Improving the Performance and Assessing the Robustness of Machine Learning Reconstruction in Neutrino Experiments

Garrett Kunkler,[1, 2] David Caratelli,[1] and Chuyue “Michaelia” Fang[1]

[1] *Department of Physics, University of California, Santa Barbara*
[2] *College of Engineering, California Polytechnic State University, San Luis Obispo*

(Dated: September 30, 2025)

Neutrinos are chargeless particles that very rarely interact with other matter, making it difficult to detect them and the oscillations between their three flavors. Many open questions remain about neutrinos, including what their masses are and whether they play a role in explaining the matter-antimatter asymmetry in the universe. To detect these particles, we use Liquid Argon Time Projection Chambers (LArTPC) which produce high resolution images of the charge deposits that result from neutrino interactions in the detector medium. Interpreting these images can be improved by implementing machine learning (ML) models, such as NuGraph2, a graph neural network (GNN) that performs categorization on the deposited charge. ML models train on simulated data, but differences between the detector simulation and real data present a challenge when applying these tools. We evaluated the impact of such differences in the detector's charge response on NuGraph2's ability to accurately categorize charge into particles. We found that NuGraph2 presents a detector modeling uncertainty comparable with traditional reconstruction methods, suggesting that this algorithm is robust and can effectively be used on experimental data collected by the detector. Additionally, we trained a prototype ML model to separate between the two photon decays of η and π^0 particles, which could improve the detection of higher-order resonances in neutrino interactions.

I. INTRODUCTION

The Standard Model (SM) of particle physics describes the universe's fundamental particles and how they interact with each other. However, it is well-known that this model is incomplete, and high energy particle physics experiments search for new physics beyond the Standard Model (BSM). The subfield of neutrino physics uses the small and uncharged neutrino to probe the anomalies which have defined this particle's experimental measurements. The neutrino was first proposed in the 1930's as an explanation for the missing energy in atomic beta decay. Once it was first detected in the 1950's, deficits in the experimental neutrino fluxes compared to theoretical prediction spurred an international effort that resulted in the discovery of the three flavors of neutrinos and the phenomenon of neutrino oscillations. Now, the next generation of detectors has been designed to constrain the parameters describing these oscillations well enough to make strong claims about the fundamental nature of neutrinos and BSM physics.

Neutrinos (ν) are chargeless and only interact with one of the fundamental forces included in the SM, the weak force.[1] So, compared to charged particles such as an electron (e) or a muon (μ), neutrinos can only be detected when they interact with matter and create other charged particles. One interesting phenomenon we find with neutrinos regards their three different flavors: the electron neutrino ν_e, the muon neutrino ν_μ, and the tau neutrino ν_τ. They are named based on their corresponding charged lepton (electron e, muon μ, tau particle τ) that often appear in interactions together with each flavor of neutrino. Surprisingly, while neutrinos propagate through space — such as from a neutrino beam to a neutrino detector — their flavor can change [1].[2]

Based on our theoretical models of neutrino oscillations, the experimental discovery of this profound phenomenon requires neutrinos to have non-zero mass. The neutrino was first included in the SM as a massless particle, suggesting some new physics is needed to explain the origin of their mass. Additionally, their masses are on the order of 10^6 times lighter than the next lightest particle, the electron. The mechanism behind these small masses is still unknown, and better understanding this fundamental particle will benefit from extremely precise measurements of neutrino oscillations. Neutrinos could also be the key to explaining the matter-antimatter asymmetry in the universe, as well as many other open questions in particle physics.

A. Neutrino Detectors

Since it isn't possible to directly "image" neutrinos due to their lack of charge, neutrino detectors rely on studying the interactions between a high-flux neutrino source and a large target medium. Liquid Argon Time Projection Chambers (LArTPC) such as the MicroBooNE

[1] The three forces are the electromagnetic, strong, and weak forces.

[2] More technically, each flavor is an observable state of a neutrino, but the mass eigenstates (ν_1, ν_2, ν_3) of the neutrino Hamiltonian are different. Each flavor is a superposition of the underlying mass states, which oscillate between each other and cause the wave function of the neutrino to change as it moves, hence allowing non-zero probabilities that it is measured as a different flavor.

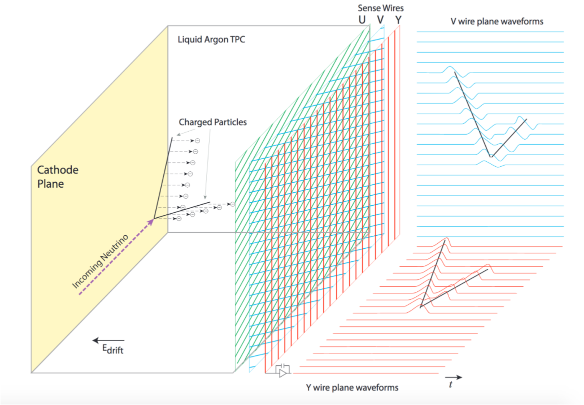

FIG. 1. Diagram of the LArTPC detector and the wire collection system. Copied from [3].

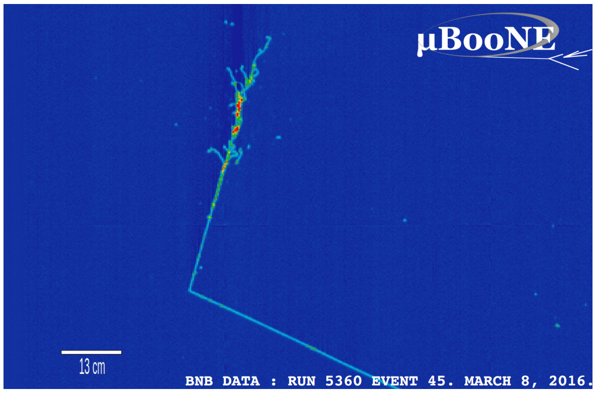

FIG. 2. Example of an electron neutrino interaction in MicroBooNE. The neutrino (unseen) comes in from the left and out of the interactions an electron shower above and a proton below. Red means more charge deposited at that location in the detector.

detector [2] contain tanks of liquid Argon that occasionally interact with the incoming neutrinos. Based on the particles that come out of the interaction, experiments are able to determine which interaction mode occurred and the flavor of the incident neutrino.

In the left part of Figure 1, the neutrino interaction releases two charged particles. As these charged particles move through the detector, they disturb the valence electrons of the Argon. This leaves behind a trail of free electrons and Argon ions in the wake of each particle. The applied electric field in the detector then separates the free electrons from the Argon ions and measures the positions of the electrons as they travel past a series of 3 wires planes. The fact that Argon is a noble gas allows the ionized electrons to be collected before reinteracting inside the detector volume. Moving charges induce current in the wires, so as an electron moves past the first two wire planes and is collected in one of the wires in the third plane, a pulse of current indicates the electron's location. Combining this with the time it takes for the electrons to travel from the initial deposition results in a 3D reconstruction of the neutrino interaction.

This method makes it possible to achieve fine spacial resolution and create detailed pictures of the neutrino interactions, such as Figure 2. With this rich data, the challenge now becomes developing algorithms that can categorize the different particles based on their signature of charge deposition and determine the kinematics of the interaction. Only then can the properties of the electrically neutral neutrino, which doesn't deposit any charge along its path, be deduced.

Great experimental progress over the last three decades has resulted in strong models about neutrino oscillations. By measuring the neutrino oscillation probabilities at different combinations of energy and distance, all but a few of the parameters describing the rate and frequency of oscillations between each flavor-pair have been determined. The DUNE experiment [4] plans to target the remaining degeneracies in these oscillation parameters: namely the mass-ordering and potential CP-violation. However, to reach the required precision to make confident conclusions about BSM physics, we can't rely on building bigger and more expensive detectors. We also need to improve our analysis techniques in order to get the most out of the available experimental data.

B. Event Reconstruction and Machine Learning

Reconstructing the particles involved in a neutrino interaction measured in a LArTPC is challenging. Every neutrino interaction looks unique, so it can be difficult to define an automated process to classify each event. However, there are patterns that we can leverage to distinguish each particle. Event reconstruction is the process of quantifying the kinematics of the reconstructed particles. By improving the tools in this analysis, we can effectively improve the detection efficiency of the different types of neutrino interactions that are recorded in the detector.

One such tool is NuGraph2 [3]. It is a graph neural network (GNN) that receives a collection of hits (localized current pulses measured in the detector) and separates them based on two different metrics. The first is a filter for background charge, namely from cosmic rays that pass through the detector during a neutrino interaction. The second task that NuGraph2 performs is classifying the remaining charge between five different types of particles based on the shape of the charge distribution.

The five different categories are: electro-magnetic (EM) shower, minimally ionizing particle (MIP), highly ionizing particle (HIP), Michel electron, and diffuse charge. EM showers are created by electrons and photons released from the original interaction vertex. For example, an electron neutrino interaction typically releases an electron. As the electron slows down in the detector, it radiates photons, which pair produce into an electron and positron pair. This process repeats and cascades into an electromagnetic shower, which looks different than the linear tracks that a muon (released in a muon neutrino interaction) create. The difference between MIPs and HIPs is the length of the tracks these particles create, rate of energy dissipation and density of charge deposition. Michel electrons come from occasional muon decay, and diffuse charges are produced from other photons.

NuGraph2 is trained on a simulated version of the MicroBooNE detector [2]. When validated on similar simulations, it has 98.0% accuracy for cosmic rejection and 94.9% accuracy for categorization [3]. However, known mismodeling of the simulated detector leads to systematic uncertainties [5]. We quantified the change in NuGraph2-aided reconstruction when certain corrections are made to correct this mismodeling.

II. DETECTOR SYSTEMATICS

A. Methods

NuGraph2 was trained on simulated neutrino events in the MicroBooNE detector. However, no simulation can replicate the physics inside a real detector exactly. In [5], a specific parametrization for some of these differences was determined. By simulating well-understood cosmic rays to match measured cosmic interactions in the detector, they were able to quantify how the response in the current-collecting-wires changed depending on the location and orientation of the cosmic rays in the detector. These changes are called wire modifications, and they can be applied to the simulations to generate more realistic samples. We call samples from the original samples the "central value", whereas the samples with the modified waveforms are called "detector variations".

To assess the difference between the central value and the detector variations, we perform a typical analysis on these different samples and look at how the output changes. The current analysis uses the modular software package Pandora [6]. NuGraph2 has been integrated into the traditional analysis framework to influence the cosmic rejection and clustering of charge into distinct particles. The output of this analysis is a large series of variables describing the different aspects of the interaction. Then, to simulate a use-case that NuGraph2 is best suited for, we performed a selection for single electron shower events, which are produced by ν_e interactions. This involves parameters such at the initial energy deposit rate at the beginning of the shower, as well as the categorization variable directly from NuGraph2. Along with other parameters, this forms a ν_e selection.

The specific limits for these variables are chosen to balance selection efficiency and selection purity in simulations. Selection efficiency is the percentage of ν_e events that actually make it through the selection. Those that don't pass the selection often look similar to background events we are trying to remove. The proportion of non-ν_e events that are selected determines the selection purity.

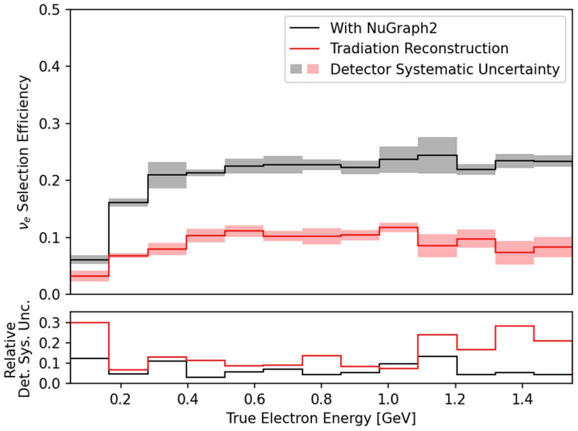

FIG. 3. The effect of NuGraph2 on selection efficiency and the detector systematic uncertainty on the selection efficiency.

It is challenging to have both a high selection efficiency and selection purity since there is a lot of variability in what these interactions look like. This represents an uncertainty in the analysis process, and after working to improve the selection, we must quantify how well we expect our analysis to perform to make accurate statements about the experiments uncertainty.

Our simulation samples contain only electron neutrinos that produce a single electron shower. This way we can accurately tell the proportion of these that pass our selection. We also used samples of neutral current π^0 interactions as a proxy of our background rejection, or selection purity. We also performed these selections using analysis done without NuGraph2 as well. In keeping the original samples the same, the only difference in the two versions of the code is the addition of NuGraph2. The rest of the selection was kept as similar as possible to isolates the effect of NuGraph2 on the uncertainty in the selection efficiency.

B. Results

We first look at the selection efficiency as a function of the electron energy. We matched the binning to the published energy distribution of the electron showers we are selecting for [7] [8]. This is because the specific binning affects the uncertainty in each bin.

In Figure 3, we see a doubling in the selection efficiency with the implementation of NuGraph2 compared to the traditional reconstruction. More importantly, we don't see a significant increase in the relative detector systematic uncertainty accompanying this improved performance. Averaging the relative uncertainty weighted across the number of events in each energy bin shows that the relative detector systematic uncertainty actually decreases from 13% to 6.7%. This means that the absolute uncertainty is remaining relatively constant while the selection efficiency is increasing. We show the different metrics for this selection summarized in Table I.

The weighted average takes the absolute or relative detector systematic uncertainty from each bin and performs an average weighted to the number of events in each bin. For the single bin, all of the events in each bin are combined before calculating the uncertainty, which results in smaller values. This is due to the indifference of this metric to drifting of events between bins and the lower statistics in each bin. However, an analysis of electron neutrinos will give values separated into these bins, so the weighted bin average provides a better estimate for the performance of NuGraph2 in a typical use-case.

C. Discussion

We see an increase in absolute detector systematic uncertainty from 1.1% to 1.4% accompanying the implementation of NuGraph2. This means that our confidence in the efficiency when applied to real data is slightly reduced. Since the selection efficiency is a necessary component of the electron neutrino calculation, an increased uncertainty weakens the precision in the final measurement. However, because NuGraph2 improves the selection from 9.2% to 21.0%, the relative uncertainty actual decreases significantly. The absolute uncertainty is still low enough for us to consider NuGraph2 robust to these detector modeling uncertainties. With its improved performance, further studies of MicroBooNE data will be able to leverage NuGraph2 to make more precise measurements with the confidence that this ML model isn't introducing too much noise into the reconstruction. All code from this project is available on Github.

III. $\eta - \pi^0$ SEPARATION

A. Motivation

Eta (η) mesons are particles that are produced in higher-order resonant interactions within the Argon nucleus and are difficult to separate from other background processes. They are electrically neutral, so their signature in a LArTPC is dependent on their decay mode. Luckily, the lifetime of the η particle is extremely short, on the order of attoseconds (10^{-18}s), so we can rely on them decaying inside the detector. The most common decay mode of the η particle is into two photons [1]. However, this decay looks very similar to the two photon decay of the neutral pion, π^0. The key distinction comes from the kinematics of the two photons, where the energy and angle between the photons can be combined to give the invariant (rest) mass of the decayed particle from Equation 1.

$$M_{\gamma\gamma} = \sqrt{2E_1E_2(1 - cos\theta_{\gamma\gamma})} \tag{1}$$

The η particle is much heavier than the π^0, so $M_{\gamma\gamma}$ can be used to selection out η particles from a background of π^0. The difficulty is that this selection requires the complete reconstruction of the photon kinematics, which introduces a lot of uncertainty into $M_{\gamma\gamma}$. The current MicroBooNE η selection efficiency is only 13.6%[9].

We would like to determine whether a neural network can learn these underlying kinematics. If successful, this would allow us to bypass the uncertainties in reconstruction and improve the current selection. To be useful, a comparison with a selection on the di-photon invariant mass would need to show that this model performs better despite the intrinsic uncertainties that come with ML. Not only would this improved η selection increase precision in probing neutrino-Argon resonances, but it would open opportunities to train other models on other selections that also currently rely on reconstructed kinematics.

B. Methods

We based our ML project on the Python-based PointNeXt model [10]. This model takes a set of 3D points and our configuration outputs a continuous scalar between 0 and 1. A binary cross entropy (BCE) loss was used to train the model that lower values (close to 0) predict that the showers come from a π^0 decay while higher values (close to 1) predict an η decay.

$$\text{BCE} = -\frac{1}{\text{N}}\sum_{i=1}^{\text{N}}[t_i log(p_i) + (1 - t_i)log(1 - p_i)] \tag{2}$$

Equation 2 is the average loss between the target t_i and the prediction p_i across N events. The target, t, is either 0 for π^0 events or 1 for η events, so only one of the terms in the sum is active for a given event. The logarithms punish incorrect classifications much more than it rewards correct classifications. This tends to result in the model predicting most events that it can't classify to be around 0.5.

As in any ML application, acquiring good data is vital for making sure that the differences between our two classes is presented in an interpretable way to the model. We don't know what these differences are on an individual parameter scale, but we can control the information we feed the model. Only the two photon showers is important for this classification, so we have filtered our data to only the charges and their positions in the two reconstructed showers.

In training, we supply the model with certain hyper parameters which indicate how many parameters the model has and how they can change in response to the training. Table II contains a summary of the hyperparameters of the model. The learning rate is adjusted throughout the

TABLE I. ν_e selection efficiency with detector systematic uncertainties with respect to true electron energy. MCC9.10 is the new version with NuGraph2 and MCC9 is the traditional reconstruction. The absolute uncertainties are direct changes in the selection efficiency, whereas the relative uncertainties are proportions of the selection efficiency. The total events are the number of events with true electron energy between 0.05 and 1.55 GeV. The weighted average is performed with respect to then number of events in each true electron energy bin in Figure 3.

Version	CV Selection	Weighted Bin Average		Single Bin	
	Efficiency	Abs. Uncertainty	Rel. Uncertainty	Abs. Uncertainty	Rel. Uncertainty
MCC9.10	21.0%	1.4%	6.7%	0.43%	2.1%
MCC9	9.2%	1.1%	13%	0.28%	3.0%

TABLE II. PointNeXt Hyperparameters

Hyperparameter	Value	Description
Learning Rate (LR)	0.001	Gradient step size
Number of Epochs	100	Iterations through the data
Number of Points	512	Fixed number of points
Noise Proportion	0.1	Percentage random points
MLPs	256, 128, 64	Multilayer Perceptrons

training using the AdamW optimizer[3] [11]. This dynamically reduces the learning rate based on the loss function in each epoch and improves training performance.

In order for PointNeXt to be agnostic to the order of input points as well as other permutations, it requires the number of points to be consistent between all events. So, we need to down-sample events with more points and up-sample events with fewer points. The down-sampling is quite simple. We remove the events with the lowest charge associated with them since they should have the least weight in the kinematics. It should be noted that this is a repeatable process that can be performed without knowledge about the truth of the sample.

Up-sampling is a bit more difficult. When adding more points, we don't want to leave behind artifacts that the model can pick up. This would leave information about the number points before up-sampling and this value is dependent on the type of particle. In Figure 4, we see that simulated π^0 events have fewer points than η events. This leads to π^0 events requiring up-sampling more often than η events. Initially we just added points with zero charge to the origin, but when reducing the the number of points, the model would pick up on the difference between up- and down-sampling. In an attempt to counterbalance this, we instead place these zero charge points uniformly random in a cube around the origin. Additionally, we require that a certain percentage of events in both up- and down-sampled events to be these random points. This "noise proportion" reduces the visual difference between the two behaviors and helps the model escape this particular local minima.

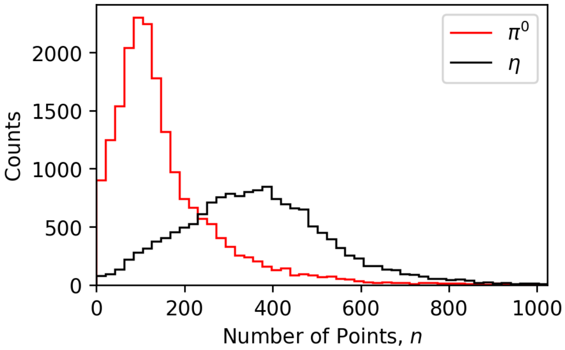

FIG. 4. Number of points in the reconstructed showers for η and π^0 events. π^0 events typically have fewer events, and this distinction can be picked up by the model if upsampling isn't done properly.

The model was given 30000 events, half for each η and π^0. This was then randomly split into training, validation, and test data. 70% is used during the training to influence the parameters of the model by gradient descent. 15% is used to validate the performance of the model during training. Since the validation data doesn't influence the learning of the model, it can be a good benchmark for how well the model will perform on future data. Validation was perform every other epoch, and the best model was determined by the minimum validation loss. After training is completed, the best and last epochs are given the remaining 15% of the data as a final test.

C. Results

Even after training, the model doesn't make very strong predictions. The BCE validation loss was best in epoch 76 at 0.6303. The confusion matrices in Table III and IV given the percentages of each category that are predicted as η $(p > 0.5)$ or π^0 $(p \leq 0.5)$.

TABLE III. Confusion matrix between η and π^0 predictions normalized by actual particle class. The η selection rate is 57% and the π^0 selection rate is 30%. A perfect selection corresponds to 100% η selection rate and 0% π^0 selection rate.

Actual	Predicted Particle Class	
Particle Class	π^0	η
π^0	69.9%	30.1%
η	42.7%	57.3%

[3] AdamW Implementation in PyTorch

TABLE IV. Confusion matrix between η and π^0 predictions normalized by predicted particle class.

Actual Particle Class	Predicted Particle Class	
	π^0	η
π^0	62.7%	35.0%
η	37.3%	65.0%

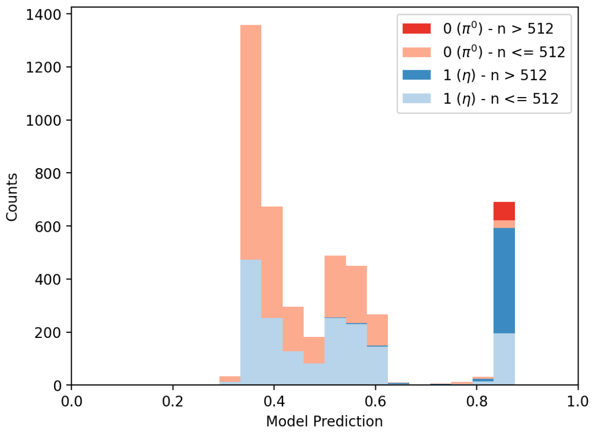

FIG. 5. Best epoch model predictions stacked by a cut on the number of points, n. Statistically, more of the high point events are likely to be η particles, but the existence of correctly predicted low n η events is indicative that the model is not incorrectly learning a simple cut on n.

The continuous predictions on the test samples are shown in Figure 5. Most of the predictions are close to 0.5, but there are some stronger predictions above 0.8 that contain mostly η, which is indicative of some learning.

The receiver operating curve (ROC) is shown in Figure 6. It shows the trade-off between η selection (the bottom right cell in Table III and the inclusion of π^0 events in the selection (the top right cell in Table III) as the threshold for a predicted η is sweeped through the range 0 to 1. The area under the ROC (AUROC) is representative of the general confidence of the model in its correct predictions independent of the default 0.5 threshold. A perfect predictor would have an AUROC of 1, so 0.68 indicates a minor amount of learning above a uniformly random prediction.

We know from the invariant mass that there is a kinematics relation that can separate η and π^0 particles. These results may show promise that some kinematics can be learned through ML. However, it may need more fine-tuning of the hyperparameters and well as a comparison to the di-photon invariant mass selection to assess its usefulness. All code from this project should be accessible on Github.

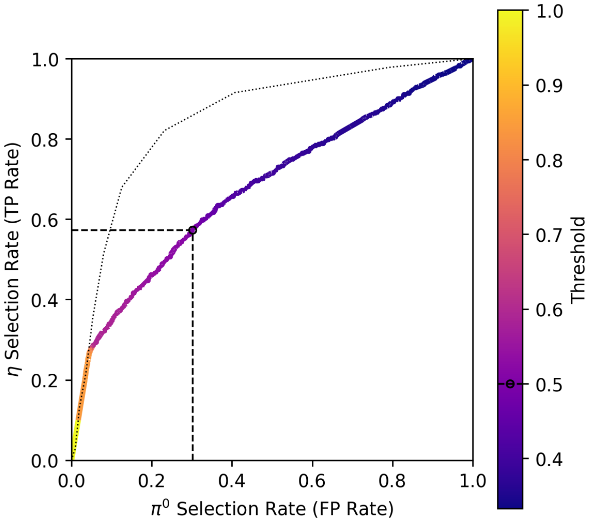

FIG. 6. The receiver operating curve (ROC) for the best epoch model at varying thresholds required for a η prediction. The circular point and dashed lines indicate the performance of the default $p = 0.5$ thresold. The AUROC (Area Under ROC) is 0.68. The dotted curve is the behavior of a cut using a simple threshold in the number of points. Since the ROC of the model isn't consistent with this baseline, the up-sampling technique appears to be working.

IV. CONCLUSION

NuGraph2 and our new η-π^0 separation model stand at opposite ends of the ML development process. NuGraph2 was already completely developed, and we just verified that differences between simulation and real data don't pose a large threat to the performance NuGraph2 offers during neutrino reconstruction. While η-π^0 separation seems like a much simpler problem on the the surface, our new model doesn't appear to have strong performance on heavily processed simulated data. However, this challenge is indicative of whether ML tools can learn to make inferences about the kinematics of an interaction. Kinematics are a very high level concept compared to the space points and charge that our model receives. This model shows promise that this type of abstraction is possible, and with further development, it could present a new direction for event reconstruction algorithms.

V. ACKNOWLEDGMENTS

I would like to give my appreciation to Dr. David Caratelli and Chuyue "Michaelia" Fang for their amazing support throughout this REU. Neither of these projects would have been possible without their help and guidance. The entire UCSB neutrino physics group was also very welcoming this summer, and I would especially like to thank Lucas Solomon for his guidance in setting up the PointNeXt machine learning model. I would also like to thank the MicroBooNE NuGraph2 group and those at FermiLab that spent time listening to my progress updates every week and giving me feedback on my detector systematics work. Finally, I would like to appreciate Dr. Guruswamy for her work organizing this REU program. This work is supported by NSF REU grant PHY-2349677.

[1] R. L. Workman *et al.* (Particle Data Group), Review of particle physics, Progress of Theoretical and Experimental Physics **2022**, 083C01 (2022), https://academic.oup.com/ptep/article-pdf/2022/8/083C01/49175539/ptac097.pdf.

[2] R. Acciarri *et al.* (MicroBooNE Collaboration), Design and construction of the microboone detector, Journal of Instrumentation **12** (02), P02017.

[3] A. Aurisano, V. Hewes, G. Cerati, J. Kowalkowski, C. S. Lee, W. Liao, D. Grzenda, K. Gumpula, and X. Zhang, Graph neural network for neutrino physics event reconstruction, Phys. Rev. D **110**, 032008 (2024).

[4] A. Falcone (DUNE Collaboration), Deep underground neutrino experiment: DUNE, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **1041**, 167217 (2022).

[5] P. Abratenko *et al.* (MicroBooNE Collaboration), Novel approach for evaluating detector-related uncertainties in a lartpc using microboone data, The European Physical Journal C **82**, 10.1140/epjc/s10052-022-10270-8 (2022).

[6] R. Acciarri *et al.* (MicroBooNE Collaboration), The pandora multi-algorithm approach to automated pattern recognition of cosmic-ray muon and neutrino events in the microboone detector (2018).

[7] P. Abratenko *et al.* (MicroBooNE Collaboration), Search for an anomalous production of charged-current ν_e interactions without visible pions across multiple kinematic observables in microboone 10.1103/x259-z6mf (2024), arXiv:2412.14407.

[8] MicroBooNE Collaboration, Search for an Anomalous Production of Charged-Current ν_e Interactions Without Visible Pions Across Multiple Kinematic Observables in MicroBooNE, HEPData (collection) (2025), `https://doi.org/10.17182/hepdata.159762`.

[9] P. Abratenko *et al.* (MicroBooNE Collaboration), First measurement of η meson production in neutrino interactions on argon with microboone, Phys. Rev. Lett. **132**, 151801 (2024).

[10] G. Qian *et al.*, Pointnext: Revisiting pointnet++ with improved training and scaling strategies (2022), arXiv:2206.04670 [cs.CV].

[11] I. Loshchilov and F. Hutter, Decoupled weight decay regularization (2019), arXiv:1711.05101 [cs.LG].