AI/ML at DUNE

Jianming Bian

for the DUNE Collaboration

University of California, Irvine

06-15-2026

NPML2026, Irvine

AI/ML at DUNE

- 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)
- Exploring advanced AI technologies: LLMs, multimodal AI, foundation models, AI agents, and transformers
- DUNE leverages and advances AI infrastructure across major U.S. national laboratory computing facilities, including NERSC, ALCF, SDCC, EAF, and S3DF.

Reconstruction Framework: Well-Cell

The DUNE Well-Cell framework is a scalable, cell-based signal processing and event reconstruction framework for LArTPC detectors, enabling efficient 3D pattern recognition and particle reconstruction.

- Scalable Inference for LArTPC Signal Processing with MobileU-Net and Overlap–Tile Chunking
  https://ml4physicalsciences.github.io/2025/files/NeurIPS_ML4PS_2025_280.pdf
- CNN (U-Net) for regions of interest (ROI) detection in DUNE signal processing algorithm WireCell, JINST
  https://iopscience.iop.org/article/10.1088/1748-0221/16/01/P01036

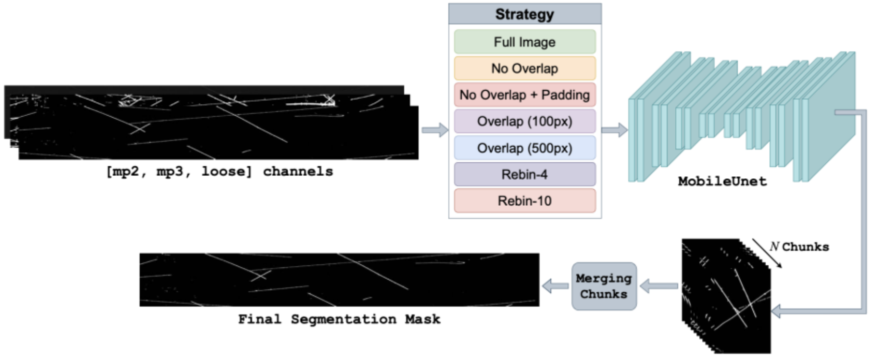


Figure 2: Schematic diagram of the proposed approach.

Table 1: Comparison of U-Net and MobileU-Net across inference strategies on the LArTPC ROI task. Metrics: Intersection over Union (IoU), Purity, Efficiency on the test set; single-instance CPU memory and inference time. The definition of each strategy can be found in Sec 2.4.

		Test Set			Single Instance	
Model	**Strategy**	**IoU**	**Purity**	**Efficiency**	**Memory (MB)**	**Time (s)**
U-Net	Full image	0.83	0.89	0.88	6193.7	106.57
	No Overlap	0.82	0.89	0.87	936.2	100.47
	No Overlap + Padding	0.83	0.83	0.88	1028.5	126.96
	Overlap (100 px)	0.83	0.89	0.87	954.4	132.71
	Overlap (500 px)	0.83	0.89	0.87	954.5	552.28
	Rebin 4	0.55	0.52	0.83	1811.8	30.33
	Rebin 4 (Retrained)	0.79	0.83	0.88	1812.8	30.64
	Rebin 10	0.31	0.34	0.48	918.9	12.17
MobileU-Net	Full image	0.84	0.89	0.88	1719.7	14.41
	No Overlap	0.77	0.82	0.87	442.4	12.90
	No Overlap + Padding	0.75	0.85	0.84	463.5	12.85
	Overlap (100 px)	0.80	0.86	0.87	442.4	12.08
	Overlap (500 px)	0.82	0.88	0.88	442.4	50.28
	Rebin 4	0.63	0.44	0.85	635.0	4.29
	Rebin 4 (Retrained)	0.83	0.86	0.88	635.0	4.32
	Rebin 10	0.48	0.33	0.61	435.5	2.39

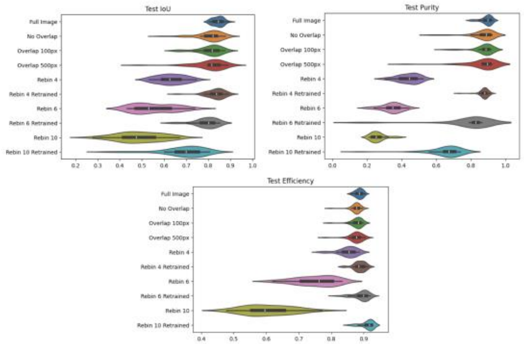
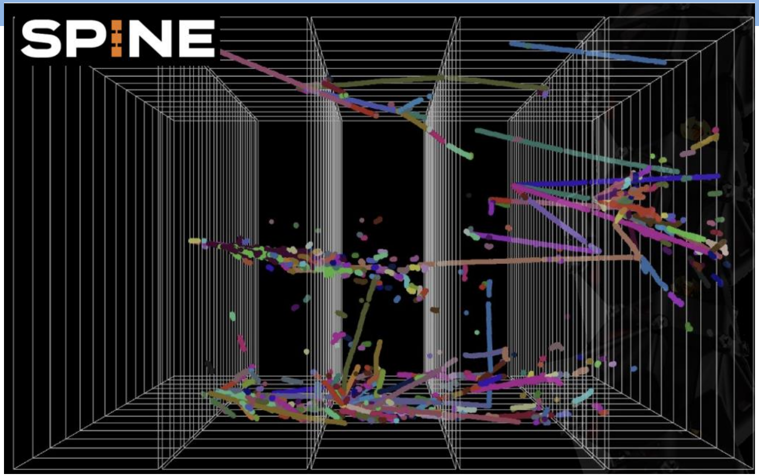


Figure 3: MobileUnet's IoU, Purity, and Efficiency distribution violin plots over the test set via multiple strategies.

Reconstruction Framework: SPINE



- A number of software frameworks use deep learning for clustering, regression and PID
- 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]()

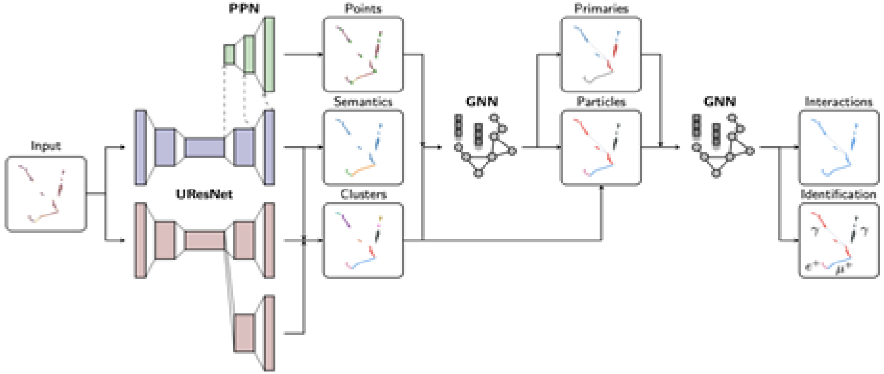
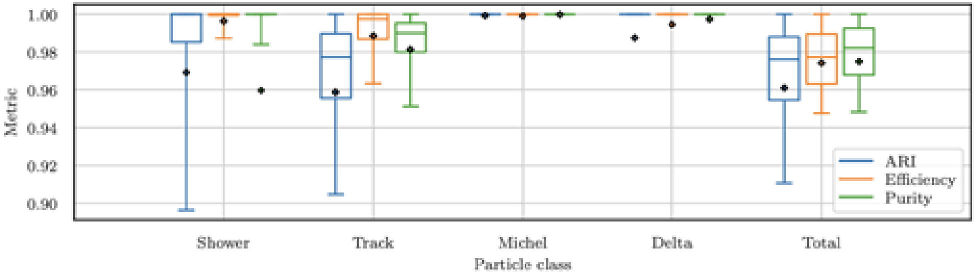


Figure 1: Schematic architecture of the end-to-end, ML-based reconstruction chain for LArTPCs.

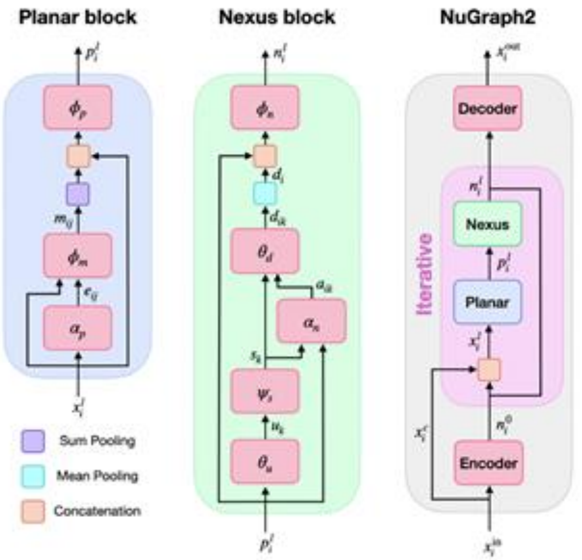
SPINE particle clustering performance



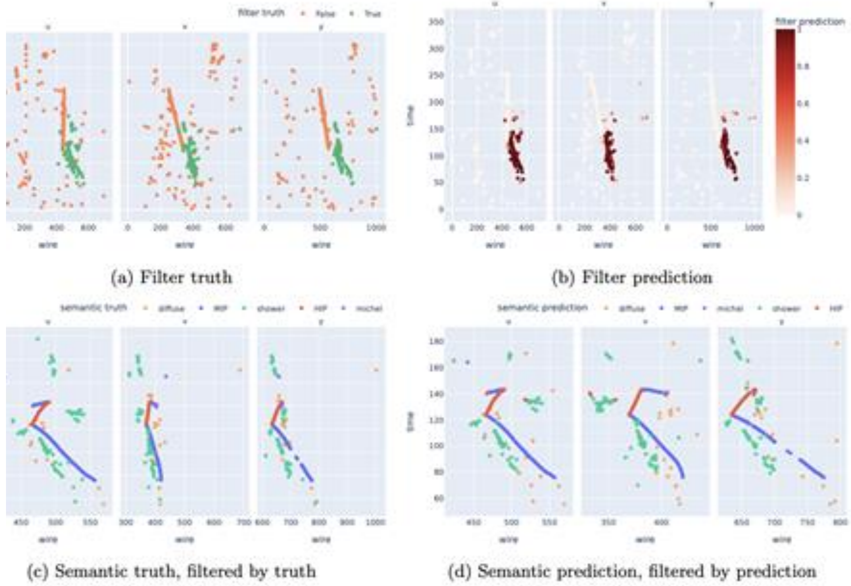
Reconstruction Framework: NuGraph

- 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

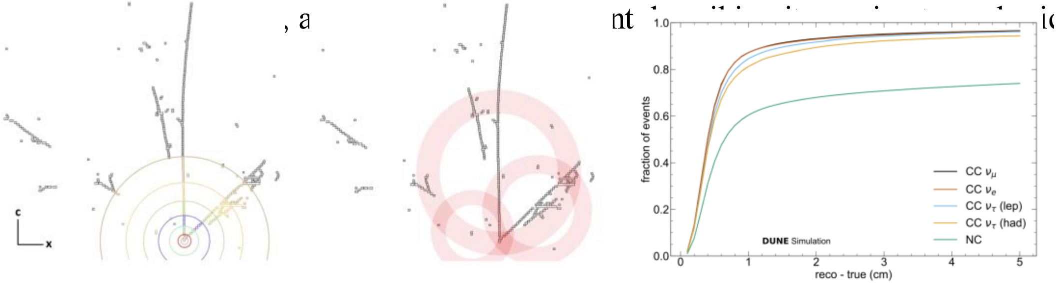


Currently dedicated for Nutau analysis, but can also for other channels

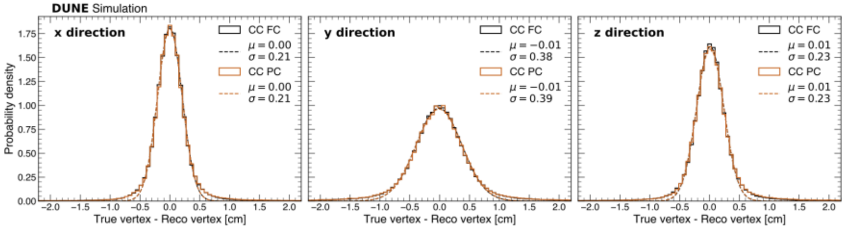
Reconstruction Framework: Pandora

Pandora: A widely used framework for particle reconstruction in LArTPC experiments, with increasing integration of AI/ML techniques in recent years.

CNN-based vertex reconstruction
- Predicts distance to the vertex position for each hit and uses ring-drawing method to predict vertex position

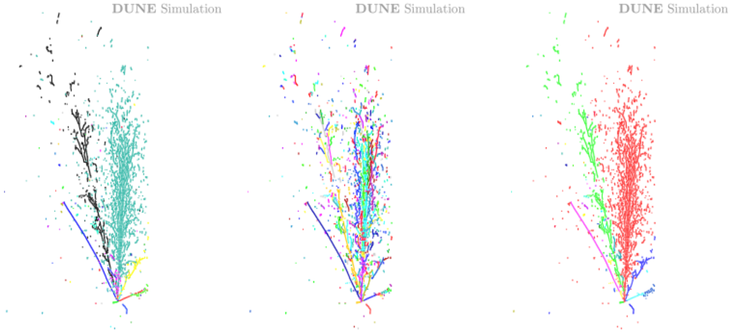


DUNE Collaboration, Neutrino interaction vertex reconstruction in DUNE with Pandora deep learning, Eur.Phys.J.C 85, 697 (2025) [Eur. Phys. J. C 85, 697 (2025)]

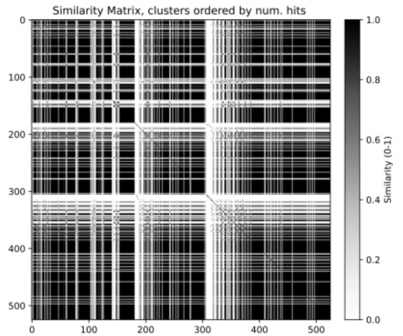


DUNE Collaboration, Reconstruction of atmospheric neutrinos in DUNE's horizontal-drift far-detector module, arXiv: 2601.05697 [hep-ex] (2026)

State-of-the-art shower reconstruction with transformers
- Significant improvement in dense environments with overlapping showers
- Predicts the similarity matrix between initial simple clusters
- Non-exhaustive list with many other exciting developments underway!



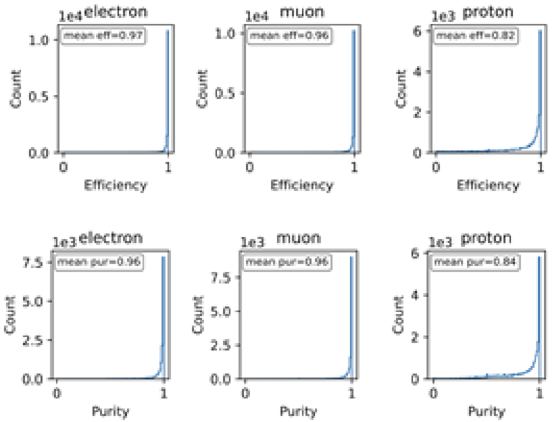
Truth   Simple clustering   Transformer shower growing

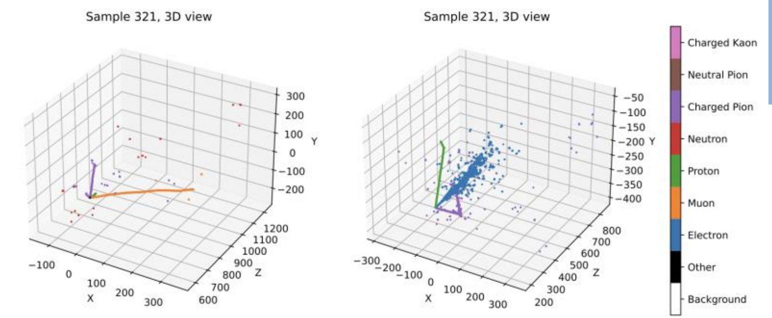


Transformers for Reconstruction

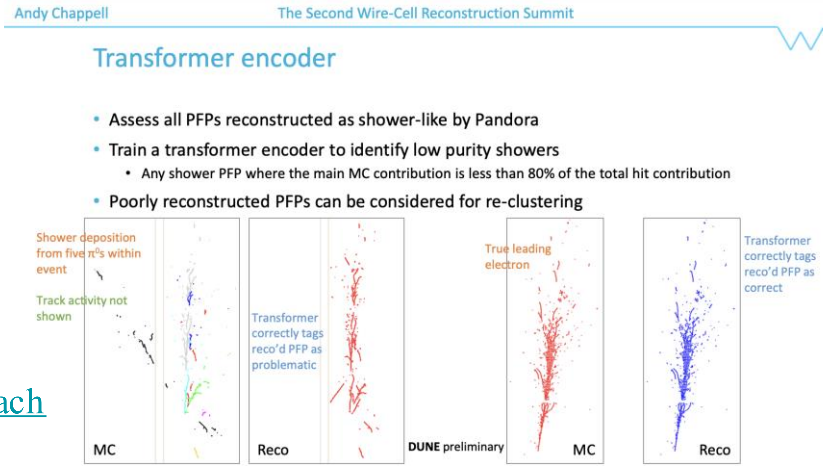
Particle Hit Clustering and Identification Using Point Set Transformers in LArTPC

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**



Pandora shower reconstruction with transformers
- Significant improvement in dense environments with overlapping showers
- Predicts the similarity matrix between initial simple clusters

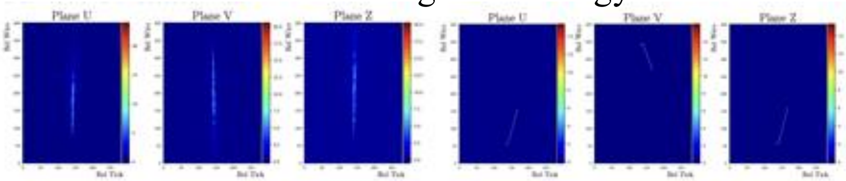
Andy Chappell
https://indico.bnl.gov/event/21492/contributions/88564/attachments/53743/91937/pandora_ml.pdf

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

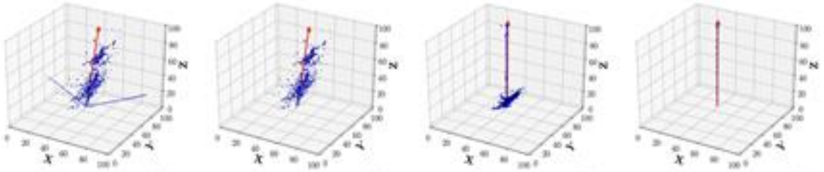
[J.Liu, J.Ott, J.Collado, B.Jargowsky, W. Wu,J.Bian, P.Baldi, NeurIPS proceedings, arXiv 2012.06181]

2-D images for energy reconstruction

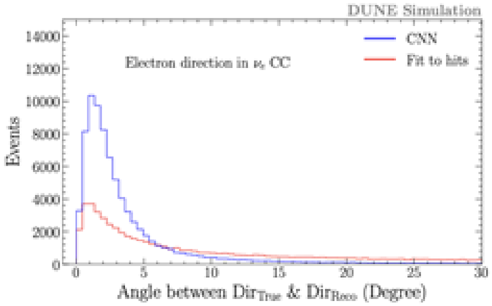


(a) Full-event ν_e CC (b) Full-event ν_μ CC

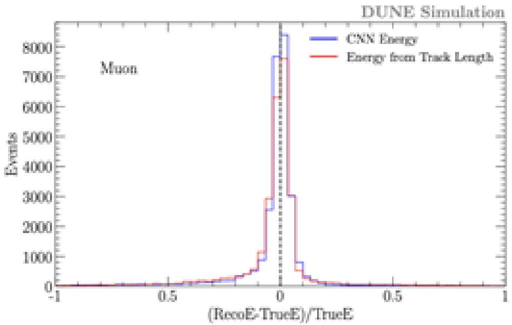
3-D images for angle reconstruction



(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

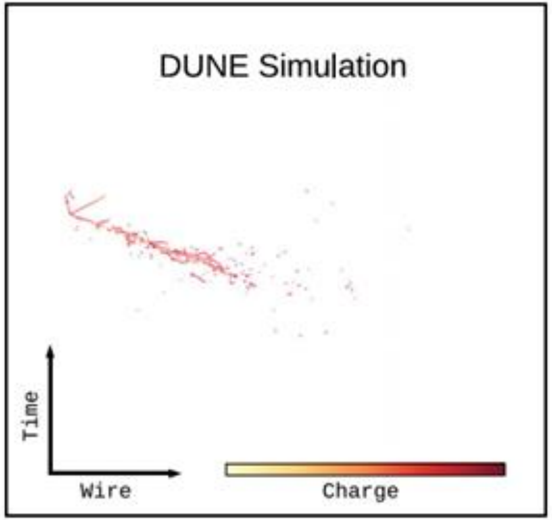


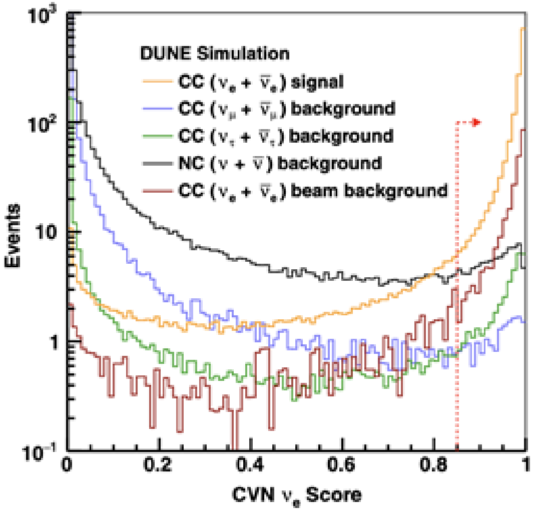
(e) Muon prong energy (contained).

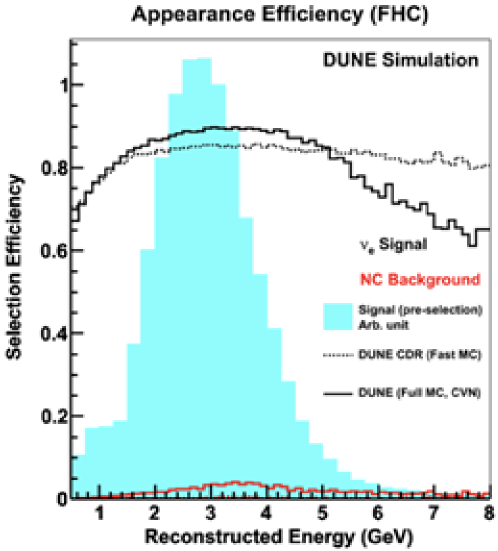
- SPINE and NuGraph also calculate kinematic variables in their workflows

Reconstruction: 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



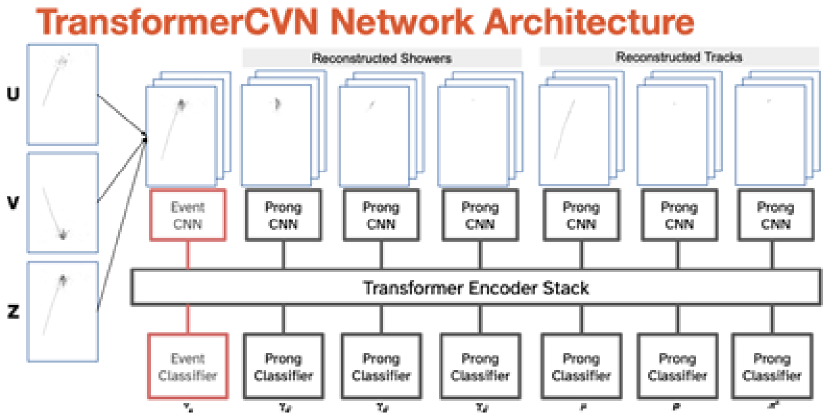


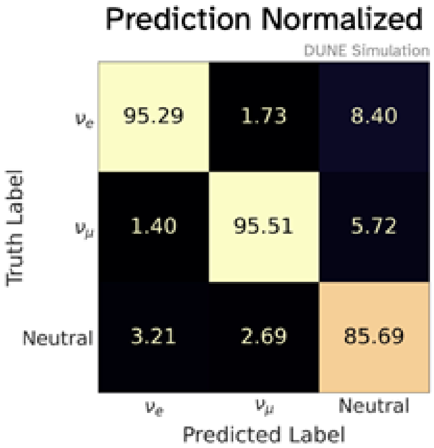


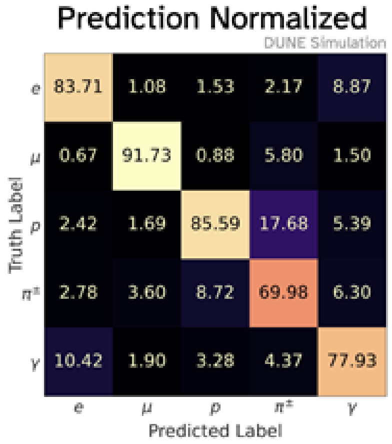
Reconstruction: Event-Classification and Particle Identification

- 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;



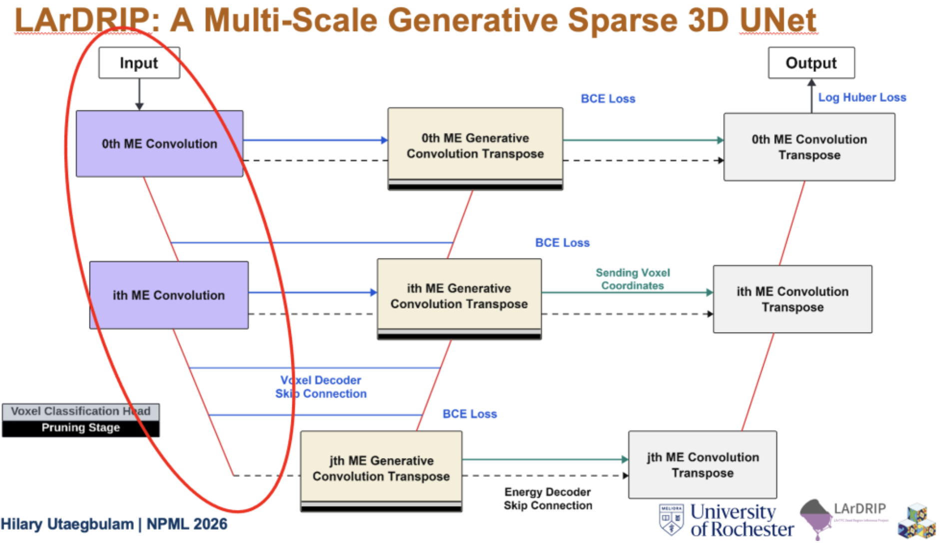


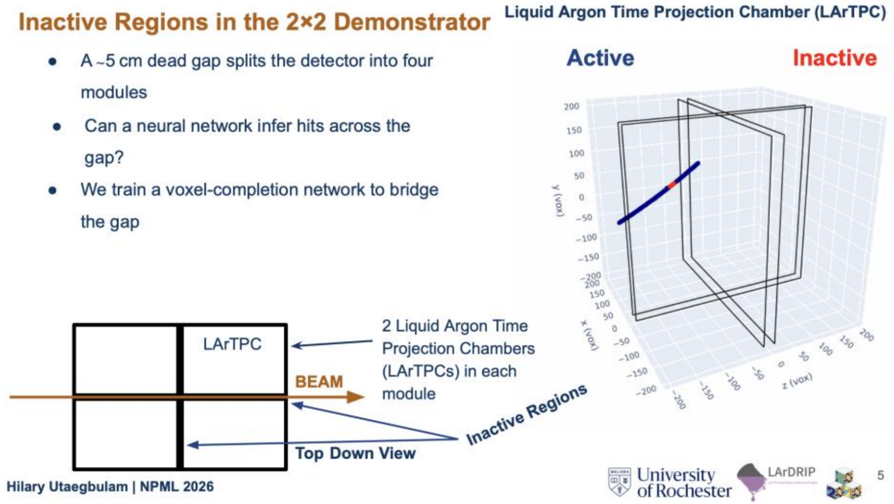


- SPINE and NuGraph also perform particle id as part of their respective workflows

Dead Region Recovery with Generative AI

- Machine Learning Track Inference in the Dead Regions of DUNE's Near Detector Prototype
- Generative AI for detector completion
- LArDRIP: A Multi-Scale Generative Sparse 3D Unet
- Recovering missing energy in detector dead regions





See Hilary Utaegbulam's talk

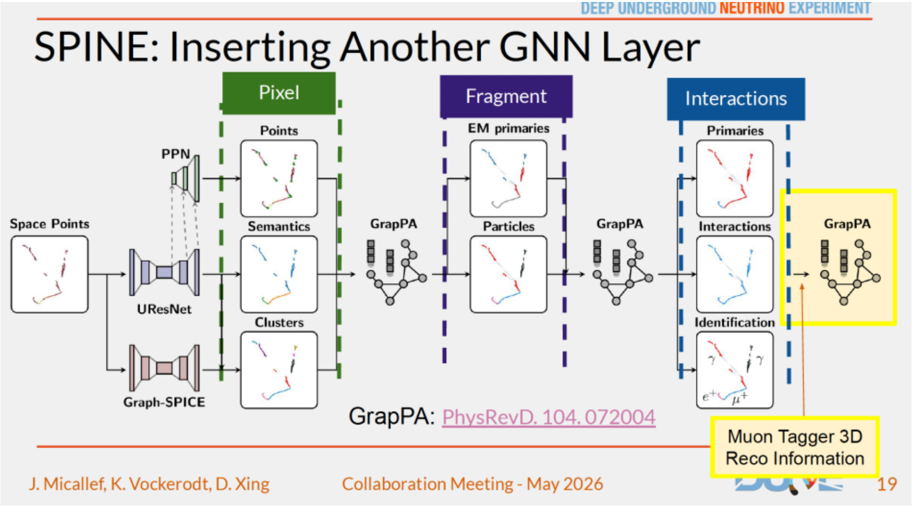
AI-Assisted Multi-Detector Event Reconstruction

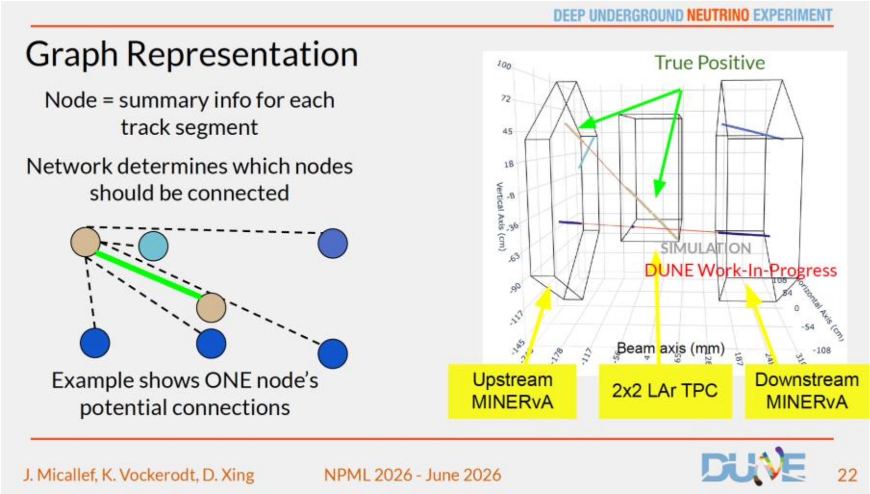
GNN-Based Track Matching Across DUNE Near Detectors

•Graph neural networks are used to connect track fragments reconstructed in ND-LAr and TMS.

•Integrated with the SPINE reconstruction framework for multi-detector event reconstruction.

•Enables robust track association in DUNE's high-rate Near Detector environment.





See J. Micahllef's talk

Simulation: Generative model for photon simulation

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]

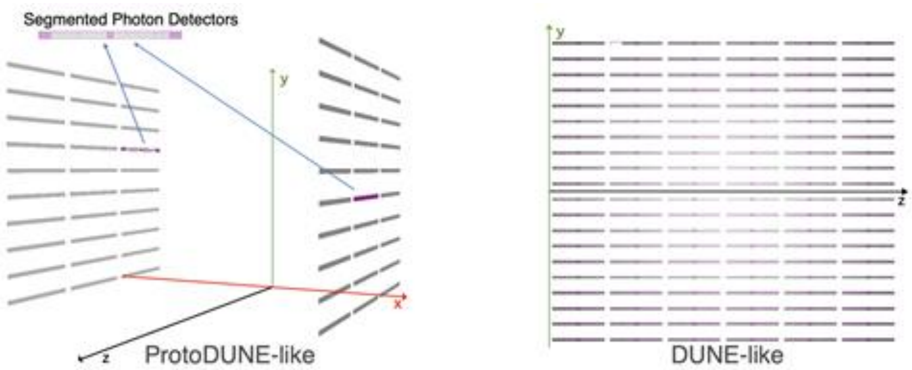


Figure 2. Photon detection systems in ProtoDUNE-like (left) and DUNE-like geometries (right). The grey rectangles are the photon detectors.

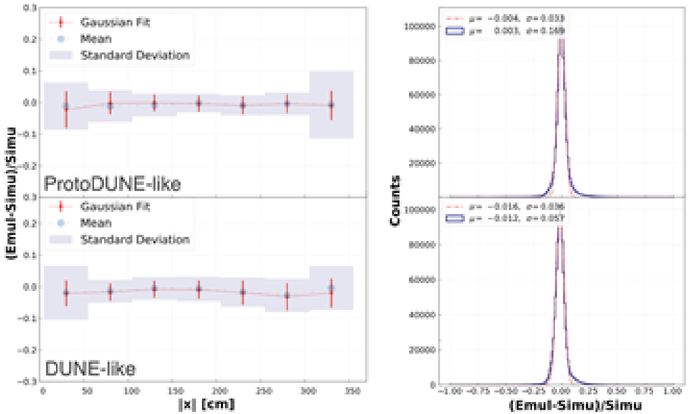


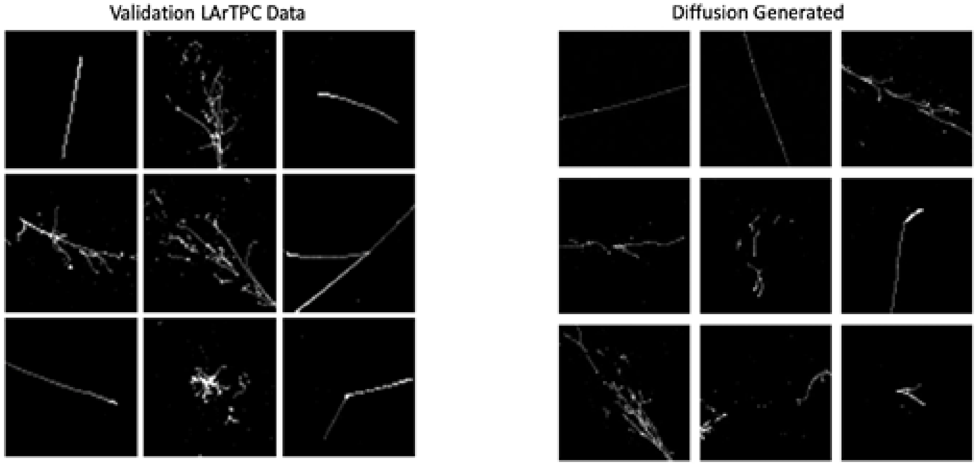
Figure 4. Discrepancy of total detected photons per scintillation vertex. Left: discrepancies variation along the x-axis. Right: overall discrepancies, where the red dashed line is a Gaussian fit. Comparison for ProtoDUNE-like geometry is on the top and DUNE-like geometry on the bottom.

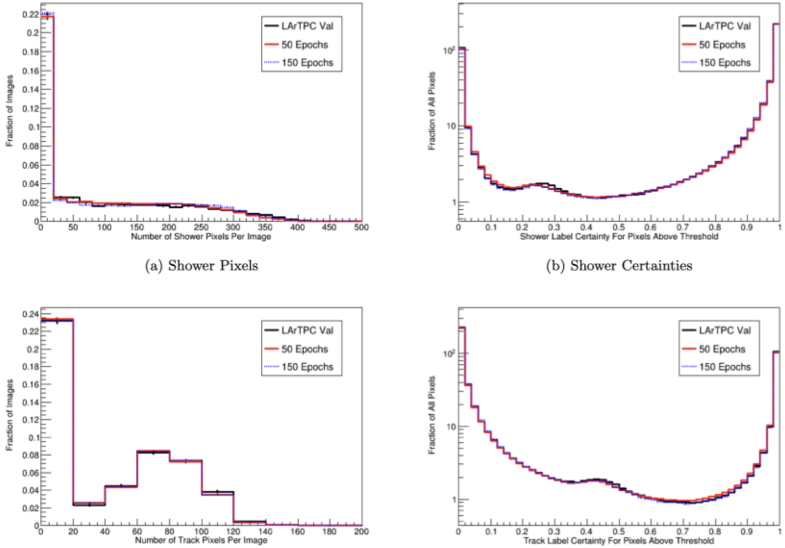
Simulation: Diffusion Models

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

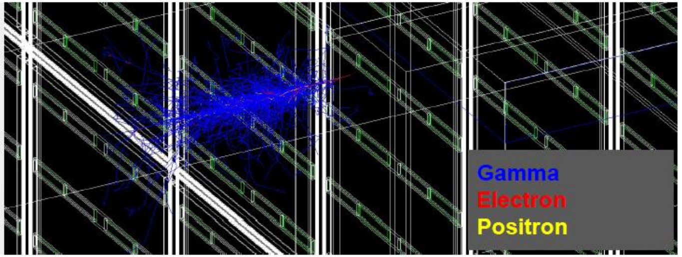
(c) Track Pixels (d) Track Certainties

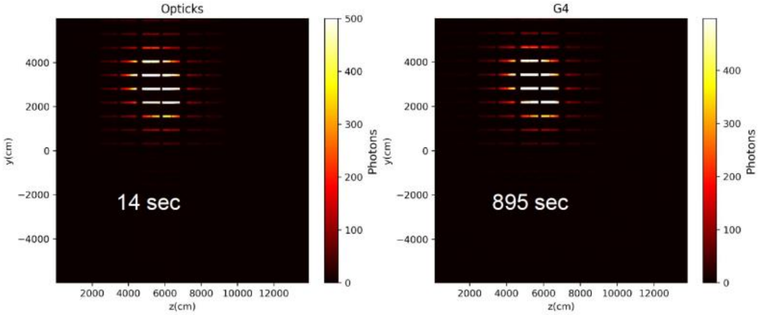
Good agreement between Simulation and generation

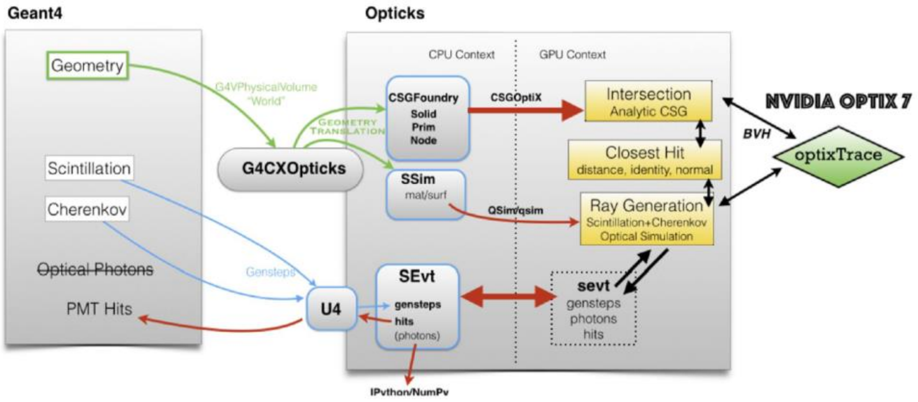
GPU-Accelerated Simulation

- Accelerated optical photon simulation via GPU ray tracing
- Opticks: open-source GPU-accelerated optical simulation using NVIDIA OptiX engine with Geant4 integration
- Having access to high-fidelity (full optical simulations) will allow us to develop and test ML approaches, without GPU acceleration this is impractical
- ~100×speedup transforms optical simulation from bottleneck to enabler for simulation-based inference and ML

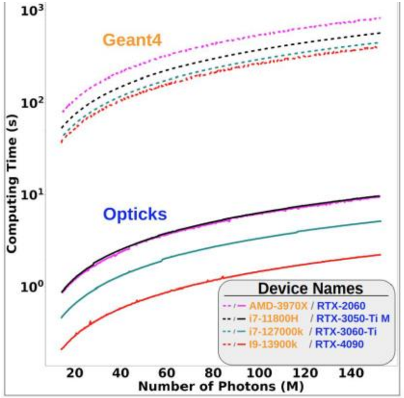
Aaron Higuera, I. Parmaksiz, etc (Rice University)







https://doi.org/10.1051/epjconf/202533701093



Eur. Phys. J. C 85 (8) 910 (2025)

AI for Detector Calibration & Systematics

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)

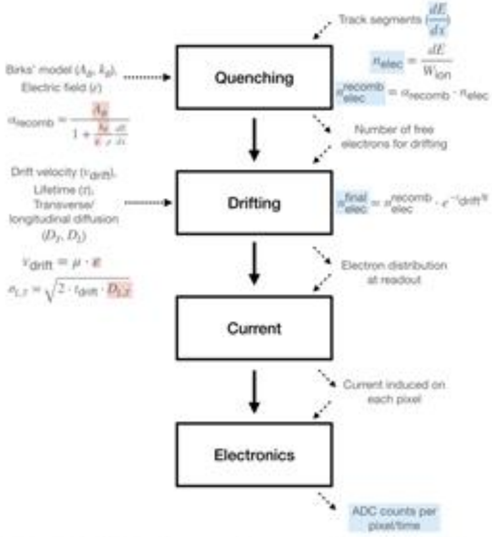


Figure 1. Flow diagram of the simulator, highlighting inputs and outputs of each stage (blue) as well as commonly calibrated model parameters (red).

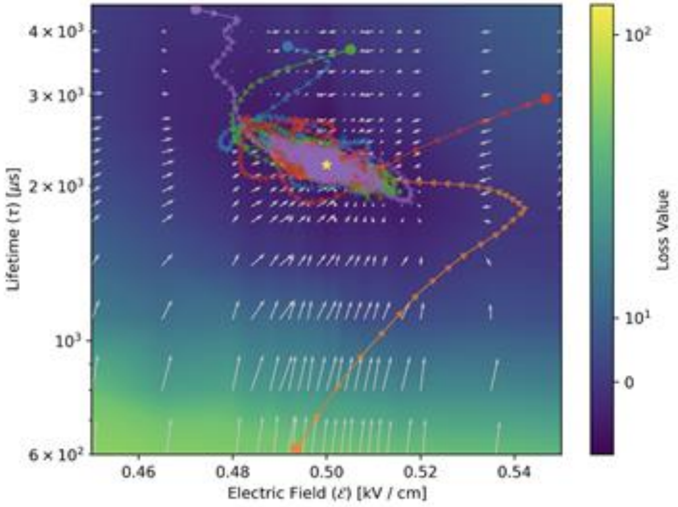
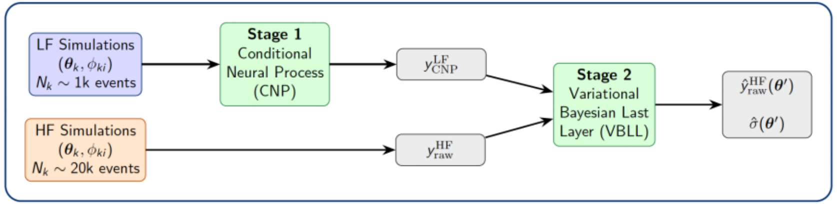


Figure 9. Loss landscape in a 2D parameter space of electric field $\mathcal{E}$ and lifetime τ, averaged across batches. The gold star labels the target parameter values. The negative gradient, shown in the white arrows, points towards the loss landscape minimum at the target. Five example fit trajectories, 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).

AI for Detector Calibration & Systematics

- DetSuM: An uncertainty-aware AI surrogate simulation framework based on the RESuM framework [arXiv:2410.03873].
- Models detector uncertainties such as electron lifetime and charge recombination parameters from sparse simulation samples.
- Reduces the computational cost of DUNE rare-event sensitivity studies.
- Demonstrated on the DUNE neutron–antineutron oscillation analysis.

DetSum Architecture: Two-Stage Pipeline



Stage 1 – CNP: event-level regression. Maps $(\theta_k, \phi_{ki}) \to \beta_{ki} \in [0, 1]$ (calibrated signal probability per event). Aggregated across all events to get $y^{\mathrm{LF}}_{\mathrm{CNP}}$.

Stage 2 – VBLL: parameter-level regression. Maps $\theta \to (\hat{y}, \hat{\sigma})$ using LF as dense prior and HF as ground truth anchor.

Uncertainty Quantification, Visualized

Visually, our band plot is:

DUNE Work in Progress

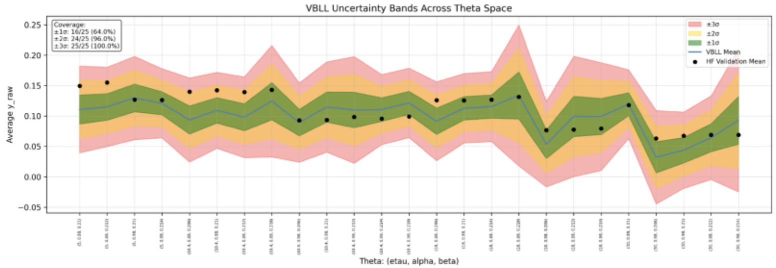
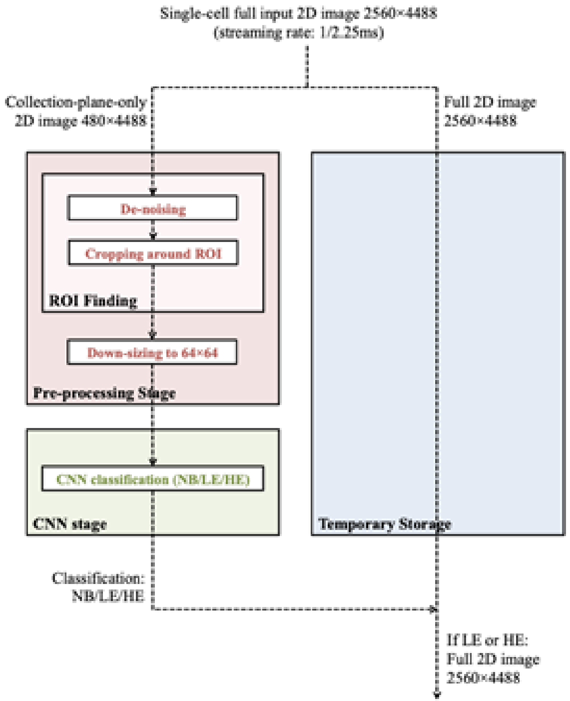


Figure: Statistical coverage of uncertainty predictions on HF validation data. Parameter space configurations θ organized by increasing β', then α, then τ_e.

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



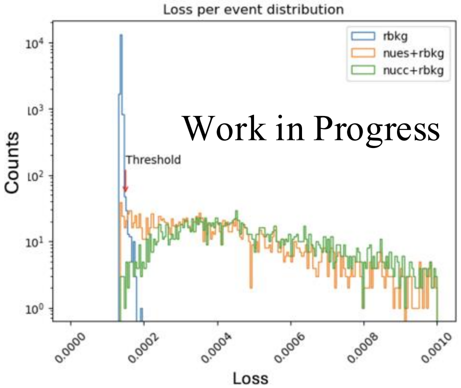
[Y.j.Jwa, G.Di Guglielmo, L.Arnold, L.Carloni and G.Karagiorgi, Front. Artif. Intell. 5, 855184 (2022), arXiv 2201.05638]

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%)

- Anomaly Detection for Supernova and BSM Triggers using a CNN-based Autoencoder
- Goal: FPGA implementation



Sunny Seo, FNAL

Trigger & DAQ

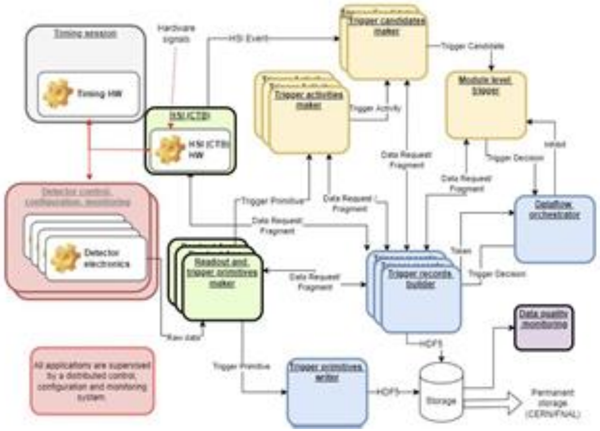
Software Trigger ML Algorithms on DAQ Cluster

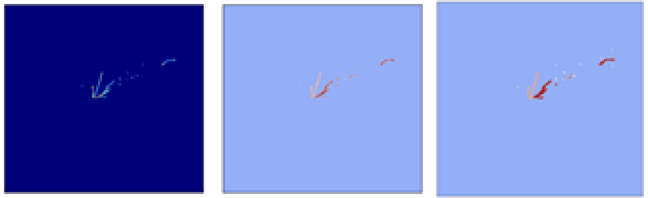
Sparse Convolutional Neural Network for DUNE Supernova Trigger :
S. Damish1 and M. Bhattacharya,
https://lss.fnal.gov/archive/2024/pub/fermilab-pub-24-0524-student.pdf

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



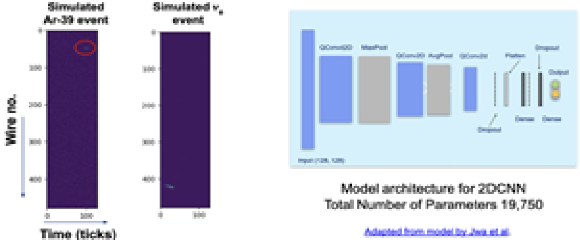


(a)ADC. (b)Ground truth. (c)Model prediction.

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

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!



Algorithm	Performance	GPU Inference Time for Optimal Batch	FPGA Inference Time and Resource Summary
2D CNN for Ar-39 background rejection	99.42% signal efficiency, 99.69% background rejection	80 us per image	850 us per event LUT: 5.22% BRAM: 18.95% Register: 4.26% DSP: 8.17*
1D AE for electronics noise reduction	~69% primary track efficiency ~99% background rejection	0.061 sec/nuCC event 0.031 sec/eES event	See J. Mitrevski's Fast ML talk from 2023.
Neutrino Interaction Identification LBP + Dense Model	AUC: 0.85 83% eES->eES	Not measured	Not implemented

Fermilab

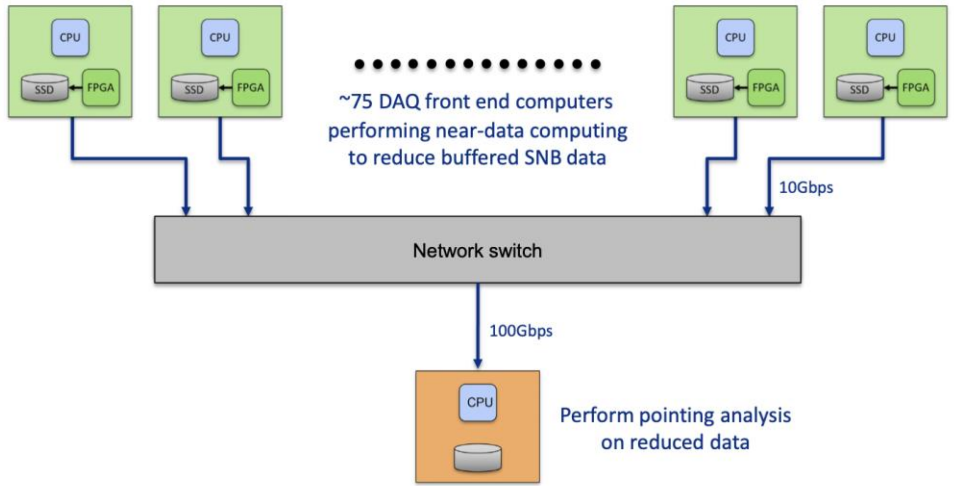
Trigger & DAQ

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

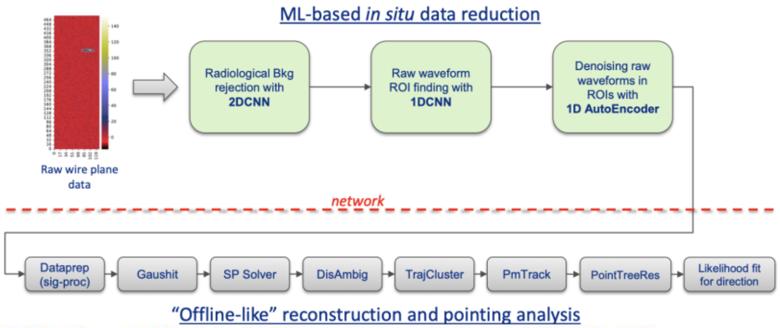
- In-storage AI deployed at the data buffering/storage layer (SSD + FPGA/AI accelerator)
- Performs early ML-based data reduction to enable early supernova pointing information

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

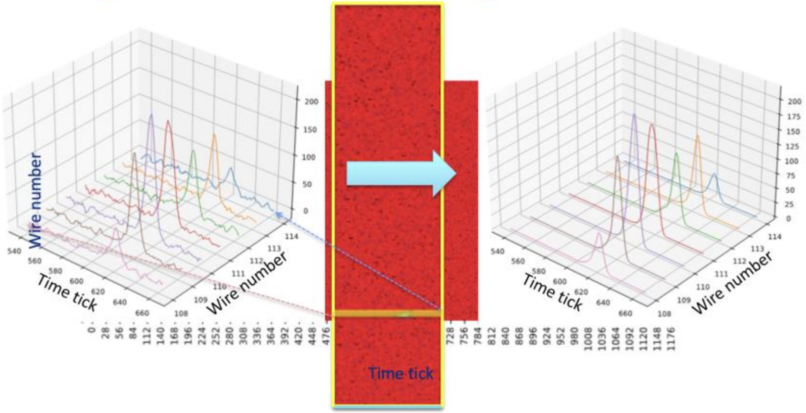
Strategy for fast pointing determination: hardware



Strategy for fast pointing determination: algorithms



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

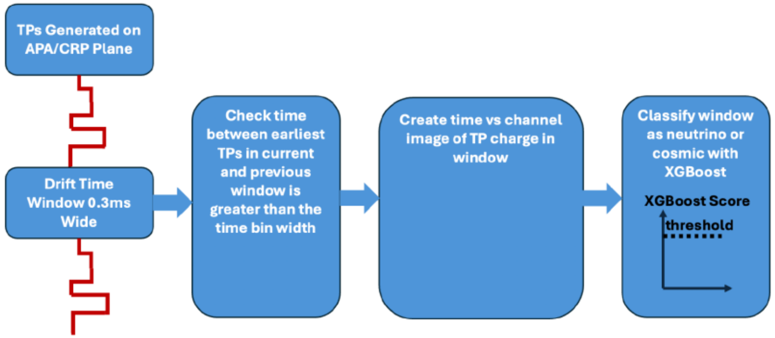


ML Trigger Application in ProtoDUNE

- XGBoost-based online neutrino/BSM event selection.
- Real-time deployment with sub-μs CPU inference.
- Improved neutrino–cosmic separation over conventional triggers.
- Demonstrates AI integration into DUNE TDAQ.

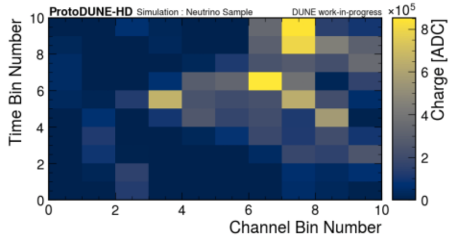
New ML-Based Trigger Algorithm for ProtoDUNE

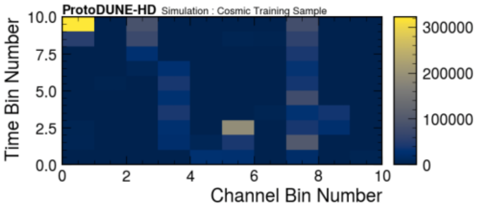
- Provide **low-resolution real-time views of activity in detector to an ML model**
- Updated algorithm: removed filtering on total charge step



Choice of model: XGBoost

- XGBoost: framework for building and training **gradient boosted decision tree models (GBDT)**
- Important consideration - **speed**!
- ProtoDUNE DAQ has strict latency requirements – **NNs slow** on CPUs
- Trained XGBoost model is **exported** (via Treelite → TL2cgen) to **compilable C++**
- XGBoost model compiled as an object in the DUNE DAQ software for fast **sub-microsecond inference**



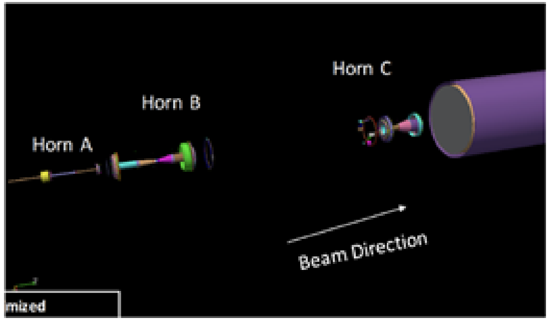


Ciaran Hasnip

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



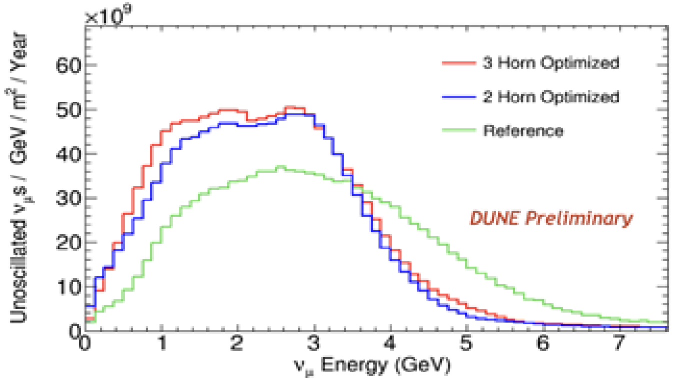
Optimized Design

Three horns, not similar to NuMI, run at 300 kA

2.2 m long carbon target

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

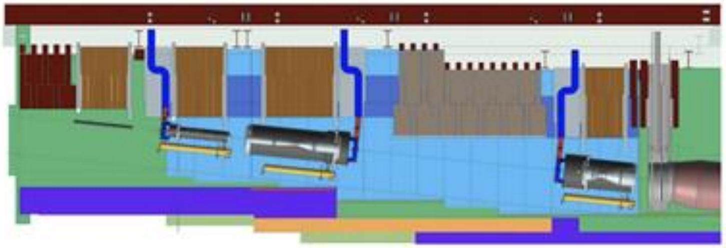


FIG. 2. An illustration of the upstream portion of the LBNF neutrino beamline. A horn-protection baffle, three focusing horns, and the decay pipe are shown inside the target chase respectively from left to right (the beam direction).

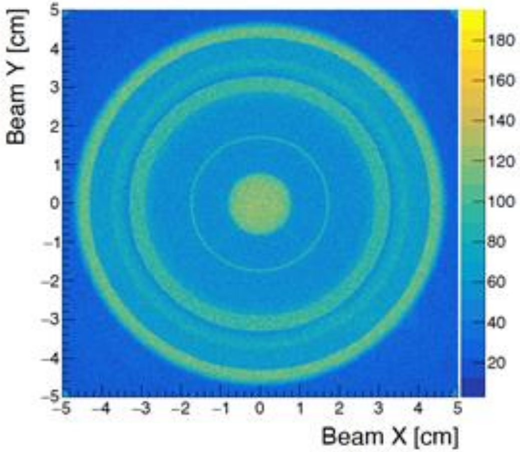


FIG. 6. Two-dimensional horizontal vs. vertical proton beam positions with uniform distribution sample for recorded beam interactions on the LBNF target that has a neutrino candidate at the downstream neutrino detector.

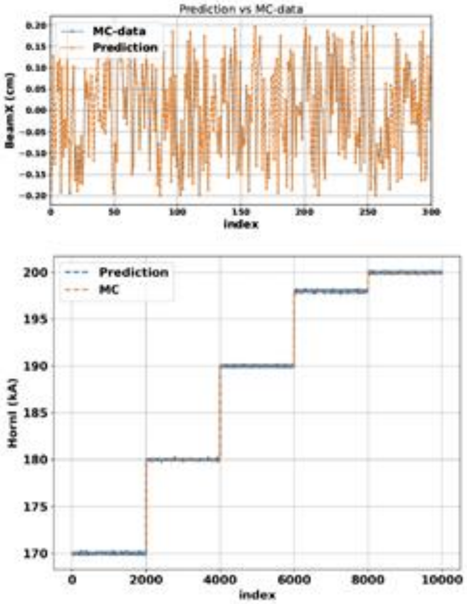


FIG. 29. Top: Horizontal beam position prediction, bottom: horn current prediction using a linear regression model.

AI-driven Real-Time Beam Monitoring

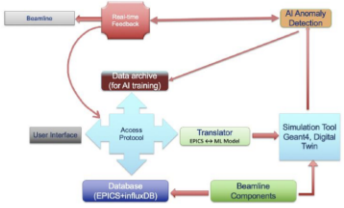
- ML-based beam diagnostics with muon monitors for spill-by-spill inference of beamline parameters.
- Physics-informed digital twin for real-time beam monitoring and anomaly detection.
- Surrogate modeling and Bayesian optimization for efficient beamline simulations.
- AI-assisted beam control and next-generation beam instrumentation for LBNF/DUNE.

- Sudeshna Ganguly, https://lss.fnal.gov/archive/2025/slides/fermilab-slides-25-0182-ad.pdf

ML Architecture and Data Sources

- **Inputs:**
  - 2D muon charge maps from MuMS (Muon Monitor System)
  - Beamline metadata (e.g., horn currents, POT, ambient sensors)
- **Targets:**
  - Predict beam centroid, width, horn current shifts
  - Infer pion focusing conditions and energy spectrum shape
- **Models Used:**
  - **Fully Connected Neural Networks (FCNN):** Fast, real-time regression
  - **Linear regression:** For interpretable, low-latency estimates
  - **Autoencoders:** Anomaly detection via reconstruction loss
- **Purpose:**
  - Infer unmeasured beam parameters from downstream observables
  - Detect instabilities or drifts
  - Enable corrective actions through feedback loop

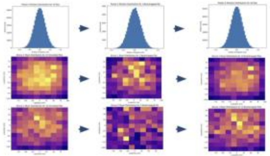
- **Physics-informed digital twin**
  Uses G4LBNF & Synergia simulations combined with real-time muon monitor data to model pion production and focusing dynamics.
- **Muon profiles reconstruct pion phase space**
  Muon monitor centroids correlate with pion kinematics; energy-dependent slopes encode focusing performance and horn current effects.
- **Spill-by-spill prediction enables correction & suppression**
  ML models (e.g., FCNN) infer beam parameters (x_0, y_0, σ_x, σ_y, I_{horn}) per spill, driving real-time feedback and reducing flux uncertainty from ~2-3% to <1%.



Beam Monitoring System Architecture: Real-time data acquisition, AI-driven anomaly detection, and feedback correction loop

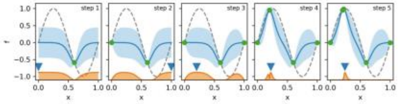
Enhancing Simulations with Bootstrapping & Bayesian Optimization

- Simulating many beamline conditions (e.g., target misalignments, horn current shifts) is computationally costly
- High-dimensional input space limits how finely we can sample
- **Bootstrapping:** Expanding Training Data
- Used to artificially generate additional training samples by resampling existing simulations
- Helps diversify the dataset and improve model generalization
- Bootstrapping helps stabilize learning in low-statistics regimes



Bridget O'Brien
CCI Summer Intern, 2025

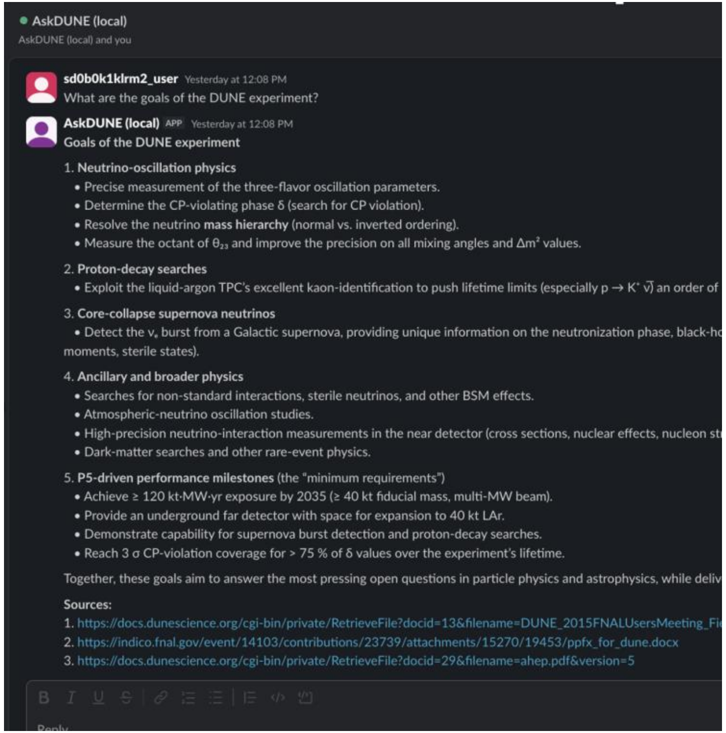
Bayesian Optimization: Prioritized Simulation Sampling
•Learns a **surrogate model** of beam response using existing simulation results
•Uses an **acquisition function** to identify the most informative next simulation point
•Balances **exploration** (high uncertainty) and **exploitation** (high model variance)
•Guides simulation to regions that improve ML training or constrain physics predictions
•**Minimizes the number of expensive simulations** while maximizing statistical impact
•Ideal for **targeted exploration** of high-dimensional input spaces (e.g., beam offset × horn current)



Jason St. John, Fermilab,
US-JAPAN Collaboration Meeting, 2025

LLM: AI-Assisted Knowledge System for DUNE (AskDUNE)

- Scalable RAG/LLM pipelines for large scientific document collections
- Automated ingestion + metadata extraction for DocDB/Indico
- Workflow optimization on leadership-class systems (Balsam, Aurora)
- Current efforts:
  - Backend service now lives on DUNEGPVM
  - AskDUNE Slackbot development for the front-end
  - Working on developing conversational interface



Aleena Rafique

Multimodal AI: Vision-Language Models for Classification and Explanation

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

https://www.nature.com/articles/s42005-026-02688-3

Prompt

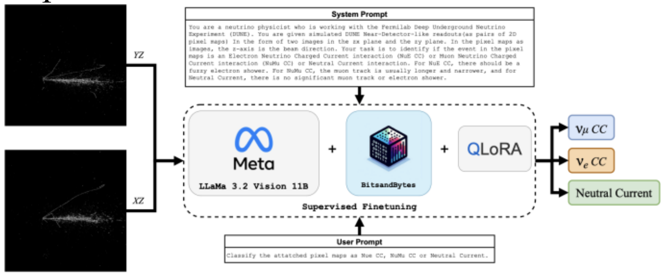


Figure 1: LLaMa 3.2 Vision finetuning pipeline.

Performance

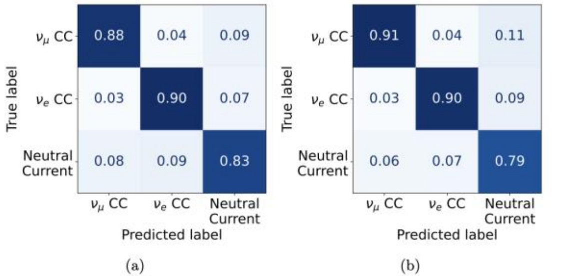


Figure 5: Finetuned LLaMa 3.2 Vision's (a) recall matrix (truth normalized) and (b) precision matrix (prediction normalized).

- ViT-level classification performance
- Robustness to different detector configuration, foundation model potential
- Natural-language explanations
- Human–AI interaction for event analysis

Dikshant Sagar, Kaiwen Yu, Alejandro Yankelevich, Jianming Bian, Pierre Baldi

Generalization

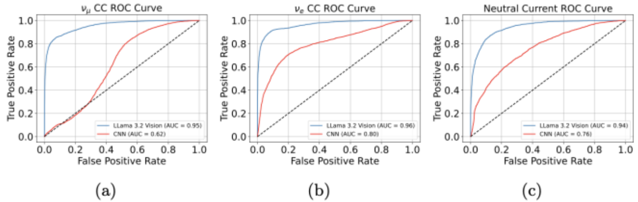


Figure 10: ROC curves for each class (a) ν_μ CC, (b) ν_e CC, and (c) NC comparing performance between the finetuned LLaMa 3.2 Vision and the CNN for generalization testing.

Interpretability

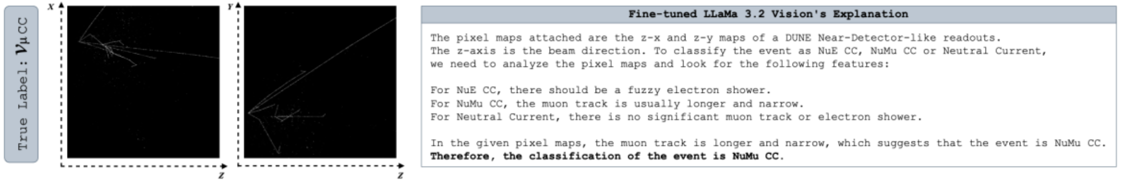


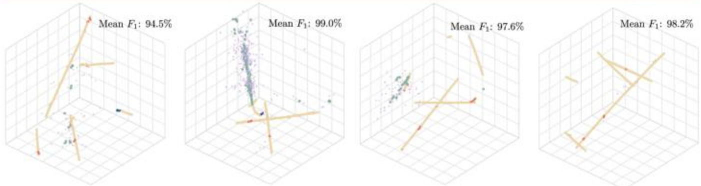
Figure 3: LLaMa 3.2 Vision Prediction Explanation.

Foundation Models

Toward Foundation Models with HEP Sensory Dataset

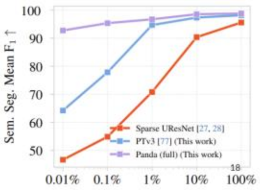
Task A: Semantic Segmentation

Samuel Young (Stanford)



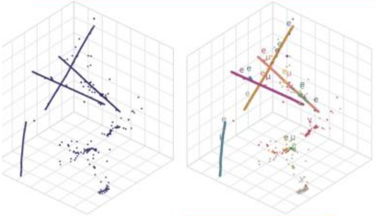
Semantic Segmentation	$PT_{K\%} + FT_{K\%}$					$PT_{100\%} + FT_{K\%}$				
Method / $K\%$	0.01%	0.1%	1%	10%	100%	0.01%	0.1%	1%	10%	100%
Supervised										
○ UResNet [27, 28]	46.6	54.8	70.8	90.4	95.6 =	46.6	54.8	70.8	90.4	95.6
• PTv3 [77]	64.2	77.8	94.7	97.3	98.2 =	64.2	77.8	94.7	97.3	98.2
Self-supervised										
• Panda (lin.)	79.1	90.1	93.2	93.7	93.9	92.2	93.6	93.8	93.9	93.9
• Panda (dec.)	83.5	92.8	95.9	97.3	98.2	92.6	95.2	96.2	97.8	98.2
• Panda (full)	**85.2**	**93.7**	**96.4**	**97.7**	**98.8**	**92.8**	**95.4**	**96.7**	**98.5**	**98.8**

○ Prev. SOTA • This work



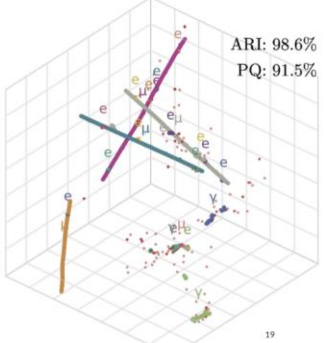
Task B: Clustering Pixels Into A Particle + Type ID

Samuel Young (Stanford)



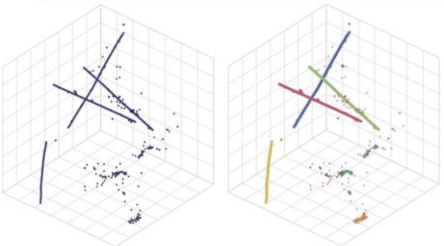
Instance Seg.	Param.	Interaction Id.		Particle Identification						
Method		PQ	ARI	PQ	ARI	γ	e	μ	π	p
Supervised										
• PTv3 [77]	95.1 M	96.3	93.7	86.9	96.6	93.5	94.9	98.6	94.2	98.1
Self-supervised										
• Panda (dec.)	4.4 M	96.6	94.5	89.5	97.3	96.2	96.3	99.1	95.7	98.4
• Panda (full)	95.1 M	**97.6**	**94.9**	**92.5**	**98.0**	**98.4**	**97.2**	**99.3**	**96.0**	**98.6**

• This work



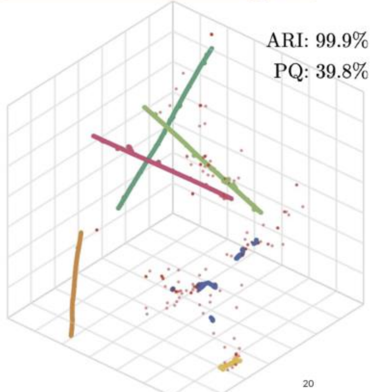
Task C: Clustering Particles Into an Interaction

Samuel Young (Stanford)



Instance Seg.	Param.	Interaction Id.		Particle Identification						
Method		PQ	ARI	PQ	ARI	γ	e	μ	π	p
Supervised										
• PTv3 [77]	95.1 M	96.3	93.7	86.9	96.6	93.5	94.9	98.6	94.2	98.1
Self-supervised										
• Panda (dec.)	4.4 M	96.6	94.5	89.5	97.3	96.2	96.3	99.1	95.7	98.4
• Panda (full)	95.1 M	**97.6**	**94.9**	**92.5**	**98.0**	**98.4**	**97.2**	**99.3**	**96.0**	**98.6**

• This work



The first step toward sensor-level FMs in HEP

- Mask-based self-supervision: PoLAr-MAE
- Autoregressive token modeling: FM4NPP

  Efforts so far focus on foundation across tasks.

  Interest/effort in cross-experimental FMs.

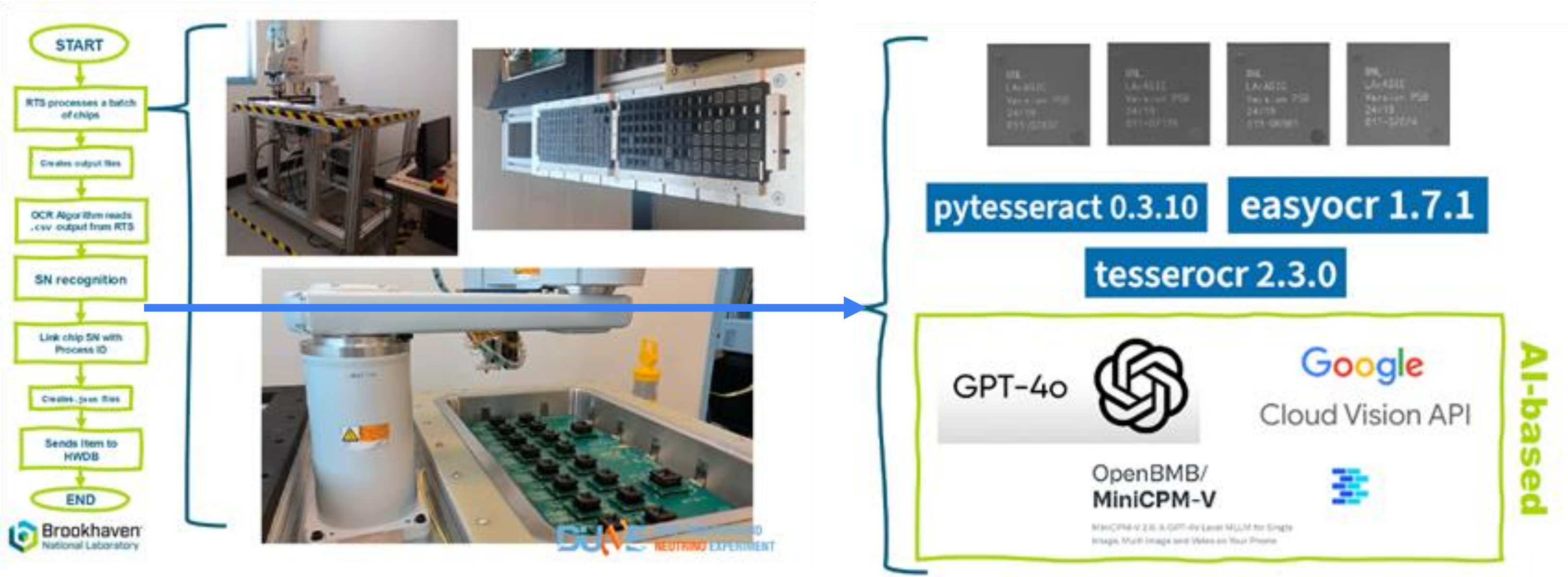
Kazuhiro Terao, SLAC

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.

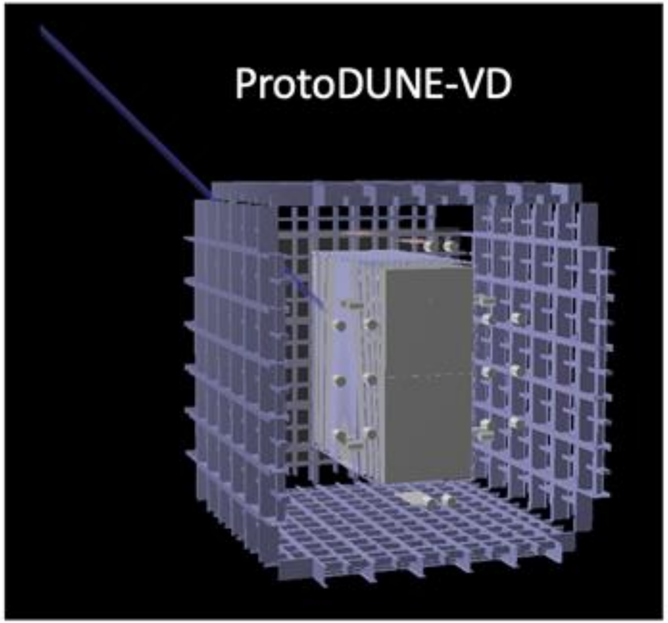


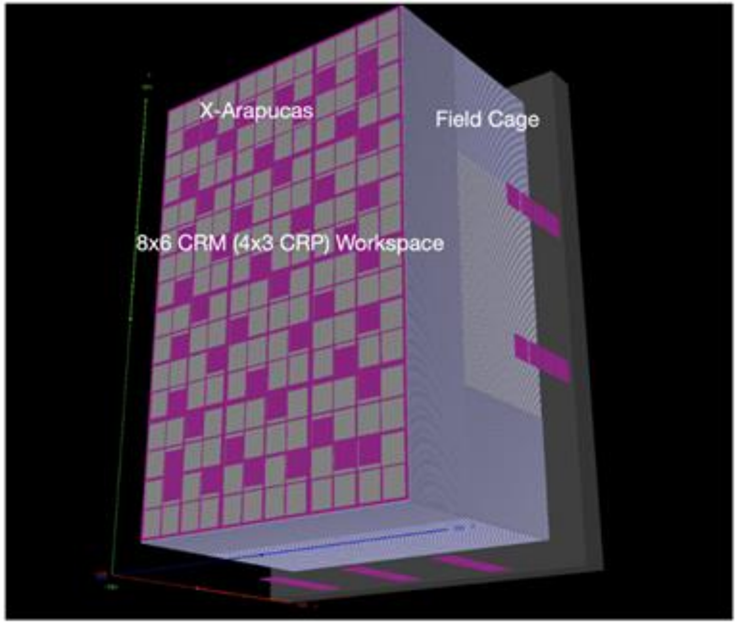
Kieran Wall, BNL

AI-assisted Code Generation

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

DUNE Tech. Prelim.





Nitish Nayak, BNL

AI Infrastructure

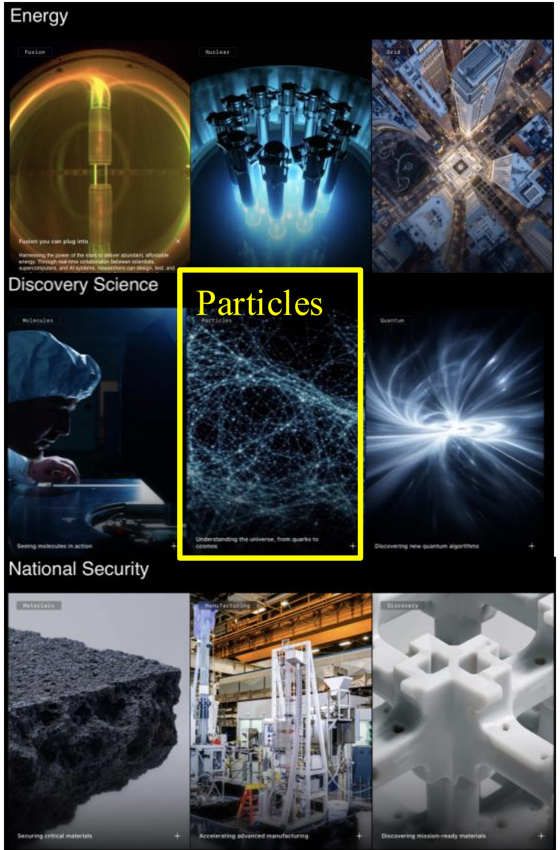
- National Lab HPC Facilities:
  - Fermilab EAF: Elastic Analysis Facility (EAF)
  - SLAC Shared Scientific Data Facility (S3DF)
  - ANL Polaris
- Distributed HTC GPU Clusters: FermiGrid, QMUL, OSG etc
- National HPC accessed via HEPCloud: NERSC/Perlmutter
- Other potential HPC Centers: ORNL, Commercial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

Genesis Mission

The Genesis Mission is a DOE led initiative aimed at coordinating and accelerating the integration of Artificial Intelligence with scientific research to drive transformative discoveries and maintain U.S. leadership in science and technology.

https://genesis.energy.gov

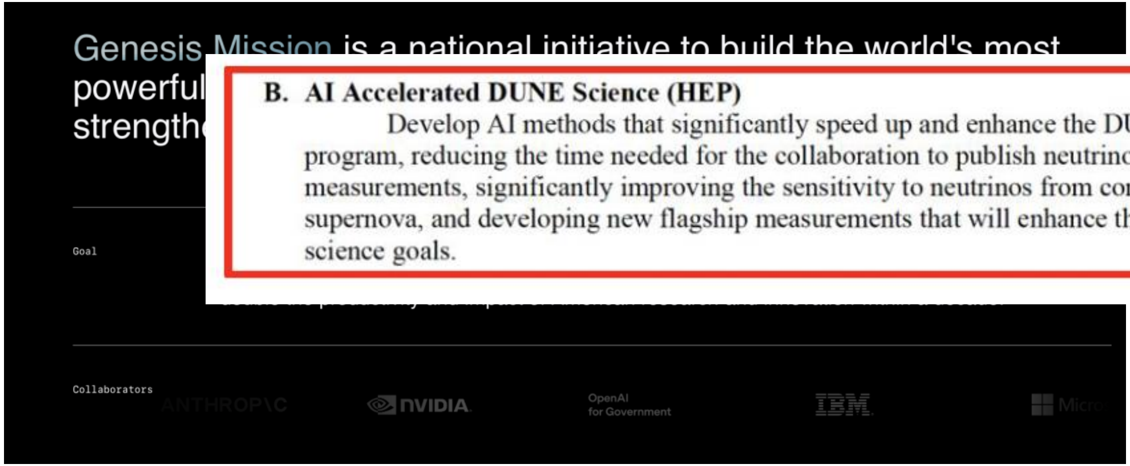


Genesis scope still under definition, with strong emphasis on rapid progress and early deliverables

Genesis Mission

The Genesis Mission is a DOE led initiative aimed at coordinating and accelerating the integration of Artificial Intelligence with scientific research to drive transformative discoveries and maintain U.S. leadership in science and technology.

https://genesis.energy.gov

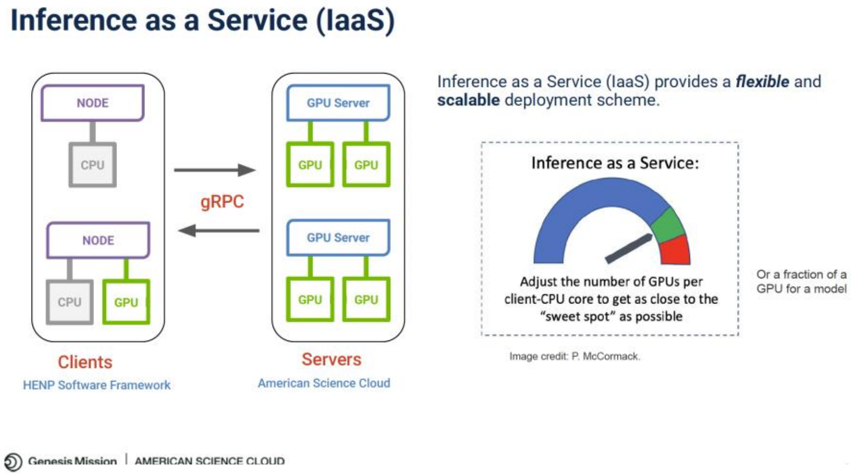


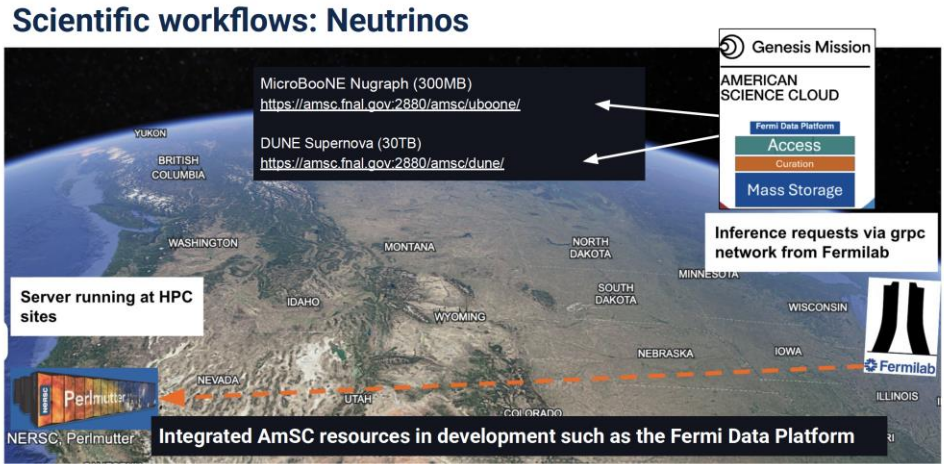
B. AI Accelerated DUNE Science (HEP)

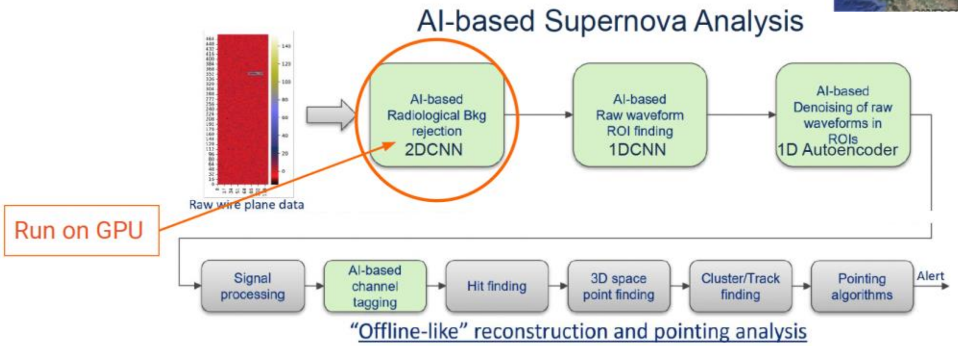
Develop AI methods that significantly speed up and enhance the DUNE science program, reducing the time needed for the collaboration to publish neutrino oscillation measurements, significantly improving the sensitivity to neutrinos from core-collapse supernova, and developing new flagship measurements that will enhance the DUNE science goals.

Genesis scope still under definition, with strong emphasis on rapid progress and early deliverables

DUNE as a demonstrator for American Science Cloud under the Genesis Mission







- **DUNE ML-based Supernova offline reconstruction**:
    - ***IaaS advantage***: *~6X faster*

Inference as a Service for HEP & NP data analysis

Meghna
on behalf of the Scientific
User Facilities IaaS team

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
- Provides a natural platform for the Genesis Mission