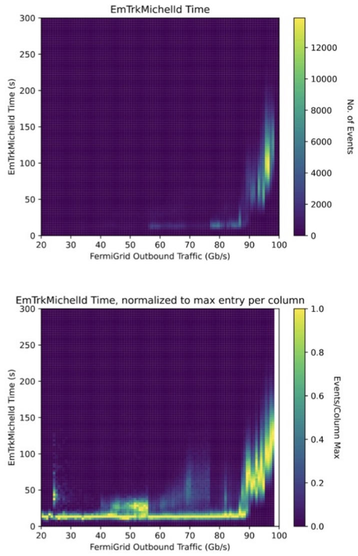


Fig. 6: The average EmTrk duration before Oct. 7 as a function of the total network traffic through the 100 Gb/s network switch at Fermilab used by the batch processing cluster. The top plot shows the real event rate. The bottom plot is the same as the top one, with each column scaled separately so the maximum amplitude is 1 for each column.

Summary

- AI/ML is widely integrated into DUNE
  - Developed in all major reconstruction components
  - Applications cover simulation, beam monitoring, trigger/DAQ, and QA/QC.
- Helps DUNE improve measurement performance, enhance physics reach, reduce computational burdens, and increase labour efficiency
- Leverage and improve National Labs and institutions' AI infrastructure
- Well aligned with priority research directions for AI/ML

感謝

thanks