Data Selection Improvement For MicroBooNE

Brayden Dillon

July 2025

Abstract

Data selection is an extremely important part of data analysis for any experiment. Finding a physics result is often the result of sifting through a massive amount of data, keeping data that we believe to be signal, and throwing out data we do not. This process is called data selection. Creating a selection algorithm is an intensive process that must balance keeping enough data to have statistics and maximizing the signal purity of that data. We also need to choose the right reconstruction method, a tool to take raw data from the detector and convert it into physics results. In this study, we used three different reconstruction tools, Pandora, WireCell, and LANTERN, for the MicroBooNE experiment in conjunction to improve the selection algorithm for analysis. For the case of this study, we look into the charged current N proton 0 pions (CCNp0π) interaction channel. This is the dominant channel for the Short Baseline Neutrino (SBN) program and is expected to be a large contributor to the Deep Underground Neutrino Experiment (DUNE). We first investigated each of the three tools to find out more about their strengths and weaknesses as reconstructions, and compared them to the truth information directly from the MicroBooNE simulation pipeline. We then put together a direct comparison of the three methods to find which method or combination of methods would return the best result for us. While the study is ongoing, we have learned a lot about data selection for the experiment and the differences between the reconstruction tools.

This manuscript has been authored by FermiForward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

1 Introduction

MicroBooNE is a neutrino detector that operated in the Booster Neutrino Beam (BNB) at Fermilab from 2015-2021. It is a Liquid Argon Time Projection Chamber (LArTPC) with the physics goal of bettering our understanding of neutrino interactions. The liquid argon tank experiences a strong transverse electric field that drifts electrons produced by the neutrino interaction toward three wire planes, where they induce a charge that is read out as data. The BNB is created by bombarding a target with 8 GeV protons and then focusing the products of the collision with a magnetic horn. These products are allowed to decay and collide with a concrete beam stop, which absorbs all particles except for the neutrinos.

The detector sees many events, and many more events are simulated in an attempt to better understand the neutrino interactions. This deluge of data contains both signal and background, differentiated by what we are interested in observing. To sort through this data, we use reconstruction tools combined with selection criteria to narrow down the data into interesting physics events, disregarding background.

Selection efficiency, defined as the number of selected signal events that are generated by Monte Carlo over the total number of signal events generated, ($\frac{N_{sel,sig}}{N_{gen,sig}}$), has been a problem with the current MicroBooNE selection scheme. While this enables us to have high data purity, ($\frac{N_{sel,sig}}{N_{gen,sel}}$), the tradeoff is a lack of statistics in our final data. To find a solution, we need to look into all available reconstruction tools to come up with a better scheme to select signal events. The three selection methods currently in use are Pandora, WireCell (WC), and LANTERN. Pandora is the most widely used of these reconstruction tools, having been adapted by many MicroBooNE analyses as well as other experiments. WC is a different approach, using timing data to form a full three-dimensional image of an event before reconstructing the data. LANTERN is a brand-new method that is being applied to data for the first time by MicroBooNE. It uses machine learning that was trained on Monte-Carlo data to reconstruct events and identify particles.

A new resource that we employed in this study is a file format for MicroBooNE data called a Super-Unified Re-Processing Really Improving Selection Efficiencies (SURPRISE) file. These files contain variables from each of the three main reconstructions, allowing for effective mixing and matching to find the best version of a data selection scheme between the reconstructions. As a first step toward optimal combined selection criteria, I studied the performance of the three reconstruction approaches using the same input neutrino events.

2 Signal Definition

The channel in which my study is interested is the CCNp0π channel. This means we are interested in observing a muon neutrino charged-current interaction producing at least one proton and zero pions. Additionally, our signal definition included limits on the momentum of the outgoing muon ([0.1, 1.2] GeV) and highest momentum, or leading, proton ([0.25, 1.0] GeV). This is motivated by requirements for efficiency, resolution, and systematic uncertainties. The lower bound on the muon momentum ensures that all accepted muon tracks are long enough that selection efficiency becomes appreciable. The upper limit on the muon is due to poor resolution at higher momentum. The momentum constraints on the leading proton are driven by the detection threshold and particle identification criteria of low-energy emission of nucleons from neutrino interactions. More detailed explanations can be found in this MicroBooNE paper [1].

True information about the particles generated in the MicroBooNE simulation chain is analyzed to determine which events meet the signal criteria. The selection criteria discussed in this paper are also applied to the reconstructed particle information to attempt to isolate the CCNp0π signal events from uninteresting background interactions. Trivially, there is also a requirement for signal that the neutrino interaction vertex lies within the detector fiducial volume; otherwise, we would simply not see the event.

3 Selection Criteria

The goal of data selection is to emulate as closely as possible the signal definition criterion using only reconstructed data, in contrast to the truth information by which the signal is defined. This is because we need a very intimate understanding of the selection performance with simulation data before we can apply it to data.

3.1 Original Criteria

The selection criteria that I started with used exclusively Pandora variables and were previously established by the MicroBooNE collaboration. [2]. These criteria can be grouped into three categories: Charged current (CC) inclusive, N protons, and zero pions. This became the template I used to create selection algorithms for the other two reconstruction methods as well. These selections also had vertex coordinate cuts, muon-momentum cuts, and proton-momentum cuts that match the signal definition. For each selection cut, I calculated the efficiency and purity of the remaining data and plotted them cut by cut to find out how each cut affected the data. (Figure 1)

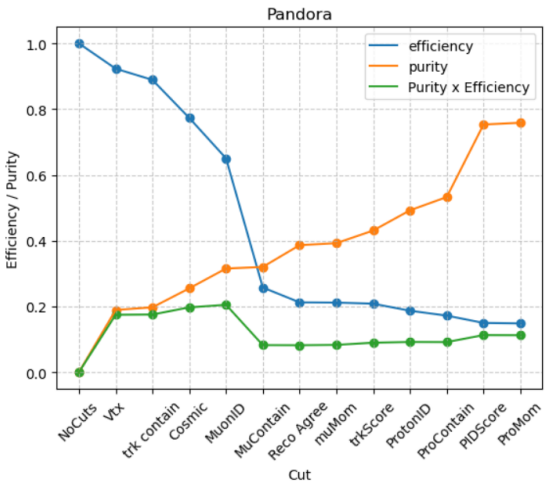

Figure 1: Efficiency and purity plot of Pandora selection cuts

3.2 WireCell Selection

The WC selection criteria are much less strict than the Pandora selection. Where Pandora has over ten cuts to narrow down the data, WC has much more general cuts taking into account only one or two variables. The branches examined for selection ensure that WC reconstructs a CC event with a muon neutrino, at least one proton track of appropriate length, the reconstructed muon track lies within the containment volume established with the Pandora cuts, and that it produces zero pions. The zero-pion cut for WC is a feature that I implemented with the help of Ben Bogart, a graduate student at Fermilab who contributed to the development of WC. The cut involves discarding events with two or more reconstructed muons, as pions are often mis-identified as an additional muon, and events where variables for π^0 meson energy, mass, and opening angle indicate a pion was reconstructed.

WC selection, because it has fewer cuts and cuts on fewer variables than Pandora, has a very different looking efficiency curve (Figure 2).

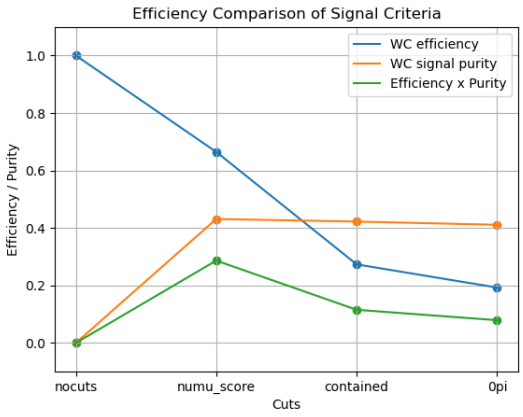

Figure 2: Efficiency and purity of WireCell selection cuts

We can see that while efficiency performs slightly better here, the purity is not good and still needs a lot of work if we want to be competitive with Pandora.

3.3 LANTERN Selection

LANTERN is a reconstruction algorithm that has not been deeply investigated because it is in its infancy. As such, the selection criteria for it are completely new, and I will go into more detail about them here.

For the most part, the cuts are inspired by those used for Pandora. However, the variables used are very different as LANTERN has a more refined particle ID algorithm.

Cut 0: LANTERN has variable `foundVertex` which contains whether or not a neutrino interaction vertex was reconstructed at all. We make sure that all passing events have a vertex accordingly.

Cut 1: Here we use the three variables, `vtx[XYZ]`, which are the Euclidean coordinates of the reconstructed vertex. To ensure that the vertex falls within the fiducial volume below.

$$21.50 \leq x \leq 234.85$$
$$-95.00 \leq y \leq 95.00$$
$$21.50 \leq z \leq 966.80$$

Cut 2: This cut is used as a second check that we are able to see the interaction. It ensures that all primary particles of the neutrino interaction have tracks that begin within the containment volume:

$$10.00 \leq x \leq 246.35$$
$$-106.5 \leq y \leq 106.5$$
$$10.00 \leq z \leq 1026.80$$

The variables used are `trackStartPos[XYZ]` for track start coordinates and `trackIsSecondary` for primaries. Making sure that `trackIsSecondary` is zero for the track ensures that the particle is a primary particle.

Cut 3: Here we apply a cut to exclude cosmic rays from our selection. LANTERN defines a variable `vtxFracHitsOnCosmic` which identifies what fraction of pixels in a track are tagged as cosmic-like. We set the cutoff at 0.1 to be thorough.

Cut 4: This stage cuts out events that do not contain a muon, using LANTERN's native PID variable, `trackPID`

Cut 5: Here we apply a cut ensuring that the end of the muon track is also contained by the containment volume described in cut 2. We use the matching variables `trackEndPos[XYZ]` with the `trackPID` to find the muon.

Cut 6: We apply the muon momentum constraints to the track identified as the muon. LANTERN does not natively have momentum information, so we need to calculate the relativistic momentum using the kinetic energy given by the `trackRecoE` variable and the muon mass: 105.65837 MeV.

Cut 7: LANTERN PID was determined to be fairly good, so we base our decision as to whether an event has a proton on the native PID.

Cut 8: The other proton cut involves the phase-space constraints on the momentum of the leading proton, or the highest momentum. Similar to the muon phase-space cut, we need to calculate the momentum using the proton mass [0.93827 GeV] and the track energy `trackRecoE`

Cut 9: The first pass we make to exclude pions is to exclude events where there is more than one reconstructed muon. As mentioned in the WC selection, it is sometimes the case that a pion will be misidentified as a second muon

Cut 10: Finally, we only pass events where all particle candidates, apart from the muon candidate, are identified as protons. We chose to cut pions in this way because the proton ID for LANTERN seems to perform slightly better than the pion ID.

Cut 11: Here we ensure that there are no showers present in the event. Lantern stores the number

of showers attached to the interaction vertex in the variable `nShowers`, we cut showers by making sure this was equal to zero.

The culmination of the cuts was a competitive performance in both purity and efficiency of the data, shown below (Figure 3). This is a very interesting result to me, as with these selection criteria we are able to bump the purity very high without sacrificing much efficiency.

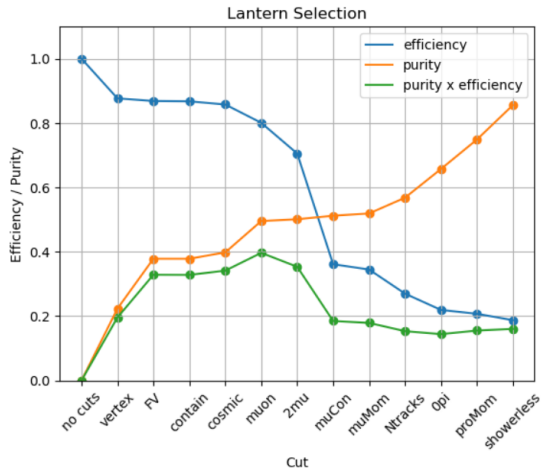

Figure 3: Efficiency and purity plot of LANTERN selection cuts

4 Results and Comparison

It is important to note that for this study, all of the different information used for selection cuts was pulled from one file. SURPRISE files enabled us to have one unified file containing information from all three of the different reconstruction methods. This allows us to construct comparisons directly between the methods and selection schemes. For example we can keep a constant signal definition and compare the way that the different selections perform in trying to isolate it. The relevant comparison between the first three figures is the value of efficiency multiplied by purity, as maximizing this value means we have both high data purity and good efficiency for uncertainty and statistics.

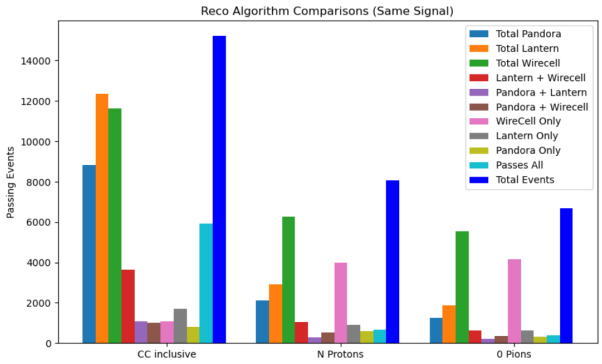

Figure 4: Chart of events exclusive to different combinations of selection schemes

This plot (Figure 4)shows us more explicitly that while LANTERN and Pandora are fairly close together in terms of the number of events that they have exclusively, WC allows for a much higher efficiency in terms of events exclusive to it. This is due to the much less strict selection criteria WC employs, so it naturally follows that there are many events exclusive to WC. We can also use this to compare the combinations of any two of the selection schemes, and possibly expand this in the future to compare selections where we mix and match reconstruction methods.

For further analysis, we ran each of the selection schemes in their entirety for simulated data, beam off, and beam on data to compare simulation to data and as a sanity check in comparison with similar plots that had already been made. To allow for a direct comparison between the data and Monte Carlo simulation, the simulation results were rescaled to match the same beam exposure (measured in protons-on-target or POT) as the data. We also used truth information to determine the category of each event to see in more detail how the interactions work and what backgrounds or signal channels are dominant.

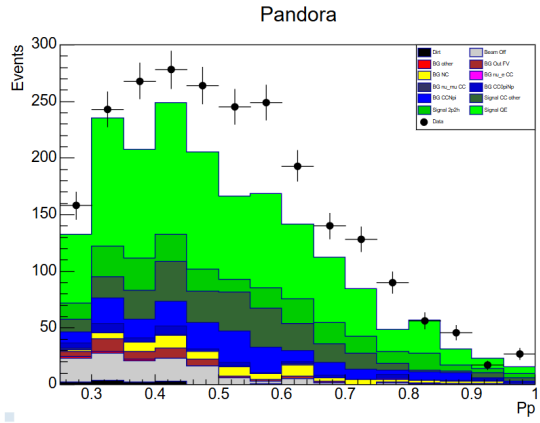

Figure 5: Proton momentum distribution for Pandora selected events

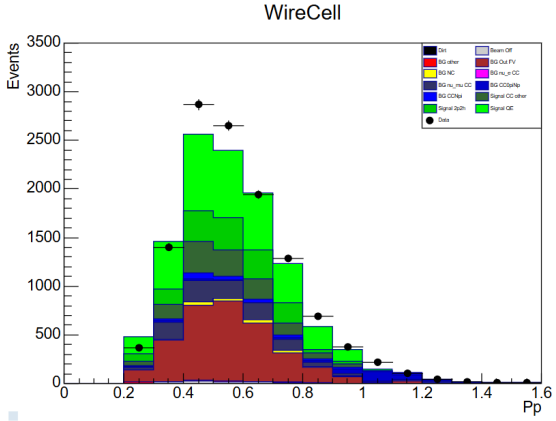

Figure 6: Proton momentum distribution for WireCell selected events

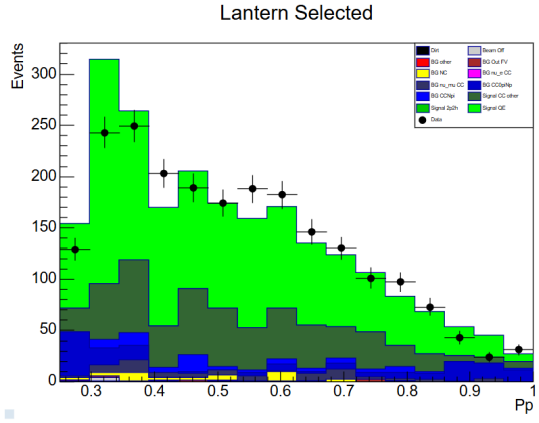

Figure 7: Proton momentum distribution for LANTERN selected events

These plots (Figure 5-7) show the leading proton momentum distribution for the three reconstructions. The different colors represent different classifications of data and allow us to see exactly what the selection schemes are letting in. We can see some obvious differences. None match the data exactly, but they all do to varying degrees with Pandora being seemingly the worst. However Pandora also has the largest ratio of signal to background of the three. Many more plots like these still need to be made before we can come to a definitive conclusion about which of these three is the best.

For a final comparison we can plot the purity and efficiency of each at different stages of the reconstruction process. (Figure 8) This allows a very direct comparison. Each uses the same signal definition in the calculations of purity and efficiency, so we are essentially comparing apples to apples. We can see how each reconstruction evolves the purity and efficiency at a different rate as we make more selection cuts. This is also another showcase of the strength of the SURPRISE files, as we can apply the same signal criteria to each of the reconstruction and selection algorithms with the same data and compare them to see how they perform relative to one another.

5 Conclusion

While no definite result has been achieved in this study so far, there has still been significant progress. We have created the SURPRISE file format and shown here that it can be used effectively and easily to compare the three reconstruction schemes available to MicroBooNE. Using variables from all three reconstructions, we have constructed selection algorithms both in relevant stages and in complete forms. This has never been done before for WireCell or LANTERN, the former having had a selection not including zero pions, and the latter made completely from scratch. This is ongoing research, and in the future we hope to leverage the power of the unified files to create the best possible selection for MicroBooNE. A best-case-scenario selection algorithm means that we will be able to increase our

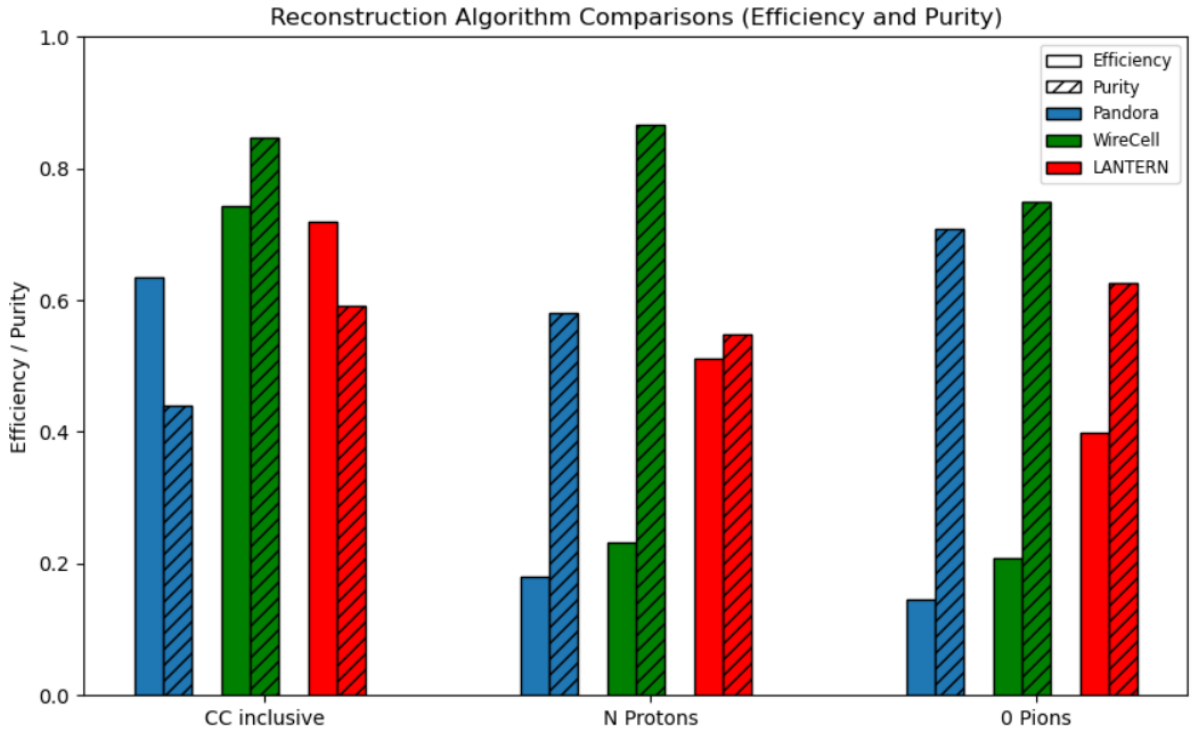

Figure 8: Efficiency and Purity performance at different points along selection process.

sensitivity to rare physics events and could possibly lead to new physics and better understanding of beyond the standard model phenomena. Previously, any analysis had to make a somewhat arbitrary choice of which reconstruction to use and stick with that choice. Now, we know a little bit more about the strengths and weaknesses of each reconstruction, and a more intelligent choice can be made. Alternatively, analysis can utilize a unified selection scheme that incorporates the strongest elements from each reconstruction. Having access to all of the data in a single file is an incredibly powerful tool that MicroBooNE and other experiments will be able to utilize in future analyses, creating a best-case-scenario selection scheme to probe for all manner of physics.

6 Acknowledgment

Thank you to Steven Gardiner, Liang Liu, and Ben Bogart for their help and guidance on this project.

References

[1] S. Gardiner et al. Double-differential measurements of mesonless charged-current 2 muon neutrino interactions on argon with final-state protons 3 using the microboone detector. *Fermi National Accelerator Laboratory*, 2025.

[2] P. Abratenko et al. Measurement of double-differential cross sections for mesonless charged-current muon neutrino interactions on argon with final-state protons using the microboone detector. *arXiv*, 2024.