Needs of the Intensity Frontier Experiments (excluding Dune)

Lisa Goodenough

FNAL AI Infrastructure Planning Mini-workshop

6 April 2023

Mu2e

- **Historically** they have used Root TMVA in many places

- **More recently:** most uses of TMVA have migrated to TensorFlow training using Python outside of art - they have been running these workflows at NERSC
  - Use ROOT SOFIE to write inference code in C++ that they can call in our code.
  - Examples: hit classification; tracker pat-rec; calibration of tracker T-to-D; track quality estimation
  - Mostly simple dense-layer networks; experimenting with conformal NN for event classification

Mu2e

- **In the future:** They expect that new opportunities for AI/ML will arise as the experiment transitions to commissioning, operations and analysis.
  - No definite plans at this time.
  - Possibilities include:
    - More powerful AI/ML where is it used now.
    - Tagging events as background candidates (both signal-like and sideband events)
    - If trigger farm hardware is refreshed during the long shutdown (starting Jan 2027), they could consider GPU-rich options to support more AI/ML and algorithms that better exploit GPUs. Limiting factor will be physicists to develop algorithms.
    - Mu2e-II community is looking at AI in the trigger and for pattern recognition.

- **Have used CPUs exclusively thus far and have no plans at this time for high volume, GPU-rich resources**

Muon g-2

- Muon g-2 has no plans for using AI in any large-scale way in their operations or production workflows.
- They do not foresee any need for specific resources for AI.

NOvA

- **Currently have two different production workflows that use AI:**
  - **Standard Reconstruction:**
    - Runs a handful (5-10) lightweight networks, MobileNet-based CNNs and LSTMs
    - These are typically run via the TensorFlow C++ interface, all on CPUs
    - Even without accelerated inference, these networks make up a minority of the runtime for reconstruction
    - Would likely not change this workflow regardless of new AI-focused resources since the time savings wouldn’t justify the effort.

  - **Cosmic Filtering:**
    - This workflow runs a larger ResNet18 CNN, and that inference task is the primary focus of the workflow.
    - Currently running on ALCF’s ThetaGPU machine so they can use GPU-accelerated inference.
    - Use a local client-server setup (communicating via FIFO pipes) where 8 GPUs handle inference for 128 simultaneous ART jobs.
    - Could conceivably move this workflow to AI-focused FNAL hardware, but it would hinge on being able to use the Balsam workflow management system on those machines.

NOvA

- **Cosmic Filtering Resource Usage**
  - Process is connected to the NOvA “freight train” workflow for prestaging several different datasets tape-by-tape
  - If it all worked perfectly smoothly, it would run continuously (with some fractional duty factor)
  - Cosmic dataset being processed is >1 PB - too large to prestage and store somewhere all at once
  - Also are filtering NUMI trigger files, which are numerous but minuscule
  - Currently using 2-3k node-hours (1 node = 8 GPUs, 128 CPUs) per 6 months - would use more if everything ran smoothly
  - Once they are done with the back processing (this summer?) it will be much less as they transition to just keep-up processing of new files.

NOvA

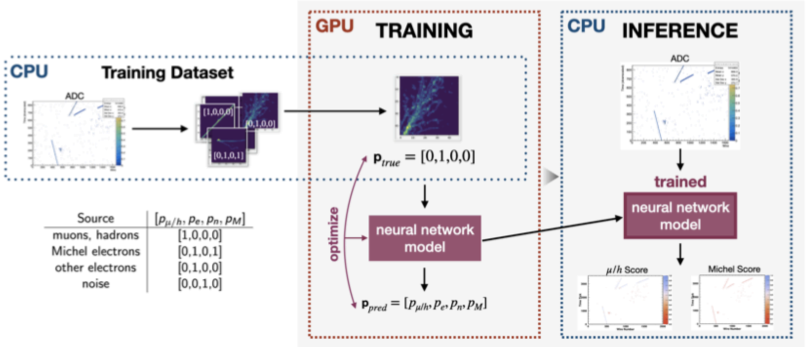
- **Future Workflows:**
  - Exploring some future architectures (sparse, graph-based, multi-function networks), which could really benefit from future FNAL hardware investments.
    - Some of the people involved are already pursuing zmq messaging to communicate with a separate python-based inference process, even if all CPUs are on the same node.
    - It’s not a big leap from there to instead communicate with a remote server using zmq.
    - The same folks are also involved in some LAr AI work, so there’s potential for synergy of tools.

SBND

- ML-based tools from computer vision and pattern recognition are very useful for LArTPC experiments and research is actively ongoing
- SBND is starting to work on integration of ML-based tools
- Currently working on porting and adapting useful ML packages from other experiments
  - "Hit Classification with CNN" from DUNE
  - ”Neutrino Interaction Classification with CVN" from DUNE
  - Additional workflows being investigated

SBND

Workflow Example



Source	$[p_{\mu/h}, p_e, p_n, p_M]$
muons, hadrons	[1,0,0,0]
Michel electrons	[0,1,0,1]
other electrons	[0,1,0,0]
noise	[0,0,1,0]

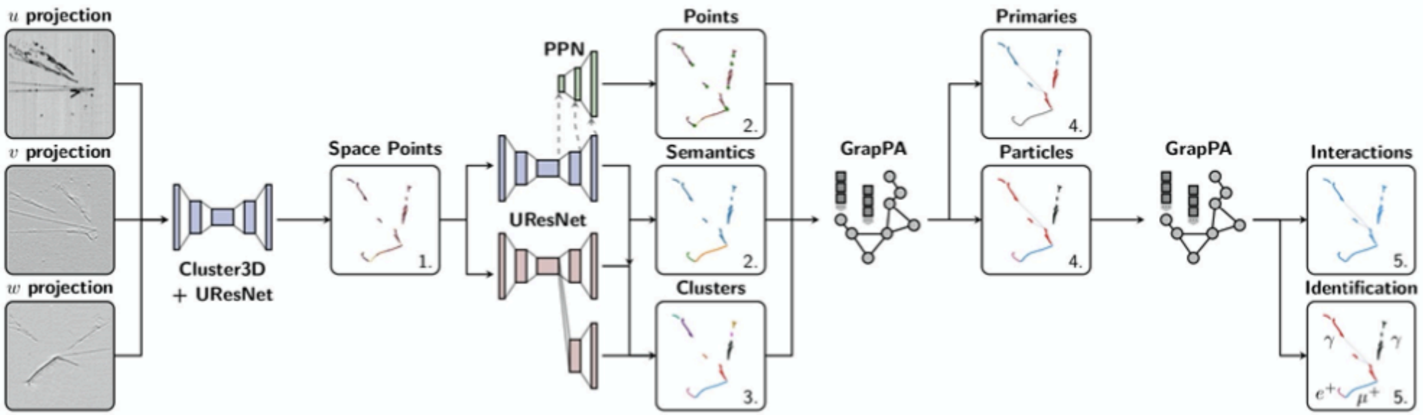
- Hit Classification with CNN Workflow
  - Package lives in LArSoft
  - Task: Classify hits according to source of energy deposition and identify hits from Michel electrons, using only local information
  - Resources: CPU on FermiGrid, GPU on EAF @ Fermilab (NVIDIA Ampere A100, 20GB memory)
  - Unique tool comes with straightforward integration:
    - code runs in LArsoft, easy to add to the current production format – ex. additional branches in a CAF file
    - Inference doesn't require GPUs - easy to add to current workflow, with an API for freezed neural network
  - Supports analyzers on PID, complementary to Pandora

SBND

- SBND is working on ML-based tools that can be added on to the existing infrastructure, as well as novel methods.
- Want to plan things out in advance for an efficient pipeline later on - they are thinking about this now!
  - Eager to learn from other experiments
  - ML in production chain is still a work in progress
  - resource requirements unknown at this time

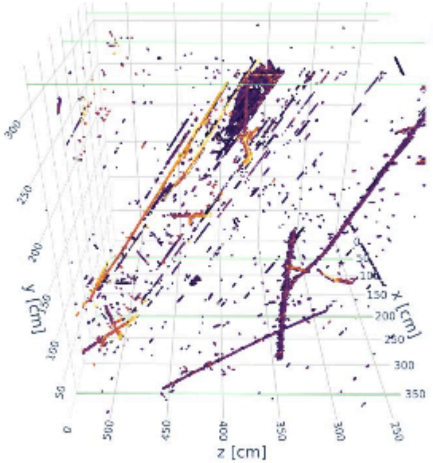
ICARUS

- ML-based reconstruction chain
  - 3D space point building (T. Usher) + artifact removal + charge rescaling (Cluster3D + **CNN**: UResNet)
  - Voxel semantic classification, point identification (**CNN**: UResNet+PPN)
  - Dense clustering (DBSCAN + **CNN**: Graph-SPICE)
  - Particle aggregation, primary identification (**GNN**: GrapPA-Track/Shower)
  - Interaction aggregation, particle identification (**GNN**: GrapPA-Interaction)

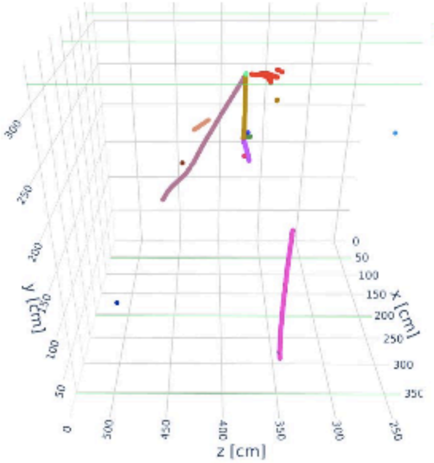


ICARUS

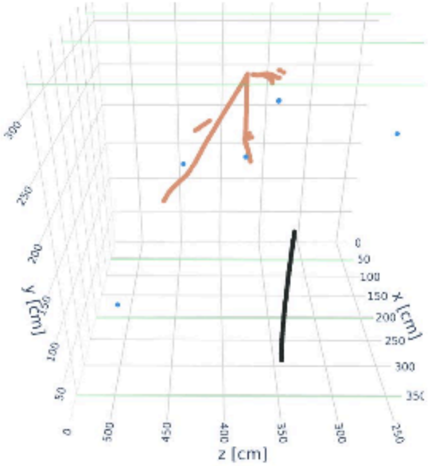
- The ML reconstruction chain outputs high-level description of LArTPC images:
  - List of **interactions** (= slices): 1 per neutrino, 1 per cosmogenic particle and its daughters
    - For each interaction: **vertex**, list of **particles**
    - For each particle: set of charge deposition voxels, particle identification, primary identification, energy



Run 7924, Event 4966, TPC EW



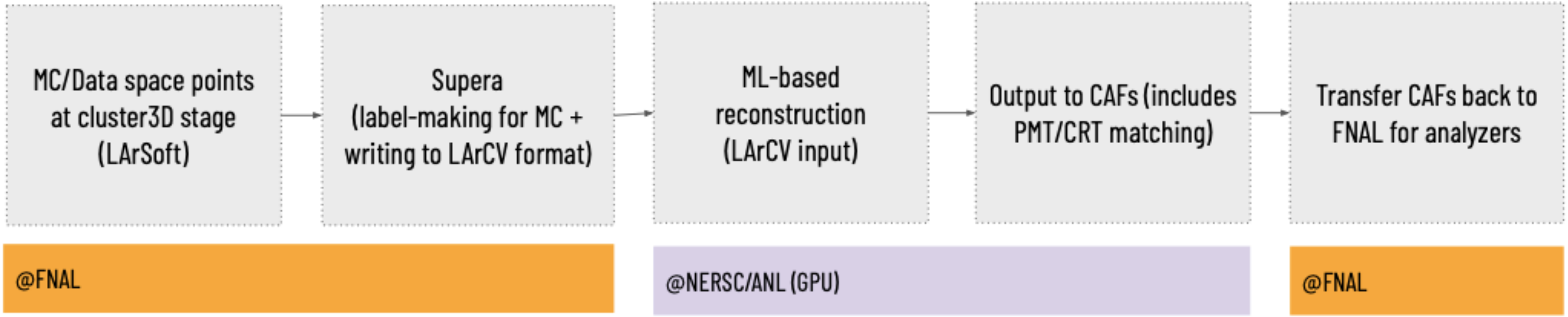
Reconstructed particles



Reconstructed interactions

ICARUS

- ML Reconstruction Strategy Goals:
  - **Share most of the pipeline** for MC production & data processing with non-ML reconstruction chain
  - Convert LArSoft space points to LArCV (input to the reconstruction chain) and make labels, **as part of stage1**
  - **Transfer LArCV files from FNAL to a GPU HPC cluster (NERSC or ANL)**
  - Run reconstruction + flash matching/CRT matching on GPU HPC cluster



MC/Data space points at cluster3D stage (LArSoft)

Supera (label-making for MC + writing to LArCV format)

ML-based reconstruction (LArCV input)

Output to CAFs (includes PMT/CRT matching)

Transfer CAFs back to FNAL for analyzers

@FNAL

@NERSC/ANL (GPU)

@FNAL

SBN Summary

- Two approaches are very different:
  - SBND is integrating the networks in LArSoft and executing them from there (on CPUs)
  - ICARUS exports information in dedicated files, transfers them offsite and runs on GPUs at HPC. Then they need to transfer back the results.
- May ultimately want the best of the two approaches:
  - a way to run any custom network from LArSoft (but without the overhead of integrating the network code)
  - and the option to run it either on CPU or GPU. So basically nuSONIC or a similar solution.

Summary for IF Experiments

My reading of information I was given:

- NOvA and SBN experiments will be drivers of AI needs in the future.
- NOvA is satisfied with current resources. Would likely do more with AI given the personnel to develop the code and workflows.
- SBN Experiments, ICARUS and SBND, are actively pushing AI usage but are unsure of needs at this time.
- Mu2e may do more with AI during and after the long shutdown starting in January of 2027.