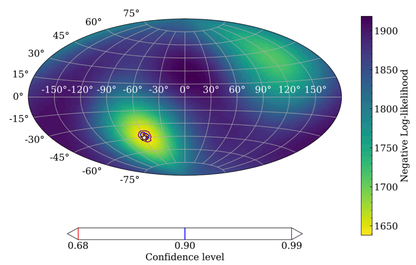
Scalable Low-Energy Event Reconstruction in DUNE Using NuGraph and Inference-as-a-Service

Meghna Bhattacharya, Michael Wang (Fermilab) for the DUNE Collaboration

FERMILAB-POSTER-26-0029-CSAID

Science motivation

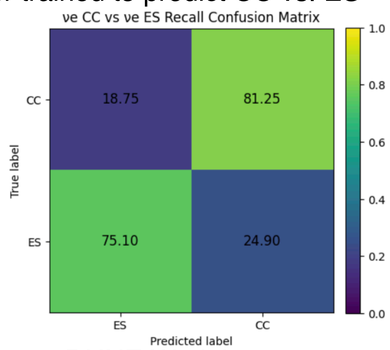
- One of the major physics goals of DUNE is to detect the next galactic supernova neutrino burst
- The challenge is to turn 140 X 4 TB data into discovery as fast as possible for optical follow-up



- An end to end ML based reconstruction pipeline for low energy neutrinos
- Reduce algorithmic complexity and suitability for real time low energy neutrino detection
  - Controlled memory and inference costs

Preliminary Performance: Nugraph

- Network training and inference performed on Fermilab's Elastic Analysis Facility (Nvidia A100 GPUs)
- NuGraph3's event decoder trained to predict CC vs. ES
- Decoder's output is a set of simple softmax probability scores for CC, ES



DUNE Work in progress

- The GNN achieves high recall for νe CC events (81%), while νe ES events (75%)

Inference as a Service

- A scalable, maintainable way to accelerate inference in production and ensure reproducibility

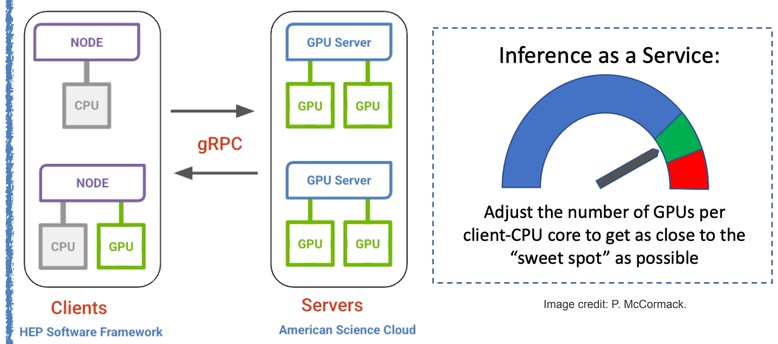
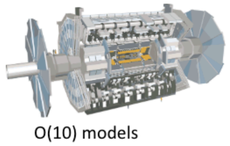


Image credit: P. McCormack.

- Inference as a Service (IaaS) provides a flexible and scalable deployment scheme

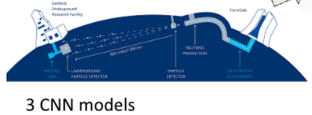
AI/ML demonstrators developed as part of American Science Cloud

Input: Detector data



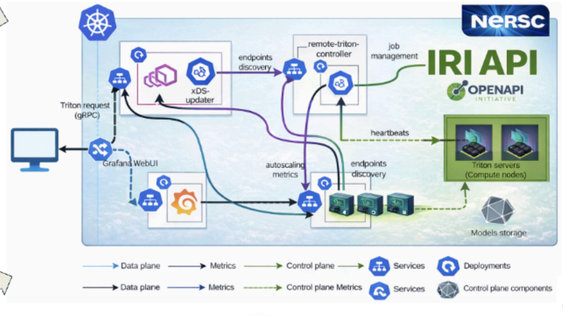
O(10) models

GNN model

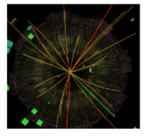


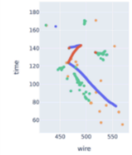
3 CNN models

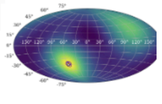
AI Model Inference: Triton Server

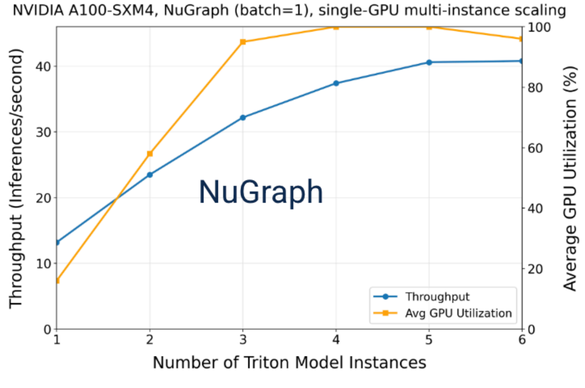


Output: Physics information









- We benchmarked hosted ML models by varying the IaaS server configurations
- The IaaS allows us to use GPUs at 100%

Next Steps

- First demonstration of NuGraph for low energy neutrinos
- Train NuGraph with TPC and PDS information to help with background rejection
- Leverage multiple user facilities such as ALCF, OLCF for IaaS
- Integrate federated identity into the remote triton controller, add authentication token to inference requests
- Optimize IaaS performance in terms of reducing latency and increasing throughput
- Register metadata of existing HENP datasets to data catalog
- Leverage data APIs to manage large AI models

This work was produced by Fermi Forward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics