Data Preservation at MINERvA

Rob Fine

Snowmass Neutrino Frontier 06
Neutrino Cross Section Data Usage and Archival Workshop

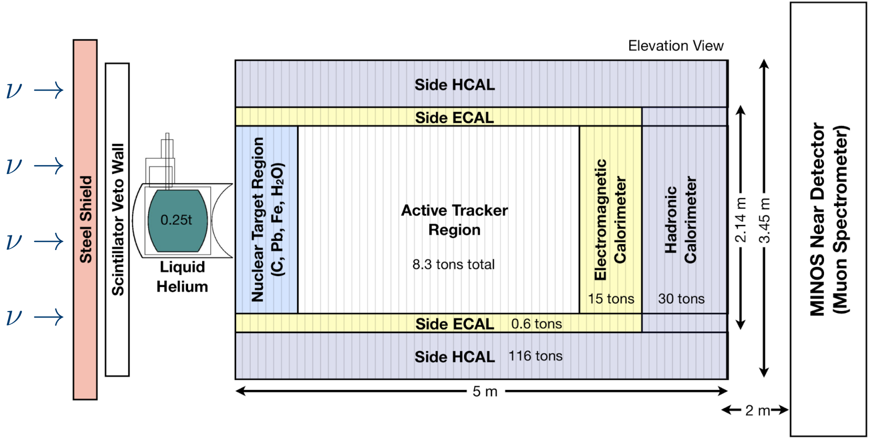
4 September 2020

UNIVERSITY of ROCHESTER

Observations from Day 1 of this Workshop

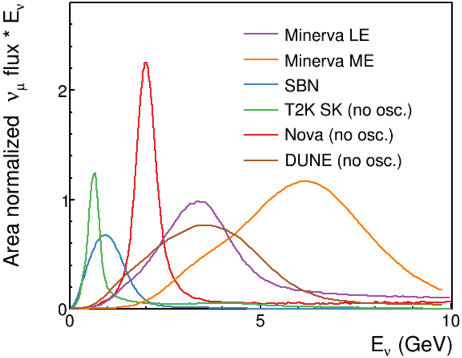
- We routinely revisit neutrino interaction data as our methods and understanding evolve
- We know from precedent that historical data remain relevant, though our ability to interpret them diminishes
  - *"Many of us have participated in physics archaeology"*
- What steps can we take now to maintain the utility of modern data 10+ years in the future?
  - *"[Our measurements] will likely be used in ways we can't predict"*
  - *"Future-proofing data is hard"*

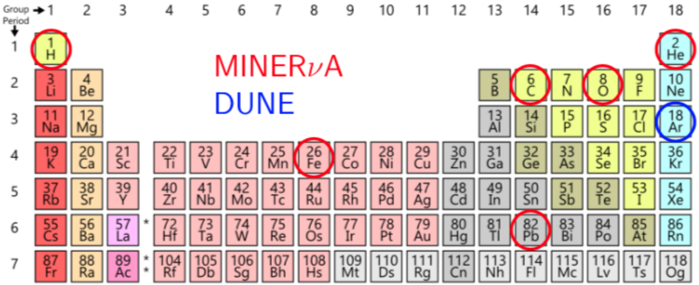
MINERνA



ν →
ν →
ν →
ν →
Steel Shield
Scintillator Veto Wall
0.25t
Liquid Helium
Elevation View
Side HCAL
Side ECAL
Nuclear Target Region (C, Pb, Fe, H_2O)
Active Tracker Region
8.3 tons total
Electromagnetic Calorimeter
15 tons
Hadronic Calorimeter
30 tons
Side ECAL 0.6 tons
Side HCAL 116 tons
5 m
2.14 m
3.45 m
2 m
MINOS Near Detector (Muon Spectrometer)

MINERvA Data are Relevant to DUNE





- High statistics at intermediate and high Q^2
- Measurements on multiple nuclear targets

Plot courtesy of L. Fields; Illustration courtesy of Wikipedia

MINERvA Data Support Many Measurements

- MINERvA data can be categorized by the configuration of NuMI:
  - "Low Energy" (LE) and "Medium Energy" (ME)
  - Forward horn current (FHC; ν_μ dominated)
  - Reverse horn current (RHC; $\bar{\nu}_\mu$ dominated)

	LE	ME
ν_μ	4.0	12.1
$\bar{\nu}_\mu$	1.7	12.4

Protons on target ($\times 10^{20}$)

	LE	ME
ν_μ	$\gtrsim 300K$	$\gtrsim 4M$
$\bar{\nu}_\mu$	$\gtrsim 50K$	$\gtrsim 2M$

CC-Inclusive Interactions

	LE	ME
ν_μ	22	$2 + ??$
$\bar{\nu}_\mu$	10	??

σ measurements

- MINERvA works to expand our understanding of neutrino interactions
  - High-dimensionality measurements (2D inclusive, 2D CCQE-like, 3D CCQE-like, $\langle ... \rangle$)
  - Cutting-edge analysis techniques (Nuclear binding energy, $\nu - e$ scattering, TKI, $\langle ... \rangle$)

This is not a talk about how MINERvA data will be useful to the community for the next 10+ years...

(But you should invite us to give that talk, too...)

This is a talk about how MINERvA is taking steps to keep our data accessible for the next 10+ years

Looking Forward

- MINERvA is no longer taking data
- The number of analyses undertaken by the collaboration will begin to exponentially decay starting in 2021
- We are working to preserve our data so that it remains usable by the community in the medium-term future
- This will also enable MINERvA to continue its analysis program in the near-term future as person-power declines

LOI submitted to Snowmass CompF07 and NF06

Data Preservation for Neutrino Physics

- In our view, there are two critical components to a successful data preservation campaign:
  - Access to the data
  - Infrastructure to analyze the data
- While historical data may technically be available, there is no precedent within modern neutrino physics of *infrastructure* to support its re-analysis

Data Preservation at MINERvA

- The MINERvA data preservation project consists of three components:
  1. Preservation of MINERvA data into a single ROOT tuple that incorporates low- and high-level reconstructed objects
  2. The MINERvA Analysis Toolkit (MAT) – a broadly applicable HEP software toolkit for calculating systematic uncertainties using tuple objects
  3. A software package built on the MAT for reproducing MINERvA published results, which includes templates for performing new analyses.
- Component (1) provides access to the data, and components (2) and (3) provide the infrastructure to analyze the data

Component 1: The Tuple

- Historically, MINERvA analyses have each employed their own tailored ROOT tuples, using a powerful, but cumbersome, framework
  - Enables parallel development of distinct reconstruction techniques
  - Decentralizes the production of analysis tuples
  - Requires storage of many times the total data set
- The MINERvA data preservation project will utilize a unified tuple
  - Summarize the reconstruction for a broad variety of final states
  - Support a large number of analyses
  - Smaller disk footprint
  - Obviates the need to maintain the tuple-producing framework
  - Includes low-level reconstruction objects that could, in principle, be used for novel reconstructions
- This will include all LE/ME, FHC/RHC data and simulation

Aside: Balancing Priorities

- We cannot perfectly preserve the ability to do *anything* with our data
- One of the original goals that won't be achieved is developing a mechanism to generate new events in our preserved tuples
- We will maintain support for event reweighting, but this will be restricted in practice to the generated phase space
- Something to think about: Which (new) corners of phase space will be important in the future?

"[Our measurements] will likely be used in ways we can't predict"

"Future-proofing data is hard"

Component 2: The MINERvA Analysis Toolkit

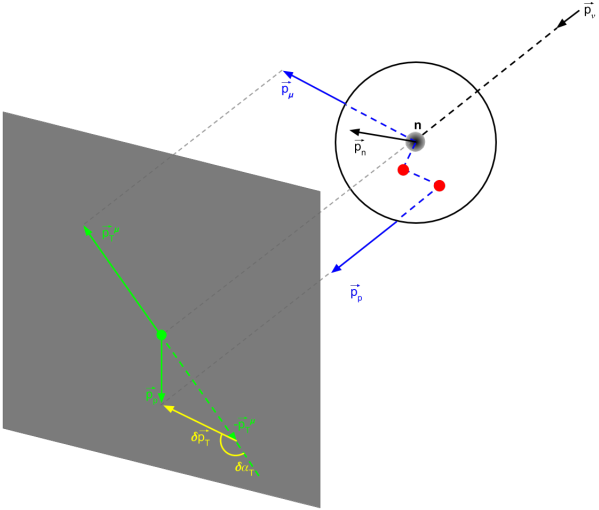
- Software toolkit for performing physics analyses
- Emphasis on handling of systematic uncertainties
  - Central role in analysis flow
  - Transparent treatment
  - Flexible, modular
  - Centralized, standardized across experiment
- Uses customized version of ROOT's TH{1,2}D classes – "MnvH{1,2}D"
  - Handles all histograms corresponding to various systematic universes
  - Propagates systematic variations through histogram manipulations
  - Facilitate straightforward extraction of systematic uncertainty, correlations, etc.

Component 3: Analysis Scripts

- We plan to provide scripts that will reproduce MINERvA analyses using only the preserved data tuples and the MAT
- Users will have the flexibility to modify any aspect of existing MINERvA analyses, or to design completely new analyses (within reason)
- Users will have automatic access to the complete suite of common systematic uncertainties
- It will be straightforward to incorporate new reweighting schemes (e.g. to test future interaction models)

Example: Transverse Kinematic Imbalance

- Relatively new technique useful for probing intranuclear effects
- Utilizes novel kinematic variables
- Evidently historical data cannot be re-analyzed at the level of calculating new variables
- We strive to support future novel analysis techniques through:
  - Access to low-level reconstruction objects
  - Complete flexibility in the design of event-loop analysis



Link to Transverse Kinematic Imbalance paper

Computational Requirements

- We expect the resources needed to analyze MINERvA data in the near-term to be relatively small
  - $\mathcal{O}(10)$ TB total disk footprint to store data
  - $\sim$ 1 hour to process entire ME FHC data set with complete systematics (using $\mathcal{O}(100)$ FermiGrid nodes)
- We likely can continue to utilize FNAL resources for future analysis, but...
  - That limits access to the data to FNAL users
  - This solution doesn't necessarily extend to other experiments with much larger disk footprints
  - The access to these resources isn't guaranteed

Timing

- The timeline for making our data available is $\sim$ late 2021
  - There is time to incorporate community feedback!
  - Is our plan consistent with your use-case?
- We expect some of our analysis tools (the MAT) to become available for early adopters in $\sim$ early 2021
  - If this is something you are interested in, reach out to us!

THANK YOU

Points of Contact:

Data Preservation strategy
Laura Fields (ljf26@fnal.gov)
Debbie Harris (dharris@fnal.gov)
Kevin McFarland (kevin@rochester.edu)

MINERvA Analysis Toolkit
Rob Fine (finer@fnal.gov)
Ben Messerly (bmesserl@fnal.gov)

BACKUP