FERMILAB-THESIS-2022-27

Searching for Clues for a Matter Dominated Universe in Liquid Argon Time Projection Chambers (LArTPCs)

Yeon-jae Jwa

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Searching for Clues for a Matter Dominated Universe in Liquid Argon Time Projection Chambers

Yeon-jae Jwa

Liquid Argon Time Projection Chambers (LArTPCs) represent one of the most widely utilized neutrino detection techniques in neutrino experiments, for instance, in the Short Baseline Neutrino (SBN) program and the future large-scale LArTPC: Deep Underground Neutrino Experiment (DUNE). The high-end technique, facilitating excellent spatial and calorimetric reconstruction resolution, also enables testing exotic Beyond Standard Model (BSM) theories, such as baryon number violation (BNV) processes (e.g., proton-decay, neutron-antineutron oscillation). At the same time, Machine Learning (ML) techniques have demonstrated their ubiquitous use in recent decades; ML techniques have also become some of the most powerful tools in high-energy physics (HEP) analyses. Furthermore, the development of algorithms to cater to the needs of problems in HEP (i.e., triggering, reconstruction, improving sensitivity, etc.) has also become an active area of research. By developing a combined approach using Convolutional Neural Network (CNN) and Boosted Decision Tree (BDT) techniques, the sensitivity of neutron-antineutron oscillation in DUNE is evaluated for a projected exposure of 400kton·years. Additionally, to meet the triggering requirement to select such rare events in DUNE, such a search is only supported with highly efficient self-triggering algorithms. An ML-based self-triggering scheme for large-scale LArTPCs, such as DUNE, is also developed with the intention of implementation on field-programmable gate arrays (FPGAs). The ML-based approach for searching for neutron-antineutron oscillation can be demonstrated and validated on the current LArTPC MicroBooNE. The analysis in MicroBooNE

represents the first-ever search for neutron-antineutron oscillation in a LArTPC. DUNE's projected 90% C.L. sensitivity to the neutron antineutron oscillation lifetime is 6.45×10^{32} years, assuming 1.327×10^{35} neutron· years, equivalent to 10 years of DUNE far detector exposure (400kton· years). For MicroBooNE, assuming 372 seconds of exposure (equivalent to 3.13×10^{26} neutron·years), the 90% C.L. lifetime sensitivity is found at 3.07×10^{25} yrs, after accounting for Monte-Carlo statistical uncertainty and systematic uncertainty from detector effects.

Table of Contents

Acknowledgements

I thank my thesis advisor Prof. Georgia Karagiorgi for her support throughout my Ph.D. I was fortunate to have been provided ample opportunities to explore my academic interests in the field of neutrino physics with her encouragement and guidance. Georgia has been a reminder to stay curious and diligent, and I wish the lessons I learned from her will last in me.

I thank Dr. Mark Ross-Lonergan and Dr. Davio Cianci for introducing me to neutrino physics during our phenomenology study at the beginning of my Ph.D. I greatly appreciate Yuyang Zhou and Dr. Vic Genty for sharing their knowledge with respect to machine learning (ML) tools.

I could not have done the DUNE neutron-antineutron oscillation study without discussions with Dr. Josh Barrow, Dr. Jeremy Hewes, and Dr. Aaron Higuera. I also would like to express my appreciation to (current/former) DUNE high-energy working group conveners Prof. Vitaly Kudryavtsev, Prof. Lisa W. Koerner, and Dr. Yun-Tse Tsai for their support of the analysis. I greatly appreciate Dr. Giuseppe Di Guglielmo and Lukas Arnold for collaborating on the ML-based triggering development for DUNE.

For MicroBooNE analysis, I cannot name all the MicroBooNE collaborators who enabled this study. I greatly appreciate (current/former) MicroBooNE APE group conveners Dr. José I. Crespo-Anadón, Dr. Pawel Guzowski, and Dr. Ivan Lepetic for their support and management of the analysis. I greatly appreciate Dr. Daisy Kalra for joining me in MicroBooNE analysis. MicroBooNE analysis was possible by collaborating with the Wire-Cell reconstruction developers and the official production by the sample production team. I would like to thank Dr. Jay Hyun Jo, Prof. Hanyu Wei, and Dr. Anyssa Navrer-Agasson.

I would like to thank Nevis neutrino group Prof. Mike Shaevits, Dr. Leslie Camilleri, Dr. Kathryn Sutton, Guanqun Ge, and Iris Ponce for providing a friendly academic environment. I thank Dr. William Seligman especially for the computing support as well.

I am sending love to my family and friends for allowing me to stand on my feet during this time; Mom, Chungmu, Jane, Hayun, Seoyun, Jiwon, Claire, Yeon-ji, Bola, Hyunseok, Sangmin, and Jungwan.

I express my appreciation to my former lab mates, Dr. Sungbin Oh and Junho Choi, for sharing our enthusiasm for particle physics, whether we are close or far.

I wish to express my special thanks to everyone with whom I have shared the joy of running, especially members of the Brooklyn Track Club and the Social Running Club.

Some of the work presented in this thesis is based upon work supported by the National Science Foundation under Grant No. NSF-1914965 and No. PHY-1753228.

Chapter 1: Introduction

The Liquid Argon Time Projection Chamber (LArTPC) technology has become the mainstream technology to perform sensitive measurements and discoveries in neutrino physics. Current LArTPCs include the Short Baseline Neutrino (SBN) program at Fermilab, Illinois. The future Deep Underground Neutrino Experiment (DUNE) represents the flagship experiment in the U.S. The detector will consist of the largest LArTPCs ever built. DUNE's primary physics goals include the search for CP violation and the determination of the ordering of neutrino mass eigenstates. At the same time, DUNE provides opportunities for astroparticle and exotic searches such as baryon number violation (BNV) processes (e.g., proton-decay and neutron-antineutron oscillation). As part of my Ph.D., I have focused on developing and demonstrating DUNE's physics reach in searching for BNV processes, especially for neutron-antineutron oscillation. In this endeavor, I have worked both on physics analysis and on triggering development for data selection of exotic high-energy interactions such as neutron-antineutron oscillation.

The growing popularity of Deep-Learning (DL) techniques' application in high energy physics (HEP), along with the LArTPCs high-resolution reconstruction, has inspired me to develop novel DL-based techniques for triggering development and physics analysis in DUNE. This approach of applying DL-based techniques on an exotic search in LArTPC can be validated in a smaller-scale detector. MicroBooNE, an 80 ton LArTPC operating since 2015, offers an excellent opportunity for such validation. However, the search for neutron-antineutron oscillation in MicroBooNE is limited by the small exposure to provide a competitive limit, compared to current existing limits from experiments with different detector techniques; MicroBooNE is also bombarded by background cosmic rays as an on-surface operating detector. Regardless, the search in MicroBooNE serves as an important development stepping stone for the future search in DUNE. I developed a

selection for $n - \bar{n}$ on MicroBooNE data. At the time of this thesis writing, this analysis is nearing the final stage of systematic uncertainty evaluation and data unblinding.

In addition to the work presented in this thesis, I have served as a MicroBooNE readout expert. The neutrino group at Columbia University has developed and has been responsible for MicroBooNE and Short-Baseline Near Detector (SBND) readout electronics. As a group member, I participated in SBND electronics testing at Nevis laboratory and commissioning at Fermilab. I have also studied the implications of 3+N sterile neutrinos on the sensitivity of DUNE's CP violation measurement using toy Monte Carlo simulations. This work was presented at ICHEP 2018 under the title "Resolving DUNE oscillation parameter ambiguities in the 3+1 sterile neutrino scenario using SBN."

Chapter 2: Theory and measurements

2.1 The Standard Model

The Standard Model (SM) is the most successful theory describing the universe thus far. It is a quantum field theory (QFT) where a Lagrangian governs the dynamics and kinematics of the theory. The current formulation was finalized in the mid-1970s. Since then, its robustness has been tested, and its predictions have been confirmed with precision throughout measurements, such as measurements on weak interactions. In recent decades, confirmation of the top quark [1, 2], the tau neutrino [3], and the Higgs boson [4, 5] have added further credence to the SM.

The theory includes fundamental interactions (strong, weak, and electromagnetic interactions), while omitting gravity in the description. The gauge group structure of the three fundamental interactions is

$$\mathcal{G}_{SM} = SU(3)_C \times SU(2)_L \times U(1)_Y, \tag{2.1}$$

consisting of the strong (color, $SU(3)_c$), left-handed weak ($SU(2)_L$), and electromagnetic ($U(1)_Y$) interactions. These interactions are mediated by gauge bosons with integer-spin; gluons for strong interaction, W and Z bosons for the weak interaction, and photons for the electromagnetic interaction. The properties of these gauge bosons arise from the nature of the gauge symmetry of each interaction. The SM includes 12 elementary fermions over three generations, four fermions within each, as shown in Fig. 2.1. These fermions are the smallest building blocks to form matter (or ordinary matter, as opposed to dark matter). There are six quarks and six leptons in total, and each generation entails two quarks and two leptons. There is also the Higgs boson; the strength of the interaction of the Higgs field and fermions determines their mass.

All of the fermions in the SM are spin-half particles; thus, they follow the Fermi-Dirac statis-

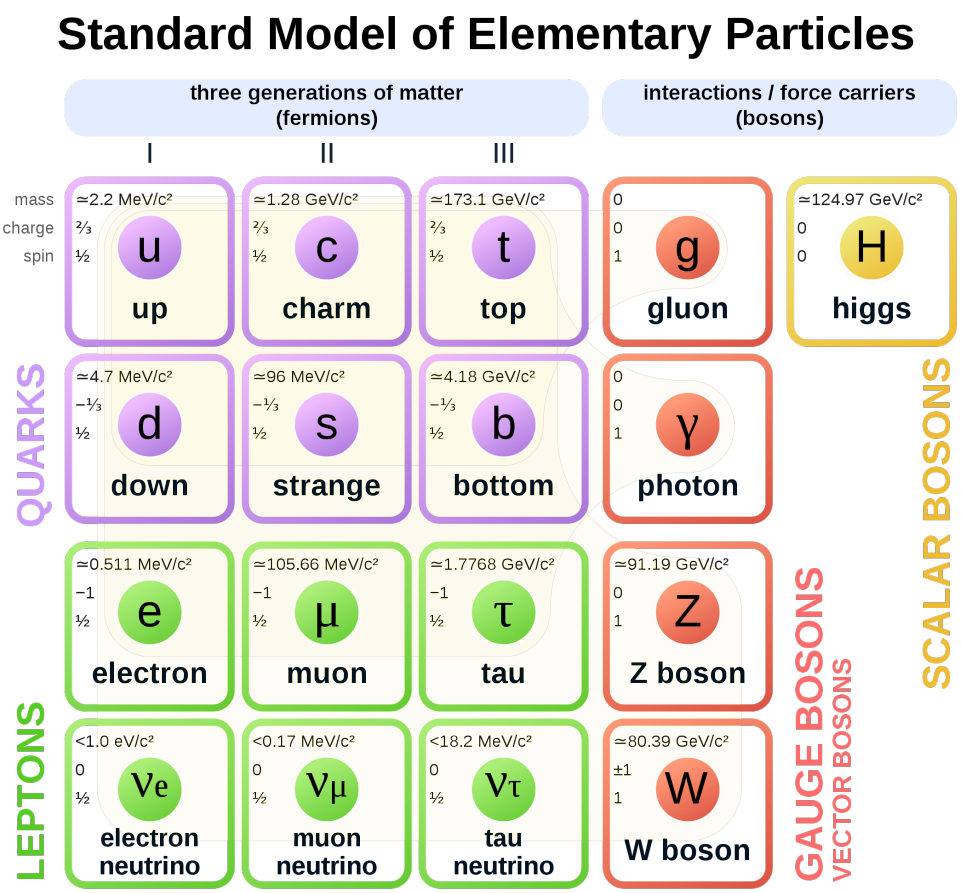


Figure 2.1: The Standard Model

tics and the Dirac equation. Also, each of them has its counterpart particle: antifermion. An antifermion has the same mass and spin as the fermion but an opposite sign for quantum numbers such as the electric charge (q), the lepton number ($\mathcal{L}$), and the baryon number ($\mathcal{B}$). These antifermions form antimatter. In the SM, $\mathcal{B}$-$\mathcal{L}$ is globally conserved. However, the conservation of $\mathcal{B}$ and $\mathcal{L}$ respectively is considered to be an "accidental" symmetry; i.e., their conservation is not related with the global symmetries in the SM, whereas charge (q) conservation arises from the gauge invariance of electromagnetic fields.

Although the SM is an extraordinarily successful theory, there are significant questions that are not answered by the SM. For instance, the SM does not explain the matter-antimatter asymmetry

quark (antiquark)	q	$\mathcal{B}$	$\mathcal{S}$	lepton (antilepton)	q	$\mathcal{L}$
$u(\bar{u})$	2/3 (-2/3)	1/3 (-1/3)	0 (0)	$e^-(e^+)$	-1 (+1)	1 (-1)
$d(\bar{d})$	1/3 (-1/3)	1/3 (-1/3)	0 (0)	$\nu_e(\bar{\nu_e})$	0 (0)	1 (-1)
$c(\bar{c})$	2/3 (-2/3)	1/3 (-1/3)	0 (0)	$\mu^-(\mu^+)$	-1 (+1)	1 (-1)
$s(\bar{s})$	1/3 (-1/3)	1/3 (-1/3)	1 (-1)	$\nu_\mu(\bar{\nu_\mu})$	0 (0)	1 (-1)
$t(\bar{t})$	2/3 (-2/3)	1/3 (-1/3)	0 (0)	$\tau^-(\tau^+)$	-1 (+1)	1 (-1)
$b(\bar{b})$	1/3 (-1/3)	1/3 (-1/3)	0 (0)	$\nu_\tau(\bar{\nu_\tau})$	0 (0)	1 (-1)

Table 2.1: Quantum numbers of quarks (antiquarks) and leptons (antileptons) in the SM. q refers to the electric charge where $q = 1$ is referenced to the charge of a positron. $\mathcal{B}$ refers to the baryon number, $\mathcal{S}$ is strangeness, $\mathcal{L}$ refers to the lepton number. S is not a conserved quantity in the SM; it is conserved through strong interaction and electromagnetic interaction, but not through the weak interaction. $\mathcal{B}$ is 0 for leptons, $\mathcal{L}$ is 0 for quarks. $\mathcal{B}$ and $\mathcal{L}$ are conserved in the effective Lagrangian of the SM.

in the universe and the nature of the non-zero, albeit very small, neutrino mass.[1] Some of these unanswered questions have been the driving force for physicists to seek beyond-Standard Model (BSM) physics.

2.2 Baryogenesis and Sakharov conditions

According to the Big Bang cosmology, the universe began from a charge-symmetric state (C-symmetry). Thus, in the very beginning, matter and antimatter existed in identical amounts. On the other hand, the CPT theorem states that when Lorentz invariance stands (i.e., the laws of physics are the same for different observers), any quantum field theory is invariant under CPT (charge, parity, time-reversal) transformation: i.e., if a particle is at a position ($\vec{x}$), time (t), the antiparticle at position ($-\vec{x}$), and time (-t) follows the same laws of physics as the original particle. In a theory where CPT theorem holds, a particle and its counterpart antiparticle decay at the same rate. Combining these ideas, one would expect the same number densities for matter and antimatter in the current universe. Instead, it is observed that the universe consists of more matter than antimatter [6, 7]. The magnitude of baryon abundance (BAU) is characterized by the asymmetry parameter

[1] It is noteworthy that the SM Lagrangian, without any extension, cannot accommodate neutrino mass. Only left-handed neutrinos are observed, and right-handed neutrinos are not. While the SM prescribes massless neutrinos, at least two non-zero mass eigenstates are confirmed through neutrino oscillation.

[8–10]:

$$\beta = \frac{n_{\mathcal{B}} - n_{\bar{\mathcal{B}}}}{n_\gamma} \approx 10^{-10}, \tag{2.2}$$

where $n_{\mathcal{B}}$ represents the baryon number density, $n_{\bar{\mathcal{B}}}$ represents the antibaryon number density, and n_γ represents the photon number density in the universe.

This asymmetry is confirmed to be isotropic across the universe because 1) cosmic microwave background (CMB) shows the isotropic temperature of the universe, and 2) no energetic γ-radiation is observed to postulate clustered antimatter abundance (e.g., if antimatter abundant region existed, the pair annihilation at the boundary between antimatter abundant region and matter abundant regions should generate strong γ-radiation such as $p\bar{p} \rightarrow \gamma\gamma$) [6].

Baryogenesis is a theory explaining such excess in baryon number density compared to antibaryon number density, assuming the rise of this asymmetry of baryon over antibaryon in the early universe after the Big Bang. In 1967, Andrei Sakharov proposed [11] the three necessary conditions for baryogenesis to occur:

- Baryon number violating process
- CP-symmetry violation
- Interactions outside of thermal equilibrium

These conditions can be illustrated by a simple example. The assumptions within this example are that the universe started with the C-symmetric state ($\mathcal{B}_i = 0$) and the current universe observed BAU ($\mathcal{B}_f > 0$). Thus,

$$|\Delta\mathcal{B}| = \mathcal{B}_f - \mathcal{B}_i = \mathcal{B}_f > 0. \tag{2.3}$$

If we consider an arbitrarily heavy particle that existed some time after the beginning of the universe, X, and its antiparticle, $\bar{X}$,

$$|\Delta\mathcal{B}| = \mathcal{B}_f = \mathcal{B}_X + \mathcal{B}_{\bar{X}}. \tag{2.4}$$

If X decays to Y_1 with branching fraction f and Y_2 with branching fraction $(1-f)$, and $\bar{X}$ decays to $\bar{Y_1}$ with branching fraction $\bar{f}$ and $\bar{Y_2}$ with branching fraction $(1-\bar{f})$,

$$\mathcal{B}_X + \mathcal{B}_{\bar{X}} = f\mathcal{B}_{Y_1} + (1-f)\mathcal{B}_{Y_2} + \bar{f}\mathcal{B}_{\bar{Y_1}} + (1-\bar{f})\mathcal{B}_{\bar{Y_2}}. \tag{2.5}$$

Since Y_1 and $\bar{Y_1}$, Y_2 and $\bar{Y_2}$ are antiparticles of each other, and the decays of X and $\bar{X}$ have to happen at the same rate by the CPT-theorem,

$$\mathcal{B}_X + \mathcal{B}_{\bar{X}} = f\mathcal{B}_{Y_1} + (1-f)\mathcal{B}_{Y_2} - \bar{f}\mathcal{B}_{Y_1} - (1-\bar{f})\mathcal{B}_{Y_2} = (f-\bar{f})(\mathcal{B}_{Y_1} - \mathcal{B}_{Y_2}) > 0. \tag{2.6}$$

As a result, 1) $f - \bar{f} \neq 0$, and 2) $\mathcal{B}_{Y_1} - \mathcal{B}_{Y_2} \neq 0$.

The first inequality means there is an asymmetry between branching fractions of a particle and an antiparticle's decay (i.e., non-conservation of CP-symmetry). The second inequality means that there must have occurred a baryon number violation at some point. These are the first two of the Sakharov conditions. The third condition is necessary by definition; if this process occurs in a thermal equilibrium state, the rate for the inverse process of above occurs at the same rate. Thus, the net change of baryon number will be canceled. Due to the rapid expansion of the early universe, the particles in this situation and their corresponding antiparticles do not achieve thermal equilibrium.

We now briefly discuss C, P, and CP-symmetries before further looking into baryon number violating processes. C-symmetry (charge conjugation symmetry) is the symmetry of the laws of physics under charge conjugation, which is a transformation that switches all particles to their corresponding antiparticles and vice versa. Electromagnetism, gravity, and the strong interaction all obey C-symmetry, but weak interactions violate C-symmetry [12]. CP-symmetry is the combined symmetry of C-symmetry and P-symmetry. P-symmetry (parity symmetry) is the symmetry of physics laws under a transformation that flips the sign of spatial coordinates. P- and CP-symmetries are found to be broken in weak interaction processes [13–16]. As mentioned earlier, the CPT theorem is assumed in QFTs. On the other hand, time-reversal symmetry (T-symmetry) alone is broken

in the macroscopic universe given the low-entropy initial state of the universe (the Big Bang); T-symmetry is conserved in strong and EM interactions. Table 2.2 summarizes these symmetries and observed violations of them.

Symmetry	Transformation	Violating interactions
C	charge conjugation ($h \rightarrow \bar{h}$)	weak interaction
P	space inversion ($\vec{x} \rightarrow -\vec{x}$)	weak interaction
T	time reversal ($t \rightarrow -t$)	macroscopic reality
CP	($h(\vec{x}) \rightarrow \bar{h}(-\vec{x})$)	weak interaction
CPT	($h(\vec{x}, t) \rightarrow \bar{h}(-\vec{x}, -t)$)	assumed conserved

Table 2.2: Discrete symmetries of physics laws

Since the departure of thermal equilibrium is observed through the rapid expansion of the early universe [6], only baryon number violation out of the three Sakharov conditions is not directly observed.

2.3 Baryon number violation

In a non-perturbative energy regime ($T = 10^2 \sim 10^{12}$ GeV), the SM can accommodate baryon number violation (BNV) processes [17, 18], where $\mathcal{B} - \mathcal{L}$ is a conserved quantum number but $\mathcal{B}$ and $\mathcal{L}$ are not. The processes associated with this breaking $\mathcal{B}$ and $\mathcal{L}$ are called sphalerons. For instance, $p^+p^+ \rightarrow p^-e^+e^+e^+$ can occur in the early universe. However, this process is far too suppressed to fully explain the observed amount of baryon asymmetry [19–21]. Therefore, it is important to search for BNV in BSM theories.

There are many possible expansions of the SM, such as grand unified theories (GUTs) [22, 23], supersymmetry theories (SUSY) [24, 25], and left-right symmetry models (LRSM) [26], where BNV processes can arise. For instance, the minimal extension of the GUT with the unified SU(5) group predicts proton decay with the selection rule $\Delta\mathcal{B} = 1$, with the implications of new physics at GUT-scale energy ($\sim 10^{16}$ GeV). Because of this implication, proton decay has historically been the major effort in the pursuit of the discovery of BNV over the past few decades. The search

for proton decay through decay channels such as $p \rightarrow e^+\pi^0$ or $p \rightarrow \mu^+\pi^0$ is currently severely constrained by the search measurements by Super-Kamiokande [27] on the order of 10^{34} years. Future searches with DUNE can confirm this result and study additional proton decay modes, including the $p \rightarrow K^+\bar{\nu}$ channel. On the other hand, the lack of experimental evidence for proton decay after decades of efforts, accompanied by the lack of viable evidence for SUSY at the Large Hadron Collider (LHC), led some physicists to focus on other BNV processes rather than the $|\Delta\mathcal{B}| = 1$ scenarios.

Neutron-antineutron oscillation ($n \rightarrow \bar{n}$ or $n - \bar{n}$) is the main target of BNV searches under the selection rule $|\Delta\mathcal{B}| = 2$, where a neutron (n) spontaneously converts itself to an antineutron ($\bar{n}$). $n-\bar{n}$ processes are predicted by many BSM theories [26, 28, 29]. The observation of such a process will provide clear evidence of BNV and will give a better understanding of the generation of the BAU. Moreover, current theoretical work shows many models that predict $n - \bar{n}$ in an observable range of improved search [30–33]. Also, the minimal GUT extension with $n-\bar{n}$ prediction implies new physics near or above the TeV scale [34, 35] which is probable through LHC measurements.

2.4 Neutron-antineutron oscillation

The oscillation of a neutral particle into another neutral particle is not a foreign physical process in modern physics, and observations of such processes, in fact, have brought fruitful insights into the community. Neutrino flavor oscillations among (ν_e, ν_μ, ν_τ) suggests at least two non-zero mass eigenstates of neutrinos [36]. A closer example to $n - \bar{n}$ oscillation is $B^0 - \bar{B^0}$ oscillation, which is really a manifestation of the broken CP-symmetry in the quark sector [37]. In fact, only the conservation of baryon number forbids a neutron from transforming into an antineutron.

Experimentally, the $n-\bar{n}$ process typically leaves a remarkable signature in laboratory settings. There are two methods in searching for $n - \bar{n}$; one is in quasi-free neutrons (Sec. 2.5.1) and the other is in nucleus-bound neutrons (Sec. 2.5.2). In these cases, an antineutron annihilates with the target material or a nearby nucleon inside a nucleus, leading to an energetic interaction with

multiple pions in the final state. In the former case, a slow neutron beam is propagated to a distant target; it is required that the neutron and antineutron energies are degenerate within the range set by the Heisenberg uncertainty principle with respect to their time-of-flight (this is the so-called "quasi-free" condition). Also, the energy difference between neutron and antineutron rises due to any ambient magnetic field, which needs to be cancelled (e.g., by applying an offset magnetic field.) In the latter case, this degeneracy is removed by the large difference in nuclear potential for neutron and antineutron.

The general formalism of $n - \bar{n}$ is described in Sec. 2.4.1, and approximations for a quasi-free condition, under a magnetic field, and a nucleus-bound condition are described in Sec. 2.4.1, Sec. 2.4.2, Sec. 2.4.3, and Sec. 2.4.4 respectively. The conversion between nucleus-bound $n - \bar{n}$ lifetime and quasi-free $n - \bar{n}$ lifetime is discussed in Sec. 2.4.5. The mathematical formalism in this chapter largely follows Ch. 2 in [29].

2.4.1 General formalism for analysis of neutron-antineutron oscillation

The oscillation between neutron and antineutron can be formulated through an effective Hamiltonian H_{eff}. The transition between two eigenstates $|n\rangle$, $|\bar{n}\rangle$ is assumed to be real,

$$\langle \bar{n}| H_{eff} |n\rangle = \langle n| H_{eff} |\bar{n}\rangle \equiv \delta m \tag{2.7}$$

The diagonal matrix elements of H_{eff} can be denoted as,

$$\langle n| H_{eff} |n\rangle = M_{11}, \langle \bar{n}| H_{eff} |\bar{n}\rangle = M_{22}, \tag{2.8}$$

with the imaginary part $\mathrm{Im}(M_{jj}) = -i\lambda/2$ for $j = 1, 2$, where $\lambda^{-1} = \tau_n = 880s$; the mean lifetime of a free neutron. The full effective Hamiltonian on the basis $(n, \bar{n})$ can be written as

$$\mathcal{M} = \begin{pmatrix} M_{11} & \delta m \\ \delta m & M_{22} \end{pmatrix}. \tag{2.9}$$

Here, we also define the difference between two diagonal components

$$\Delta M \equiv M_{11} - M_{22}. \tag{2.10}$$

Then, the Hamiltonian matrix in Eq. 2.9 can be written again as

$$\mathcal{M} = \begin{pmatrix} \frac{1}{2}(M_{11} + M_{22}) & 0 \\ 0 & \frac{1}{2}(M_{11} + M_{22}) \end{pmatrix} + \begin{pmatrix} \frac{\Delta M}{2} & \delta m \\ \delta m & -\frac{\Delta M}{2} \end{pmatrix}, \tag{2.11}$$

or, more precisely, as

$$\mathcal{M} = \frac{1}{2}(M_{11} + M_{22})I + \frac{1}{2}\Delta M K, \tag{2.12}$$

where I is the identity matrix and K is denoted as

$$K = \begin{pmatrix} 1 & \frac{2\delta m}{\Delta M} \\ \frac{2\delta m}{\Delta M} & -1 \end{pmatrix}. \tag{2.13}$$

The operator K has its own eigenvalues $\kappa_{1,2}$ defined as

$$K \left|n_{1,2}\right\rangle = \kappa_{1,2} \left|n_{1,2}\right\rangle \quad . \tag{2.14}$$

Then, the energy eigenvalues can be expressed as

$$E_{1,2} = \frac{1}{2}(M_{11} + M_{22}) + \frac{1}{2}\Delta M \kappa_{1,2}. \tag{2.15}$$

Then, we can introduce an angle θ as,

$$tan(2\theta) = \frac{2\delta m}{\Delta M}. \tag{2.16}$$

By substituting Eq. 2.16 into Eq. 2.13, K is expressed as

$$K = \begin{pmatrix} 1 & tan(2\theta) \\ tan(2\theta) & -1 \end{pmatrix}. \tag{2.17}$$

Then, the eigenvalues of K are found using the characteristic equation.

$$Det[K - \kappa I] = \kappa^2 - 1 - tan^2(2\theta) = 0. \tag{2.18}$$

Solving this equation yields the eigenvalues of K,

$$\kappa_{1,2} = \pm\sqrt{1 + tan^2(2\theta)} = \pm\sqrt{1 + (\frac{2\delta m}{\Delta M})^2} = \pm\frac{\sqrt{(\Delta M)^2 + 4(\delta m)^2}}{\Delta M}. \tag{2.19}$$

We can then substitute Eq. 2.19 into Eq. 2.15 to yield energy eigenvalues,

$$E_{1,2} = \frac{1}{2}[M_{11} + M_{22} \pm \sqrt{(\Delta M)^2 + 4(\delta m)^2}], \tag{2.20}$$

with the energy difference between the two states

$$\Delta E = E_1 - E_2 = \sqrt{(\Delta M)^2 + 4(\delta m)^2}. \tag{2.21}$$

The probability of oscillation from $|n\rangle$ to $|\bar{n}\rangle$ as a function of time can be written as

$$\begin{aligned} P(n(t) = \bar{n}) = |\ \langle \bar{n}|n(t)\rangle\ |^2 &= sin^2(2\theta) sin^2[(\Delta E)t/2] e^{-\lambda t} \\ &= \left[\frac{(\delta m)^2}{(\Delta M/2)^2 + (\delta m)^2}\right] sin^2\left[\sqrt{(\Delta M/2)^2 + (\delta m)^2} t\right] e^{-\lambda t}. \end{aligned} \tag{2.22}$$

Here, ΔM accommodates any interaction effects that are different for the neutron and the antineutron. The magnitude of the off-diagonal term, $|\delta m|$, in the $n - \bar{n}$ effective Hamiltonian is known to be smaller than 10^{-29} MeV [29], which is approximately 32 orders of magnitude smaller than the neutron mass. Thus, when $|\delta m|$ can be assumed to be multiple orders of magnitude smaller than

ΔM and $\Delta M \approx \Delta E$ in any experimental condition, Eq. 2.22 can be condensed to

$$P(n(t) = \bar{n}) = \left(\frac{2\delta m}{\Delta E}\right)^2 sin^2\left(\frac{\Delta E \cdot t}{2}\right) e^{-\lambda t}. \tag{2.23}$$

2.4.2 Neutron-antineutron oscillation in field-free vacuum

In vacuum that is free of electromagnetic fields, where CPT-symmetry is assumed, the energy expectation values are

$$M_{11} = M_{22} = m_n - \frac{i\lambda}{2}, \tag{2.24}$$

without any energy difference between neutron and antineutron. Here, m_n is the rest mass of a neutron. Thus, the Hamiltonian in the $(n, \bar{n})$ basis is

$$\mathcal{M}_f = \begin{pmatrix} m_n - i\lambda/2 & \delta m \\ \delta m & m_n - i\lambda/2 \end{pmatrix} \tag{2.25}$$

with the subscript f of $\mathcal{M}_f$ representing "free".

Solving the eigen equation, $\mathcal{M}_f$ yields the mass eigenstates

$$|n_\pm\rangle = \frac{1}{\sqrt{2}}(|n\rangle \pm |\bar{n}\rangle) \tag{2.26}$$

with the energy eigenvalues

$$E_{1,2} = (m_n \pm \delta m) - \frac{i\lambda}{2}. \tag{2.27}$$

Thus, the energy difference the two states becomes

$$\Delta E = E_1 - E_2 = 2\delta m. \tag{2.28}$$

Inserting Eq. 2.28 into Eq. 2.23, the oscillation probability as a function of time can be simplified

to

$$P(n(t) = \bar{n}) = sin^2\left(\frac{t}{\frac{1}{\delta m}}\right) e^{-\lambda t} \approx \left(\frac{t}{\tau_{n-\bar{n}}}\right)^2 e^{-\lambda t}, \tag{2.29}$$

where $\tau_{n-\bar{n}}$ is the free $n-\bar{n}$ oscillation lifetime

$$\tau_{n-\bar{n}} = \frac{1}{\delta m}. \tag{2.30}$$

Since $\delta m < 10^{-29}$ MeV, $\tau_{n-\bar{n}} \gtrsim 10^5$ s. Sec. 2.5.1 describes experimental searches for the $n-\bar{n}$ oscillation using free neutrons from neutron beams.

2.4.3 Neutron-antineutron oscillation in magnetic fields

Under a magnetic field $\vec{B}$, the opposite signs of the magnetic dipole moments of the neutron and antineutron $\mu_{n,\bar{n}}$ introduce additional terms in the Hamiltonian matrix, compared to the free Hamiltonian in Eq. 2.25. Using $\vec{\mu}_{\bar{n}} = -\vec{\mu}_n$, the Hamiltonian matrix under a magnetic field $\vec{B}$ becomes

$$\mathcal{M}_{\mathcal{B}} = \begin{pmatrix} m_n - \vec{\mu}_n \cdot \vec{B} - i\lambda/2 & \delta m \\ \delta m & m_n + \vec{\mu}_n \cdot \vec{B} - i\lambda/2 \end{pmatrix}. \tag{2.31}$$

The difference between the two diagonal mass terms, using the expression in Eq. 2.10, becomes

$$|\Delta M| = 2\vec{\mu}_n \cdot \vec{B}. \tag{2.32}$$

Then, the energy eigenvalues can be expressed by substituting M_{11}, M_{22} and ΔM into Eq. 2.20,

$$E_{1,2} = m_n \pm \sqrt{(\vec{\mu}_n \cdot \vec{B})^2 + (\delta m)^2} - \frac{i\lambda}{2}. \tag{2.33}$$

For free neutrons, the presence of a magnetic field suppresses the rate of occurrence of $n-\bar{n}$ oscillations. If the magnetic field is sufficiently small ($\vec{\mu}_n \cdot \vec{B} \ll$ 1/time-of-flight), and if the time of flight of the free neutron is much shorter than the decay lifetime of a neutron, the $n-\bar{n}$ oscillation

probability is independent of the magnetic field. Thus, experimental searches using neutron beams must use magnetic shielding to counteract any ambient magnetic fields to prevent any associated suppression of $n - \bar{n}$ oscillation.

2.4.4 Neutron-antineutron oscillation in the nucleus

The $n - \bar{n}$ search in liquid argon time projection chambers (LArTPCs), which is the main work discussed in this thesis, is in the branch of experimental searches for bound $n - \bar{n}$ oscillation. The mechanism of $n - \bar{n}$ oscillation for neutrons bound inside the nucleus is entirely different from the field-free and magnetic field cases. Most significantly, the effective masses of the neutron and antineutron are modified by the potentials they experience inside the nucleus in which they are bound. We can introduce a real potential for a neutron, $V_n = V_{nR}$, but the antineutron potential has an imaginary component, associated with its ability to annihilate other nucleons in the nucleus, and is written as $V_{\bar{n}} = V_{\bar{n}R} - iV_{\bar{n}I}$. We can use these potentials to define effective masses for the bound neutron and antineutron,

$$m_{n,eff} = m_n + V_n, \quad m_{\bar{n},eff} = m_n + V_{\bar{n}}. \tag{2.34}$$

Then, the effective Hamiltonian matrix for the $n - \bar{n}$ oscillation in matter can be constructed using these effective masses as

$$\mathcal{M}_m = \begin{pmatrix} m_{n,eff} & \delta m \\ \delta m & m_{\bar{n},eff} \end{pmatrix}. \tag{2.35}$$

The potentials $V_{nR}, V_{\bar{n}R}$ and $V_{\bar{n}I}$ are all $O(100)$ MeV, and therefore $n - \bar{n}$ mixing is significantly suppressed, since δm is much smaller than $O(100)$ MeV. In this case, the mixing angle θ is defined as

$$tan(2\theta) = \frac{2\delta m}{|m_{n,eff} - m_{\bar{n},eff}|} = \frac{2\delta m}{\sqrt{(V_{nR} - V_{\bar{n}R})^2 + V_{\bar{n}I}^2}} \ll 1. \tag{2.36}$$

The energy eigenvalues in this case are obtained by solving the Eigen equation for $\mathcal{M}_m$, as

$$E_{1,2} = \frac{1}{2}\left[m_{n,eff} + m_{\bar{n},eff} \pm \sqrt{(m_{n,eff} - m_{\bar{n},eff})^2 + 4(\delta m)^2}\right]. \tag{2.37}$$

By substituting Eq. 2.34 into Eq. 2.37, E_1 can be written as

$$E_1 \approx m_n + V_n - i\frac{(\delta m)^2 V_{\bar{n}I}}{(V_{nR} - V_{\bar{n}R})^2 + V_{\bar{n}I}^2}. \tag{2.38}$$

In the case of $n - \bar{n}$ oscillation in a nucleus, the nucleus decays to a lighter nucleus due to the pair annihilation of the resulting antineutron and one of the surrounding nucleons ($(\bar{n}, p)$ or $(\bar{n}, n)$). The instability of the initial nucleus should be accounted for; the imaginary component of this expression is responsible for generating this nuclear instability. The rate of this process, particularly in which the neutron spontaneously oscillates and subsequently annihilates to produce pions, is

$$\Gamma_m = \frac{1}{\tau_m} = \frac{2(\delta m)^2 |V_{\bar{n}I}|}{(V_{nR} - V_{\bar{n}R})^2 + V_{\bar{n}I}^2}. \tag{2.39}$$

Here, we can make the opposite assumption from the quasi-free case Eq. 2.29; $\Delta E \cdot t \gg 1$, due to large differences in neutron and antineutron energy states and long lifetimes. This allows us to assume that the $\sin^2$ term in Eq. 2.23 averages to 1/2; then the oscillation probability is reduced to

$$P(n(t) = \bar{n}) \approx \frac{1}{2}\left(\frac{2\delta m}{\Delta E}\right)^2 e^{-\lambda t}. \tag{2.40}$$

2.4.5 Bound to free neutron-antineutron oscillation lifetime conversion

From Eq. 2.39, the $n - \bar{n}$ oscillation rate in the matter (Γ_m) is proportional to δm^2,

$$\Gamma_m = \frac{1}{\tau_m} \propto \delta m^2. \tag{2.41}$$

Using Eq. 2.30, this expression can be written again as,

$$\frac{1}{\tau_m} \propto \delta m^2 \approx \frac{1}{\tau^2_{n-\bar{n}}}, \tag{2.42}$$

Thus, $n-\bar{n}$ oscillation lifetime in free neutron ($\tau_{n-\bar{n}}$) and bound neutron (τ_m) are related through a coefficient R,

$$\tau_m = R\tau^2_{n-\bar{n}}. \tag{2.43}$$

The value of R varies by nucleus, where $R \approx 10^2$ MeV is equivalent to

$$R \simeq 10^{23} s^{-1}. \tag{2.44}$$

Detailed calculations for ^{56}F obtained $R \approx 1\times10^{23}s^{-1}$ and $R \approx 0.5\times10^{23}s^{-1}$ for ^{16}O, respectively [38]. Using this R, one can take the lower bound on $\tau_{n-\bar{n}}$ from $n-\bar{n}$ searches in free neutron experiments to estimate a lower bound on τ_m and vice versa. When expressed numerically,

$$\tau_m > (1.6\times10^{31} yr)\left(\frac{\tau_{n-\bar{n}}}{10^8 s}\right)^2\left(\frac{R}{0.5\times10^{23}s^{-1}}\right). \tag{2.45}$$

This expression with $R \sim 0.5\times10^{23}s^{-1}$, and the lower limit $\tau_{n-\bar{n}} > 0.86\times10^8 s$ from the ILL experiment [39] results in $\tau_m \gtrsim 10^{31}$ year. Recently this conversion factor for ^{40}Ar nucleus was calculated to be $5.6^{22}s^{-1}$ [40], followed by the interest in searches for $n-\bar{n}$ oscillation in LArTPCs.

2.5 Experimental measurements

As discussed in earlier sections (Sec. 2.4.2, Sec. 2.4.3, Sec. 2.4.4) of this chapter, $n-\bar{n}$ oscillation manifests differently under different circumstances: the lifetime of free neutrons oscillating in a magnetic field is vastly different from bound neutrons oscillating inside a nucleus. This enables searches using different experiment setups – namely, searches in free neutron beams and in large-mass detectors with nucleus-bound neutrons.

Accounting for theory-derived conversion factors between free and bound lifetime limits, both experimental approaches offer competitive sensitivity despite vastly differing techniques. The following sections describe all $n - \bar{n}$ lifetime experimental searches to date, and all limits discussed are lower limits at the 90% confidence level (C.L.). The current best limit on the lifetime of free $n - \bar{n}$ oscillation is 4.7×10^8 s, set by the Super-Kamiokande (SK) experiment in 2021, converted from an ^{16}O nucleus-bound lifetime of 3.6×10^{32} years, using a factor derived from theory [41].

2.5.1 Searches in free neutrons

One experimental approach of searching for $n-\bar{n}$ oscillation is using a neutron beam propagated to an annihilation target. A beam of slow neutrons is projected to a distant annihilation target. If a neutron in the beam has transformed into an antineutron during the propagation, pair annihilation is expected at the target, so that the surrounding detector captures the pair-annihilation signature. The current best limit of $n - \bar{n}$ lifetime using this approach comes from the search at the Institut Laue-Langevin (ILL) [39] in Grenoble.

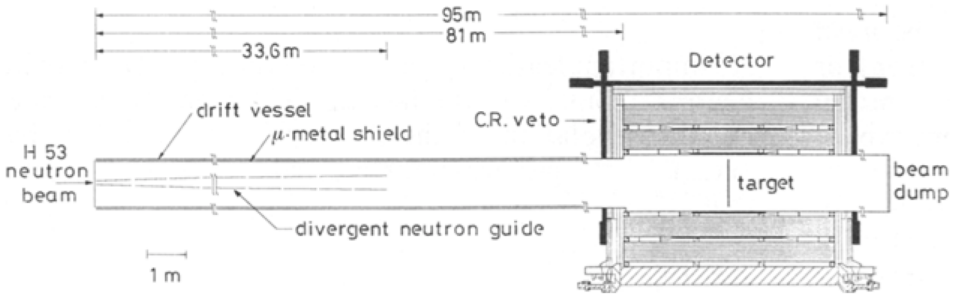


Figure 2.2: Experimental setup for the neutron-antineutron oscillation search conducted with a cold neutron beam at the Institut Laue-Langevin [39].

The ILL experiment uses a cold neutron beam from their 58 MW research reactor with a neutron current 1.25×10^{11}n/s on the annihilation target. A carbon foil is used as the annihilation target. The setup of the ILL experiment is shown in Fig. 2.2 schematically. This search set the

current best 90% C. L. limit on $n - \bar{n}$ lifetime using quasi-free neutron at $8.6 \times 10^7 s$. The NNBAR experiment at the European Spallation Source (ESS) [42] plans to search for $n - \bar{n}$ with $\sim$ 1000 times improved sensitivity using a higher-intensity pulsed neutron beam in the near future.

2.5.2 Searches in bound neutrons

Neutrons in a nucleus-bound state can also spontaneously oscillate into antineutrons, but this oscillation probability is heavily suppressed compared to the quasi-free case. On the other hand, detectors with large mass contained within their volume can contain many orders of magnitude more neutrons than a quasi-free neutron beam; this suppressed oscillation can therefore be offset with a far greater neutron number density. Thus, an $n - \bar{n}$ search with bound neutrons can offer complementary searches, and lifetime limits competitive with quasi-free $n - \bar{n}$ oscillation experiments.

Several detection techniques across multiple experiments have been applied for nuclear-bound $n - \bar{n}$ oscillation searches, including water Cherenkov detectors and iron calorimeters. These searches have set $n - \bar{n}$ oscillation lifetime limits, where these limits can be expressed in terms of equivalent free neutron lifetimes, using the conversion relation described in Sec. 2.4.5. Different conversion factors are applied for different nuclei. Table 2.3 summarizes the experimental results in chronological order.

The first search for bound $n - \bar{n}$ oscillation was carried out at the Homestake mine in 1983 [43], shortly before the first free search at ILL. This search set a free-equivalent $n - \bar{n}$ lifetime limit at $2 \times 10^7 s$ using a 300-ton water Cherenkov detector. Then, the search by the Irvine-Michigan-Brookhaven Collaboration in 1984 [44] surpassed this limit using a water Cherenkov detector with a much larger fiducial mass of 3300 tons. The new free-equivalent limit was set at 1.1×10^8s. Then in 1986, the search by the Kamiokande collaboration set the free-equivalent lifetime limit at 1.2×10^8s again using a water Cherenkov detection technique. In 1990, the first non-Cherenkov bound $n - \bar{n}$ oscillation search was carried out in the Frèjus fine-grained tracker detector [48]. The Frèjus detector was a 900-ton iron calorimeter, using ^{56}Fe as its target nucleus, at a depth of 1.8

Experiment	Year	Type	Nucleus	Detector Technique	R [$10^{23} s^{-1}$]	τ_{free}limit [s]
Homestake [43]	1983	Bound	^{16}O	Water Cherenkov	1.0	2×10^7
IMB [44]	1984	Bound	^{16}O	Water Cherenkov	1.0	1.1×10^8
ILL [45]	1985	Free	–	–	–	1.0×10^6
Kamiokande [46]	1988	Bound	^{16}O	Water Cherenkov	1.0	1.2×10^8
Triga Mk. II [47]	1989	Free	–	–	–	4.9×10^5
Frejus [48]	1990	Bound	^{56}Fe	Iron calorimeter	1.4	1.2×10^8
ILL [39]	1994	Free	–	–	–	8.6×10^7
Soudan 2 [49]	2002	Bound	^{56}Fe	Iron calorimeter	1.4	1.3×10^8
Super-Kamiokande [50]	2015	Bound	^{16}O	Water Cherenkov	1.0	2.7×10^8
SNO [51]	2017	Bound	$^2H/^{16}O$	Heavy water Cherenkov	0.25/1.0	1.2×10^8
Super-Kamiokande [41]	2021	Bound	^{16}O	Water Cherenkov	1.0	4.7×10^8

Table 2.3: List of searches for $n - \bar{n}$ oscillation, and the limits they set on the free lifetime of the process. For searches in bound nuclei, the target nucleus and corresponding rate suppression factor.

km underground. This measurement set a free-equivalent lifetime limit of $1.2 \times 10^8 s$. In 2002, the Soudan 2 collaboration also used the iron calorimeter technique. The measurement using a 770 ton fiducial volume set a lifetime limit at 1.3×10^8s [49]. A decade later, in 2015, the Super-Kamiokande collaboration performed a search for $n-\bar{n}$ oscillation with a water Cherenkov detector with a 22.5 kton fiducial mass over 1489 days, and a limit of 2.7×10^8 s was set. In 2017, the SNO collaboration carried out a search using a heavy water Cherenkov detector, considering ^2D and ^{16}O [51], setting a free-equivalent limit of 1.23×10^8s. The Super-Kamiokande collaboration updated their limit in 2021 using a greater exposure setting a free-equivalent limit at 4.7×10^8s [41]. To date, this limit is the strictest lifetime limit set on $n - \bar{n}$ oscillation.

Chapter 3: Liquid Argon Time Projection Chambers

The Liquid Argon Time Projection Chamber (LArTPC) detector technology is a technology that recently gained significant popularity for neutrino detection in neutrino experiments. The Icarus experiment[1] at the Gran Sasso laboratory was the first neutrino experiment to use a LArTPC [52]. In the U. S., the Argoneut TPC [53] used LArTPC technology for the first time (the TPC was later used as part of the LArIAT [54] test beam experiment). MicroBooNE, an 80 ton active liquid argon volume LArTPC located on the Fermi National Accelerator Laboratory (Fermilab) Booster Neutrino Beam (BNB), is the first large-scale LArTPC to run in the U. S. [55]. MicroBooNE began operating in 2015 with the main physics objective of investigating the MiniBooNE low energy excess (LEE) [56]. MicroBooNE and has produced physics results including neutrino cross-section measurements [57–59], LEE investigation [60–64], astrophysics & exotics searches [65, 66], and also R&D for future LArTPCs [67, 68]. The Deep Underground Neutrino Experiment (DUNE) is the future large-scale LArTPC in the U. S. DUNE's near detector (ND) will be located at Fermilab, and its far detector will be located at Sanford Underground Research Facility (SURF). DUNE's four FD modules are 55.6 kton active volume in total [69]. DUNE FD will be the largest ever LArTPC once built.

In this chapter, the general working principles of LArTPCs and the methodology for the signal collection and processing of MicroBooNE & DUNE detector data relevant for the $n - \bar{n}$ oscillation search are discussed.

[1] The detector was later moved to the U. S. and will be operating as the far detector of the Short-Baseline Neutrino (SBN) program.

3.1 LArTPC detector principles

LArTPCs are high-resolution imaging detectors. Particle interactions within LArTPCs can be measured by collecting ionization electrons across a large volume of liquid argon. When a charged particle travels through liquid argon, it leaves a trace of ionization electrons along its trajectory. A uniform electric field drifts the ionization electrons towards the anode where the signal is recorded by a set of wire planes, as shown in Fig. 3.1. If ionization electrons are produced near the cathode

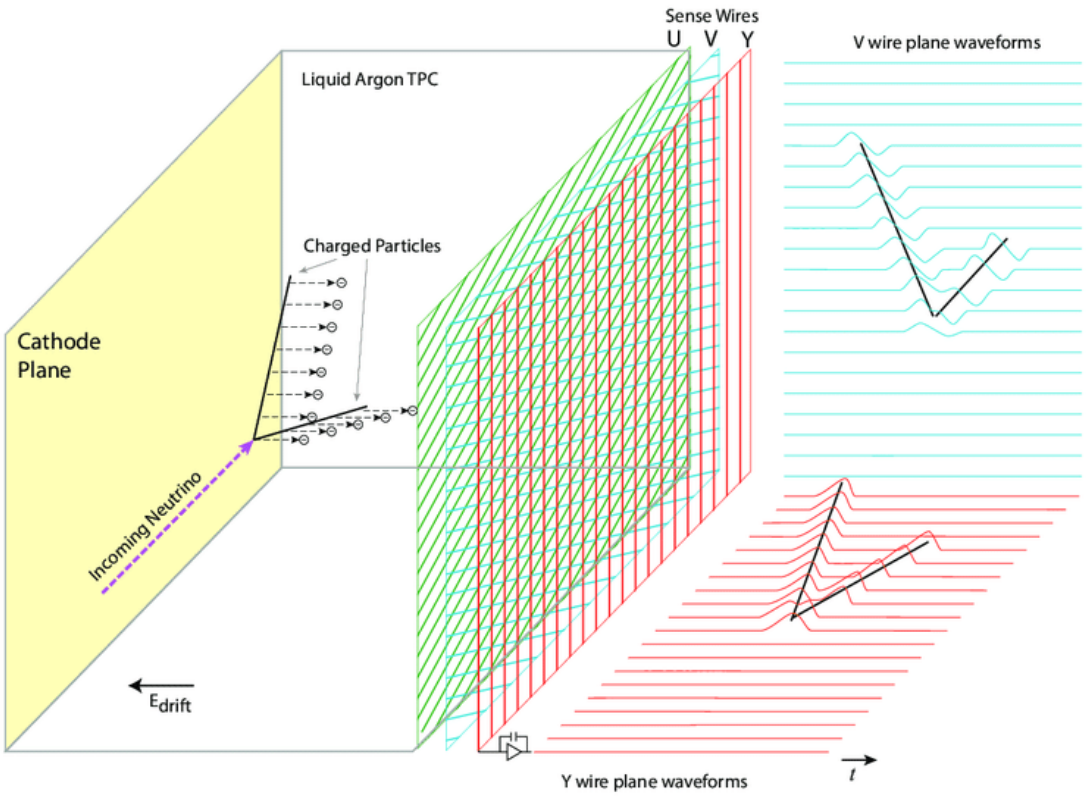


Figure 3.1: A diagram of the operating principle of a LArTPC. Charged particles traveling in the argon volume ionize and excite the argon. A uniform electric drift field is applied, which drifts the ionization electrons to a set of wire planes where the signal is recorded (anode plane). These wire planes are oriented at different angles to produce an image of the interaction in different projections. Image credit: [55]

plane, the longest drift can be a few ms. Scintillation light is also produced inside the detector from the excitation and ionization of the argon caused by the charged particles. This light is collected by photomultiplier tubes (PMTs). Since this light signal arrives with a ≈ns delay, it provides prompt information on the particle interaction time. By combining the wire signals and the light signals, a particle interaction can be measured and reconstructed in a LArTPC in 3D.

3.2 The MicroBooNE detector

3.2.1 MicroBooNE TPC

MicroBooNE is a LArTPC based at Fermilab, operating on the surface without significant overburden. MicroBooNE's TPC is a rectangular prism, 2.56 meters wide along the drift direction (horizontally, and perpendicularly to the neutrino beam direction), 2.32 meters tall, and 10.36 meters long along the neutrino beam direction, with an active volume of 85 tonnes of liquid argon. The cylindrical cryostat system and the MicroBooNE TPC field cage are shown in Fig. 3.2. In Fig. 3.3, the rendering of the TPC and the actual TPC after assembly is shown. In this figure, the right-hand side wall of the rectangular volume is the anode plane where the X position in the detector coordinate system is zero; the cathode plane is 2.56 meters down to the left, as shown in purple in the schematic diagram. The ionization electron between the cathode and the anode plane is drifted with a drift voltage of 273 V / cm. [2]

The anode plane consists of 3 wire planes, where the wire planes are parallel and lie on top of each other. The three wire planes are the "induction0" plane (U), the "induction1" plane (V) and the "collection" plane (Y), The wires are spaced with a spacing of 3 mm, the wires in the Y plane are vertical, and the wires in the U (V) plane are rotated by 60°(-60°) with respect to the vertical. The Y-plane contains 3456 wires and the U and V-plane contain 2400 wires each. Behind the wire planes, there are 32 photomultiplier tubes (PMTs) for light collection.

[2]MicroBooNE was initially designed and tested to withstand a 500 V/cm electric field across the TPC active volume. However, it has been operating at 273 V / cm during data collection.

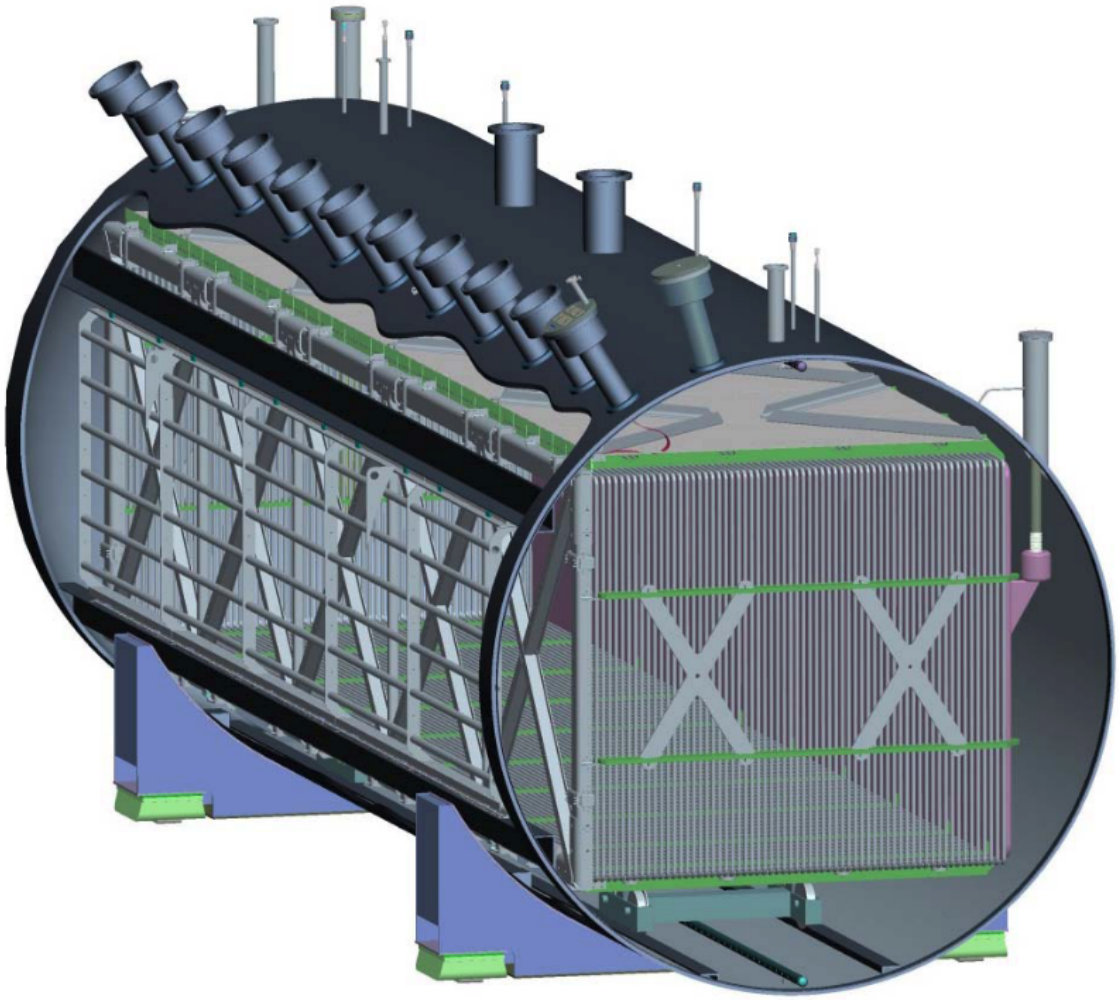


Figure 3.2: A schematic diagram demonstrating the MicroBooNE cryostat and the rectangular field cage within. The TPC is approximately 10 meters long, and approximately 2.3 meters tall and wide. Image credit: [55]

When charged particles propagate and leave trailing ionization electrons in the MicroBooNE TPC, as illustrated in Fig. 3.1, these electrons are pulled to the anode plane. These electrons first pass through the induction planes (U- and V-planes), generating bipolar signal waveforms on the wire planes as they approach and move away from the wire planes. Then, they are collected by the collection plane (Y-plane), generating unipolar signal waveforms on the collection plane wires. These signal waveforms on the wire planes are shown in Fig. 3.4.

The charge received by each wire corresponds to the energy deposited by the inciting ionizing

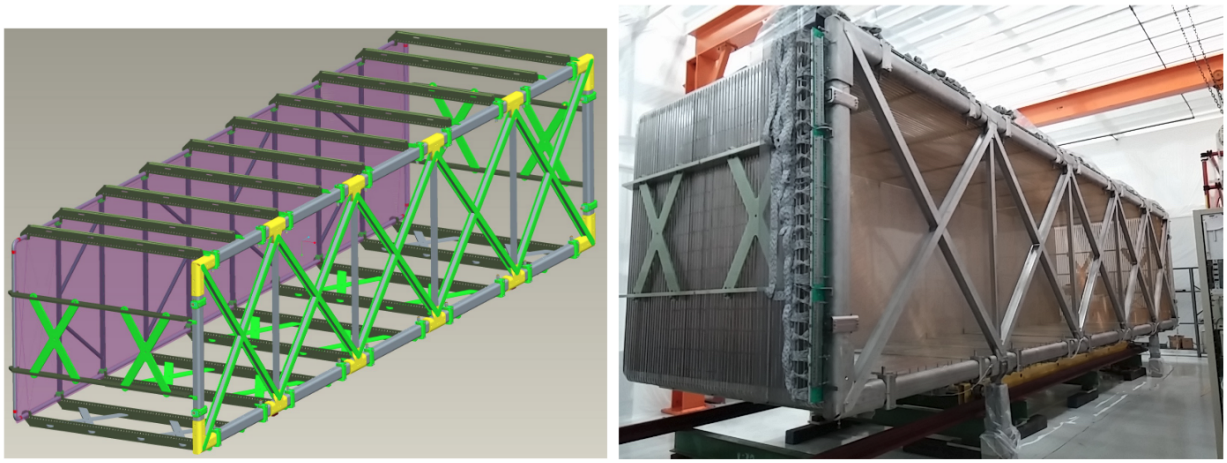


Figure 3.3: (Left) A schematic diagram of TPC and the field cage. (Right) Assembled LArTPC after wire and electronics installation. Image credit: [55]

particle, and by combining information about which wires observed particular waveforms, one can reconstruct the 2D coordinates of each ionized electron in the detector (on the YZ plane). If the timing of the initiation of an interaction is known (t_0), the X direction (drift direction) spatial location drift can be found. Typically, t_0 is measured by the light collection system (discussed in Sec. 3.2.3.1), since the light collection system has a time resolution of the order of ns, whereas the TPC drift can be as long as 2.3 ms for MicroBooNE.

3.2.2 TPC readout

The readout system for the TPC signal in MicroBooNE is composed of cold (inside the cryostat) and warm (outside the cryostat) electronics. Charge information on wires in both the collection and induction planes in the anode plane is read through electronics chips known as ASICs (application-specific integrated circuits) attached to the ends of the TPC wires. Operating at low temperature, the ASICs act as low-noise pre-amplifiers. After pre-amplification, cold cables carry the analog signals outside the cryostat. Then, outside the cryostat, the analog signals are converted to digital with a set of analog-to-digital converter front-end module (ADC/FEM) electronics boards that come from the warm electronics. Here the analog signal is digitized by a 16 MHz clock with

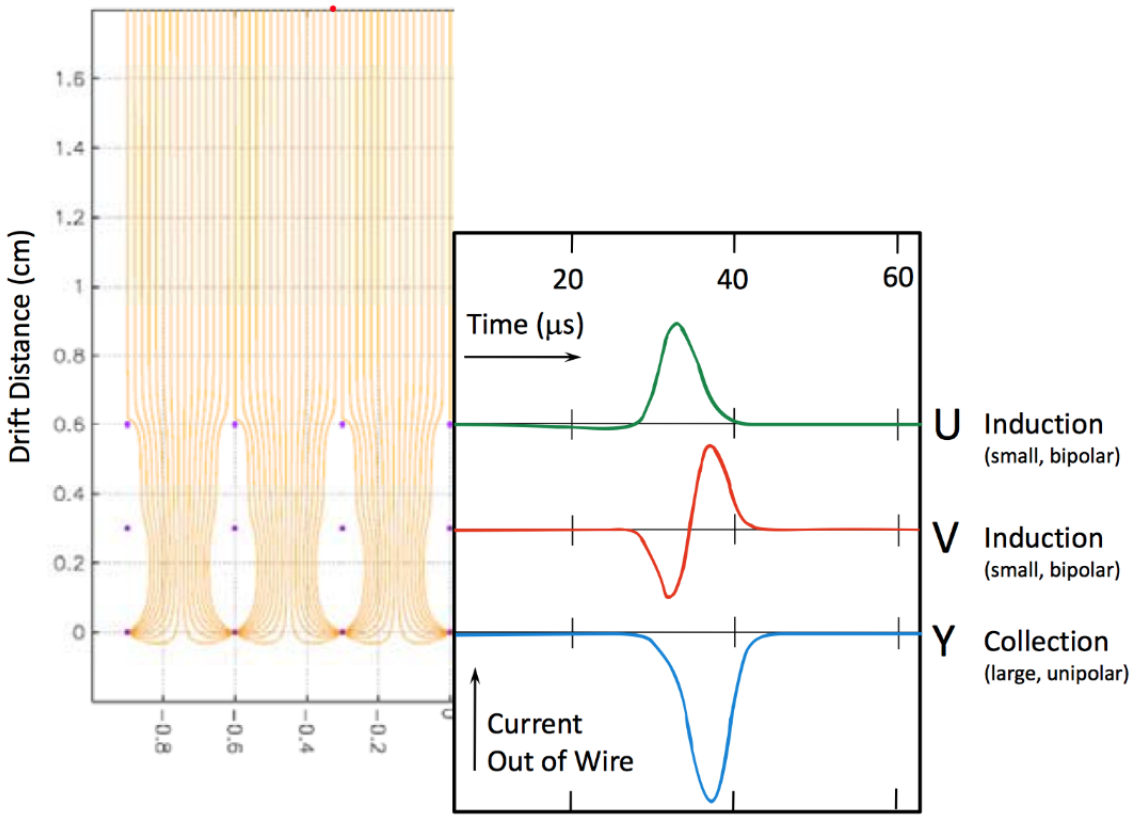


Figure 3.4: Example of the signal waveforms induced on all three wire planes by ionization electrons. The x-axis on the left figure shows the distance (cm) parallel to the collection plane, while the y-axis shows the distance from the collection plane (cm). These wires are shown as purple dots; U-plane wires are located at y=0.6 cm, V-plane wires are located at y=0.3 cm, Y-plane wires at y=0.0cm. The lines in the left plot represent electric field lines. The right plot shows an example of the waveforms produced on each wire plane. A bipolar but largely positive signal appears on the first induction plane, as electrons approach, a more evenly bipolar signal appears on the second induction plane, as the electrons pass by. Then, a unipolar negative signal on the collection plane appears, as the electrons are collected. Image credit: [70]

12 bit (0-4095) ADC. This 16 MHz sampling rate is downsampled to 2 MHz; one time-tick is 500 ns length. When TPC signals are digitized at 2 MHz rates, each 4.8 ms length of TPC information is treated as a unit of a readout for any given trigger; this is the equivalent of 9600 time-tick samplings. Event displays of particle interactions in LArTPC data typically represent the 2D imageries formed with time-tick sampling on one axis and wire numbers within a given plane in another axis,

where the color represents the ADC values.

3.2.3 Light detection and triggering

3.2.3.1 Light detection

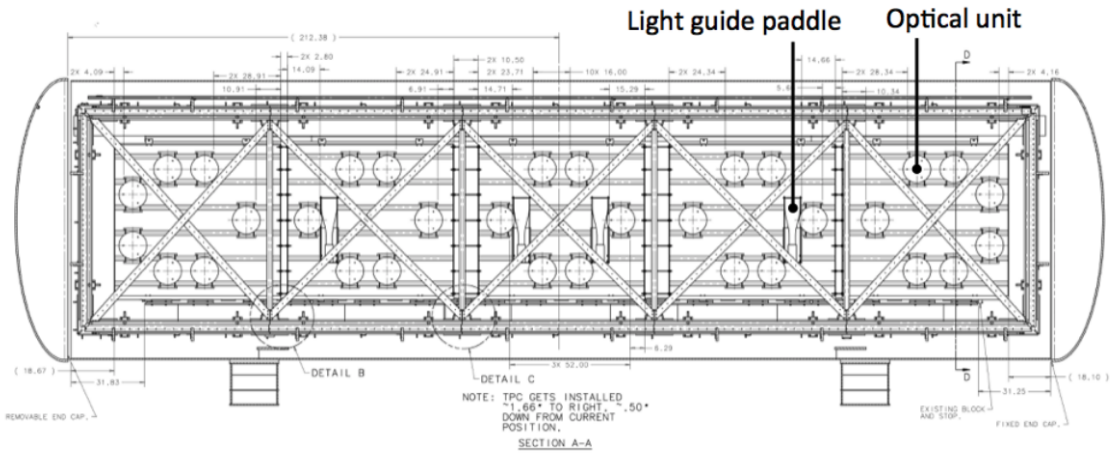


Figure 3.5: The MicroBooNE light collection system consists of a primary system of 32 optical units (PMTs) and a secondary optical system of four light guide paddles. The light collection system of MicroBooNE is placed behind the anode wire planes such that the view is not obscured by structural cross bars of the LArTPC. Image credit: [55]

MicroBooNEs light detection system includes 32 Hamamatsu R912-02mod photomultipliers (PMTs) mounted behind the anode plane. The deployment of PMTs on the side of the TPC volume is shown in Fig. 3.5. The delay due to the travel time in MicroBooNE is on the order of ns for PMT signals, while it can be a few ms for TPC signals due to the different speeds of drift electrons and photons. Therefore, PMT information plays a critical role in identifying the interaction time (t_0) and the time and position of a neutrino interaction.

3.2.3.2 Triggering

To record neutrino interactions (“BNB” data) or to take off-beam data (“EXT” data a.k.a. beam-external data), MicroBooNE uses a dedicated trigger system. For BNB data, MicroBooNE’s

Data Acquisition (DAQ) system receives queues from the BNB accelerator when the neutrino beam spill occurs. The length of this spill is 1.6 μs, with a frequency of 15 Hz. Additionally, a software trigger [71] is applied to BNB data, requiring a coincidental PMT signal with the beam spill [72]. EXT data is external to BNB, avoiding the beam spills. EXT data is collected using trigger signals given by a pulse generator with a given frequency 4-7Hz without a PMT trigger requirement. The timing of the pulse generator becomes the trigger time (t_0) for the EXT data. In many MicroBooNE analyses, including recent LEE analyses [60–64], BNB signal simulation adopted an "overlay" scheme [73], overlaying signal Monte Carlo (MC) simulation on top of the EXT data. For the $n - \bar{n}$ search in MicroBooNE, the same method is adopted. Signal simulation and background simulation is further discussed in detail in Sec. 7.1.1, 7.1.3 and 7.1.4.

Typically, the combined information from 4.8 ms of TPC readout and the coincident 6.4 ms of PMT readout defines "one event" in MicroBooNE [74]. In the $n - \bar{n}$ search analysis in MicroBooNE, which is described in Ch. 7, "one event" is defined as a time interval between the trigger time (t_0) and 2.3 ms after the trigger time (t_0+2.3 ms) in order to account for the full-drift active volume in the MicroBooNE TPC.

3.2.4 Signal deconvolution & hit finding

3.2.4.1 Deconvolution

Electron signals on the MicroBooNE TPC wires are processed through deconvolution of various detector effects that can alter or skew the TPC wire signals [75, 76]. Deconvolution aims to solve for the actual number of ionization electrons in the detector from the wire signals. Thus, it is a vital step before any higher-level reconstruction is performed using TPC information.

In practice, deconvolving possible detector effects on the signal means solving the below equation where $M_i(t_0)$ is the measured electric signal on a wire i at a time t_0 and $S(t)$ is the true, real

electrical signal (ionization electron) drifted to that wire.

$$M_i(t_0) = \int_{-\infty}^{\infty} R_i(t, t_0) S(t) dt \tag{3.1}$$

$R_i(t, t_0)$ represents the detector response, describing how a signal changes over a time interval $t - t_0$. Solving this equation works well using the collection plane only, retrieving the collected charges. This is referred to as 1D deconvolution.

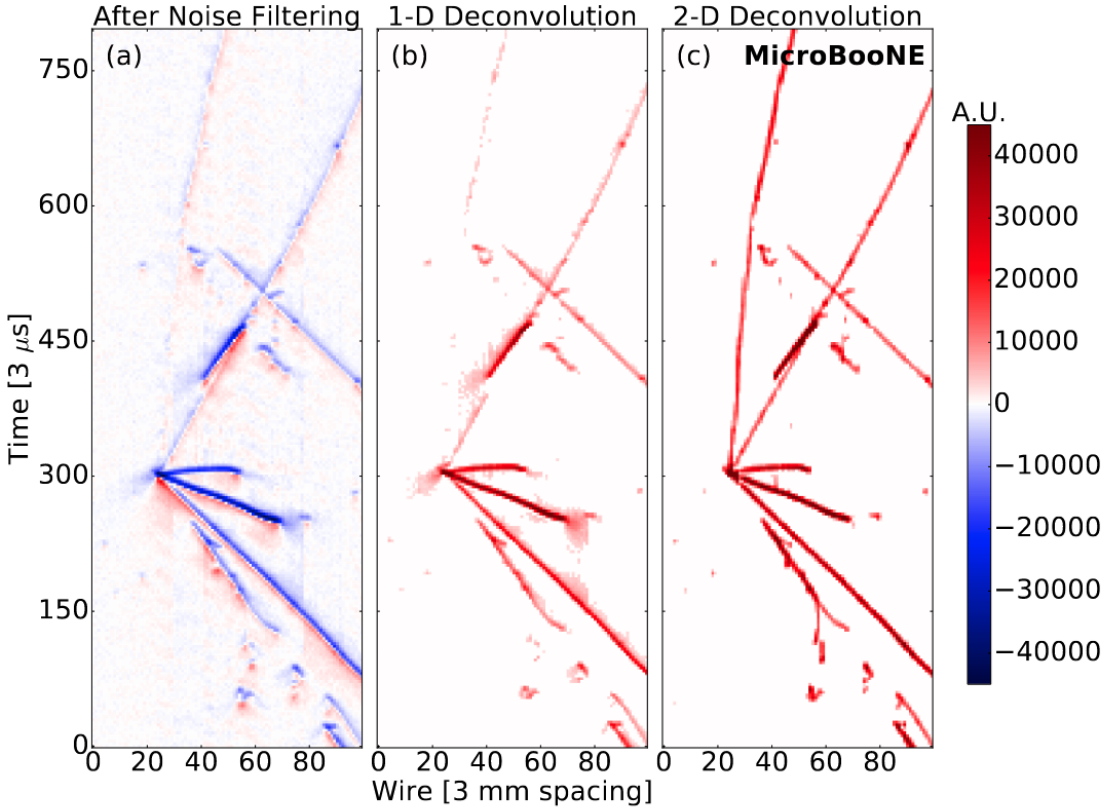


Figure 3.6: A neutrino candidate event from MicroBooNE data (event 41075, run 3493) measured on the U-plane. (a) Raw waveform after noise filtering. (b) Charge spectrum after signal processing with 1D deconvolution. (c) Charge spectrum after signal processing with 2D deconvolution. Image credit: [75]

When the induction wires are taken into account, one electron leaves multiple signals on the wires, as electrons pass the induction planes before arriving on the collection plane. Therefore, this Eq. 3.1, can expand to integrate over both time and wire dimensions. This is referred to as 2D

deconvolution; the expanded equation is below,

$$M_i(t_0) = \int_{-\infty}^{\infty} (... + R_1(t, t_0)S_{i-1}(t) + R_0(t, t_0)S_i(t) + R_1(t, t_0)S_{i+1}(t) + ...)dt \tag{3.2}$$

where M_i represents the measured signal from the wire i. S_{i-1}, S_i and S_{i+1} represent the real signal within the boundaries of the wire i and its adjacent neighbors. R_0 represents an average response function for an ionization charge passing through the wire region of interest. The average is taken over all possible drift paths through the wire region. Similarly, R_i represents the average response function for an ionization charge drifting past and in the next adjacent wire region. One can expand this definition to the n number of neighbors by introducing terms up to R_n. Equation 3.2 assumes translational invariance in the response function (that is, R does not depend on the actual location of the wire). The solutions for Eq. 3.1 and Eq. 3.2 are obtained by solving the Fourier transformed equations. The results of the 1D convolution and the 2D convolution are shown in Fig. 3.6. A clear improvement in the clarity of the track signatures is demonstrated in the TPC signals.

3.2.4.2 Hit finding

After applying 2D deconvolution to wire signals, the signal waveforms are fit to the Gaussian function around the peak. From the fit, the area of the Gaussian is interpreted as the amount of charge deposition at a given time. Combining the information from the Gaussian fit and the wire plane and wire number of the signal, a "hit" is reconstructed. These Gaussian hits are also called ROIs (region of interest). Now, the information over the 4.8 ms length readout is replaced with reconstructed hits, which are the basis of reconstruction for MicroBooNE analyses. This hit information is passed to higher-level reconstruction paradigms such as Deep-Learning [77], Pandora [78], and Wire-Cell [79] reconstruction. The Wire-Cell reconstruction method used for seeding and validating the $n - \bar{n}$ search analysis is discussed in Sec. 7.2.

3.3 The DUNE detector

DUNE is a proposed future large-scale LArTPC in the U.S. consisting of a near detector (ND) at Fermilab and a far detector (FD) at SURF, 800 miles (1300 km) from the neutrino beam source at Fermilab, as shown in Fig. 3.7. DUNE's physics objectives include [80, 81]:

- Performing a comprehensive program of neutrino oscillation measurements using ν_μ and $\bar{\nu}_\mu$ beams from Fermilab. This includes measurements of the charge parity (CP) violating phase, determination of the neutrino mass ordering (the sign of $\Delta m^2_{31} \equiv m^2_3 - m^2_1$), and more. Especially, the search for CPV in neutrino oscillations has the potential of offering an insight into the origin of the matter-antimatter asymmetry.

- Searching for baryon number violation processes such as proton decay, neutron-antineutron oscillation, etc. The observation of BNV would be a ground-breaking discovery in physics.

- Detecting and measuring the ν_e flux from a core-collapse supernova within our galaxy. Such a measurement would provide useful insight into the early stages of core-collapse and neutrino properties.

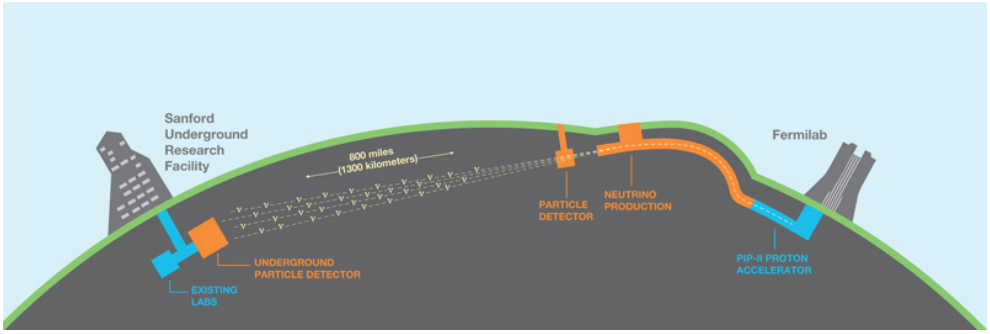


Figure 3.7: Neutrino beam from Fermilab in Illinois to Sanford Underground Research Facility (SURF) in South Dakota. The near detector (ND) is located at 575 meters from the neutrino production. The far detector (FD) is located at 1300 km from the neutrino production.

When these physics objectives are accomplished, each of them will provide significant insights into the high-energy physics community. These goals are supported by detectors (both ND and

FD) and DAQ designs for DUNE. In subsequent sections, the FD and DAQ designs are discussed, since those are critical components for a sensitive $n - \bar{n}$ search in DUNE.

3.3.1 DUNE Far Detector

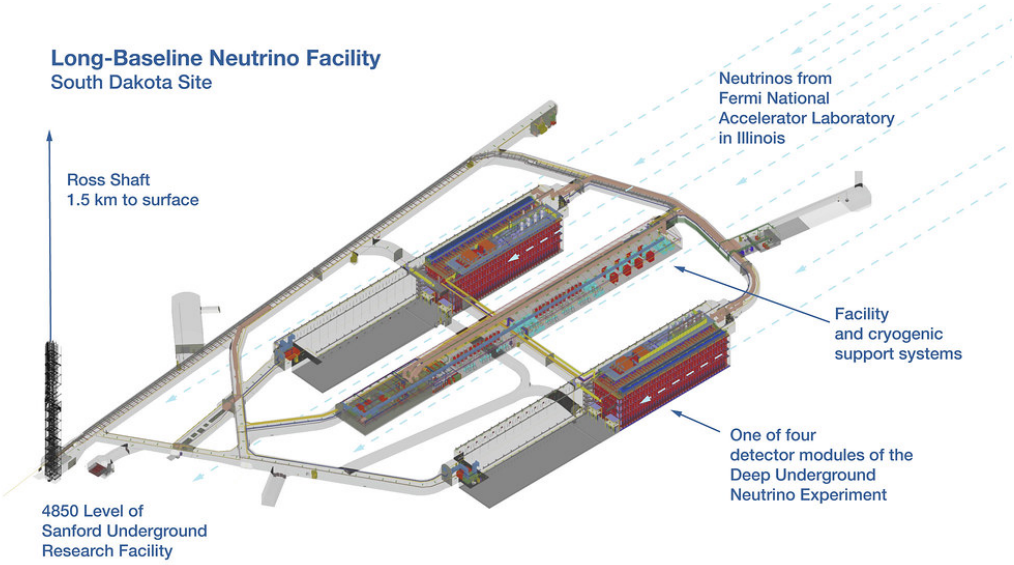


Figure 3.8: Four DUNE FD module will be located 1.5 km underground in South Dakota.

The DUNE FD will consist of four LArTPC detector modules, each with a fiducial LAr mass of at least 10 kt, reaching 40 kt fiducial volume. All FD modules will be installed approximately 1.5 km underground. DUNE is planning and currently developing two LArTPC technologies;

- Single-phase horizontal drift (HD) technology: The drift field is applied in the horizontal direction in the LAr volume.
- Single-phase vertical drift (VD) technology: The drift field is applied in the vertical direction in the LAr volume.

For the case of horizontal drift, the working principle of the LArTPC is shown in Fig. 3.1, and is similar to that of the MicroBooNE detector. The work presented in this thesis that uses DUNE simulation (Ch. 5, Ch. 6) assumes horizontal drift (or similar performance) for all four modules.

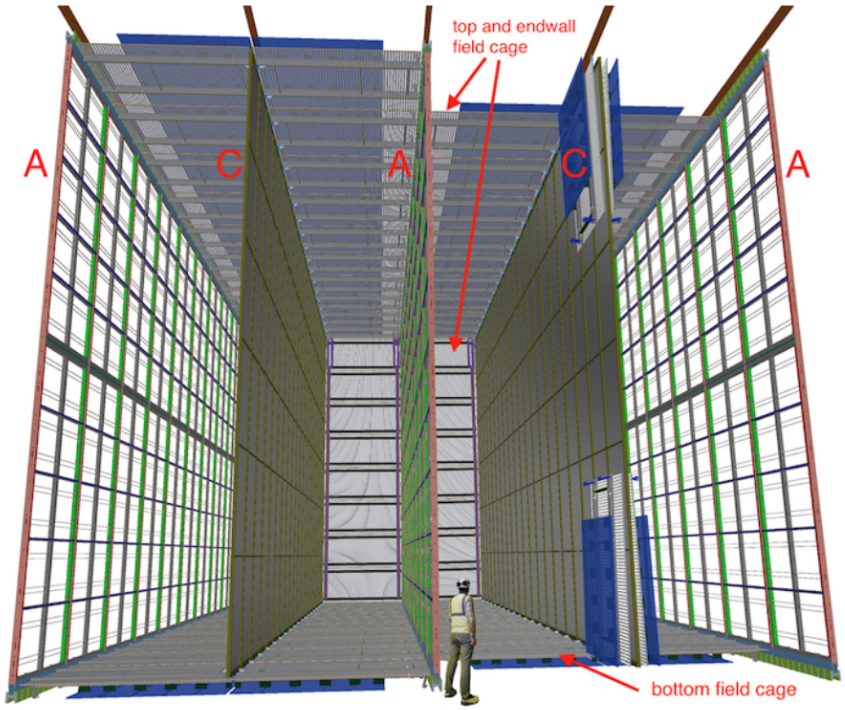


Figure 3.9: DUNE FD module TPC. The module is 58.2 meters deep, 14.0 meters wide, and 12.0 meters high. The anode plane and the cathode plane are vertically oriented and arranged in the alternating order (A-C-A-C-A). One FD module can be sectioned by four drift volumes, where a drift volume is defined as the contiguous volume between the anode plane and the cathode plane. Image credit: [69].

Compared to MicroBooNE TPC, one DUNE FD module is about 100 times larger in terms of active volume. The TPC volume in DUNE FD is designed to be partitioned into four drift volumes by alternating anode planes and cathode planes as shown in Fig. 3.9. A strong drift field of 500 V/cm is applied in the drift volumes. The maximum drift length is 3.5 m, which is the length between an anode plane and a cathode plane, corresponding to ≈2.5 ms of drift time.

An FD module includes three module length (58.2 meters) anode planes. An anode plane is constructed from 6 meters high by 2.3 meters wide anode plane assemblies (APAs), stacked two APAs high and 25 wide, for 50 APAs per plane, and 150 for one FD module. Each APA consists of two induction wire planes (U- and V-plane) and one collection wire plane (W-plane). The wire signals on these wire planes are similar to those shown in MicroBooNE detector, see Fig. 3.4.

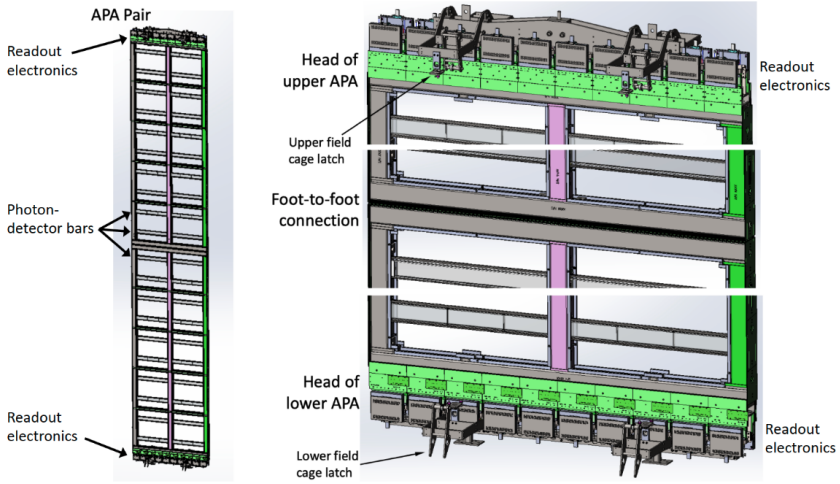


Figure 3.10: Left: two APAs vertically stacked together. PD bars can be seen installed across the width of the APAs. Right: a zoom into the top and bottom ends of the APA stack showing the connected readout electronics, and the center where the APAs are linked. Image credit: [69].

Also, one drift volume can be partitioned as 50 "drift regions" when one drift region corresponds to 1 APA within the drift volume. On the other hand, if we look at the entire module, as the middle drift regions share an anode plane, these APAs in the middle are responsible for two drift regions each. Thus there are 200 drift regions for one FD module. A drift region is a useful unit for the $n-\bar{n}$ analysis in DUNE and DUNE's trigger development because it can define the amount of data with a relatively localized signature, and the DAQ unit is shared within the entire drift region of the FD.

An APA also includes photon detectors (PDs) for light information collection as well as the wire planes for electron signals. An APA includes 10 PDs, which are bar-shaped light guides coupled to silicon photomultipliers (SiPMs) [69]. In Fig. 3.10, a pair of APAs is shown. Here, the PD bars and readout electronics attached to the APAs are shown as well. The light information from the PDs and the waveforms from the wire planes are handed to the readout electronics within the DUNE Data Acquisition (DAQ) system.

3.3.2 DUNE FD DAQ

The FD DAQ includes an upstream DAQ located underground connected to the readout from the APAs, and a downstream DAQ back-end subsystem (DAQ BE) located on the ground at SURF. The upstream DAQ preprocesses data and sends information to the trigger/data selection system (also underground), and the received data is buffered in upstream DAQ until the DAQ BE signals it is ready to receive data. The data rate to tape is aimed at no more than 30 PB/year for the entire FD. Compared to the raw data rate of 5 TB/s, this places stringent regulations on the data selection system. DUNE FD plans on collecting data continuously year-round during its operation, primarily not to risk missing a supernova burst (SNB) within our galaxy and ensuring maximal search exposure for BNV searches. In order to achieve DUNE's scientific goals, a suitable and efficient self-triggering scheme is needed, processing the continuous DUNE FD readout. For all FD readouts, the upstream DAQ triggering requires about 10^4 reduction factor for APA readout information, to meet the data rate limitation. Thus, the continuous self-triggering system in the upstream DAQ must meet the data rate requirement while providing a very high trigger efficiency for SNBs and BNV processes such as proton decay and neutron-antineutron oscillation.

DUNE FD DAQ architecture is based on the Front-End Link eXchange (FELIX) system, which is designed at CERN and used for ATLAS experiment [80]. One DAQ readout unit (DAQ RU) corresponds to a FELIX board with an onboard CPU and FPGA. For 150 APAs, 75 DAQ RUs are assigned. Thus the information from 2 APAs is read out and processed in a shared DAQ RU. The wire readouts are pre-processed in the FPGA on FELIX. In the nominal DAQ RU design, the FPGA performs noise filtering and threshold-base hit-finding based on the digitized waveforms. The CPU on FELIX assembles these found hits to form "trigger candidates," then trigger candidates from DAQ RUs are gathered for an APA-level and subsequently module-level trigger decision. Nominally, this trigger decision makes use of simple hit clustering and thresholding techniques. In Ch. 6, an alternative triggering scheme using CNN implementation on FPGAs is introduced, and the viability of that scheme is explored.

Chapter 4: Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a tool born out of the field of computer vision, where computer vision aims to identify, analyze, and process visual data efficiently. This area of study has become extremely important since the widespread use of the digital camera in the 2000s, when visual data became the most significant part of all digital data on the internet. Computer scientists had the ambition to classify every image on the internet automatically, and ImageNET was the fruition of that vision [82]. ImageNET includes more than 14 million images hand-annotated for more than 20,000 categories of pictures of animals, objects, etc. Since 2010, the annual ImageNET challenge to select high-performing models to classify 1,000 categories from ImageNET has become the standard to measure a model's learning power.

Some of the renowned CNN architectures such as AlexNet [83], VGG net [84], and ResNet [85] gained their fame by winning the ImageNET challenge. The VGG16 architecture is used as the baseline model for the CNN classifiers in event selections in Ch. 5 and Ch. 7. These benchmarking CNN architectures and many other CNN networks typically have layers of convolutions, Rectified Linear Unit (ReLU), pooling, and softmax [1]. In Sec. 4.1, the working principle of CNNs and the roles of these layers are briefly discussed.

In high-energy physics, particle interactions can often be represented in a 3D space with particle trajectories. The technology of capturing the interaction signatures and accurately reconstructing them has been a crucial part of the HEP area of research. As CNNs have demonstrated the excellence of learning visual data classifications, the exploration of ML/DL techniques by physicists is growing at an unprecedented rate [86]. This approach is, of course, prominent in physics using LArTPCs. DL-based reconstruction in MicroBooNE has also shown its strength as an analy-

[1] Techniques using hidden neural layers (such as CNNs) are Deep-Learning (DL) techniques, whereas Machine-Learning (ML) includes broader algorithmic approaches such as Boosted Decision Trees (BDTs).

sis tool [63, 87, 88]. On the other hand, Machine-learning (ML) based higher-level event selection has been demonstrated on the recent $n - \bar{n}$ search by Super-Kamiokande [41]. This is an improved search from their previous $n - \bar{n}$ search [50] with a smaller dataset, using a traditional reconstruction and a cut-based event selection. The new search shows improved analysis using multivariate analysis and a greater exposure [41]. In addition to reconstruction and data analysis applications, CNNs can be alternatives to traditional data-selection (triggering) schemes in LArTPCs, which is discussed in Ch. 6.

These specific uses of CNNs can be catered to uniquely address the goals of various selection algorithms. In Sec. 4.3.1, *Minkowski Engine* [89], the tool to perform sparse network training, is described. In Sec. 4.3.2, *hls4ml* [90], the tool for implementing CNNs on a Field Programmable Gate Array (FPGA) is described.

4.1 CNN working principles

Fig. 4.1 shows the schematic figure of a typical CNN architecture with multiple convolutions. The figure flows from the left to the right, from the input to the score layer. The input here is shown as three channels of 2D images; after a convolution, the input is mapped to smaller and deeper layers of features. For the next convolution, these feature layers from the previous convolution become input layers; then the convolution generates smaller and deeper feature layers. A Rectified Linear Unit (ReLU) and a pooling layer are typically followed by a convolutional layer. At the very last layer, the feature layers are flattened to a 1D array and connected to a categorical score, typically represented by the softmax function.

4.1.1 Convolution and activation layers

A convolution layer takes a local portion of the input and convolves it through a linear mapping between the input and the feature. This linear mapping containing weights ($\vec{w}$) and biases (b) is called a filter and typically initialized randomly. The convolution sweeps the input with localized

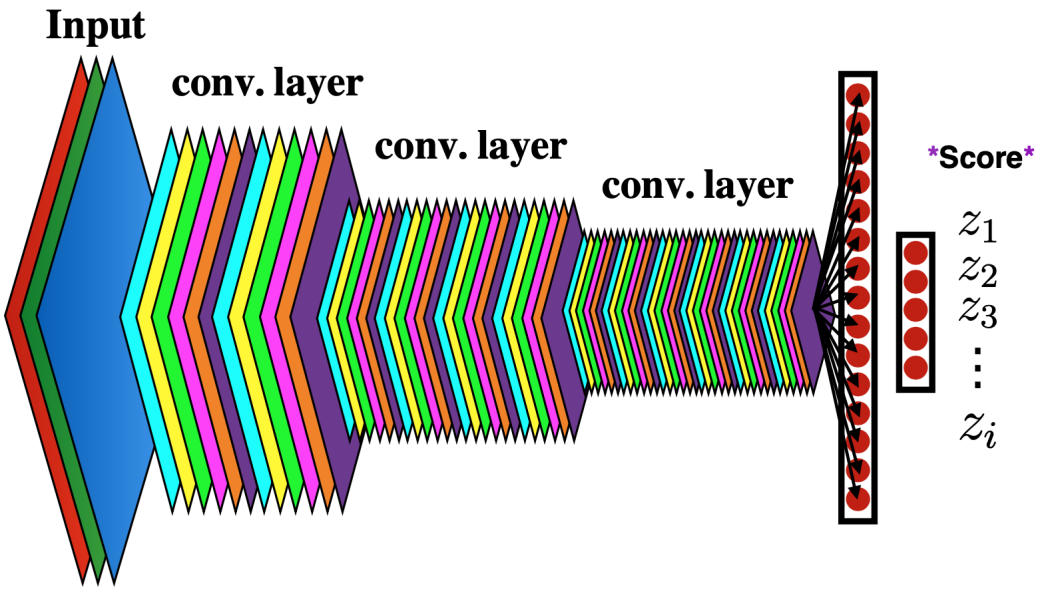


Figure 4.1: A schematic structure of a typical CNN architecture. The input representation is 2D image with three channels. The convolution is denoted as 'conv. layer'. The input reduces in its 2D size and grows in its depth over convolutions, until it is flattened at the score (softmax) layer.

filters; thus, the feature preserves the locality inherited from the input. Weights and biases are the learnable parameters in CNNs. During a CNN training, the weights and biases are updated through gradient descent, optimized to reduce the error in the prediction through the network. After a convolution, activation through the Rectified Linear Unit (ReLU) is followed; the ReLU function returns 0 if the value is negative, and returns the input value if the value is positive. The ReLU can be expressed as

$$\sigma(x) = \max(x, 0). \tag{4.1}$$

ReLU layer has the important purpose of introducing non-linearity in the CNN. Since the convolution itself is a linear mapping, when convolution layers are followed by one another, multiple layers of convolution essentially become one convolution, since multiple linear transformations can be rewritten as one linear transformation. Without non-linearity introduced by the ReLU, the depth of CNN constructed by multiple convolution layers would be meaningless. A ReLU layer after a convolution layer introduces non-linearity after a linear transformation; thus CNNs can build depth

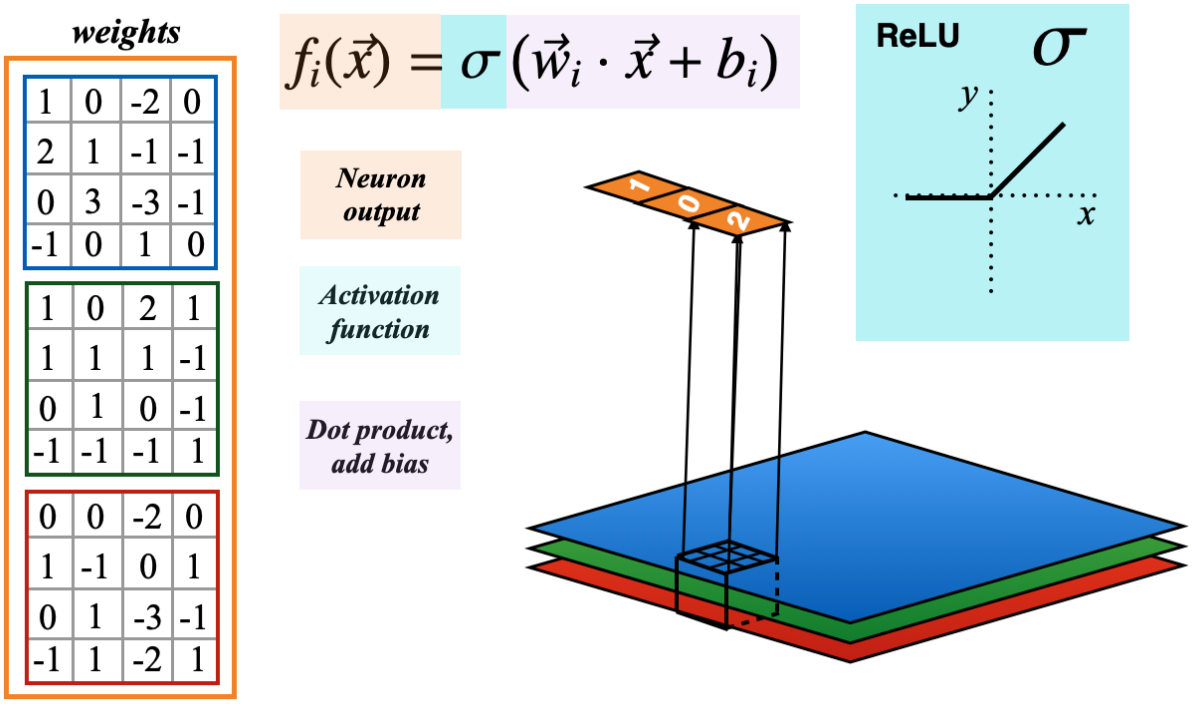


Figure 4.2: An illustration of a convolution and an activation. The localized convolution and activation extracts the features from inputs ($\vec{x}$) to feature map (f) by convolving with learnable parameters ($\vec{w}, b$) followed by ReLU activation (σ).

to their models. By using non-linear activation, CNNs can do "deep-learning."

4.1.2 Pooling layer

A pooling layer in a CNN architecture performs down-sampling of the previous layer. When a 2×2 dimension max-pooling is performed, the layer is down-sampled by a factor of 2 on each dimension and the maximum value of the original 2×2 pixel taken. Such a procedure is useful to manage the number of learnable parameters and scale down the complexity of CNNs.

4.1.3 Softmax (score) layer

The final layer in a CNN architecture represents categorical probability. This probability is normalized to be summed up to 1; this normalized probability is also called "softmax," or simply score. Equation 4.2 shows the softmax for k-th category

$$p_k = \frac{e^{a_k}}{\sum_{k'}^{K} e^{a_{k'}}}, \tag{4.2}$$

where a^k is the output of the final layer for the k-th category out of total K categories.

4.2 Network training and optimization

4.2.1 Batching and iteration

During network training, the learnable parameters (weights and biases) are updated at each training iteration. The training dataset is batched with a certain batching size (4, 8, 16, 32, ...). One batch is used per iteration, and accuracy and loss are evaluated at the end of an iteration; accuracy and loss are described in Sec. 4.2.2. Iteration and epoch are units that measure the duration of CNN training. One epoch equals the number of iterations to use the entire training dataset. (e.g., if a dataset has 64,000 inputs and batch size is 64, 1,000 iterations equals 1 epoch.) The learnable parameters are updated to minimize the loss function. The optimization for loss function minimization is enabled through the gradient descent technique supported by the back-propagation method. The optimization is further discussed in Sec. 4.2.3.

4.2.2 Prediction, accuracy, and loss

The softmax layer gives the categorical prediction (e.g., $\hat{y} = [0.1, 0.2, 0.7]$, for the case of the classification of 3 categories). If the prediction matches the input label (e.g., $y = [0, 0, 1]$), the input is accurately predicted. Accuracy is defined as the ratio between the number of correct

predictions and the number of inputs.

The loss function is another metric for the prediction performance; the value quantifies the number of mistakes. Successful training should show an increase in accuracy and a decrease in loss function. Choosing a loss function is not trivial; a popular baseline choice is a categorical cross-entropy loss (e.g., 'CategoricalCrossentropy' in PyTorch [91] framework). Categorical cross-entropy loss is defined with the prediction and the truth label as below,

$$Loss = -\sum_{k}^{K} y_k \cdot log\hat{y_k}, \tag{4.3}$$

where y_k is the input label and $\hat{y_k}$ is the prediction for the k th category and K is the total number of categories. The loss function for a batch with a batch size (N) during training iterations is defined the mean of the loss function of each input:

$$Loss_{batch} = -\frac{\sum_{i}^{N} Loss_i}{N}, \tag{4.4}$$

where $Loss_i$ is the loss for the $i-$th input in the batch.

4.2.3 Optimization

Loss minimization is achieved by a technique called gradient descent [86]. This technique is used broadly across the areas of ML techniques where there are optimizations (loss minimization in this case); it is the core concept of "how machines learn". The goal of machine learning in many cases is to teach the model to describe the dataset.[2] The learnable parameters $(\vec{w}, b)$ are easily over millions in number for a CNN/DL architecture; it is nearly impossible to accommodate an optimization for such an architecture through an analytical gradient method. Gradient descent performs iterative and numerical optimization by using

- The chain rule of derivatives

[2]This is the essential difference between parameter fitting techniques and ML techniques. Parameter fitting aims to optimize parameters for given models, while ML aims to learn and form the model rather entirely.

- Local gradients
- Back-propagation.

At every iteration (t), a local Jacobian ($\nabla_\theta J(\theta_{t,i})$) of a loss function ($J$) with respect to learnable variables ($\theta_{t+1,i}$) is calculated. The update rule for learnable parameters can be written as below,

$$\theta_{t+1,i} = \theta_{t,i} - \eta \cdot \nabla_\theta J(\theta_{t,i}), \tag{4.5}$$

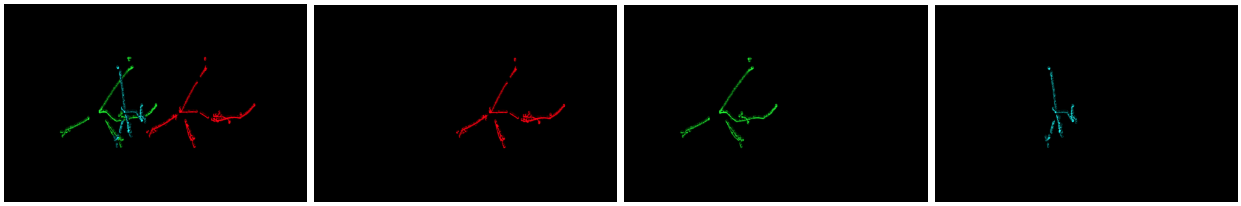
where η is a hyperparameter called the learning rate, which determines the size of the step in the minimization iterations.

This local Jacobian is calculated by a series of multiplications of local derivatives (using the chain rule and partial derivatives at the local point) going backward (i.e., back-propagation) from the final loss function to the learnable parameters. The parameter in the next iteration ($\theta_{t+1,i}$) is thus shifted by the amount of multiplication of the step size (η) and the local Jacobian. This way of updating learnable parameters has a strong computational benefit. By calculating local gradients numerically as the ingredients to form the local Jacobian at each iteration, the complex derivatives through multiple layers are broken down to mere multiplications using the chain rule. This also enables ML techniques to be accommodated for parallelization-computing on processors like GPUs, CPUs, and FPGAs. Various gradient descent techniques have been developed, such as SGD [92] and Adam [93].

4.3 Tools

This section discusses the specific recent developments used in the studies described in this thesis: generalized sparse convolution using *MinKowski Engine* [89] and CNN implementation on FPGAs using *hls4ml* [90].

(a) dense RGB image



(b) sparse image from LArTPC simulated data

Figure 4.3: (top) RGB color image with pixel resolution. Every pixel carries information. (bottom) LArTPC simulated data image on U, V, Y planes. Only pixels with physics interactions carry meaningful information.

4.3.1 Generalized convolution for sparse CNN using *MinKowski Engine*

LArTPC data from three wire planes has many similarities with pixelated images with RBG color channels. There are three 2D planes (channels) from one input (3D) image with pixel resolutions. While these similarities are strong motivation to apply DL-techniques on LArTPC data analyses, the differences between LArTPC data and pixelated color images give more insight into the practical application of DL-techniques on LArTPC data. Compared to regular RGB color-formatted images, the percentage of pixels in LArTPC data that carry meaningful information is very small; in other words, regular images are dense, and LArTPC data images are sparse. Disadvantages, such as inefficient memory use, arise when sparse data is treated in the same way as dense data. Figure 4.3a shows an example of a dense image, and Fig. 4.3b shows an example of a sparse image from LArTPC data. Comparing these example images, it is clear that the pixels in the former image are mostly equally important to form the information of the image, but in the latter,

we are usually interested in the information in the pixels containing possible physics interaction signatures.

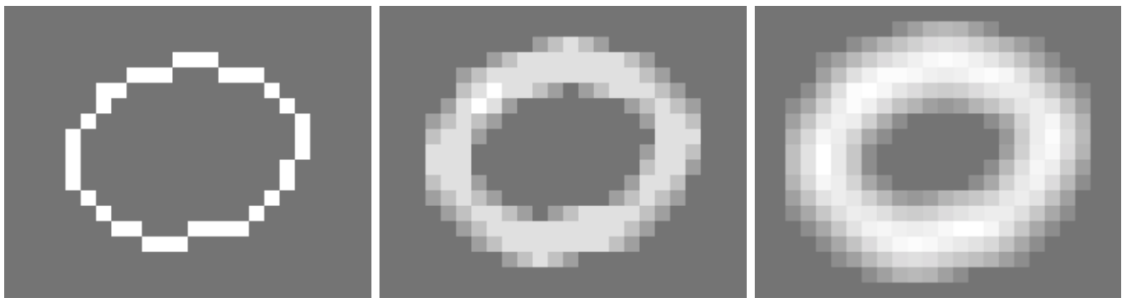


Figure 4.4: Example of submanifold dilation. Left: Original 2D input. Middle: Result of applying a regular 3×3 convolution. Right: Result of applying the same convolution again. Image Credit: [94].

The conventional method of convolution can now be called dense convolution, compared to a "sparse convolution". Another advantage of dense convolution is "submanifold dilation" [94]. This effect can be illustrated by Fig. 4.4. Over dense convolutions, the features lose their sparsity very quickly. To avoid this, using localized activations of convolution is proposed in SSCN [94]. This method only allows convolutions near the feature while masking other areas in the feature map. Through this procedure, the sparsity of the model is achieved. This procedure still operates on dense tensors while the network is sparse.

A more recent development in such an approach is "generalized sparse convolution" using sparse tensor input. This method allows for sparse tensor input using COOrdinate list (COO) format [96] for input sparsification, effectively representing only informative pixels. A generalized convolution operates sparse convolutions on a sparse tensor input, as illustrated in Fig. 4.5. This method is used in signal selection described in Sec. 7.4. The input sparsification on MicroBooNE LArTPC data effectively reduces the memory for storing inputs by a factor of 1,000 without any loss in input resolution, thus enabling quality CNN training with over 2M inputs.

4.3.2 CNN implementation on FPGAs using *hls4ml*

"High Level Synthesis for Machine Learning" (*hls4ml*) [90] is a package that translates ML models supported by Python API-based frameworks (like Keras [97] and Pytorch [91]) to high-level synthesis (HLS) to run on Field Programmable Gate Arrays (FPGAs). The HLS framework (such as *vivado* [98]) transcribes high-level C-synthesis to logic synthesis represented in the Register-Transfer Level (RTL) to be implemented on FPGAs. ML/DL algorithms have been shown to be powerful tools for many "offline" data processing applications after data collection. On the other hand, for large detector systems such as in LHC experiments and the future LArTPC DUNE, "real-time" online data processing is a major challenge. Restrictions come from both ends of the data storage limit and the data processing time (latency); i.e., the implementation of ML/DL algorithms for real-time or online execution is subject to resource constraints available within given hardware for executions that should be completed within a fixed time (depending also on data

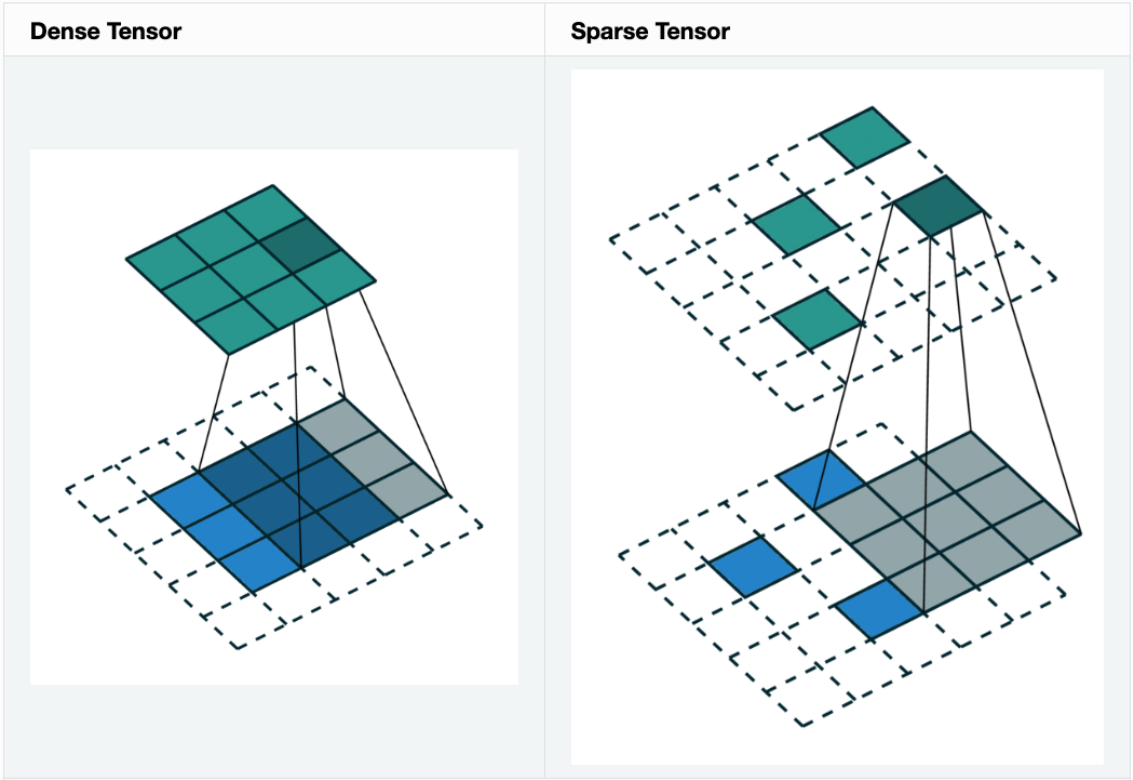


Figure 4.5: Generalized convolution using sparse tensor input. Image credit: [95].

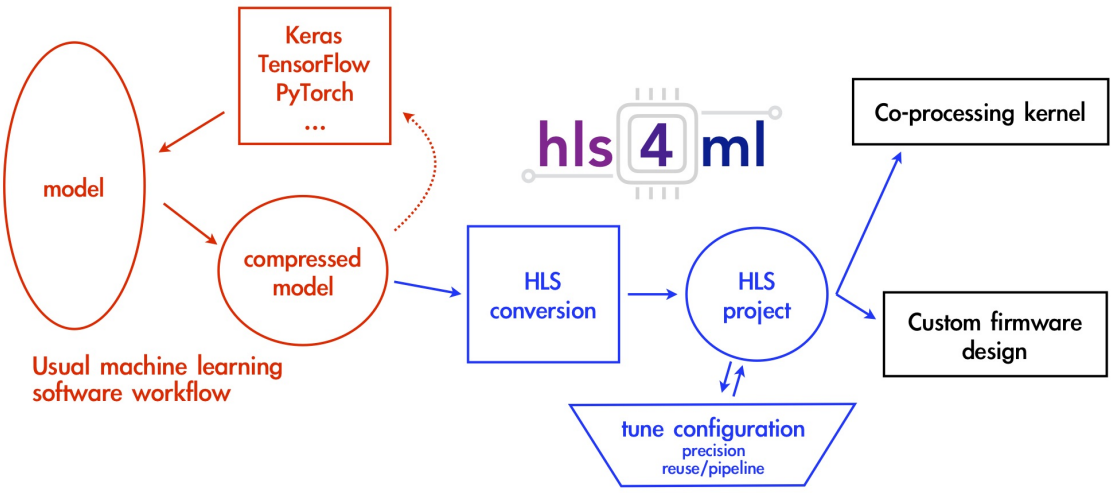


Figure 4.6: A schematic diagram showing the interfaces of *hls4ml*.

throughput). As discussed in Sec. 3.3.2, the readout unit of the DUNE FD includes an FPGA and a CPU, due to the lower power consumption of the FPGAs. CNN implementation on the target FPGA for the DUNE DAQ system, including a resource usage analysis, is explored as part of this thesis, and the feasibility of a CNN-based real-time data selection in DUNE is discussed in Ch. 6. *hls4ml* interfaces with ML frameworks and high-level synthesis (HLS) frameworks as shown in Fig. 4.6. The effort to implement the ML / DL algorithms in FPGAs used *hls4ml* tools.

Chapter 5: Neutron-antineutron oscillation search in DUNE

In this chapter, the method followed for a search for argon-bound $n-\bar{n}$ oscillation in the DUNE Far Detector (FD) is presented along with a sensitivity evaluation for a 400 kton·yr exposure. The Monte Carlo (MC) simulations used to evaluate DUNE's sensitivity to this process are described in Sec. 5.1. The search combines traditional hit- and cluster-based particle reconstruction methods, described briefly in Sec. 5.2 together with image classification techniques employing convolutional neural networks (CNNs), described in more detail in Sec. 5.3. A method combining CNN classification and Boosted Decision Tree (BDT) techniques is presented in Sec. 5.4. The projected sensitivity of DUNE to the $n-\bar{n}$ oscillation for a total exposure of ten years with the full 40 kton FD volume is evaluated in Sec. 5.5, followed by discussions for future improvements.

5.1 Signal and background simulations

5.1.1 $n-\bar{n}$ oscillation signal generation in the DUNE FD

The GENIE Neutrino Monte Carlo Generator (GENIE) v.2.12. via dunetpc[1] release v.06.82.00 is used for the generation of $n-\bar{n}$ oscillation events. The description in this section is a summary of the more detailed description provided by the original simulation developer [100]. The implementation of this process in GENIE makes use of GENIE's existing modeling of Fermi momentum and binding energy for both the oscillating neutron and the nucleon with which the resulting antineutron annihilates. The ^{40}Ar nucleus is assumed to be at rest upon $n-\bar{n}$ oscillation. The position of a neutron (to be transitioning to antineutron) within the nucleus is simulated using GENIE's density

[1] dunetpc is a package in the C++ based framework LArSoft [99]. dunetpc is used for DUNE simulation, reconstruction, and analysis.

profile of nucleons (Woods-Saxon distribution [101]),

$$\rho(r) = \frac{\rho_0}{1 + e^{\frac{r-R_0}{a}}}\ , \tag{5.1}$$

where r is the radial position inside the nucleus, $R_0 = r_0 A^{\frac{1}{3}}$ is the nuclear radius, with r_0 defined as 1.4 fm in GENIE. ρ_0 is normalized to express the nuclear density as a probability distribution, and a is a parameter that describes the surface thickness of the nucleus, set to $a = 0.54$ fm. For the nucleon Fermi momentum and binding energy, the Bodek-Ritchie Fermi gas model [102] is used. As one of the 22 neutrons inside the ^{40}Ar nucleus transitions into an antineutron ($\bar{n}$), it pairs with a nearby nucleon: either a neutron (n) or proton (p). Thus, the chances of forming $(\bar{n}, n)$ and $(\bar{n}, p)$ pairs become 21/39 and 18/39 each. The products of the annihilation process follow the branching fractions measured in low-energy antiproton annihilation on hydrogen [50], biased as necessary by conservation of total available energy in the annihilation. The pair annihilation of these annihilation pairs is simulated using the branching ratios shown in Tab.5.1. Since the annihilation products are produced inside the nucleus, GENIE further models the re-interactions of those products as they propagate in the nucleus (until they escape the nucleus).

$\bar{n} + p$		$\bar{n} + n$	
Channel	Branching ratio	Channel	Branching ratio
$\pi^+\pi^0$	1.2%	$\pi^+\pi^-$	2.0%
$\pi^+2\pi^0$	9.5%	$2\pi^0$	1.5%
$\pi^+3\pi^0$	11.9%	$\pi^+\pi^-\pi^0$	6.5%
$2\pi^+\pi^-\pi^0$	26.2%	$\pi^+\pi^-2\pi^0$	11.0%
$2\pi^+\pi^-2\pi^0$	42.8%	$\pi^+\pi^-3\pi^0$	28.0%
$2\pi^+\pi^-2\omega$	0.003%	$2\pi^+2\pi^-$	7.1%
$3\pi^+2\pi^-\pi^0$	8.4%	$2\pi^+2\pi^-\pi^0$	24.0%
		$\pi^+\pi^-\omega$	10.0%
		$2\pi^+2\pi^-2\pi^0$	10.0%

Table 5.1: Effective branching ratios for antineutron annihilation in ^{40}Ar, implemented in GENIE. The branching ratios are adapted from Super-Kamiokande's analysis and are derived from antiproton annihilation measurements, with a phase-space approximation [100].

The products of the two-nucleon annihilation propagate inside the nucleus in which the $n - \bar{n}$

oscillation occurs until they escape the nucleus. As they can travel as far as 8 fm in the argon nucleus, the chance of reinteracting is high. The final state interactions (FSI) are simulated using INTRANUKE, which is a GENIE subpackage. The branching fractions and reinteraction probabilities in INTRANUKE are tuned to bubble-chamber data on hydrogen and deuterium targets. INTRANUKE uses the free cross section to estimate the likelihood of reinteraction, by defining the mean free path as

$$\lambda(E, r) = \frac{1}{\sigma_{hN,tot} \times \rho(r)} , \tag{5.2}$$

where $\sigma_{hN,tot}$ is the cross-section and $\rho(r)$ is the nuclear density.

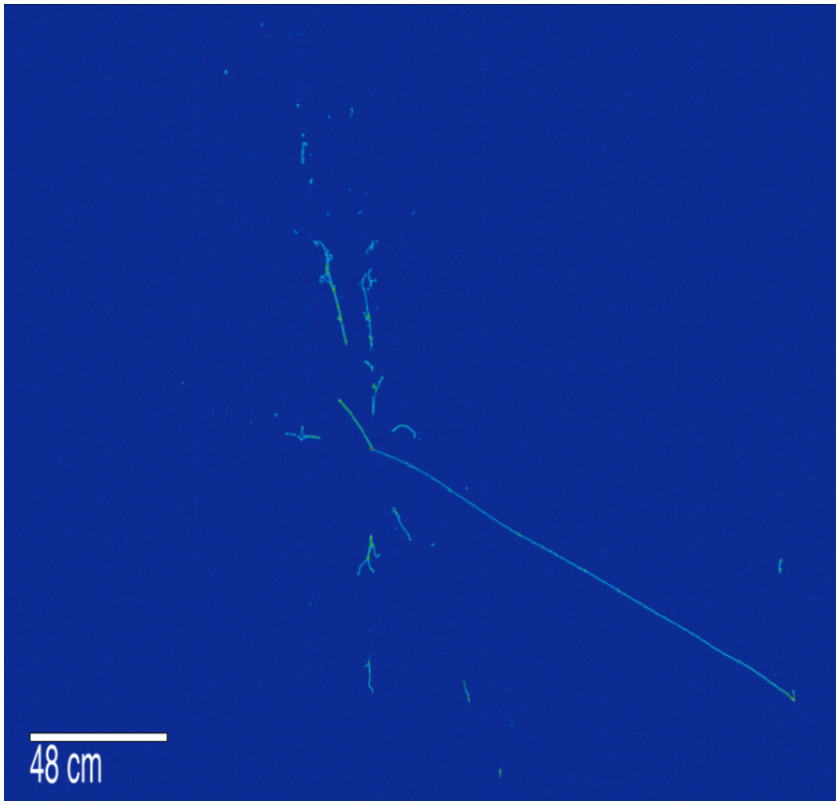


Figure 5.1: An example event display of the $n - \bar{n}$ interaction simulated through GENIE ($n\bar{n} \rightarrow \pi^+\pi^-3\pi^0$.) During the final state interaction, the π^+ is absorbed and a proton is ejected. Multiple electromagnetic showers from $\pi^0 \rightarrow \gamma\gamma$ are shown. The long track is produced from the π^-, while the short track is produced from the proton. Image Credit: [100]

For this work, an empirical data-driven method hA with version $hA2015$ in GENIE v2 is used. hA is modeled as one effective interaction, as opposed to a full cascade. This approach uses the total cross-section data available for each possible scattering process for pions and nucleons as a function of energy [103]. Thus, it is called hA. The available scattering processes include pion elastic and inelastic scattering, charge exchange, and absorption. Alternatively, one can choose the hN method, which uses a full cascade model [103].

An example signal $n - \bar{n}$ event produced by GENIE is shown in Fig. 5.1. The signature topology of the interaction is often referred to as "star-like" or "pion-star," with the multiple tracks and showers stemming from the center vertex. Kinematically, the final state particles have net momentum close to zero without a strong directionality, and the released energy peaks around two-nucleon mass.

5.1.2 Atmospheric neutrino background generation in DUNE FD

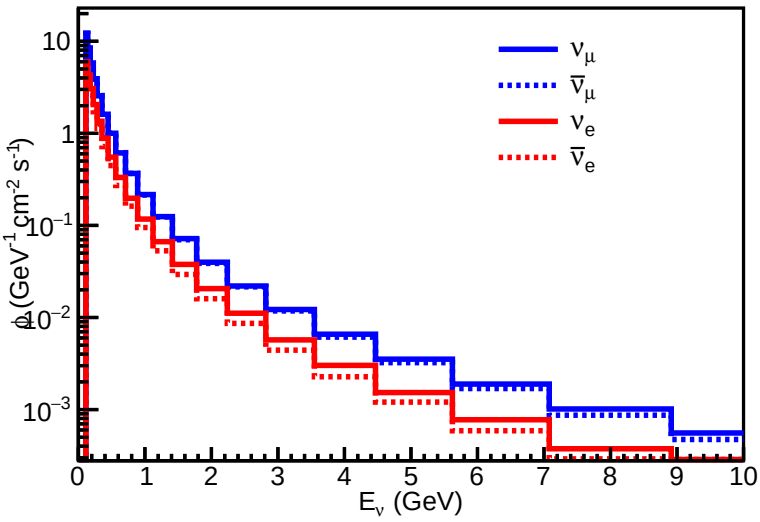


Figure 5.2: Atmospheric flux (Bartol at Soudan). Courtesy of A. Higuera.

The primary background process in the search for bound $n - \bar{n}$ oscillation in DUNE is assumed to be atmospheric neutrino interactions in the detector. Calculation of the background from atmospherics starts with simulations of the interactions of the primary cosmic rays with the atmosphere.

Figure 5.2 shows the atmospheric flux prediction for different neutrino species. The GENIE generator v.2.12.10 is also used for the generation of atmospheric neutrino background. This analysis uses the Bartol flux [104] at the Soudan Mine location, which is a calculation of atmospheric neutrino fluxes using a Monte Carlo (MC) simulated sample.[2] To estimate the atmospheric neutrino event rate for a given exposure, we integrate the flux with the cross-section. Table 5.2 shows the event rate for different neutrino species for exposure of 10kton-years, without neutrino oscillation effects included.

10kton-year	CC	NC	Total
ν_μ	1038	398	1436
$\bar{\nu}_\mu$	280	169	449
ν_e	597	206	83
$\bar{\nu}_e$	126	72	198
Total	2014	845	2886

Table 5.2: Expected rate of neutrino interactions in ^{40}Ar for a 10 kton-year exposure, before oscillation. [106, 107]

The Prob3 package [108] was used to include the effect of neutrino oscillation for neutrinos traveling between the atmosphere and the DUNE FD. Prob3 is a software package for computing three-flavor neutrino oscillation probabilities based on [109]. Prob3 accounts for matter effects in the Earth treated as several layers of uniform matter density; however, the version used here assumes an on-surface detector. To calculate the three flavor neutrino oscillation probabilities, the normal neutrino mass hierarchy with $\delta_{cp} = 0$ is assumed. Figure 5.3 shows the ratio of oscillated/unoscillated ν_μ and $\bar{\nu}_\mu$ fluxes.

5.1.3 Atmospheric neutrino interactions

The flux of atmospheric neutrinos has a wide energy range, as shown in Fig. 5.2. Different neutrino interactions are possible in this energy range. Figure 5.4 shows the cross sections of the

[2] A more recent simulation of atmospheric neutirnos in DUNE more appropriately makes use of the Honda flux prediction exactly at the DUNE FD site at SURF [105]. This simulation was not available at the time of this study, and the use of the Bartol flux at the Soudan Mine was considered a reasonable approximation.

charged current (CC) interaction for quasi-elastic scattering (QE), resonant production (RES) and deep inelastic scattering (DIS).

Where neutrino energy is above a few GeV, DIS occurs and becomes more dominant in the higher energy region. In this region, a neutrino is energetic enough to break the nucleus to generate multiple hadrons in the final state (hadronic cascade). Figure 5.5 shows the charged current (CC) and the neutral current (NC) interactions of neutrinos. The CC interaction results in a charged lepton in the final state, whereas the NC interaction does not. NC DIS interactions can leave no charged lepton in sight and multiple hadrons in the final state. Thus, NC DIS can be a critical background to $n-\bar{n}$ oscillation mimicking the signal final state including multiple hadrons without a charged lepton.

An example of such an energetic neutrino NC interaction in argon TPC can be found in Argoneut neutral current π^0 production measurement [111]. Figure 5.6 shows MC simulated event display of NC interaction with 3.6 GeV antineutrino. This event generates multiple protons and pions in the final state. However, compared to the $n-\bar{n}$ event display shown in Fig. 5.1, the signature shows a strong directionality from left to right in the event display, as the tracks from the

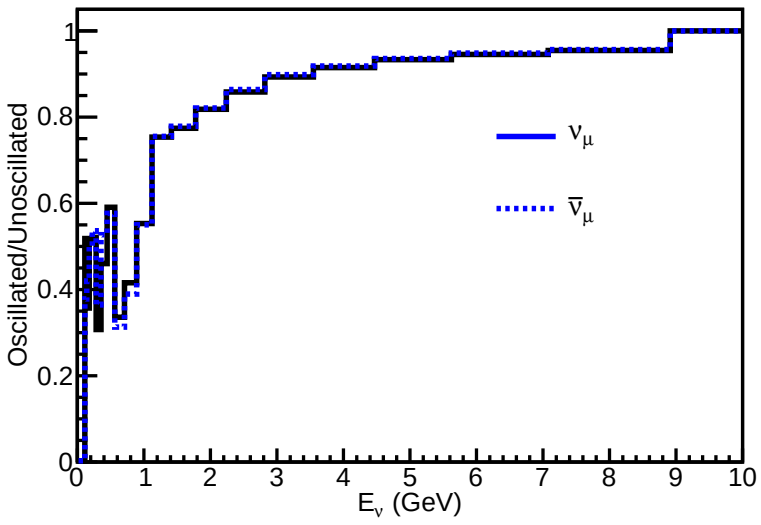


Figure 5.3: Atmospheric ν_μ and $\bar{\nu}_\mu$ ratio of oscillated/unoscillated fluxes in the DUNE FD (assumed to be on-surface). The normal hierarchy and $\delta_{cp} = 0$ are assumed. Courtesy of A. Higuera.

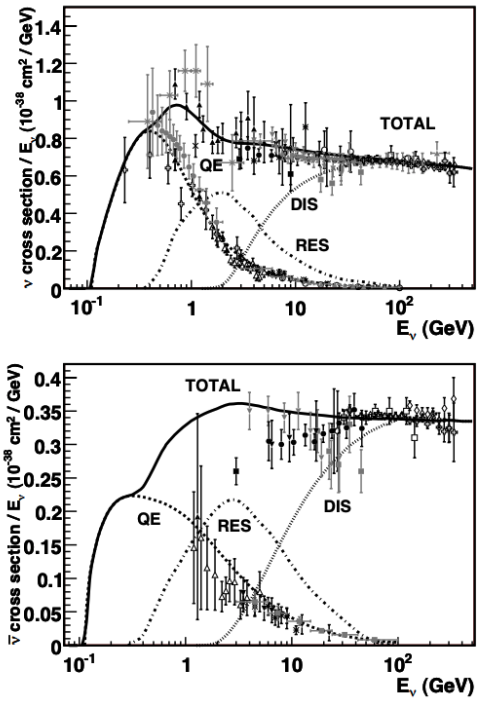


Figure 5.4: Total charged current (CC) cross sections for neutrino (top) and antineutrino (bottom) per nucleon divided by neutrino energy. This plot summarizes the current simulations and measurements. The references to each data can be found in [110]. Image Credit: [110].

vertex are mostly outgoing toward the right.

5.1.4 Detector simulation

Once $n - \bar{n}$ oscillation or atmospheric neutrino interaction is simulated to the final state, the final state particles are tracked through a full detector simulation of a $1 \times 2 \times 6$ APA geometry in the DUNE FD using GEANT [112]. Detector response simulation through dunetpc includes detector effects such as recombination, electron attenuation, and diffusion. Particle energy losses through ionization and scintillation processes are simulated, as well as the detector response to those losses. Once the TPC waveforms are obtained after detector simulation, they are run through reconstruction.

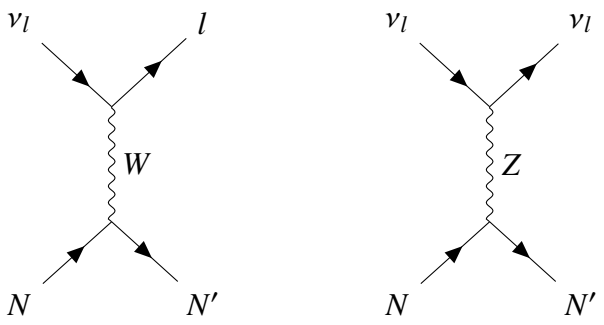


Figure 5.5: Feynman diagrams of (left) charged current (CC) neutrino interaction, (right) neutral current (NC) neutrino interaction. For the case of deep inelastic scattering, N' may include hadronic cascade.

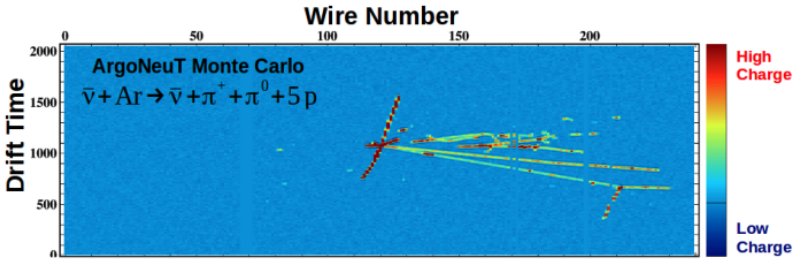


Figure 5.6: Simulated NC π^0 production event in the ArgoNeuT detector. Image credit: [111]

5.2 Reconstruction and event selection

For this analysis, both traditional reconstruction and CNN-based image classification are used to construct a combined BDT classifier, where the input variables of the BDT are kinematic variables from the traditional reconstruction and the CNN score. dunetpc v06.82.00 is used for the DUNE simulation and the reconstruction. For the CNN-based image classification, reconstructed wire signals, as a part of standard reconstruction, on the collection plane wires only are used; collection plane information is often more reliable than induction plane's due to the unipolar nature of the signals on the collection plane. The 2D image is formed as the collection plane wire index

vs. the time tick of the collection plane.

The sequence of the reconstruction algorithms starts with digitized TPC waveforms, signal processing (deconvolution), Gaussian hit finding, disambiguation, and 2D cluster formation. In cluster formation, the *linecluster* algorithm is used. After forming 2D clusters, 3D track and vertex reconstruction is done by Projection Matching algorithms (PMA) [113]. 3D tracks and eletromagnetic (EM) showers are reconstructed by PMA with their calorimetric information. These reconstructed variables are fed into a BDT, together with CNN inference from each event. Alternatively, an analysis only involving image classification for event selection can be performed. The CNN image classification is discussed in Sec. 5.3. The BDT is described in Sec. 5.4.

5.3 CNN image classification

Image classification by using convolutional neural networks (CNNs) can be carried out on 2D images formatted from DUNE's recob::wire simulation objects after the reconstruction process. The classification score output is utilized as a metric to discriminate $n - \bar{n}$ oscillation signal from the background. The next few sections describe the CNN image classification methodology.

5.3.1 Data processing

Input 2D images can be formed from TPC waveforms after hit reconstruction at DUNE. A single drift region corresponds to a one-directional full-drift on a single APA, as discussed in Sec. 3.3.2. For a single drift region, digitized and reconstructed wire signals on collection plane recob::wire objects are mapped to one 2D image. With the 2D image from 960 recob::wire objects with 4488 time-ticks, a 2D region of interest (ROI) finding is applied on this 2D image by finding the smallest rectangular area containing all pixels with baseline-subtracted ADC values over 20 (RMS noise on LArTPC collection plane is about 3.4 ADC [114]; one minimum ionizing particle (MIP) leaves around 60 ADC on a pixel according to the energy deposit from dE/dx). Once the ROI is defined, initial down-sizing by a factor of 4 on time-tick axis is performed on the 2D image.

If the image is still larger than 600× 600, a down-sizing is performed by the factor of two on both axes, and then the image is placed in a 600× 600 pixel frame. Example images with 600×600-pixel dimension are shown in Fig. 5.7. For the signal $n - \bar{n}$ oscillation and the atmospheric backgrounds, the input images are required to contain the true interaction vertices, regardless of whether the final products are fully contained in the frame or not.

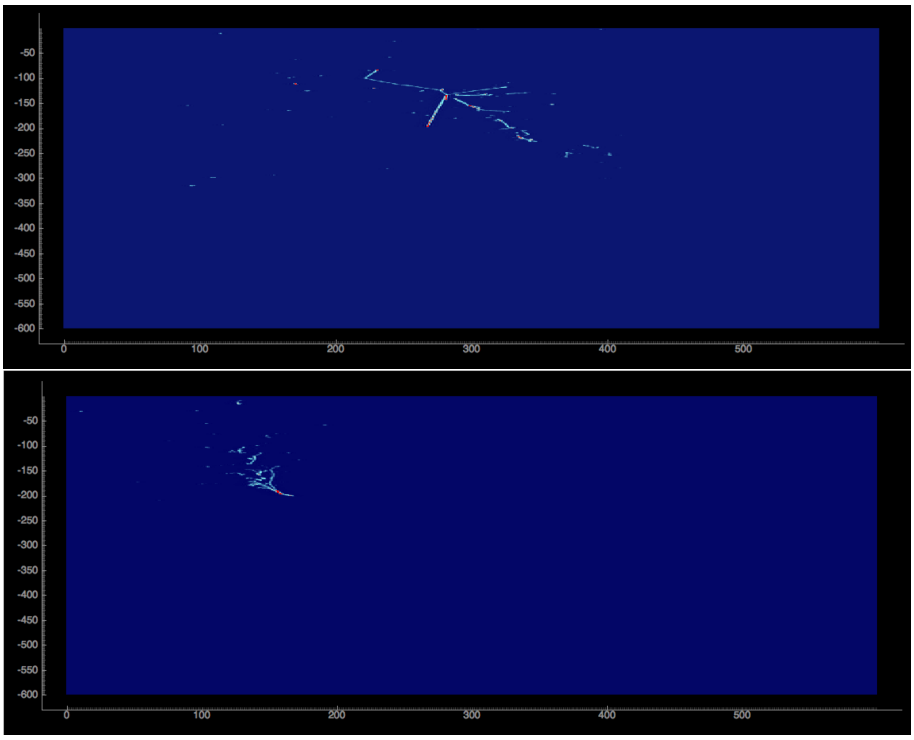


Figure 5.7: (top) Input event display for the $n - \bar{n}$ oscillation event in an APA. (bottom) Input event display for atmospheric neutrino event on the collection plane for a single APA. The x-axis is along the wire channel direction and the y-axis is along the time-tick direction. The input has been reshaped to a 600×600-pixel dimension.

5.3.2 Training, validation, inference

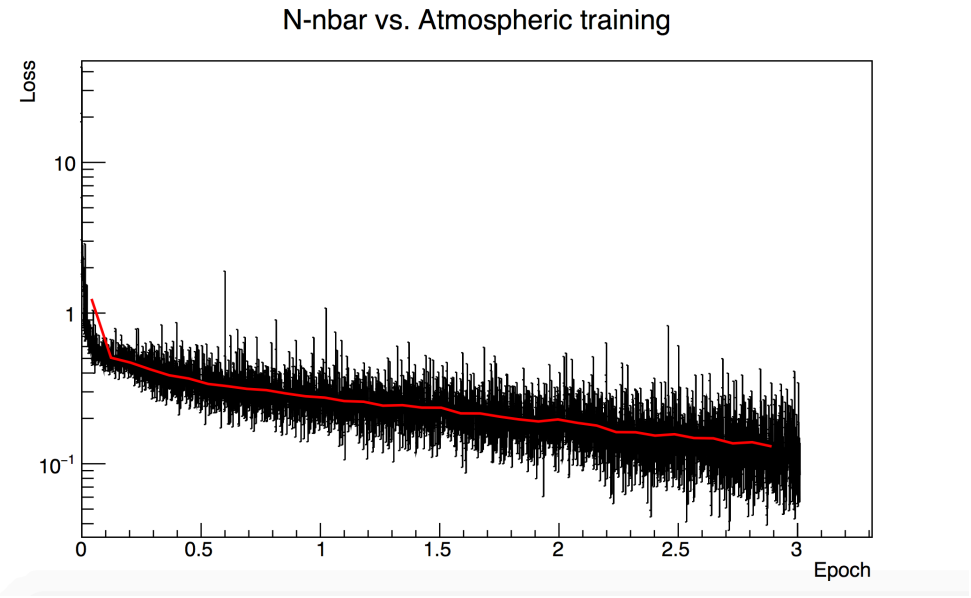
The network architecture is constructed with the reference baseline VGG16 [84], with fully connected layers removed from the original architecture. The weights are not initialized to a specific model. The network architecture contains input layers, convolution layers, Rectified Linear

Network layer	Kernel size	Filter depth
Input (Collection plane image)		
Convolution	3×3	64
Convolution	3×3	64
Max pool	2×2	
Convolution	3×3	128
Convolution	3×3	128
Max pool	2×2	
Convolution	3×3	256
Convolution	3×3	256
Max pool	2×2	
Convolution	3×3	512
Convolution	3×3	512
Convolution	3×3	512
Max pool	2×2	
Convolution	3×3	512
Convolution	3×3	512
Convolution	3×3	512
Max pool	2×2	
Fully connected		

Table 5.3: VGG16 layer structure

Unit (ReLU) layers, pooling layers, fully connected layers, and an output layer. The network architecture is shown in Tab. 5.3. The training loss and validation accuracy are monitored during the training iterations, as shown in Fig. 5.8. The training of the network is done over iterations, minimizing the loss function by updating the weight during iterations. The Caffe framework [115] is used for training. The training hyperparameters are configured through the 'solver' in the Caffe framework. The solver configuration used is described in Sec. 5.3.2.1. Approximately 38,000 images are used for training, each for the signal and the background.

Validation of the trained network is done by calculating the prediction accuracies from a separate set of inputs. Once the model with the highest accuracy is chosen from validation, the inference is performed on a statistically independent set of $n - \bar{n}$ oscillation events and atmospheric background events. The CNN's output softmax (as discussed in Sec. 4.1.3) indicates the categorical probabilities of the input to be in the classification category (class). For one input event, it obtains



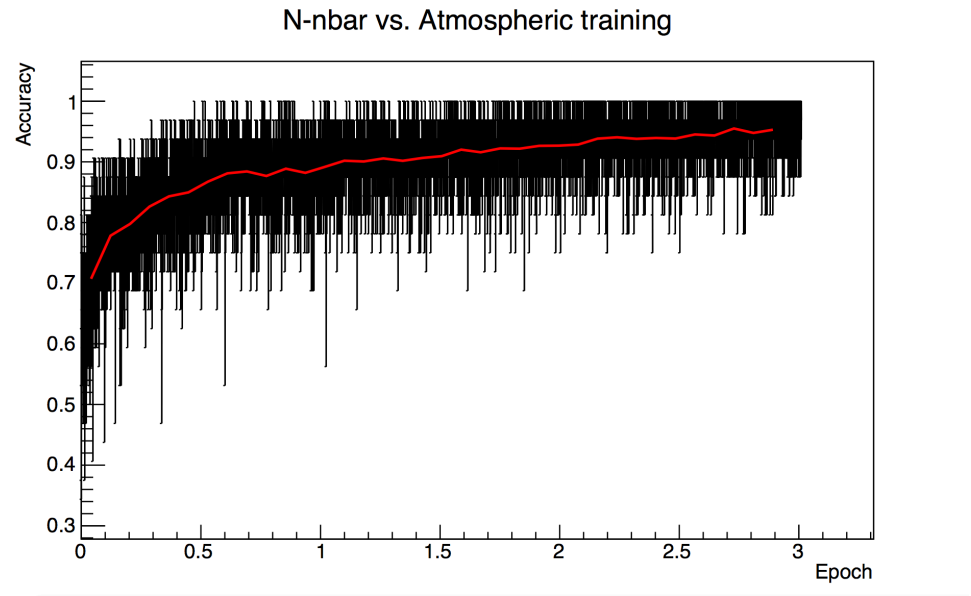


Figure 5.8: (left) Training loss and (right) training accuracy for $n - \bar{n}$ oscillation and atmospheric event classification. The averaging of 200 iterations is overlaid with the solid red line.

softmax values for the signal and the background.

5.3.2.1 Solver

As the network training goes on over iterations, the network extracts the features from the input, and the loss function is minimized, as discussed in Sec. 4.2.2. The hyperparameters for training and the optimizer settings are configured with the solver in the Caffe [115] framework. The solver configuration used for the analysis is shown below, where the "train_net" defines the network corresponding to Tab. 5.3.

```
train_net: "jhewes_vgg16b.prototxt"
iter_size: 4
base_lr: 0.00002
momentum: 0.0
weight_decay: 0.001
lr_policy: "inv"
gamma: 0.00001
power: 0.75
display: 1
max_iter: 100000
snapshot: 250
snapshot_prefix: "vgg16b"
snapshot_format: HDF5
solver_mode: GPU
type: "RMSProp"
rms_decay: 0.90
```

5.3.2.2 CNN performance

Applying a cut on the signal softmax is used to evaluate the performance of event selection with CNN only. The signal selection efficiencies and background rejection rates for different cut values are shown in Tab. 5.4. By continuously varying this cut, the background rejection rate at a given signal efficiency is obtained and shown in Fig. 5.10. For the inference, ~100k events for $n - \bar{n}$ signal images and ~380k events for background images are used.

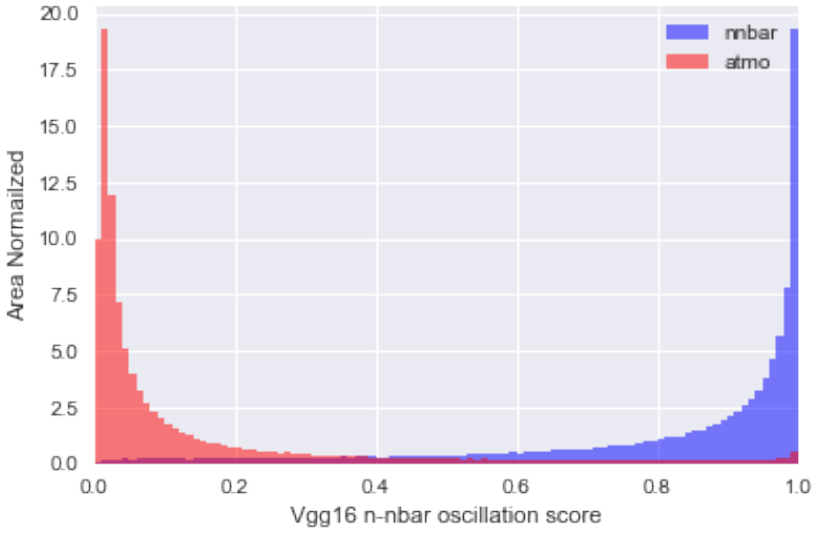


Figure 5.9: Normalized signal score distribution for the signal ($n - \bar{n}$ oscillation) and atmospheric background.

Score cut	Signal efficiency	Background rejection
0.99986	2.22 %	99.931 %
0.9999	1.86 %	99.943 %
0.99995	1.316 %	99.959 %
0.9999	0.614 %	99.980 %
0.99995	0.442 %	99.984 %
0.99999	0.2085 %	99.991 %

Table 5.4: CNN performance on $n - \bar{n}$ oscillations and atmospheric backgrounds. These are example cuts which can sufficiently remove background events, but in the analysis the CNN score is actually fed into a BDT (see Sec. 5.4).

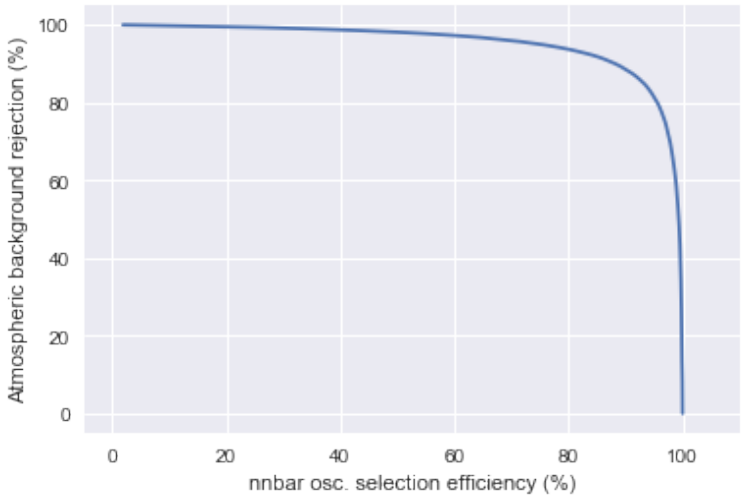


Figure 5.10: Background rejection rate vs. $n - \bar{n}$ oscillation signal efficiency curve.

This CNN training has plenty of room for improvement. First of all, the CNN training is performed on rather low statistics samples (38k). Also, the training is only based on the collection plane information, not the entire three wire planes. Both of those factors are known to have significant impacts on the quality of network training. Possible improvements of this analysis are further discussed in Sec. 5.7.

5.4 Multivariable BDT analysis

5.4.1 BDT input variables

A Boosted Decision Tree (BDT) classifier is used to combine the result from the traditional reconstruction with the PMA reconstruction and CNN-based classification. The software package Toolkit for Multivariate Data Analysis with ROOT (TMVA4) [116] was used with AdaBoost [117] as the boosting algorithm.

The kinematic variables through PMA are obtained in the form of reconstructed tracks and electromagnetic (EM) showers. The visible energy for tracks and EM showers are calculated by adding the reconstructed calorimetric energy of associated hits to the tracks or showers. As muons,

charged pions, and protons are likely to be reconstructed as tracks, a particle identification (PID) algorithm for tracks that distinguishes muons, charged pions, and protons can be useful. PIDA is a parameter useful for PID for reconstructed tracks from PMA. For a given point on a reconstructed track, dE/dx can be found at the point, as well as the residual length R from the point to the end of the track. Combining dE/dx and R for a i-th point on a track, $(\frac{dE}{dx})_i R_i^{0.42}$, PIDA is defined as the median of all track points i for which the residual range R_i is less than 30 cm

$$PIDA = \left\langle (\frac{dE}{dx})_i R_i^{0.42} \right\rangle. \tag{5.3}$$

PIDA of the longest track is used as the BDT input to reject muons in background events. The average dE/dx for the most energetic shower is also used as the BDT input, as well as the CNN score obtained through the method is Sec. 5.3. The entire BDT input variables are listed below. Figure 5.11 shows the distributions of the variables for signal ($n - \bar{n}$) and background (atmospheric neutrino) events.

- Number of tracks: total number of reconstructed tracks.
- Number of showers: total number of reconstructed showers.
- Track-like visible energy: sum of track-like hits energy.
- EM-like visible energy: sum of electromagnetic shower-like hits energy.
- Visible energy: total visible energy.
- PIDA of the longest track: PIDA value of the longest track in the event.
- Momentum of the longest track: Reconstructed momentum of the longest track.
- Electromagnetic-like visible energy fraction: EM shower-like visible energy/visible energy.
- dE/dx of the most energetic shower.
- CNN score: CNN image classification score.

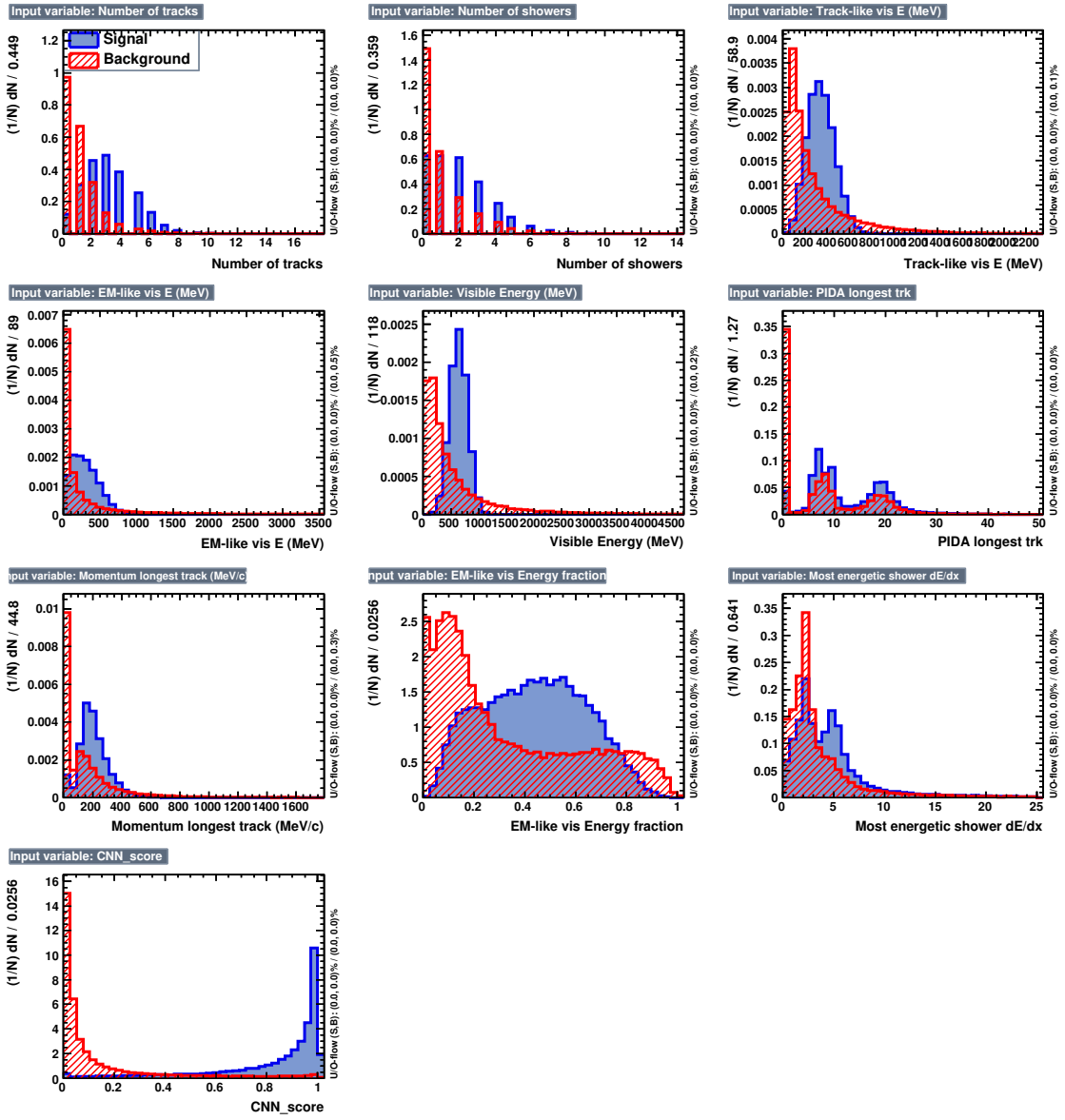


Figure 5.11: BDT input variables for signal (blue) and background (red). Reconstructed visible energy is correlated with the energy released from the signal's two-nucleon annihilation (around 1.8-1.9 GeV). Here, the reconstructed visible energy for signal peaks lower than that due to the inefficiency in reconstruction.

5.4.2 BDT performance

The BDT consists of 1,000 trees with a maximum depth of eight nodes. The BDT response is shown in Fig. 5.12. For the BDT training, 50k signal and background events were used. For inference results, an equivalent of ~200 kton-year exposure was used for the background, and approximately 100,000 events were used for the signal. Figure 5.13 shows the background and signal efficiency as a function of BDT response.

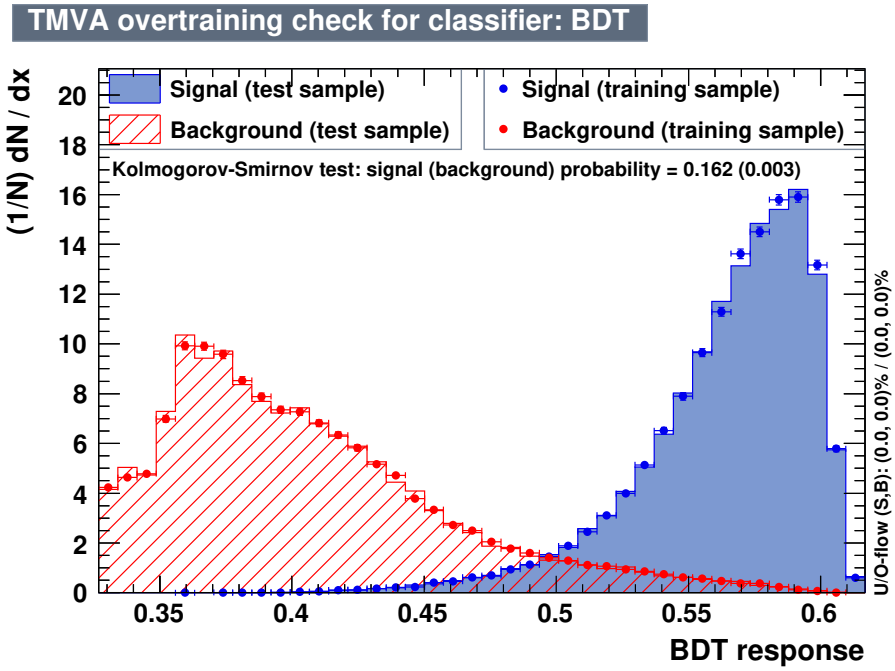


Figure 5.12: BDT response for training and testing samples, shown for the signal (blue) and background (red).

Fig. 5.14 shows a $n-\bar{n}$ event with high BDT response value (0.592). Some features are visible in the event display. EM showers from π^0 are shown in red, blue, yellow, and light green. The reconstructed $\pi^{+/-}$ tracks are shown as green and maroon lines. The topology of this event is consistent with multiple $\pi^{+/-}$ production and π^0 production.

The plot on the left side of Fig. 5.15 shows a neutral current background event $\nu_e + n \rightarrow \nu_e + p + p$

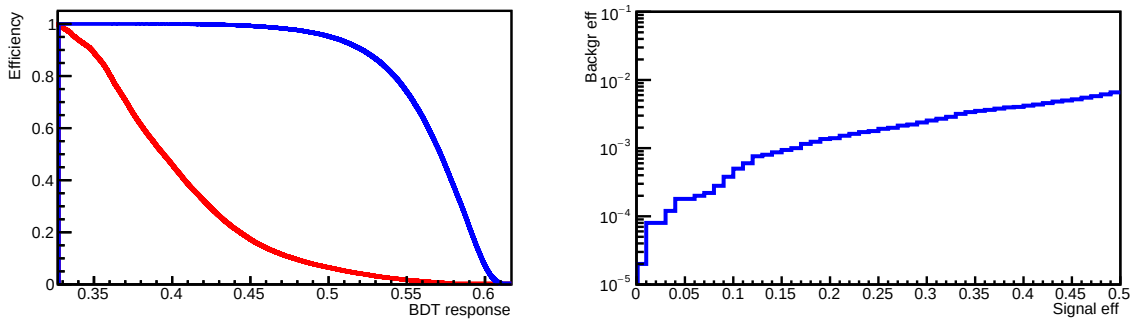


Figure 5.13: (left) Signal (represented by the blue line) and background (represented by the red line) efficiency as a function of BDT response. (right) Signal vs. background efficiency.

with low BDT response value (0.388). The two protons from the NC interaction are reconstructed as tracks and no shower activity is present. The plot on the right side of Fig. 5.15 displays a charged current background event $\nu_e + n \rightarrow e^- + p + \pi + p$ with a high BDT response value (0.598). This background event mimics the signal topology by having multi-particle production and an EM shower. Further improvements in shower reconstruction, especially dE/dx, should help further reject this type of background event in the future, since the electron shower dE/dx is different from a gamma dE/dx (pi-zero). As shown in Fig. 5.11, the dE/dx of the most energetic shower in the event is consistent with an electron shower for background events. On the other hand, there is a long tail that overlaps with dE/dx of a photon. For the case of signal events, dE/dx should have high selection efficiency; however, there is a bi-peak distribution, which is due to the misreconstruction of the gamma shower.

To summarize the performance of the BDT, the top four ranked variables are the CNN score, the total visible energy (which is highly correlated with an invariant mass measurement), the number of tracks (particle multiplicity), and the shower dE/dx (gamma vs. electron shower). The final efficiency using combined BDT is evaluated at 8% for the signal, and 0.02% for the background sample at the chosen operating point. This efficiency is propagated to calculate the projected sensitivity in DUNE.

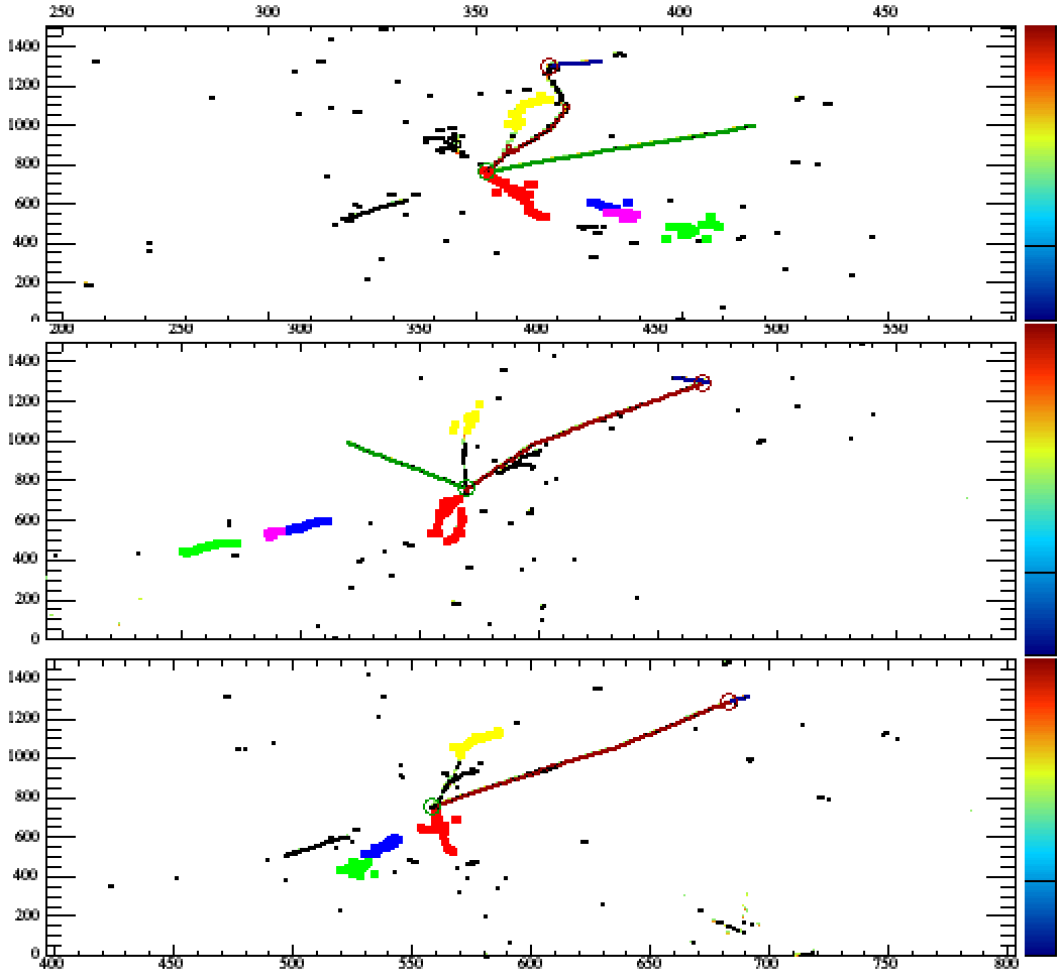


Figure 5.14: Signal event with high BDT response (0.592) value. The x-axis represents wires on a single APA where each subplot represents the first induction wire plane (top), the second induction wire plane (middle), and the collection wire plane (bottom); the y-axis represents the time-tick. The reconstructed electromagnetic (EM) showers from π^0 are shown in red, blue, yellow, and light green; the reconstructed $\pi^{+/-}$ tracks are shown as green and maroon lines.

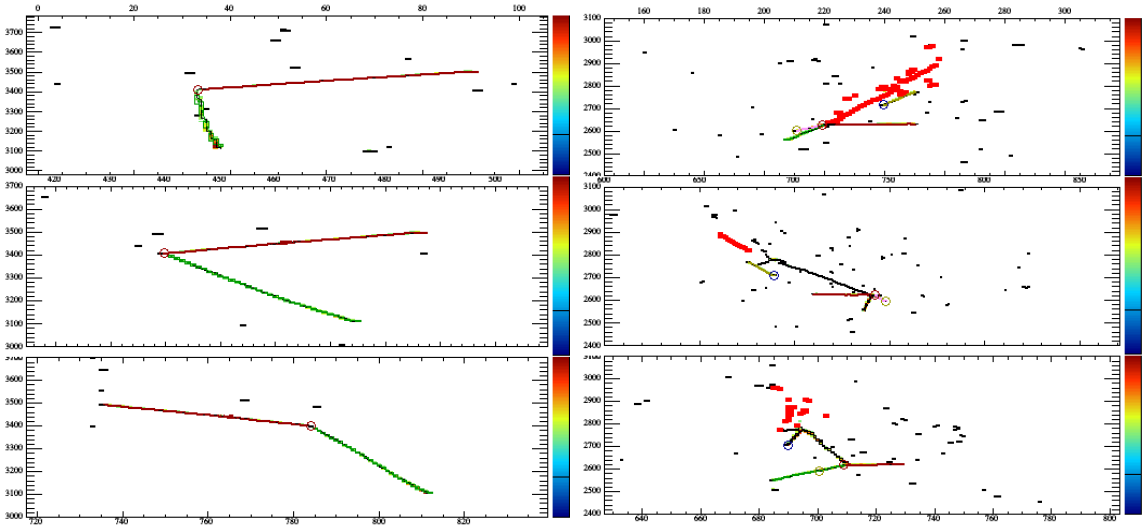


Figure 5.15: (left) Background event with low BDT response value. (right) Background event with high BDT response value.

5.5 Sensitivity calculation

The sensitivity to the $n - \bar{n}$ oscillation lifetime can be calculated with a given exposure, the efficiency of selecting signal events, and the background rate along their uncertainties. The lifetime sensitivity is obtained to 90% confidence level (C.L.) for the bound neutron. Then, the lifetime sensitivity for a free neutron is calculated using the conversion from nucleus-bound neutron to free neutron $n - \bar{n}$ oscillation [40, 118].

5.5.1 Bayesian statistical method

The Bayesian statistical method is utilized to express conditional probability when there is given observation.[3] The conditional probability of A when B is given is constructed as,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{5.4}$$

[3]This bayesian approach is chosen over a frequentist approach (e.g., Feldman-Cousins method) to use the same sensitivity calculation method used in the previous $n - \bar{n}$ search in Super-Kamiokande [50], for the purpose of the comparison of the results.

where $P(A)$ and $P(B)$ are individual probabilities of A and B.

The occurrence of events follows a Poisson distribution

$$p(n, \lambda) = \frac{e^{-\lambda}\lambda^n}{n!} , \tag{5.5}$$

where λ is the mean of the distribution that corresponds to the expected event rate, and n is the number of observed events. The expected number of events is expressed as

$$\lambda = \Gamma E \epsilon + b , \tag{5.6}$$

where $\Gamma(year^{-1})$ is the rate of $n - \bar{n}$ oscillation bound in nucleus, $E(year \times N_{neutron})$ is the total exposure of the DUNE's far detector, ϵ is the signal selection efficiency, and b is the expected number of background events during the exposure. For the evaluation of $n - \bar{n}$ oscillation lifetime sensitivity, no observation of signal events is hypothesized, i.e. the number of observed events is set to the expected background rate.

Substituting Eq. 5.5 and Eq. 5.6 into Eq. 5.4 gives the integral

$$P(\Gamma|n_{obs}) = A \int\int\int \frac{e^{-(\Gamma\epsilon E+b)}(\Gamma\epsilon E + b)^{n_{obs}}}{n_{obs}!} P(\Gamma)P(E)P(\epsilon)P(b)dEd\epsilon db , \tag{5.7}$$

where $P(\Gamma)$ is the Heavyside step function ($P(\Gamma) = 1_{\Gamma>0}$), and $P(E)$, $P(\epsilon)$, and $P(b)$ are Gaussian priors with the uncertainties of E, ϵ, and b. The systematic uncertainties of E, ϵ, and b are used for the Gaussian sigma values in the prior distributions and are given in the next section.

5.5.2 Systematic uncertainties of E, ϵ, and b

The preliminary systematic uncertainties on exposure, signal and background efficiency are estimated based on the $n - \bar{n}$ oscillation search from the Super-Kamiokande collaboration [50].

	Systematic Uncertainty
Exposure	3 %
Signal selection efficiency	25 %
Background rate	25 %

Table 5.5: Systematic uncertainties are estimated according the Super-Kamiokande $n-\bar{n}$ oscillation search.

5.5.3 Limit on bound $n - \bar{n}$ oscillation lifetime

The lifetime sensitivity to 90% confidence level of $n - \bar{n}$ oscillation process is found by evaluating Γ to the point where the integral of Eq. 5.7 makes 90% of the integral over the entire Γ domain. Thus, $\Gamma_{0.9}$ is acquired from

$$90\% = \frac{\int_{\Gamma=0}^{\Gamma_{0.9}} P(\Gamma|n_{obs})d\Gamma}{\int_{\Gamma=0}^{\infty} P(\Gamma|n_{obs})d\Gamma} \ . \tag{5.8}$$

In calculating the integrals, an open-source C package, Cubature is used [119].

As the decay width Γ is in a reciprocal relation with the lifetime T, the 90% limit for the lifetime is derived from the $\Gamma_{0.9}$ as

$$T_{n-\bar{n}} = \frac{1}{\Gamma_{0.9}}. \tag{5.9}$$

For the calculation of the sensitivity with the no-signal hypothesis, the observed number of events is set to the background estimate ($n = b$).

5.5.4 Sensitivity on free $n - \bar{n}$ oscillation lifetime

Free $n - \bar{n}$ oscillation lifetime $\tau_{n-\bar{n}}$ and bounded $n - \bar{n}$ oscillation lifetime are related with each other through the suppression factor R as

$$\tau_{n-\bar{n}}^2 = \frac{T_{n-\bar{n}}}{R} \ . \tag{5.10}$$

The suppression factor R varies for different nuclei. The calculation of this suppression factor R was done for ^{16}O and ^{56}Fe [118]. The R for ^{56}Fe, $6.66 \times 10^{22}\ s^{-1}$, has been used in previous $n - \bar{n}$ oscillation lifetime sensitivity calculations in ^{40}Ar [81, 100, 120]. Recently, the suppression factor for ^{40}Ar nuclei was calculated [40] to be $5.6 \times 10^{22}\ s^{-1}$, in the prospect of the $n - \bar{n}$ search in DUNE. The sensitivity in DUNE is calculated using both values.

5.6 DUNE sensitivity to neutron-antineutron oscillation lifetime

5.6.1 CNN-only analysis

The lifetime sensitivities of the $n - \bar{n}$ oscillation in DUNE for a 400 kton·year exposure are calculated using different CNN classification score cuts. The 90% C.L. lifetime sensitivities are shown in Tab. 5.6, with the best sensitivity at $6.563{\times}10^{31}$ years for the ^{40}Ar-bound neutron obtained when using only a CNN-based selection. The ^{40}Ar-bound sensitivity is translated to $n - \bar{n}$ oscillation sensitivity for free neutrons using Eq. 5.10. The conversion is shown both using the suppression factor (R) calculated for ^{40}Ar and ^{56}Fe, to allow for comparisons with previous work [100] which used the ^{56}Fe suppression factor.

Score cut	$T_{n-\bar{n}}$	$\tau_{n-\bar{n}}$ ($R = 5.6e^{22}s^{-1}$)	$\tau_{n-\bar{n}}$ ($R = 6.66e^{22}s^{-1}$)
0.99986	6.563×10^{31} yrs	1.923×10^{8}	1.763×10^{8} s
0.9999	6.430×10^{31} yrs	1.903×10^{8}	1.745×10^{8} s
0.99995	6.011×10^{31} yrs	1.840×10^{8}	1.687×10^{8} s
0.99999	4.986×10^{31} yrs	1.676×10^{8}	1.537×10^{8} s

Table 5.6: 90% C.L. lifetime sensitivity on argon-bound $n-\bar{n}$ oscillations using CNN classification with respect to 400 kton·year exposure.

5.6.2 BDT analysis

Combining image recognition with standard reconstruction into a multivariate analysis (BDT), the lifetime sensitivity for a bound neutron can be calculated by following the procedures in

Sec. 5.5 with the assumption on systematic uncertainties from Tab. 5.5. The best bound neutron lifetime sensitivity with the BDT analysis is achieved using the signal efficiency of 8.0% at the background rejection rate of 99.98%. These efficiencies were obtained by optimizing the sensitivity as a function of the BDT score cut. The 90% C.L. limit of a bound neutron lifetime is 6.45×10^{32} years. The corresponding limit for the oscillation time of free neutrons is calculated to be 6.03×10^{8} seconds using $R = 5.6\times10^{22}s^{-1}$, and 5.53×10^{8} seconds using $R = 6.66\times10^{22}s^{-1}$, for a 400 kton-year exposure. This demonstrates a better sensitivity than the current best limit [41].

5.7 Discussion

A number of ways to improve this analysis have been discussed and planned for the future. Here the discussion entails (1) improving CNN classification performance, (2) investigation of BDT variables, (3) accurately modeling argon-bound $n-\bar{n}$ oscillation, (4) properly estimating the systematic uncertainties.

There are approaches that we can take to improve CNN classification performance. First of all, the network training can be done on samples with higher statistics. The current network training uses $\sim$ 38k images for each signal and background, which may not be enough for the network to fully learn the features of the $n-\bar{n}$ oscillation signal and the atmospheric background from their reconstructed images. Secondly, CNN analysis can be performed using all three plane images rather than only the collection plane. Using three planes was previously explored in [100]. However, the performance was not better than the collection plane only case, which is counterintuitive. CNN classification performance using all planes is worth investigating again, since using all three planes is thought to provide more information about the events to the network. Moreover, a truth level study to better understand the case of background events with high signal score can be done. Also, the current method relies on down-sizing up to a factor of eight on the time-tick axis, which harms the resolution of the image. A new definition of input frame without heavy down-sizing can be applied. Further on, the 960 collection plane wires of a single APA should be separated into two

images with 480 wires each, since each set views a different drift volume from the other side of the APA. If adopted successfully, all of these approaches will benefit from the background rejection of this analysis.

The BDT classification using current input variables can provide a good enough separation between the signal and background to achieve a stronger sensitivity than the current best limit. However, further exploration of BDT variables possibly using various reconstruction paradigms such as Pandora and WireCell [79, 121] may improve the BDT classification even further. Also, truth level information for the corresponding input variable should be shown as well to validate the reconstruction paradigm in the signal reconstruction. For example, the truth level visible energy should be shown along the reconstructed visible energy shown in Fig. 5.11. One can also consider adding more variables in the BDT such as momentum balance.

The argon-bound $n - \bar{n}$ oscillation modeling can be updated.

The current modeling uses the neutron density distribution to model annihilation; however, the annihilation density distribution is known to be different from the neutron density distribution because antineutron absorption occurs at the periphery of the nucleus with a higher probability. Thus, the annihilation density tends to be more peripheral, while the neutron density is more concentrated near the center [40, 122]. The nuclear potential difference between neutron and antineutron should also be taken into account.

In the long term, the systematic uncertainties for the relevant variables such as the signal efficiency and background efficiency should be estimated rigorously. The systematic uncertainties in Tab. 5.5 should be considered as temporary placeholder values. The detector design specifics (e.g. horizontal vs. vertical drift modules) and far detector staged installation plans should be taken into account. Also, the stability of CNN classification should be estimated by further studies. The validation of this analysis method on MicroBooNE data is in progress, and discussed in Ch. 7 of this thesis. Lastly, the assumption that the atmospheric neutrino event is the major and only background of this analysis should be confirmed. For instance, the effects from cosmic ray hadronic showers in the vicinity of the detector may need in-depth study.

Chapter 6: Self-triggering in large-scale LArTPCs

As discussed in the previous chapter, the offline analysis searching for $n - \bar{n}$ oscillation in DUNE shows an already promising sensitivity projection and gives multiple pointers for further improvements. The result of this analysis does not consider any possible loss of signal events due to the efficiency of online data selection (trigger) in DUNE FD. In other words, 100% triggering efficiency is assumed for such a highly energetic interaction like $n - \bar{n}$ oscillation. This assumption applies to other interactions of BSM searches and off-beam physics topics in DUNE [81, 120], such as proton-decay, atmospheric interactions, and cosmogenic interactions. On the other hand, supernova neutrinos from supernova bursts (SNBs) have low energy around the MeV scale. Still, multiple supernova neutrinos from one SNB over 10 seconds are expected to be observable by DUNE. Thus, triggering SNBs and high-energy interactions versus intrinsic impurity activities, which are the background, is crucial to fulfilling DUNE's physics objectives.

To record high-energy interactions and low-energy interactions from SNBs, DUNE FD plans on a self-triggering DAQ using continuous TPC readout. One FD module of DUNE corresponds to 384,000 TPC wire channels readout and digitized continuously with the sampling frequency of 2MHz. In order to meet the data requirement (30PB/year) in DUNE, a 10^4 reduction factor for background is necessary. While the nominal DUNE DAQ design uses algorithmic triggering using low-level trigger candidate reconstruction, CNN-based real-time triggering is explored in this chapter as an alternative.

The following two papers demonstrate the work I have been leading in the effort of inspecting a CNN-based triggering scheme in a large LArTPC. This includes investigating the prospect of the implementation of CNNs on FPGAs to be accommodated in the DUNE DAQ system. "Accelerating Deep Neural Networks for Real-time Data Selection for High-resolution Imag-

ing Particle Detectors" was published as proceedings to the New York Scientific Data Summit (NYSDS) meeting in 2019 [123]. "Real-time Inference with 2D Convolutional Neural Networks on Field Programmable Gate Arrays for High-rate Particle Imaging Detectors" has been submitted to arXiv [124], and is currently under review for publication in Frontiers of AI.

6.1 Accelerating CNNs for real-time data selection

In the following paper, CNN inference for the data selection scheme for DUNE is demonstrated on GPUs, and CNNs are further implemented on FPGAs using custom-designed hardware acceleration. The 2D input from simulated DUNE's raw digit is used for this study. Various CNN architectures and 2D input sizes were explored on GPUs. The performance of CNNs on GPUs was evaluated by the classification accuracy and the latency of the inference. However, GPUs are not the ideal option for a DAQ system located underground, such as DUNE, due to their power consumption and heat dissipation; this motivated the implementation of CNN inference algorithms on FPGAs. The paper also demonstrated the hardware acceleration design to implement one of the CNNs on FPGAs, showing resource analysis.

Accelerating Deep Neural Networks for Real-time Data Selection for High-resolution Imaging Particle Detectors

Yeon-jae Jwa
Dept. of Physics
Columbia University
New York, NY, USA
yj2429@nevis.columbia.edu

Giuseppe Di Guglielmo
Dept. of Computer Science
Columbia University
New York, NY, USA
giuseppe@cs.columbia.edu

Luca P. Carloni
Dept. of Computer Science
Columbia University
New York, NY, USA
luca@cs.columbia.edu

Georgia Karagiorgi
Dept. of Physics
Columbia University
New York, NY, USA
georgia@nevis.columbia.edu

***Abstract*—This paper presents the custom implementation, optimization, and performance evaluation of convolutional neural networks on field programmable gate arrays, for the purposes of accelerating deep neural network inference on large, two-dimensional image inputs. The targeted application is that of data selection for high-resolution particle imaging detectors, and in particular liquid argon time projection chamber detectors, such as that employed by the future Deep Underground Neutrino Experiment. We motivate this particular application based on the excellent performance of deep neural networks on classifying simulated raw data from the DUNE LArTPC, combined with the need for power-efficient data processing in the case of remote, long-term, and limited-access operating detector conditions.**

***Index Terms*—convolutional neural network, deep neural network, hardware acceleration, LArTPC, particle detector**

I. Introduction

Liquid Argon Time Projection Chambers (LArTPCs) represent a particle detector technology that has been widely adopted in the field of high energy physics. Over the last two decades, LArTPCs have been increasingly used for studying neutrino-argon interactions with high calorimetric (energy) and spatial resolution. LArTPCs are already in use for a number of detectors; the most recent of these detectors, MicroBooNE [1] and ProtoDUNE [2], represent a significant R&D effort which is underway to scale up the LArTPC detector technology by up to two orders of magnitude in physical detector size. This phasing approach is necessary in order to realize the future Deep Underground Neutrino Experiment (DUNE) [3], [4], which will feature the largest LArTPC detector to be ever constructed and operated at a deep underground location in Lead, South Dakota, in the United States, starting in $\sim$2025.

LArTPCs, including DUNE, work by imaging particle tracks and other signatures imprinted in a large, uniform detector volume by particles produced in neutrino or other rare physics interactions. Different interactions yield distinct image topologies that are identifiable and differentiatable by their spatial extent, shape, and pixel intensity, when viewed as two-dimensional projections of a three-dimensional detector region. Furthermore, the format of the detector-generated raw data represents exactly two-dimensional projections of the activity inside the detector; as such, a potentially advantageous solution for real-time data processing and data selection (triggering) on interesting detector activity is image analysis with hardware-accelerated Deep Neural Networks (DNNs).

DNNs are already being applied successfully for the offline analysis of data recorded by existing high energy physics experiments [5], including operating LArTPCs. In the case of the latter, MicroBooNE is pioneering the use of deep learning for neutrino physics analyses (see, e.g., [6], [7]), and similar DNN-based methodologies have now been adopted for several analyses planned with the future DUNE experiment [4]. Machine learning approaches to LArTPC data analysis are gaining increasing traction (see, e.g. [8]); meanwhile, new techniques are continually being considered to improve data processing latency and resource requirements, with promising results [9].

At the same time, the success of DNNs more generally has motivated the research and development of many specialized system architectures and accelerators both in academia and in industry. An excellent overview of the challenges of accelerating DNNs in hardware and a comprehensive survey of many techniques and frameworks that have been proposed so far in the literature is provided in [10]. In terms of implementation, DNN frameworks mainly target CPUs and GPUs. In particular, GPUs offer high computational density and high level of programmability; this simplifies the interface with operating systems while providing access to powerful computational platforms for data-parallel algorithms and dense floating-point operations. GPU performance, however, comes with high power dissipation, making a GPU-based solution unsustainable for many high-performance embedded systems that require major power efficiency. Thanks to their hardware reconfigurability, Field Programmable Gate Arrays (FPGAs) are a valid alternative solution as power-aware platforms for DNN acceleration [10]. In addition to hardware developments, frameworks such as Caffe [11] and Tensorflow [12] allow a much larger user base for modern DNNs.

In this paper, we investigate the viability of DNN implementations in a variety of architecture systems, including

GPU for online data processing, and FPGA or mixed FPGA-CPU architecture systems for real-time data processing, both for the purposes of data selection (triggering) for a high-resolution and high-rate imaging detector. The application we specifically target is that of DUNE, which involves real-time streaming of data rates of the order of tens of terabits per second. The proposed data selection schemes, however, may be applicable to any LArTPC, sharing the same technology as DUNE, and particularly viable for smaller-scale ones. We note that the application of machine learning algorithms for triggering purposes has been considered for other types of particle detectors (see, e.g. [13]). However, the application proposed here for LArTPCs is a new effort, and it deals with a unique set of challenges: specifically, LArTPC triggering is governed by a much larger input (image) size, but also benefits from relaxed latency constraints due to a much slower detector response than other types of particle detectors. The targeted DUNE application and DUNE detector design are presented in Sec. II.

To motivate the application of DNNs for DUNE data selection purposes, we train and investigate the performance of a number of DNNs on simulated LArTPC raw data images. Results obtained on GPUs are presented in Sec. III, and demonstrate high efficiency in selecting rare physics interactions of interest, while maintaining a sufficiently low selection rate from background interactions and detector noise. Latency and power dissipation considerations, however, motivate the investigation of inference on FPGA or mixed FPGA-CPU systems, which have been shown to achieve significant speedup [14]. As such, in Sec. IV, we present several contributions for designing hardware acceleration of Convolutional Neural Network (CNN) inference algorithms on resource-constrained platforms like FPGAs. By using a customizable and efficient hardware accelerator design for the various layers, we show that the flexibility of the accelerator design together with the possibility of leveraging the knobs provided by High Level Synthesis (HLS) tools enable the design of high-performance accelerators that can greatly benefit the deployment of DNN models. Finally, in Sec. V, we identify DNNs which would satisfy DUNE physics and latency requirements, considering also resource utilization on an FPGA with specifications that might be suitable for DUNE readout.

II. Application Use Case: Deep Underground Neutrino Experiment

DUNE is an international particle physics experiment that aims to study neutrinos and their oscillation patterns with unprecedented sensitivity as well as search for other rare particle interaction signatures that will inform our understanding of nature at the most fundamental level. In particular, DUNE measurements aim to elucidate the underlying mechanism responsible for the prevalence of matter over antimatter in our observable universe. To accomplish these physics goals, the DUNE far detector will employ four LArTPC modules, each holding 10 kilotons of liquid argon in total fiducial detector mass, and will operate for more than a decade in a deep underground location at Sanford Labs, in Lead, South Dakota, beginning in the middle of the next decade.

TABLE I
Expected rates of rare off-beam events and other off-beam signatures in a 10 kton (fiducial mass) DUNE far detector module [4].

Interaction Type	Event Type	Expected Rate
Rare off-beam events		
Proton decay	High Energy (HE)	< 1 / year
Neutron-antineutron oscillation	High Energy (HE)	< 1 / year
Galactic supernova burst[a]	Low Energy (LE)	< 1 / year
Other off-beam events		
Atmospheric neutrinos	High Energy (HE)	1200 / year
Cosmic ray muons	High Energy (HE)	1.3×10^6 / year

[a]A galactic supernova burst is expected at a rate of roughly once per century. The latest galactic supernova burst was observed in 1604 [15].

To study neutrino oscillations, DUNE must detect interactions of neutrinos from a high-intensity pulsed beam from Fermi National Accelerator Lab, in Batavia, Illinois. Selecting and recording these interactions is straightforward since they are all expected to arrive only during a relatively short time dictated by the beam pulse structure; the latter is precisely known due to external beam timing signals informing the trigger decision. To study other rare, off-beam events such as proton decay events, neutron-antineutron oscillation events, and interactions of neutrinos from galactic supernova bursts, however, DUNE must continually process its data in order to make a data-driven decision to select and record these signatures. This is because these signatures are random in nature, and no prompt external timing signal is available to independently inform the data selection decision. The expected rate of rare off-beam events, and other off-beam interactions of interest, in DUNE is provided in Tab. I.

The DUNE system responsible for data selection must, in the end, only allow for effectively 30 petabytes of data per year to be diverted to permanent storage offline [4]. As such, given the multiple tens of terabits per second raw data rate of the DUNE far detector, a factor of 10^4 data reduction must effectively be achieved by the system, without compromising efficiency for selecting rare events of interest. Generally, a trigger efficiency of $>99\%$ is required for high energy events, including atmospheric neutrino interactions, proton decay events, neutron-antineutron oscillation events, and cosmic ray muon events in the detector. Similar trigger efficiency is also required for selecting aggregates of multiple low energy supernova neutrino interactions that are expected to occur in case of a galactic supernova burst. In that case, the trigger efficiency requirement on any individual supernova neutrino interaction can be relaxed, and a multiplicity condition can be used to boost efficiency for coincident interactions[1].

The main challenge, specific to the supernova burst trigger, is that individual supernova events are characterized by low energy deposition in the detector; as such, their observable

[1]In the case of a supernova at the edge of our galaxy, for example, approximately 50 supernova neutrino interactions are expected over the span of ten seconds in each DUNE 10 kton module.

signature is similar to that of intrinsic radiological backgrounds and electronics noise in the detector, which are the dominant contributor to observable signals in the DUNE data. Consequently, in order to achieve the desired data reduction factor, significant noise and radiological background rejection is needed.

Two distinct detector designs are in development for the DUNE far detector modules. We restrict the discussion and studies presented in this paper to the so-called "single phase" LArTPC module technology, described in the following subsection, following [4]. However, both the single phase and "dual phase" technology operate on high-resolution imaging principles; we therefore expect that comparable challenges, solutions, and performance would be achievable for the "dual phase" technology for DUNE as well.

A. DUNE Single Phase Detector Design

In the case of the DUNE far detector "single phase" design, each DUNE 10 kton far detector module is segmented into 150 individual "cells" (rectangular volumes) of liquid argon, which are imaged by sensor-wire arrays, called an Anode Plane Arrays (APAs). An APA is positioned in the middle of each cell, and it consists of multiple planes of parallel wires oriented in three distinct directions relative to the vertical direction. The wires sense ionization charge (electrons) liberated by charged particles along the charged particles' paths as they traverse the liquid argon volume enclosed in the cell; the ionization charge drifts toward the wire planes under the influence of a strong, uniform electric field applied across each cell, on either side of the APA. Given the arrival time of the ionization charge, relative to the time of the interaction (identified and recorded by detecting the prompt scintillation light produced at the time of the interaction, using a dedicated photon detection system), the drift coordinate of the event can be reconstructed. The ionization signals recorded as a function of wire number across each wire plane, and as a function of time, can then be mapped into a two-dimensional projected view of the cell, for a given time; this makes it possible to reconstruct a three-dimensional view of any interaction inside a cell by matching signals across the three stereoscopic views (one per plane).

The studies in this paper involve only signals from vertically oriented wire planes. One such plane exists on each side of the APA, and makes up a so called charge "collection" wire plane. Due to the electric field configuration and readout electronics response, recorded signals on collection wires are unipolar; as such, their amplitudes and integrals, in particular, correlate highly with the amount of ionization charge arriving at each wire. We refer to channel vs. time data which spans the equivalent of a collection plane times drift time (drift length on one side of the APA divided by wire signal sampling rate) as an "APA-frame". For the DUNE APA cell physical dimensions and nominal electric field configuration, the APA-frame drift time corresponds to 2.25 ms.

Simulations of APA-frames representative of several topologies of interest, from Tab. I, are show in Fig. 1. APA-frames are simulated using the LArSoft framework [16], [17] and DUNE Monte Carlo generation tools [18]. DUNE Monte Carlo generation configuration parameters are set to the dunetpc v07_13_00 default values, except for the electronics noise RMS levels, which are artificially enhanced for conservatism; specifically, in our simulations, we increase the collection plane electronics noise RMS by 40% relative to the default value. All APA-frames with topologies of interest also include default radiological background and electronics noise.

B. DUNE Data Acquisition System Design

DUNE will have to operate continually, for more than a decade, streaming data out of its LArTPC detectors at a total rate of multiple tens of terabits per second. For reference, MicroBooNE [1] and ProtoDUNE [2], the two largest currently operating LArTPCs, stream images continually at a data rate of greater than 260 and 490 gigabits per second, respectively. Unlike DUNE, these experiments do not have a rare event search physics scope. Data reduction for these detectors is therefore achieved through a combination of external trigger signals informing when to record a small subset of that data, and additional real-time compression, filtering and/or zero-suppression carried out in FPGA and/or CPU (see, e.g. [1], [19]).

Differently from these detectors, the DUNE detector must be capable of processing its data in real time, or, in an online fashion, in order to make data-driven decisions to record what might be rare physics events. DUNE's data acquisition system (DAQ), and in particular its data selection (sub)system, must do so with negligible dead-time, to maximize the detector's physics sensitivity to rare signatures. An additional constraint is power distribution limitations at the (underground) detector site. Specifically, the DUNE far detector DAQ is limited to 500 kVA of power underground, or 125 kVA per 10 kton module, plus an additional 50 kVA of power available on the surface for back-end DAQ [4].

The baseline DUNE DAQ design is documented in detail in [4]. It employs a multi-level data selection system. First, a low-level data selection decision is achieved on a combination of CPU and FPGA resources. This level of data selection is executed independently on a per-APA basis, while the second level, to first order, aggregates low-level information from all APAs in a single module to make a module-level trigger decision. The module-level trigger decision is executed on CPU resources, and its latency is limited to a few seconds. When formed, a module-level trigger decision instructs the readout of several milliseconds worth of continuous data from all 150 APAs in the module, or, in the case of a supernova burst trigger decision, 100 seconds worth of continuous data from all 150 APAs. Non-supernova burst trigger decision rates of up to O(1) Hz are possible, while supernova burst trigger decision rates are limited to one per month. These upper limits on trigger rates include fake triggers on accidental backgrounds and noise; therefore, background noise considerations are especially important in the case of supernova burst triggers. Additional data down-selection can be achieved by the use

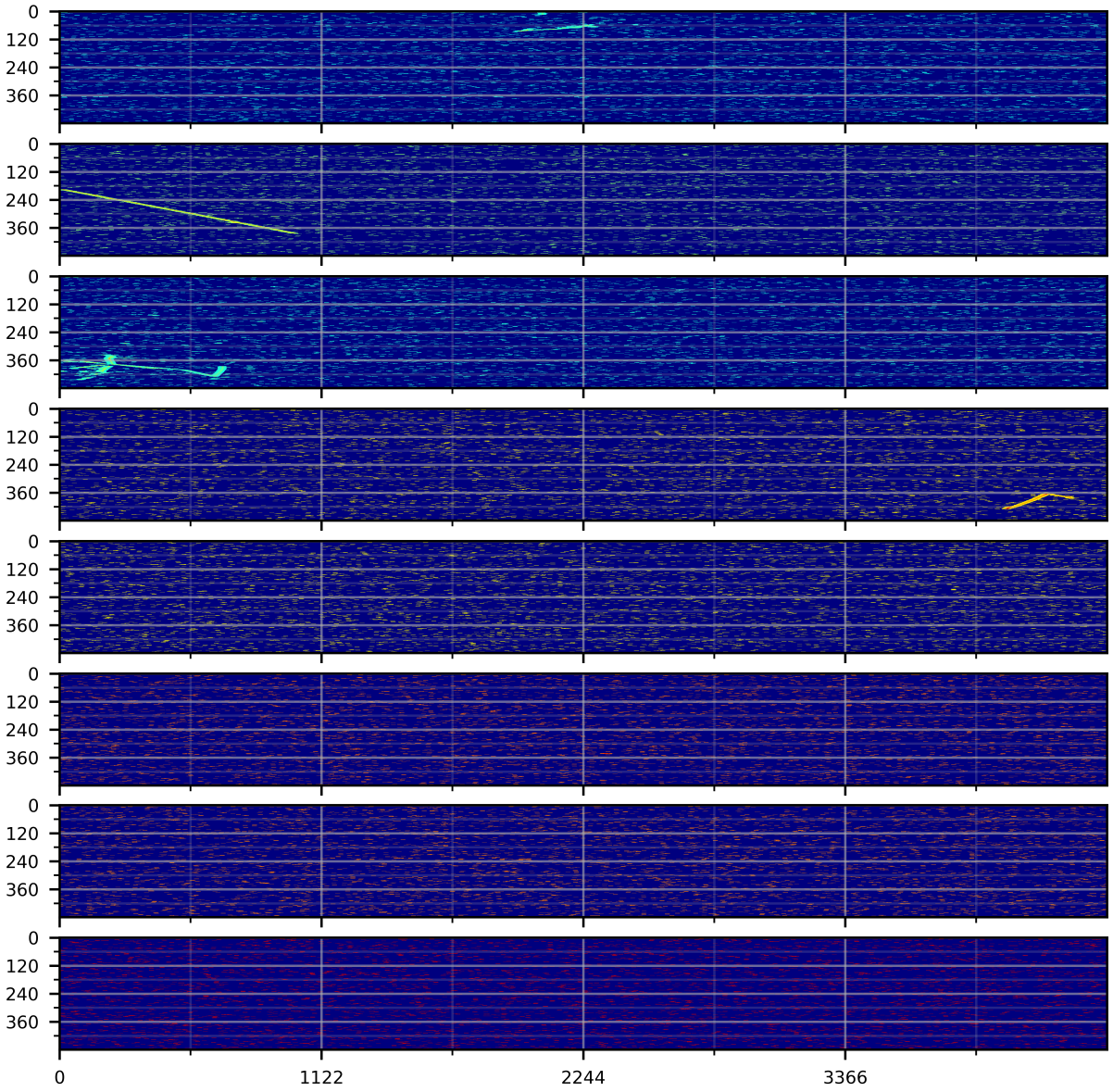


Fig. 1. Simulated APA-frames, representative of three main types of signatures of interest. The top four frames correspond to high energy events; the lower four frames correspond to low energy supernova neutrino events (first two) and empty events including only background noise (bottom two). APA-frames are defined according to the APA drift volume in which the interaction originates. The y axis of each frame corresponds to collection plane channel number; the x axis corresponds to time tick (2 MHz) across a full 2.25 ms readout.

of a high-level filter farm, which is envisioned to employ data selection techniques similar to the ones presented in this work.

III. DNN-based LArTPC Data Selection

To motivate DNN-based LArTPC data selection, we have studied a number of DNNs in terms of their performance on classifying simulated DUNE far detector single phase APA-frames. We have considered a multi-class data classification scheme, where the different classes represent different types of off-beam physics events of interest that can occur in the DUNE far detector, as well as non-physics events (intrinsic to the detector materials radiological backgrounds and electronics noise backgrounds).

The methodology we followed assumes that a two-level data selection system is used to (1) first generate a low-level data selection decision, specifically the classification of APA-frames according to their content with the use of a DNN, and to (2) subsequently process those decisions further in order to

make a module-level data selection decision. More specifically, the module-level data selection stage keeps track of information[2] from APA-frames that have been tagged as a certain type of interaction over the entire 10 kton detector module, over a given time interval. In this way, for example, a supernova burst trigger decision can be generated at the module level if multiple APA-frames are tagged by the "low-level" trigger as containing supernova neutrino interactions over a short amount of time (typically on the order of seconds). Our studies focus particularly on the low-level stage of processing.

The APA-frames stream continually from each DUNE detector module, at a rate of 200 frames (one for each drift volume) per 2.25 ms. Each frame is 480 channels wide by (2.25 ms)×(2 MHz)=4500 samples[3] wide, corresponding to a total of 4.15 megapixels, with 12-bit color resolution. Because of the large APA-frame size (3.2 GB), significant down-sizing is necessary in order to fit APA-frames into image sizes typically processed by DNNs. Down-sizing is also applied in anticipation of the limited resources available on FPGAs that the DUNE far detector data selection system will employ for low-level data selection [4], which we consider to be a candidate hardware platform for DNN deployment.

Two methods were followed to pre-process APA-frames in preparation for DNN classification; classification was carried out with a VGG16b network [20] trained and tested independently for each method on a GPU:

- Method 1: In the first method, noise removal was minimally applied to each APA-frame, and the resulting image was re-sized by down-sampling it into a 600×600 image, to be used for DNN inference.
- Method 2: In the second method, aggressive noise removal was applied to each APA-frame before down-sizing the image for inference, followed by cropping around a signal "region of interest" (ROI), and re-sizing the resulting ROI (by down-sampling or up-sampling) into a 64×64 image. The noise removal and ROI finding were informed by studying the ADC distributions of simulated APA-frames of different signatures, as shown in Fig. 2. Examples of ROIs are shown in Fig. 3.

For both methods, the images were used to train a customized VGG16b network, and the resulting network was tested on a statistically independent sample of images, prepared in the same way, for accuracy and inference speed. The tests were performed on a single GPU, NVIDIA GeForce GTX 1080 Ti.

Inference results on GPU from each method for VGG16b are summarized in Tabs. II and III. The tables show the number of ROI images used for training and testing for each sample; and resulting accuracy, identified in terms of the fraction of input images in the testing case which get classified under each label: background noise (NB), low energy supernova neutrino interaction (LE), or high energy interaction (HE). The given fractions are inclusive of all event energies. Finally, per-

[2]Spatial coordinate, type of interaction, etc.
[3]More specifically, 4488 samples are used for simulation purposes.

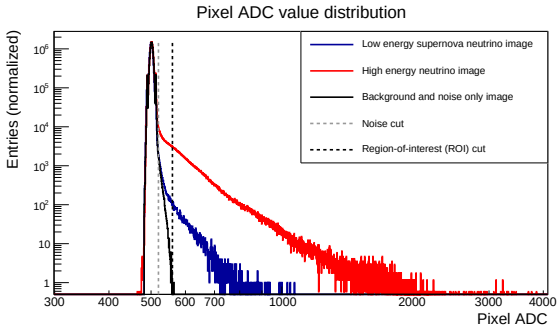


Fig. 2. Pixel ADC distributions of frame images for the three classes in consideration: background noise (black), low energy neutrino (blue), and high energy neutrino (red) images. The pixel ADC values range from 0 to 4095 (12-bit ADC). The distributions are absolutely normalized to 100 images with 480×4488 pixels each. The background noise distribution peaks below 520 ADC for all frames. The dashed vertical lines indicate cuts that were used in pre-processing input images for the networks, in order to de-noise the raw images and to select regions with candidate physics interactions. Based on these distributions, a noise removal cut (indicated by the dashed gray line) and an ROI cut (indicated by the dashed black line) was set to 520 and 560 ADC, respectively.

TABLE II
GPU INFERENCE RESULTS USING METHOD 1, OBTAINED WITH A WITH VGG16B NETWORK (TRAINING FOR 2 EPOCHS AND LEARNING RATE SET TO $2 \times 10^{-}4$).

Sample	Train Size	Test Size	Accuracy (%) ϵ_{NB}	ϵ_{LE}	ϵ_{HE}	Inference Time (ms)
NB	51,100	99,000	91.45	8.49	0.06	
LE	44,900	29,800	3.17	96.83	0	27.7±8.6
HE	52,828	67,178	6.03	3.48	90.48	

APA-frame inference times are provided, in milliseconds, and include image input i/o from host (GPU server) memory. The key table parameters are the correct classification rates of low energy and high energy frames, both of which are required to be high by DUNE physics performance requirements, as well as the mis-classification rate of noise frames as low energy frames, which should be as low as possible by data reduction requirement considerations. Both methods are found to yield comparable results in terms of classification accuracy. More specifically, the networks are able to select high energy and low energy frames with efficiencies in excess of 95% and 90%, respectively. Required efficiency for high energy frames should be $> 99\%$ for interactions with visible energy in excess of 100 MeV. The obtained efficiencies are integrated over all energies (which extend below 100 MeV); it is expected that a HE efficiency calculated relative to interactions with visible energy in excess of 100 MeV would be higher. While signal efficiency performance is comparable for the two methods, Method 2 performs much better with respect to mis-classification rates for background noise frames as LE frames, where a false pass rate of 0.35% is achieved.

Inference latency for the two methods is also comparable, although Method 2 inference is faster by more than a factor of five, due to the reduced size of the input image. Latency

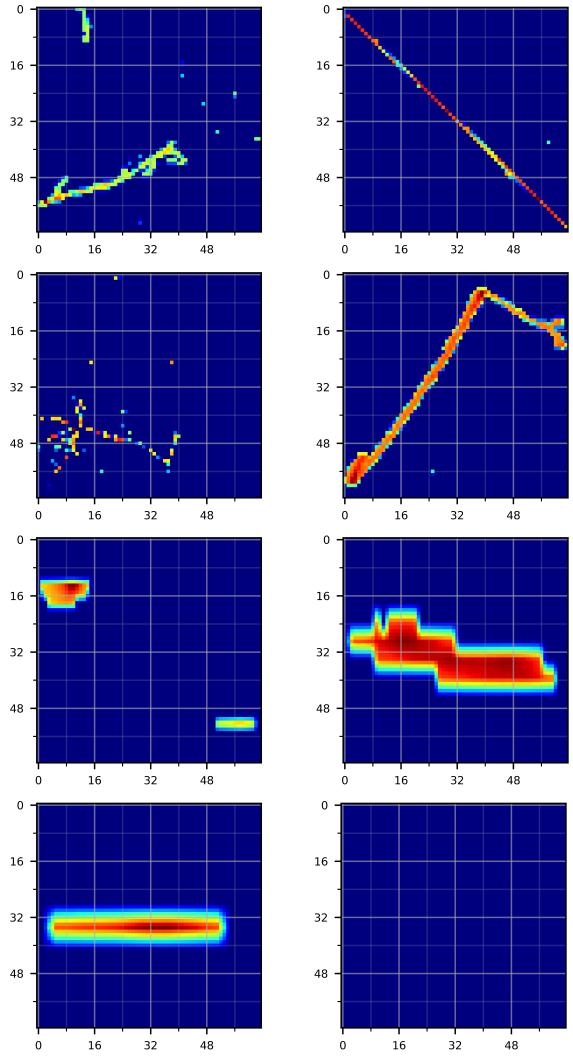


Fig. 3. ROIs extracted using Method 2 for the simulated frames shown in Fig. 1. The y axis represents channel space; the x axis represents time space. The top four panels correspond to high energy interactions; the subsequent two correspond to low energy interactions; the bottom two correspond to background noise (typically empty frames, after noise removal, or noise artifacts). Noise removal is achieved by zero-suppressing pixels with ADC values below 520 ADC; an ROI is defined by first finding the smallest contiguous rectangular region in a frame that contains at least one pixel value exceeding 560 ADC, padded by five (5) additional pixels in each direction (left, right, top, or bottom); the resulting region is down-sized or up-sized by down-sampling or up-sampling to fit into a 64×64 image, as shown here, defined as an ROI, and is then fed into a DNN for inference.

considerations determine whether frame-by-frame inference can be applied during the low-level data selection stage of the DUNE far detector DAQ system; such application would have to keep up with the frame rate of 66.6×10^3 fsps. In the case of Method 1, if we required that every frame go through image classification, the observed latency of 27.7 ms (an order of magnitude off 2.25 ms even with a 150-fold parallelization)

TABLE III
GPU INFERENCE RESULTS USING METHOD 2, OBTAINED WITH THE VGG16B NETWORK (TRAINING FOR 13 EPOCHS AND LEARNING RATE SET TO 10^-4). NB* CORRESPONDS TO EXPLICITLY NON-EMPTY BACKGROUND NOISE ROIS, CONTAINING NOISE ARTIFACTS, WHICH REPRESENT APPROXIMATELY 2% OF THE REGIONS FOUND AFTER NOISE REMOVAL.

Sample	Train Size	Test Size	Accuracy (%)			Inference Time (ms)
			ϵ_{NB}	ϵ_{LE}	ϵ_{HE}	
NB	12,023	4,027	99.65	0.35	0	
NB*	12,023	293	79.9	19.8	0.34	5.0±0.3
LE	12,050	3,970	3.78	95.04	1.18	
HE	10,137	3,417	2.99	6.88	90.14	

TABLE IV
GPU INFERENCE RESULTS USING METHOD 2, OBTAINED WITH THE CNN_S NETWORK (TRAINING FOR 48 EPOCHS AND LEARNING RATE SET TO 2×10^-3).

Sample	Train Size	Test Size	Accuracy (%)			Inference Time (ms)
			ϵ_{NB}	ϵ_{LE}	ϵ_{HE}	
NB	12,023	4,027	99.53	0.47	0.12	
LE	12,050	3,970	4.01	94.48	1.51	1.6±0.1
HE	10,137	3,417	3.63	6.15	90.22	

would preclude such application during low-level data selection, unless a more-than-10-fold parallelization of frame-by-frame processing were to be implemented; application at high-level filter stage, however, is viable, because a relatively low module-level trigger rate (for example of order 1 Hz readout of 200 APA-frames) would make data rate handling more manageable. In the case of Method 2, the inference latency (comparable to APA-frame length of 2.25 ms) is far more promising for a frame-by-frame online low-level data selection implementation; furthermore, the processing time requirement for this method can be relaxed further based on the additional reduction of frame rate gained by the aggressive noise removal and ROI formation pre-processing stage. We have found that after noise removal and ROI finding, only 2% of the 2.25 ms-long background noise frames survive. Considering that most APA-frames that DUNE will be reading out will contain only background noise, we expect that the average frame rate reduction factor gained will be close to that of the background noise reduction factor. Hence, directing only ROIs containing non-zero pixels to network inference, for example, could relax the processing time requirement by a factor of 50.

Additional fake (background noise) trigger reduction is possible at the module-level data selection stage, by aggregating APA-frames classified as LE and considering their coincidence over the anticipated duration (10 seconds) of a supernova burst, following the methodology for supernova burst triggering in [4]. Findings from preliminary studies [4], [21] support the successful application of the coincidence-based methodology fed by CNN-based (using a VGG16b network) low-level information.

The promise of Method 2 for online application for low-level data selection further motivates the use of smaller networks, and, for the purposes of further acceleration on FPGA,

TABLE V
GPU INFERENCE RESULTS USING METHOD 2, OBTAINED WITH THE MLP_1 NETWORK (TRAINING FOR 65 EPOCHS AND LEARNING RATE SET TO $2 \times 10^{-}4$).

Sample	Train Size	Test Size	Accuracy (%) ϵ_{NB}	ϵ_{LE}	ϵ_{HE}	Inference Time (ms)
NB	12,023	4,027	99.50	0.45	0.05	
LE	12,050	3,970	4.48	89.70	5.82	1.0±0.08
HE	10,137	3,417	7.29	13.08	79.63	

TABLE VI
GPU INFERENCE RESULTS USING METHOD 2, OBTAINED WITH THE RESNET50 NETWORK (TRAINING FOR 30 EPOCHS AND LEARNING RATE SET TO $10^{-}5$).

Sample	Train Size	Test Size	Accuracy (%) ϵ_{NB}	ϵ_{LE}	ϵ_{HE}	Inference Time (ms)
NB	12,023	4,027	99.28	0.55	0.17	
LE	12,050	3,970	3.55	88.89	7.56	15.3±1.2
HE	10,137	3,417	2.84	15.13	82.03	

smaller input images. The second method was therefore further explored for a number of other customizable networks, besides VGG16b [20], including a smaller, simpler CNN than VGG16b, referred to as CNN_s [22], a Multi-Layer Perceptron (MLP) network [23], and a ResNet14b network [24]. Results from the three additional networks are provided in Tabs. IV through VI, to be considered in comparison with VGG16b results in Table III. The best performance is obtained with VGG16b and CNN_s. The simple CNN (CNN_s) performs comparably with VGG16b in terms of the accuracy, albeit with with slightly higher pass rate ($\sim$0.5%) on background noise ROIs. MLP and ResNet14b also have comparable pass rates ($\sim$0.5%) for background noise, but the accuracies for low energy and high energy ROIs are not as high as those for VGG16b or CNN_s. Inference times with CNN_s (on a single GPU card) are an order of magnitude lower than for VGG16b, due to the reduced number of layers and convolutions per layer.

Finally, we note that lower background noise pass rates could be achievable using a variation of a CNN-based selection. For example, in [21], Method 1 is used to train against six classes: NB, LE, plus the four subclasses of the HE class including atmospheric neutrino interactions (atm), nucleon decay (ndk), neutron antineutron oscillation (nnbar), and cosmic interactions (cosmic). Rather than classifying frames in terms of the six labels according to the label returning the highest score, a cut on the NB classification score is applied in order to reject frames with high enough NB scores, and select all surviving APA-frames. Results based on this classification scheme are summarized in Tab. VII as a function of NB score cut. The number of ROI images used for training and testing for each sample in Tab. VII correspond to those given in Tab. II. The main difference relative to Tab. II is that accuracy is identified in terms of the fraction of input images in the testing case which have NB score lower than what is indicated on the left column. (Here, too, fractions are inclusive of all energies.) The average inference time is comparable to that presented in Tab. II, and includes image input i/o from host (GPU server) memory.

IV. CNN IMPLEMENTATION IN FPGA

The accuracy performance of CNN_s obtained with reduced-size raw data images, combined with the reduced size of the network relative to VGG16b, motivate studies for further hardware acceleration. Hardware accelerators can be designed according to two main different approaches [25]: the designer can tightly couple the hardware functional unit inside the pipeline of a processor core or choose a loose out-of-core coupling architecture. Loosely-coupled accelerators (LCA) are hardware accelerators capable of performing Direct-Memory Access (DMA) to external main memory. LCAs are located outside the processor cores, for example on the FPGA fabric, and interact with the rest of the chip through on-chip interconnects. They can implement coarse-grain operations with dedicated datapaths that can accelerate a complete application functionality (e.g. the convolutional layers in the case of a CNN). To implement our accelerators we adopted the LCA approach, as it represents a perfect match in terms of reconfigurability and flexibility with FPGAs and embedded SoCs.

The bottleneck for the performance of inference of CNNs are the convolutional layers, which alone are responsible for more than 90% of the computations performed on networks like VGG16b. Thus we chose to specifically design a *convolutional* LCA for our CNNs.

While the workloads of many accelerators described in literature are fixed and known at design time [26], a convolutional layer has a number of parameters that are known only at run time (input dimensions, number of input channels, number of filters per layer, etc.). Different configurations of these hyperparameters lead to drastic changes in memory requirements and computational capabilities. Thus, we chose to design a LCA that is *configurable* at run-time.

We used High-Level Synthesis (HLS) to obtain the FPGA implementation starting from specifications made in C/C++ [27]. Current HLS tools enable an effective exploration of the design space of an accelerator to obtain many alternative implementations which are trade-offs of resource/power requirements and performance [28]–[30].

A. Accelerator Architecture

Figure 4 illustrates the main components and memories of our configurable convolutional LCA. It embeds three private local memories (for storing the input and output features and the filter weights), a patch extractor (for data reordering), and several multiply-and-accumulate engines which are the core of the convolution operations. The accelerator communicates with the rest of the chip through AXI4 interconnects [31].

Private Local Memory. Custom hardware accelerators allow designers to tune the microarchitecture and enable higher level of optimization to meet a specific configuration and workload, providing high performance and energy efficiency [32].

TABLE VII
GPU INFERENCE RESULTS USING METHOD 1, OBTAINED WITH THE VGG16B NETWORK (TRAINING FOR 2 EPOCHS AND LEARNING RATE SET TO $2 \times 10^{-}4$), TRAINED ON SIX CLASS LABELS. SEE TEXT FOR MORE DETAILS.

	Accuracy (%)						
NB cut	ϵ_{NB}	ϵ_{LE}	ϵ_{HE}	$\epsilon_{HE:nnbar}$	$\epsilon_{HE:ndk}$	$\epsilon_{HE:atm}$	$\epsilon_{HE:cosmic}$
0.1	0.73	88.18	96.12	99.98	99.29	92.24	92.57
0.01	0.14	83.27	95.68	99.98	99.18	91.01	92.46
0.001	0.033	77.11	95.21	99.98	99.05	89.76	92.23
0.0001	0.011	69.74	94.61	99.97	98.74	88.39	91.71
0.00001	0.002	60.73	93.79	99.95	98.22	86.61	90.97

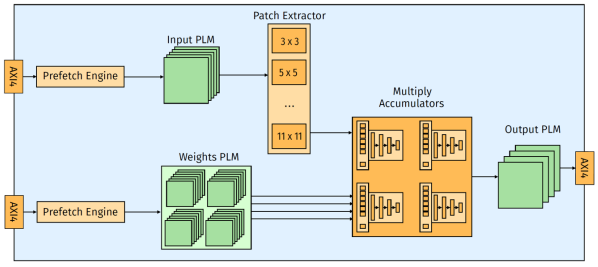


Fig. 4. Overview of the configurable loosely-coupled accelerator.

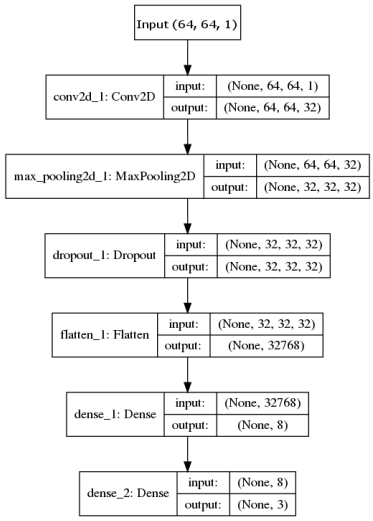


Fig. 5. Overview of our customized CNN, CNN_s.

General purpose processors (CPUs) leverage the hierarchy of caches and memory to provide the best solution in terms of bandwidth and latency across a variety of applications. Similarly, GPUs offer very high bandwidth and massive availability of parallel computational cores (CUDA cores for GPUs NVIDIA). When implementing custom hardware accelerator on FPGAs, resource utilization and allocation is an important design constraint. The designer should carefully optimize the accelerator to reuse data as much as possible, thus balancing communication versus computation and reducing expensive memory transfers from the off-chip main memory. This requires the use of private local memories (PLMs), which offer low latency, high bandwidth memory and customizable word widths. They do so, by providing many banks and ports that are exclusively accessed by the datapath logic of the LCA that embeds them [25]. Careful design and tailoring of these structures for input/output ports, partitioning, and resource allocation is essential to constantly provide data to be fed to all the high-performance computational engines.

Patch Extractor. The patch extractor is an optimized module for retrieving the portion of the input features where the filters are applied. This operation is highly dependent on the choice of hyperparameters. Due to the irregular access pattern that this module performs while fetching data from the Input PLM, we decided to have various implementations for the most common cases, from the smallest 3×3 filters up to bigger 11×11 filters. At run time, accordingly to the settings of the convolutional layer, the accelerator would choose and enable the correct patch extractor.

Multiply-and-Accumulate (MAC). The computational core of convolutional layers lies in the MAC operation. The amount of MAC per input image added up quickly from few thousands for LeNet network [33] up to tens billions for VGG16 network [20]. To meet this computation requirement, our accelerator embeds several MAC engines. Each of these works on an independent input filter, allowing the parallelization of the computation of the output activation map across multiple filters. Internally, each MAC is implemented with a set of multipliers and accumulators. Changing the number of those components directly affects the degree of parallelism.

B. Performance and Power Analysis

We ran our tests on a Xilinx Embedded FPGA (Zynq UltraScale+ XCZU9EG MPSoC) that combines both an ARM Cortex-A53 64 bits multi-core processor and FPGA fabric fabricated in 16 nm technology. Overall, it represents a state-of-the-art *embedded* platform for a fair evaluation between FPGA acceleration of deep-learning inference tasks and pure software execution. We implemented a customized CNN, CNN_s (*DUNE-CNN-01*), in C language as a reference for our performance and power analysis. Figure 5 provides an overview of CNN_s.

Table VIII summarizes the results. The inference time of our customized CNN_s for a single image is 0.0855 seconds when executed as software on the ARM Cortex-A53 CPU. The

TABLE VIII
PERFORMANCE AND POWER ANALYSIS RESULTS ON THE EMBEDDED FPGA (ZYNQ ULTRASCALE+ XCZU9EG MPSOC).

Platform	Model	Time (s)	Power (W)	Energy Efficiency (img/s/W)
ARM C-A53	CNN_s	0.0855	2.871	4.074
FPGA	CNN_s	0.0511	1.110	17.630

CPU runs at 1.2 GHz. The inference time of the same network when it leverages the FPGA-acceleration is 0.0511 seconds. The accelerator runs at 100MHz on the FPGA fabric. The total power for the processing system (CPU) and for the FPGA accelerator are 2.871 Watts and 1.110 Watts respectively, as reported in Vivado Power Analysis. The energy efficiency of the FPGA implementation is more than 4 times better than the embedded CPU.

V. VIABILITY OF DNN APPLICATION FOR DUNE DATA SELECTION

Our studies demonstrate that DNNs in general can meet trigger efficiency requirements for selecting off-beam rare events in the DUNE far detector. In addition, for several CNNs (e.g. VGG16b), sufficiently low fake trigger rates can be met, such that the required data reduction factor of 10^4 can be achieved for high energy triggering and potentially also for low energy triggering with a subsequent module-level data selection stage; the latter is the subject of future investigations.

For the case of an online data selection implementation where inference is carried out exclusively in GPUs, out of the four DNNs considered, CNN_s is identified as the most viable option for GPU deployment at the low-level data selection stage. We assume that the necessary pre-processing from preparing the ROIs, which consists of operations which are commonly done in FPGA, can keep up with the raw detector APA-frame rate, and consider only the inference stage latency for the purposes of this discussion. Given that the inference time for an ROI with CNN_s is comparable to the APA-frame length (2.25 ms), CNN_s should on average keep up with frame-by-frame selection, with each APA's frames processed in a separate GPU card; this, however, implies that a 200-fold parallelization would be needed (across 200 GPU cards) to facilitate low-level data selection for a 10 kton module; this is unfeasible given power restrictions underground at the far detector location. On the other hand, a factor of 50 reduction in required GPU processing would be possible if a pre-processing step were to be added to remove empty ROIs before the inference stage. Such a step would remove all but 2% of the background noise ROIs from the inference stage, allowing for, on average, 112.5 ms per ROI for inference. The same scheme would make VGG16b viable for online inference no GPU as well, which yields characteristically higher efficiency for all rare events of interest.

In the case of FPGA inference we find that a factor of four (4) increase in energy efficiency (img/s/W) is possible over a software implementation in CPU of the same (CNN_s) algorithm, motivating consideration of deployment of CNNs for low-level data selection on FPGA. The performance improvement over a software implementation is comparable for both inference speed (factor of 1.7) and power efficiency (factor of 2.6). Furthermore, we find that for smaller networks, such as for CNN_s, the resource allocation requirements for a full network implementation processing ROIs of 64×64 size are comparable with those available in state-of-the-art FPGAs, a desirable feature for simplified parallelization and for minimizing costs.

VI. SUMMARY

Acceleration of DNNs for real-time data selection is motivated by a number of up and coming high-resolution imaging particle detectors, in particular LArTPCs which work by imaging particle traces that are identifiable by their distinct topologies (spatial extent, shape, and pixel intensity) in two-dimensional view projections of three-dimensional detector regions. We have investigated the viability of DNN application for the purposes of real-time or online data selection (triggering) for such detectors, with a particular focus on the future DUNE experiment. Data selection is achieved by frame-by-frame classification of raw data streamed in channel vs. time space from 200 independent, self-contained regions of one of four DUNE far detector modules, assuming a single phase design.

Using simulated DUNE raw data images (APA-frames), we have found that such techniques yield promising results in terms of image classification accuracy, for a large variety (in terms of depth and size) of networks. Sufficiently high trigger efficiencies are achieved for selection of APA-frames with high energy interactions; lower trigger efficiencies are achieved for APA-frames with low energy interactions. However, supernova burst trigger efficiency can be optimized further by exploiting a higher-level decision which aggregates selected APA-frames over time, following the approach in [4].

We have further shown that latency and power considerations make the implementation of DNNs on GPUs for online inference viable for smaller networks and with significantly re-sized and down-selected ROI image inputs. Larger networks with re-sized full-frame information are viable only for the high level filter stage, at this time.

Finally, we have shown that implementation of DNNs on FPGAs for real-time inference at the low-level stage is promising, and have provided a viable path for development and optimization.

ACKNOWLEDGMENT

The authors thank J. Hewes, Y. Zhou, and S. Koo for early contributions to the development of CNNs and simulation tools used in the GPU studies, S. Rossi for a preliminary analysis of the FPGA implementation of the CNN, and K. Terao for valuable input and feedback to this work. This material is based upon work supported by the National Science Foundation under Grant No. PHY-1753228, and work supported in

part by the Research Initiatives in Science and Engineering (RISE) program of Columbia University.

REFERENCES

[1] R. Acciarri *et al.* [MicroBooNE Collaboration], "Design and construction of the MicroBooNE detector," JINST, vol. 12, no. 02, 2017.

[2] B. Abi *et al.* [DUNE Collaboration], "The Single-Phase ProtoDUNE technical design report," FERMILAB-DESIGN-2017-02.

[3] B. Abi *et al.* [DUNE Collaboration], "The DUNE far detector interim design report, Volume 2: Single-phase module," FERMILAB-DESIGN-2018-03.

[4] DUNE Collaboration, "Deep Underground Neutrino Experiment (DUNE) technical design report," in preparation.

[5] A. Radovic *et al.*, "Machine learning at the energy and intensity frontiers of particle physics," Nature, vol. 560, no. 7716, pages 41-48, 2018.

[6] C. Adams *et al.* [MicroBooNE Collaboration], "A deep neural network for pixel-level electromagnetic particle identification in the MicroBooNE liquid argon time projection chamber," FERMILAB-PUB-18-231-ND.

[7] R. Acciarri *et al.* [MicroBooNE Collaboration], "Convolutional neural networks applied to neutrino events in a liquid argon time projection chamber," JINST, vol. 12, no. 03, 2017.

[8] L. Domine and K. Terao, "Scalable deep convolutional neural networks for sparse, locally dense liquid argon time projection chamber Data," arXiv:1903.05663.

[9] V. Sze *et al.*, "Efficient processing of deep neural networks: A tutorial and survey," Proc. of the IEEE, vol. 105, no. 12, pages 2295-2329, 2017.

[10] B. Falsafi *et al.*, "FPGAs versus GPUs in data centers," IEEE Micro, vol. 37, no. 1, pages 60-72, Jan. 2017.

[11] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," Proc. Intl. Conf. on Multimedia, pages 675-678, Nov. 2014.

[12] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," In OSDI, vol. 16, pages 265-283, 2016.

[13] J. Duarte *et al.*, "FPGA-accelerated machine learning inference as a service for particle physics computing," arXiv:1904.08986 [physics.data-an].

[14] P. Coussy *et al.*, "An introduction to high-level synthesis," IEEE Design & Test of Computers, 26(4):8-17, 2009.

[15] Clark, D. H. and Stephenson, F. R., "The Historical Supernovae," Supernovae: A survey of current research; Proceedings of the Advanced Study Institute. Cambridge, England: Dordrecht, D. Reidel Publishing Co., pages 355-370, 1982.

[16] LArSoft release redmine 06.60.00 https://cdcvs.fnal.gov/redmine/projects/larsoft/wiki/ReleaseNotes066000.

[17] LArSoft release redmine 07.13.00 https://cdcvs.fnal.gov/redmine/projects/larsoft/wiki/ReleaseNotes071300.

[18] dunetpc release redmine 07.13.00 https://cdcvs.fnal.gov/redmine/projects/dunetpc

[19] MicroBooNE Collaboration, "The Continuous Readout Stream of the MicroBooNE Liquid Argon Time Projection Chamber for Detection of Supernova Neutrinos," MicroBooNE Public Note, MICROBOONE-NOTE-1030-PUB, 2019.

[20] Simonyan, Karen and Zisserman, Andrew, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556.

[21] G. Ge, Y. Jwa and G. Karagiorgi, "ML-based triggering for DUNE", DUNE docd-id-11311, 2018.

[22] Krizhevsky, Alex and Sutskever, Ilya and E. Hinton, Geoffrey, "ImageNet classification with deep convolutional neural networks," Neural Information Processing Systems, vol. 25, 2012.

[23] Frank Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," Cornell Aeronautical Laboratory, Psychological Review, vol. 65, No. 6, 1958.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," arXiv:1512.03385.

[25] Cota, Emilio G. *et al.*, "An analysis of accelerator coupling in heterogeneous architecture," ACM/EDAC/IEEE Design Automation Conference (DAC), 2015.

[26] Piccolboni, Luca *et al.*,"Broadening the exploration of the accelerator design space in embedded scalable platform," IEEE High Performance Extreme Computing Conference (HPEC), 2017.

[27] R. Nane *et al.*, "A survey and evaluation of FPGA high-level synthesis tools," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2016.

[28] A. Prost-Boucle *et al.*, "A fast and autonomous HLS methodology for hardware accelerator generation under resource constraints," In Euromicro Conf. on Digital System Design, pages 201-208, 2013.

[29] S. Rossi, "Accelerating convolutional neural networks with high-level synthesis," M.S. Thesis, Columbia University in the city of New York, New York (USA), 2018.

[30] X. Liu *et al.* "High level synthesis of complex applications: An H.264 video decoder," In Proc. Intl. Symp. on Field-Programmable Gate Arrays, pages 224-233, 2016.

[31] ARM AMBA, "AXI and ACE Protocol Specification," 2011.

[32] Borkar, Shekhar and Andrew A. Chien, "The future of microprocessors," Communications of the ACM, vol. 54, no. 5, 2011.

[33] LeCun, Yann *et al.* "Gradient-based learning applied to document recognitio," Proceedings of the IEEE, vol. 86, no. 11, 1998.

6.2 Real-time inference with 2D CNNs on FPGAs

In the following paper, 2D CNN optimization and implementation simulation on FPGAs are demonstrated using *hls4ml* tools, described in Sec. 4.3.2. Multiple 2D CNN architectures were explored in terms of classification accuracy and the number of learnable parameters. The hyperparameters for a chosen CNN architecture were scanned to achieve optimal accuracy. The CNNs, initially written in a programming language, were translated to logic synthesis for FPGAs via *hls4ml*. The performance of CNNs in logic synthesis was evaluated through implementation simulation. The usage of FPGA resources was also estimated, targeting the FPGA used for the DUNE FD readout unit. The viability of applying CNNs for real-time data selection for DUNE with respect to the DUNE DAQ specifications is discussed.

Real-time Inference with 2D Convolutional Neural Networks on Field Programmable Gate Arrays for High-rate Particle Imaging Detectors

Y. Jwa,[a,*] **G. Di Guglielmo,**[a] **L. Arnold,**[a] **L. Carloni,**[a] **G. Karagiorgi**[a]

[a] *Columbia University, New York, NY, 10027, USA*

E-mail: yj2429@columbia.edu

Abstract: We present a custom implementation of a 2D Convolutional Neural Network (CNN) as a viable application for real-time data selection in high-resolution and high-rate particle imaging detectors, making use of hardware acceleration in high-end Field Programmable Gate Arrays (FPGAs). To meet FPGA resource constraints, a two-layer CNN is optimized for accuracy and latency with KerasTuner, and network *quantization* is further used to minimize the computing resource utilization of the network. We use "High Level Synthesis for Machine Learning" (*hls4ml*) tools to test CNN deployment on a Xilinx UltraScale+ FPGA, which is an FPGA technology proposed for use in the front-end readout system of the future Deep Underground Neutrino Experiment (DUNE) particle detector. We evaluate network accuracy and estimate latency and hardware resource usage, and comment on the feasibility of applying CNNs for real-time data selection within the currently planned DUNE data acquisition system. This represents the first-ever exploration of employing 2D CNNs on FPGAs for DUNE.

Keywords: Hardware acceleration; Machine learning; Real-time machine learning; Fast machine learning; Neutrino detectors; Data acquisition; Trigger; Data selection; Data filter; Particle imaging; Liquid argon time projection chamber

[1]Corresponding author.

Contents

1 Introduction

Modern-day particle physics experiments produce vast amounts of data that must be processed to down-select interesting (and usually rare) signals for further physics study and scientific discovery. This process of data selection is applied across several stages of the data processing pipeline. In recent years, such pipelines have increasingly made use of deep learning (DL) [1, 2]. Additionally, as data rates grow, there is increased need to accurately and efficiently execute data selection in real time, i.e. at a rate commensurate with data generation throughput and with low latency, by employing "triggers". These are real-time data-driven decisions, which translate physical measures—quantities calculated based on the incoming data itself and/or other external signals—into instructions on which data to keep or permanently discard.

Driven in part by the need to increase accuracy in selecting high-dimensional and highly-detailed data from modern-day particle detectors, machine learning (ML) algorithms based on both supervised and unsupervised learning have been proposed and shown to be capable of effectively triggering on incoming physics data, proving to be a promising solution for the upcoming data challenges of future experiments. Implementing ML algorithms into dedicated hardware for triggering, such as GPUs, or FPGAs, can potentially guarantee fast execution of the algorithm while taking advantage of the algorithm's accuracy in selecting data of interest with maximal signal efficiency and signal purity. Additionally, software toolkit development projects such as *hls4ml* [3] are providing suitable and user-friendly frameworks for easily employing ML algorithms into hardware for application-specific usage (see, e.g. [4, 5]).

Further motivated by a widely used particle imaging detector technology—liquid argon time projection chambers (LArTPCs)—we explore the applicability of algorithms commonly used in

image analysis for LArTPC triggering purposes, following [6]. LArTPCs work by continuously imaging a large and homogeneous 3D detector volume, wherein electrically charged particles are visible through the trails of ionization they leave along their trajectories. This type of technology is employed in searches of rare events such as interactions of dark matter particles or supernova core-collapse neutrinos with the detector medium. More so than for other particle detector technologies, LArTPC data are well-suited for image analysis given that neutrino or other rare event signals are translationally invariant within a generally sparse 2D or 3D image of the detector volume. In past work [6], we have shown that 2D convolutional neural networks (CNNs) tested on simulated raw data from a LArTPC can yield sufficient accuracy and can be implemented onto parallelized data processing pipelines using GPUs to perform data selection in a straightforward way, while meeting the physics performance and latency requirements of future LArTPC experiments.

The need to improve the long-term operation reliability and power utilization of such data processing pipelines motivates the exploration of alternate implementations of CNN-based data selection, specifically implementations on Field Programmable Gate Arrays (FPGAs). FPGAs are low-power digital microelectronics devices commonly used for signal processing and data acquisition applications. They are commonly used in front-end readout electronics systems for particle physics experiments; their on-device nature (often capable of receiving the full-rate of detector-generated data prior to any data filtering or reduction) and their reliability for long-term operation make them attractive for data processing algorithm implementation, especially if only minor pre-processing is necessary in the data pipeline. In general, algorithm implementation into a front-end device is advantageous as it makes large data movement unnecessary, reduces power consumption and trigger latency, and increases reliability. More recently, there has been a growing interest in using FPGAs as accelerators for deep neural networks [7].

A number of ML algorithms have already been explored for particle triggering and suitability for FPGA applications; see, e.g. [8–16]. Explored algorithm implementations range from Artificial Neural Networks, to Boosted Decision Trees, Graph Neural Networks, to Autoencoders, etc. In this paper, we investigate, for the first time, the implementation of a relatively small 2D CNN onto an FPGA, targeted for use in the front-end readout electronics of the future Deep Underground Neutrino Experiment (DUNE) [17–20], motivated by previous exploration and findings in [6]. While the use of CNNs for image classification applications has been established for well over a decade [21], their specific use in fast-inference applications in particle physics has been restricted to non-LArTPC applications [22, 23]. On the other hand, in the case of LArTPCs, CNNs have been used successfully for *offline* data analysis and physics measurements (see, e.g. [24–29]). Keeping in mind the 2D nature and high resolution of LArTPC raw data, we explore and evaluate techniques to reduce the computational resource usage of CNN inference on FPGAs. We focus on the DUNE case, and show that we can meet the technical specifications of the DUNE readout system, while still satisfying the physics accuracy requirements of the experiment. We add that other DL algorithms have also been studied for offline data analysis of LArTPC data [30, 31], and would also be worth exploring for FPGA implementation for LArTPC trigger applications.

In Sec. 2, we describe the DUNE Far Detector (FD) LArTPC in more detail, including its operating principle, and the technical specifications and requirements of its readout and data selection (trigger) system. In Sec. 3 we explore different CNN architectures, and explore their accuracy in selecting data containing rare signal events, paying attention to the overall size of the network,

in anticipation of minimal computational resource availability in the DUNE FD readout system. Subsection 3.1 describes how simulated raw data from the DUNE FD are prepared as input to the CNN; Subsec. 3.2 describes some CNN architectures and the classification accuracy performance on simulated input images; in Subsec. 3.3, we further optimize the network architecture and hyper-parameters in an automated way, using the KerasTuner package [32, 33], and compare classification accuracy of the automatically optimized network to the non-optimized ones. Throughout all sub-sections, we also present network accuracy results using "HLS-simulated" versions of the CNNs, produced using the *hls4ml* package [3]. One key feature of *hls4ml* is a reduction in accuracy due to quantization of the network, which we avoid by employing quantization-aware training, following [34, 35], as discussed in Subsec. 3.4. Finally, in Sec. 4, we provide estimates of FPGA resource usage of the optimized networks (with and without quantization-aware training), using an *hls4ml* synthesized design for a targeted FPGA hardware implementation. We demonstrate that the use of 2D CNNs for real-time data selection in the future DUNE is viable, and advantageous, given the currently envisioned front-end readout system design.

2 Application Case: Real-time Data Selection for the Future DUNE LArTPC

LArTPCs are a state-of-the-art charged-particle detector technology with broad applications in the field of particle physics, astro-particle physics, nuclear physics, and beyond. This high-rate imaging detector technology has been adopted by multiple particle physics experiments, including the current MicroBooNE experiment [36], two additional detectors that are part of the upcoming Short-Baseline Neutrino (SBN) program [37], as well as the next-generation DUNE experiment [17–20], and it is also proposed for future-generation astro-particle physics experiments such as GRAMS [38]. LArTPCs work by imaging ionization electrons produced along the paths of charged particles, as they travel through a large (multiple cubic meters) volume of liquid argon. Charged particle ionization trails drift uniformly toward sensor arrays with the use of a uniform electric field applied throughout the liquid argon volume, and are subsequently read out in digital format as part of 2D projected views of the 3D argon volume. This is illustrated in Fig. 1. Densely packed sensor arrays sample the drifted ionization charge at a high rate, typically using a 12-bit, 2 MHz Analog to Digital Converter (ADC) system recording the amount of ionization charge per sensor per time-sample, thus imaging charge deposition across 2D projections of the argon volume with millimeter-scale resolution. Typically, digitized image frames of O(10) megabytes each are streamed out of these detectors in real time and at a rate of up to hundreds of thousands of frames per second, amounting to raw data rates of multiple gigabytes to several terabytes (TB) per second.

The future DUNE experiment represents a special case, with the most stringent data processing requirements among all currently running or planned LArTPC experiments. DUNE consists of a near and a far detector complex, which will be located at Fermi National Accelerator Laboratory (Fermilab) in Illinois and at the Sanford Underground Research Facility (SURF) in South Dakota, respectively. The far detector (FD) complex will be located 1 mile deep under ground, and will comprise the largest LArTPC ever to be constructed, with an anticipated raw data rate for its first of four LArTPC modules of 1.175 TB/s. This first detector module will be operated continually, and for at least ten years, with subsequent modules coming online before the end of the current decade. The DUNE FD will therefore be constructed with a readout and data selection system that

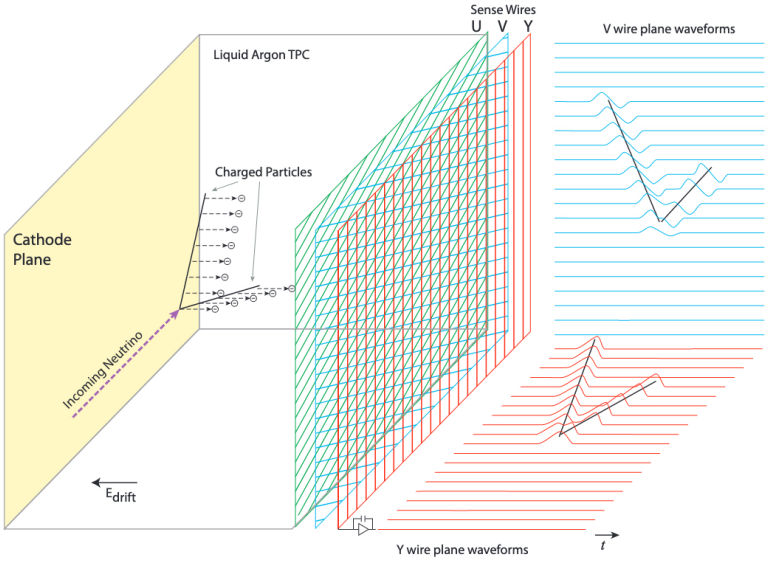


Figure 1. Operating principle of a LArTPC. The ionization electrons are drifted toward sensor arrays, e.g. planes of ionization charge sensor wires. Each wire is connected to an analog amplifier/shaper, followed by an ADC, and its resulting digital waveform is read out continually. Waveforms of adjacent wires appended together form 2D images. Image credit: [36].

is required to receive and process an overall raw data rate of 4×1.175 TB/s, achieve a factor of 10^4 data reduction, and maintain > 99% efficiency to particle interactions of interest that are predicted to be as rare as once per century [19].

The scientific goals of DUNE include, but are not limited to, observing neutrinos from rare (once per century) galactic supernova bursts (SNBs) [20, 39], searching for rare baryon number violation processes such as argon-bound proton decay and argon-bound neutron-antineutron oscillation, and studying interactions of neutrinos that are produced in cosmic ray air showers in the Earth's atmosphere [20, 40]. From the data acquisition (DAQ) and data selection (trigger) point of view, these rare physics searches and in particular the requirement to be > 99% efficient to a galactic SNBs with a less than once per month false positive SNB detection rate, cast particularly stringent technical requirements.

More specifically, in order to select these "events", which take place randomly and unpredictably, the DUNE DAQ and trigger system must scan *all* detector data continuously and with zero dead time, and identify rare physics signatures of interest in a "self-triggering" mode—without relying on any external signals prompting data readout. Furthermore, a self-triggering scheme reaching nearly perfect (100%) efficiency for rare physics events is needed in order for DUNE to achieve its full physics reach. This further requires temporarily buffering large amounts of data while this processing takes place. In the case of DUNE, buffering constraints translate into a sub-second latency requirement for the trigger decision. Additionally, the trigger decision needs to achieve an overall 10^4 data rate reduction, and with high signal selection efficiency, corresponding to an

average of >60% efficiency on individual supernova neutrino interactions, and >90% efficiency to other rare interactions including atmospheric neutrino interactions and baryon number violating events.

The first DUNE FD module will image charged particle trajectories within 200 independent but contiguous liquid argon volume regions ("cells"). Charged particle trajectories within each cell will be read out by sensor wires arranged in three planes: one charge-collection wire planes, plus two charge-induction wire planes. Each plane readout corresponds to a particular 2D projected view of the 3D cell volume, and the combination of induction and collection plane information allows for 3D stereoscopic imaging and reconstruction of a given interaction within the 3D cell volume. In total, the first FD module will consist of 384,000 wire sensors, each read out independently; this outnumbers current LArTPC neutrino experiments by more than a factor of 500 (e.g., MicroBooNE makes use of 8,256 wire sensors). For this work, we focus exclusively on charge-collection wire readout. Charge-collection wires give rise to signals which are unipolar in nature (as opposed to charge-induced signals, which are bipolar in nature, and therefore susceptible to cancellation effects). As such, charge-collection readout waveforms preserve sensitivity to charge deposition even for extended charge distributions. Since particle identification (and subsequent data selection decision making) relies on quantifying the amount of charge deposition per unit length of a charged particle track, charge-collection waveform information is anticipated to provide better particle identification performance.

The 200 cells of the first DUNE FD module will be read out in parallel, by 75 "upstream DAQ" readout units. Each unit makes use of a Front-End LInk eXchange (FELIX) PCIe 3.0 card [19, 41] holding a Xilinx UltraScale+ FPGA to read out digitized waveforms, and pre-process the data. In the nominal DUNE readout unit design, the FPGA processes continuous waveforms in order to perform noise filtering and hit-finding; hit-finding summaries are then sent for additional processing to a FELIX-host CPU system, in order to form trigger candidates (particle interaction candidates); the latter inform a subsequent module-wide trigger decision. An alternate potential solution, and the scope of this work, is to apply more advanced data processing and triggering algorithms within the available FPGA resources on-board the FELIX card, such as CNNs, which can intelligently classify a collection of waveforms representing activity across the entire cell volume in real time. This would eliminate the need of subsequent CPU host (or GPU) processing, potentially increase trigger efficiency and purity (through the use of more intelligent algorithms), and potentially further minimize power consumption needs. It is worth noting that, since most interactions of interest have a spatial extent which is smaller than the cell volume, a per-cell parallelization of triggering algorithms is appropriate, and it is therefore sufficient to focus trigger studies to a per-cell level, ignoring cell volume boundary effects.

3 CNN Design and Optimization for Real-time LArTPC Data Selection

In recent years, DL algorithms such as CNNs have been shown to achieve very high signal selection efficiencies when employed in offline physics analyses of LArTPC data. MicroBooNE is leading the development and application of DL techniques, including CNNs, for LArTPC data reconstruction [24–27], and CNN-based analyses and DL-based reconstruction are actively being developed for SBN and for DUNE [28, 29].

In a previous study [6], we have also shown that sufficiently high efficiencies can be reached by processing raw collection plane data from any given DUNE FD cell, prior to removing any detector effects or applying data reconstruction. As such, we proposed a CNN-based triggering scheme using streaming raw 2D image frames, whereby the images are pre-processed, downsized, and run through CNN inference to select ones containing SNB neutrino interactions or other rare interactions of interest on a frame-by-frame basis. The data pre-processing and CNN-based selection method demonstrated that target signal selection efficiency while reaching the needed 10^4 background rejection could be achieved, given sufficient parallelization in GPUs. As the DUNE FD DAQ and trigger design is subject to stringent power limitations and limited accessibility in the underground detector cavern, a particularly attractive option is to fully implement this pre-processing and CNN-based inference on FPGAs, in particular ones that will be part of the DUNE upstream DAQ readout unit design. We examine the viability of this option in this work.

Specifically, we explore the accuracy of relatively small CNNs in classifying streaming DUNE FD LArTPC cell data, and proceed to employ network optimization in an effort to reduce its computational resource footprint while preserving network accuracy. The following subsections describe the CNN input image preparation (Sec. 3.1), CNN performance without (Sec. 3.2) and with (Sec. 3.3) network optimization, and with quantization-aware training (Sec. 3.4).

3.1 Input Image Pre-processing

Because of the parallelism in the DUNE FD DAQ and trigger design, we only consider a single cell's worth of data at a time, and focus exclusively on raw collection plane waveforms. Following [6], collection plane waveforms for a single cell in the DUNE FD are simulated in the LArSoft framework [42, 43], using the default configuration of the *dunetpc* software, and using an enhanced electronics noise level configuration, to be conservative. Besides electronics noise, the simulation includes radiological impurity background interactions that are intrinsic to the liquid argon volume. The radiological background interactions (predominantly from ^{39}Ar decay) are expected to occur at a rate of 10^7 Hz per FD module, and they are considered as likely backgrounds particularly to supernova neutrino interactions. Signal waveforms from interactions of interest, including low-energy supernova neutrino interactions or other high-energy interactions (proton decay, neutron-antineutron oscillation, atmospheric neutrino interactions, cosmogenic background interactions), are overlaid on top of intrinsic radiological background and electronics noise waveforms.

Given the physical dimension of a cell along the ionization charge drift direction, and the known ionization charge drift velocity, 2.25 ms worth of continuous data from the entire collection plane represents a 2D image exposure of the full cell volume. As such, we define a 2D image in terms of 480 collection plane wire channels spanning the length of the cell volume, times the 2.25 ms drift direction sampled at 2 MHz (4488 samples) spanning the width of the cell volume. This corresponds to a 2.1 megapixel image, with 12-bit ADC resolution governing the range of pixel values, dictating the amount of ionization charge collected by each wire, and indicating the energy deposit within the 3D volume across the given 2D projection.

For network training purposes, the 2.1 megapixel input images are labeled as containing either electronics noise and radiological background only (NB), or low-energy supernova neutrino interactions (LE), or high-energy interactions (HE), each superimposed with electronics noise and radiological background, according to the simulation truth. Figure 2 shows example input 2D images

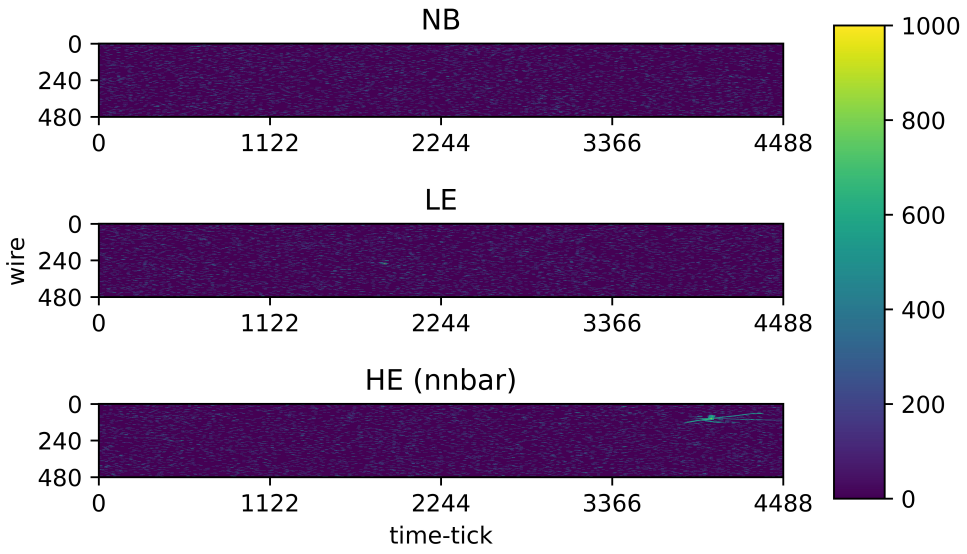


Figure 2. Examples of 2D images formed from one full drift (2.25ms) of 480 collection plane wires in one DUNE FD cell. Top: Image containing electronics noise and radiological background only (NB). Middle: Image containing one low-energy supernova neutrino interaction (LE) superimposed with electronics noise and radiological background. Bottom: Image containing one high-energy interaction (HE), specifically from neutron-antineutron oscillation (nnbar), superimposed with electronics noise and radiological background. The color z axis indicates the wire waveform ADC value (simulated with 12-bit resolution). These images are pre-processed prior to CNN processing.

before pre-processing steps. We note the sparsity of these images, mostly containing uniformly distributed low-energy activity from noise and radiological backgrounds. While it is possible to train a CNN with 2.1 megapixel images, it is not memory-efficient, and it may furthermore not be an efficient way to propel a CNN to learn the different features between the three event classes (NB, LE, and HE). Following [6], we adopt pre-processing steps that include de-noising (zero-supression), cropping around the region-of-interest (ROI), and resizing the ROI through down- or up-sampling. The de-noising step uses a configurable threshold for the pixel ADC value and zero-suppresses pixel values below this threshold; a threshold of 520 ADC (absolute scale) was used in these studies, where $\sim$ 500 ADC represents the baseline. ROI cropping was performed by finding a contiguous rectangular region containing pixels with values over 560 ADC. The most extreme image coordinates (smallest and largest channel number, as well as smallest and largest time tick) with pixel values greater than the lower threshold of 560 ADC were used to determine the ROI boundaries. Once an ROI was found, the ROI region was resized (through up-sampling or down-sampling) to occupy exactly 64×64 pixels, as shown in Fig. 3. Resulting image pixel values were then re-normalized to a range between 0-1 prior to CNN processing.

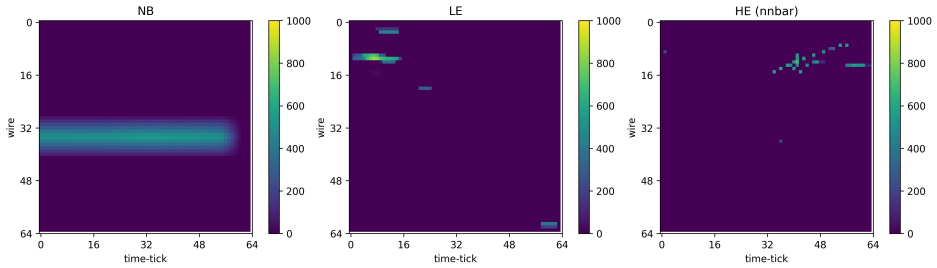


Figure 3. Example ROIs formed after pre-processing. Left: Image containing electronics noise and radiological background only (NB). Middle: Image containing one low-energy supernova neutrino interaction (LE) superimposed with electronics noise and radiological background. Right: Image containing one high-energy interaction (HE), specifically from neutron-antineutron oscillation (nnbar), superimposed with electronics noise and radiological background. The color z axis indicates ADC values after down- or up-sampling. These images are renormalized with pixel values ranging from 0-1 prior to inputting to a CNN for subsequent processing (data selection).

Resized ROIs were generated for each of the three categories indicated in Tab. 1, with comparable statistics, and used for network training and testing for all studies presented in the subsequent sections. As we are investigating the viability of CNN implementation on FPGAs, computational resource utilization by the CNN is a key concern, and we therefore begin by investigating performance for relatively small-sized CNNs. For this task, we are working with a relatively small data set, split into training (60%), validation (20%) and test set (20%). Table 1 shows the statistics for only the training and testing data sets.

Table 1. Number of ROIs, according to truth label, used for training and testing of CNNs. Another statistically independent sample of images with similar statistics to the test sample was used for validation purposes during training. A total of 45,624 ROIs were used in the study.

	Label: NB	Label: LE	Label: HE
Training set size	12,023	12,050	10,137
Testing set size	4,027	3,970	3,417

The overall data processing and data selection scheme proposed and examined in this study is summarized in Fig. 4.

3.2 Performance of CNN-based Data Selection

Targeting FPGA implementations, we designed and tested custom CNN architectures with only one or two convolutional layers: **CNN01**, **CNN02**, and a downsized version of the latter, **CNN02-DS**. These networks have far simpler architectures than some of the more popular CNN architectures commonly used in image classification tasks (e.g. VGG [44] or ResNet [45] network architectures), by design, as they are targeted for implementation in computational-resource-constrained systems.

The network architecture for **CNN01** is shown in Fig. 5. **CNN01** has one convolutional layer, with convolutional width kernel dimension (3,3,32), and one max-pooling layer. One fully

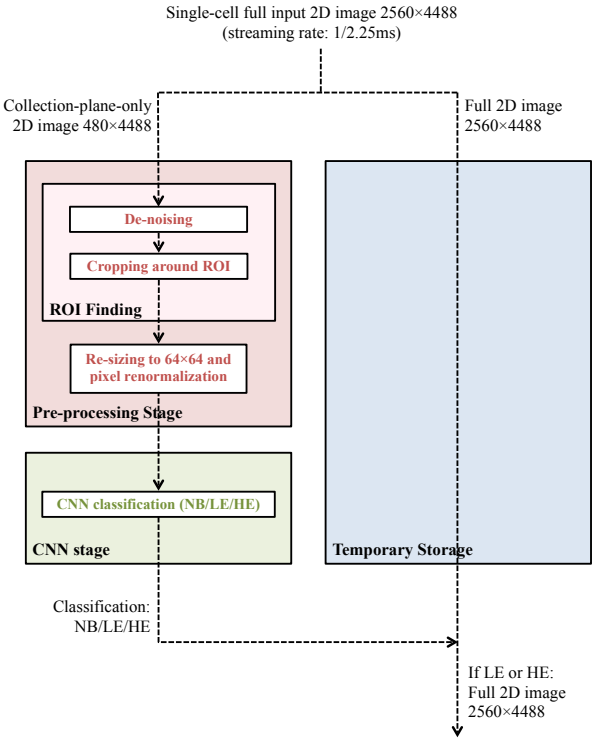


Figure 4. The data processing and data selection scheme under study for potential implementation in the upstream DAQ readout units of the future DUNE FD. The streaming 2D input images contain, > 99.9% of the time, NB data. This overall scheme should select true HE and LE images with > 90% accuracy, and true NB images with > 99.99% accuracy, in order to meet the DUNE FD physics requirements. Additionally, the pre-processing and CNN inference algorithms should meet the computational resources of the DUNE FD upstream DAQ readout units, and the algorithm execution latency should meet the data throughput requirements of the experiment.

connected layer follows at the end. In contrast, **CNN02** has two convolutional layers, and one max-pooling layer after each convolution. Also, here, one fully connected layer follows at the end. Finally, **CNN02-DS** is a downsized version of **CNN02**, where the convolution depth is significantly reduced. All three custom network architectures are summarized in Tab. 2.

Table 3 shows the classification performance of the three networks, for a GPU or CPU implementation using Keras [46]. The performance of these three networks is comparable. For all three networks, the false positive identification rates (which affect data reduction capability) are comparable, and the (correct) classification accuracy is over 99% for NB labeled ROIs, over 93% for LE labeled ROIs, and over 90% for HE labeled ROIs. Despite the difference in architecture (one vs. two convolution layers) and number of trainable parameters, no clear impact on classification performance is observed.

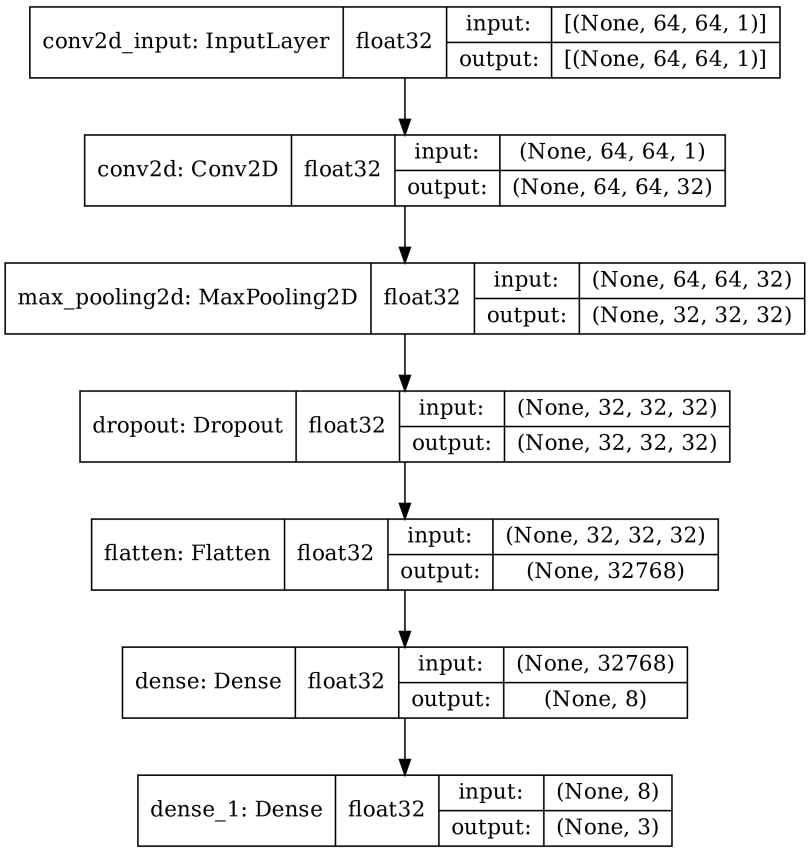


Figure 5. Network architecture of **CNN01**.

Table 2. Summary of explored CNN architectures.

	CNN01	**CNN02**	**CNN02-DS**
Number of convolution layers	1	2	2
Convolution kernel dimension (first conv.)	3×3×32	3×3×32	3×3×4
Convolution kernel dimension (second conv.)	N/A	3×3×64	3×3×8
Number of max-pooling layers	1	2	2
Max-pooling dimension (first max-pool)	2×2	2×2	4×4
Max-pooling dimension (second max-pool)	N/A	2×2	4×4
Number of trainable parameters	262,499	149,923	1,395

While accuracy results meet signal efficiency requirements[1], the high false positive rate (in

[1]In this study, accuracy is defined identically to signal efficiency, i.e. as a true positive classification rate given a set of true labels.

Table 3. Classification accuracy comparison for **CNN01**, **CNN02**, and **CNN02-DS** on GPU or CPU.

CNN01	NB	LE	HE
True NB	99.4%	0.55%	0%
True LE	3.8%	94.2%	1.9%
True HE	3.4%	6.1%	90.5%
CNN02	NB	LE	HE
True NB	99.5%	0.50%	0%
True LE	4.0%	93.2%	2.8%
True HE	3.2%	6.6%	90.2%
CNN02-DS	NB	LE	HE
True NB	99.5%	0.52%	0%
True LE	3.7%	94.4%	1.9%
True HE	3.0%	6.5%	90.5%

particular for true NB ROIs to be mis-classified as LE events at a rate of 0.5%) suggests a steady-state data reduction factor for a frame-by-frame data selection implementation that is a factor of 50 lower than the required reduction factor of 10^4. This is because the overwhelming majority (>99.9%) of the streaming ROIs in DUNE are expected to be truly NB ROIs, and therefore a 0.5% mis-classification rate would result in approximately one in 200 ROIs being (falsely) selected, as opposed to the targeted one in 10,000. Additional data reduction, however, can be provided by an ROI pre-selection stage, as motivated in [6]; specifically, approximately only one in 50 2D true NB images are expected to be non-empty after ROI finding (see Fig. 4) and therefore 98% of the ROIs can be discarded prior to CNN processing.[2] This suggests that an overall factor of 10^4 is achievable.

In this work, the ML models were trained and tested on GPUs with single-precision floating-point arithmetic (standard IEEE 754), and then post-training quantization (PTQ) was performed with the aim of running ML inference on FPGA. It is worth noting that FPGAs support integer, floating-point, and fixed-point arithmetic. An FPGA implementation may require orders of magnitude higher resources, besides higher latency and power costs, when compared with a finely-tuned fixed-point implementation of the same algorithm [47]. Predictably, PTQ impacts ML classification performance, although the profiling tools in *hls4ml* help the designer decide the appropriate model precision [48]. The resulting accuracy values for PTQ networks targeted for FPGA (with fixed-point precision) are shown in Tab. 4, and contrasted to those with floating-point precision in Tab. 5. We adopted quantization-aware training (QAT) to address this accuracy drop, as discussed in Sec. 3.4.

[2]Note that the CNN studies presented in this paper are performed exclusively on non-empty ROIs. For images containing LE and HE events, ROI-finding does not cause any additional reduction in efficiency, and the ROI classification accuracy represents the signal efficiency. For images containing only NB, only one in approximately 50 images is kept after ROI-finding.

Table 4. Classification accuracy comparison for **CNN01**, **CNN02**, and **CNN02-DS**, using post-training quantization (PTQ).

CNN01	NB	LE	HE
True NB	98.1%	1.8%	0.02%
True LE	6.6%	89.1%	4.3%
True HE	19.4%	37.7%	42.9%
CNN02	NB	LE	HE
True NB	98.1%	0.25%	1.9%
True LE	22.8%	10.6%	66.6%
True HE	21.5%	3.7%	74.7%
CNN02-DS	NB	LE	HE
True NB	99.5%	0.47%	0%
True LE	4.9%	93.1%	1.9%
True HE	21.2%	40.1%	38.7%

Table 5. Combined classification accuracy for true NB, LE, and HE ROIs for floating-point vs. PTQ fixed-point implementations of the trained networks. The combined classification accuracy is evaluated collectively on all of the testing set ROIs in Tab. 1, combined.

	CNN01	**CNN02**	**CNN02-DS**
Floating-point accuracy	94.9%	94.5%	95.0%
Fixed-point accuracy (PTQ)	78.5%	60.7%	79.1%

3.3 Automatized CNN Hyperparameter Optimization using KerasTuner

In the initial network performance comparison presented in Sec. 3.2, the classification performance does not appear to be highly sensitive to the network architecture and number of trainable parameters. In general, the choice of network hyperparameters such as the dimensions of hidden layers, and learning parameters, changes the number of trainable variables. Thus, the quality of training can be modulated by tuning the hyperparameters using the training and validation samples. This can be cumbersome to optimize, but further optimization of networks with respect to a large phase-space of hyperparameters can be performed methodically and in an automated way using open-source tools such as KerasTuner [32, 33].

We used KerasTuner for hyperparameter optimization for the baseline network architecture **CNN02-DS**. The scanning range and granularity of the hyperparameters explored is shown in Tab. 6. A total of twenty combinations were randomly sampled from the hyperparameter scanning region. The optimized network **CNN02-DS-OP** with the (marginally) highest classification accuracy, found at 95.22%, corresponds to a network with a first convolution depth of 8, second convolution depth of 16, dense layer size of 12 , and learning rate of 2.9×10^{-3}.

Table 6. Scanning range and granularity of the hyperparameters explored during automated network optimization using KerasTuner.

Hyperparameter	Range	Default Value
First convolution depth (conv1)	[4, 8, 16]	4
Second convolution depth (conv2)	[8, 16, 32]	8
Dense layer size (fc)	[8, 12, 16, 20, 24]	12
Learning rate (lr), logarithmic sampling	$[2 \times 10^{-4}, 2 \times 10^{-2}]$	2×10^{-3}

Table 7. Classification accuracy for the five top-performing and default (**CNN02-DS**) hyperparameter configurations. Note that the default accuracy obtained during hyperparameter optimization slightly differs from that in Tab. 5, due to differences in (random) initialization of the network weights before training, and randomness during the training.

	conv1	conv2	fc	lr	Accuracy
First-best (**CNN02-DS-OP**)	8	16	12	2.9×10^{-3}	95.22%
Second-best	16	32	12	4.9×10^{-4}	95.21%
Third-best	4	16	20	6.0×10^{-4}	95.21%
Fourth-best	16	8	16	7.0×10^{-4}	95.19%
Fifth-best	16	8	12	1.9×10^{-3}	95.19%
Default	4	8	12	2×10^{-3}	95.09%

3.4 Network Quantization in CNN-based Data Selection

The cost reduction and performance improvement of fixed-point arithmetic with HLS is highly encouraged when designing ML algorithms for FPGA deployment. Typically, when a trained network within an ML framework (e.g. Keras) on CPU or GPU is translated to HLS, the floating-point precision is reduced to the fixed-point precision of a given configuration. As a consequence, generally, network quantization resulting from fixed-point precision effectively reduces the precision of the calculations for weights, bias, and/or inputs, resulting in lower inference accuracy performance than what would otherwise be possible with floating-point precision. This is evident in Tab. 5.

In principle, one cannot achieve the flexibility and accuracy of a floating-point precision with any fixed-point representation. However, if accuracy can be maintained with an optimized choice of fixed-point precision, one can benefit from the inherent advantage of reduced computing resource utilization. Maintaining of accuracy therefore can be achieved with quantization-aware network training [34, 35, 49–51].

Quantization-aware training (QAT), achieved by committing calculations in ML algorithms with already-reduced fixed-point representation as part of network training, can prevent reduction in inference accuracy. The QKeras package [52] supports quantization-aware training by quantizing any given network using Qlayers. The quantized network derived from a given network architecture can be constructed by replacing the layers in the initial network to Qlayers. We refer to the

quantized version of **CNN02-DS-OP** obtained with QKeras as **Q-CNN02-DS-OP**. The precision configuration of **Q-CNN02-DS-OP** is shown in Fig. 6. The precision configuration of the reference **CNN02-DS-OP** is shown is shown in Fig. 7.

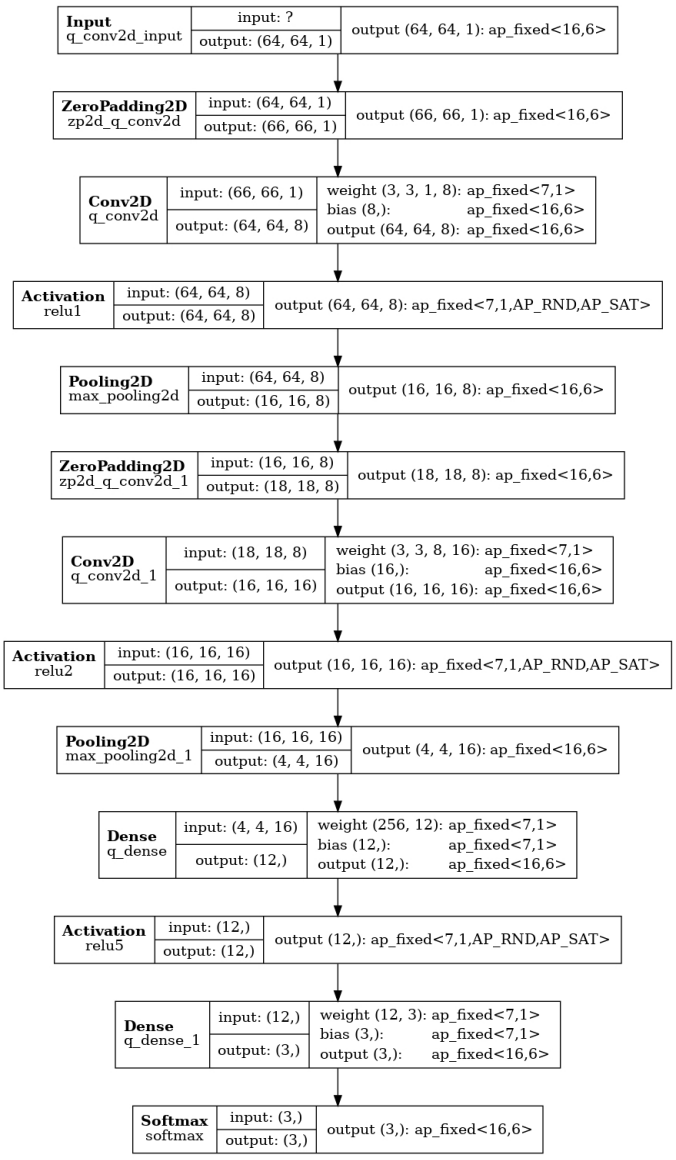


Figure 6. Precision configuration of layers in **Q-CNN02-DS-OP**. The precision configuration of the reference **CNN02-DS-OP** can be found in Fig. 7. Note that FPGA resource utilization is generally reduced with smaller ap_fixed values.

The classification results obtained using the reference **CNN02-DS-OP** network with and without PTQ are shown in Tab. 8; the corresponding results obtained with the quantization-aware trained

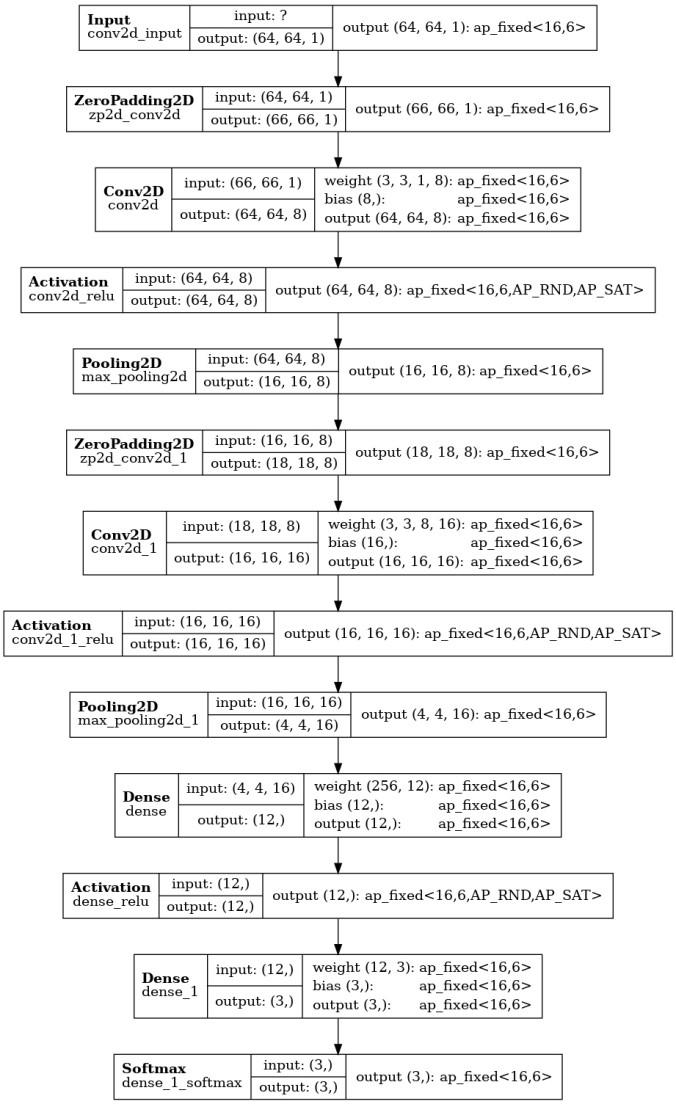


Figure 7. Precision configuration of layers in the reference **CNN02-DS-OP**.

(QAT) **Q-CNN02-DS-OP** are shown in Tab. 9.

For the network trained without QAT, **CNN02-DS-OP**, the overall classification accuracy for the entire testing sample (superset of three truth labels) drops significantly with PTQ, from 95.4% to 72.4%. For the network trained with QAT, **Q-CNN02-DS-OP**, however, the overall classification accuracy is maintained for what would be an equivalent FPGA implementation (with PTQ), at 95.2% and 95.2%. This demonstrates that a relatively small CNN, applied on a frame-by-frame basis, and trained with quantization that is consistent with FPGA fixed-point precision, can achieve the accuracy (signal efficiency and target data reduction factor) required for the DUNE FD.

Table 8. Optimized performance for **CNN02-DS-OP**, without quantization-aware network training. The difference in total accuracy for the floating-point case compared to that reported in Tab. 7 is due to retraining.

Floating-point	NB	LE	HE
True NB	99.5%	0.50%	0%
True LE	3.5%	95.1%	1.4%
True HE	2.9%	6.1%	91.0%
Total Accuracy		95.4%	
Fixed-point (PTQ)	NB	LE	HE
True NB	99.8%	0.17%	0%
True LE	6.3%	88.9%	4.8%
True HE	26.9%	52.2%	20.9%
Total Accuracy		72.4%	

Table 9. Optimized performance for **Q-CNN02-DS-OP**, with quantization-aware network training.

Floating-point (QAT)	NB	LE	HE
True NB	99.6%	0.40%	0%
True LE	3.8%	94.0%	2.2%
True HE	3.2%	5.4%	91.4%
Total Accuracy		95.2%	
Fixed-point (QAT)	NB	LE	HE
True NB	99.7%	0.32%	0%
True LE	3.9%	94.7%	1.4%
True HE	3.2%	6.4%	90.4%
Total Accuracy		95.2%	

4 Estimation of FPGA Resource Usage

In this section, we estimate FPGA resource usage and examine whether a Xilinx Virtex-7 UltraScale+ FPGA can accommodate a pre-trained CNN that meets the accuracy as well as resource and latency specifications of the DUNE FD DAQ and trigger system.

The estimated hardware usage for the quantized inference block of each of the optimized CNNs (**Q-CNN02-DS-OP** and **CNN02-DS-OP**) from Vivado HLS is shown in Tab. 10. The hardware usage of the discussed inference shows that the target FPGA, a high-end device, is well fit for implementing either the **Q-CNN02-DS-OP** or the **CNN02-DS-OP** network. As expected, the **Q-CNN02-DS-OP** network uses significantly lower FPGA resources. It is worth noting that, in addition to using more resources, **CNN02-DS-OP** (PTQ) has a lower accuracy than **Q-CNN02-DS-OP** (QAT), at 72.4% vs. 95.2%, illustrating the advantages of QAT.

Assuming a clock cycle of 5.00 ns, we find that the design is expected to meet timing requirements, with an inference latency of 4680 clock-cycles, corresponding to 23.4 μs. This is

well below the exposure time corresponding to a single input image of 2.25 ms; thus, assuming sufficient parallelization (i.e. at least two input 2D images processed in parallel by each FELIX unit), frame-by-frame real-time data selection based on collection plane-only image analysis with CNNs is a viable solution for the DUNE FD. Note that this does not consider additional resource utilization or latency associated with image pre-processing (ROI finding and down-sizing).

Table 10. Estimated resource utilization from Vivado HLS for CNN inference on a Xilinx UltraScale+ (XCKU115) FPGA. *Block RAM* refers to these types of memory elements, digital signal processors (DSPs) are elements dedicated to fast operations in signal processing (such as floating-point multiplication), Flip Flops and Look-up tables are standard

	Block RAM	**DSP Units**	**Flip Flops**	**Look-up Tables**
Available	4320	5520	1326720	663360
CNN02-DS-OP (PQT)	331 (7%)	4309 (78%)	226982 (17%)	163460 (24%)
Q-CNN02-DS-OP (QAT)	187 (4%)	2106 (38%)	142128 (10%)	138715 (20%)

We note that, in the current stage, the ML-based FPGA design has been synthesized, but it has not been implemented yet into the hardware; this is the focus of continuing development efforts.

5 Summary

In recent years, ML algorithms such as CNNs have shown tremendous growth of their use in high energy physics, including physics analysis with LArTPCs [1, 2]. In particular, CNNs have been shown to achieve very high signal selection efficiencies especially when employed in offline physics analyses of LArTPC data. MicroBooNE is leading the development and application of ML techniques, including CNNs, for event reconstruction and physics analysis as an operating LArTPC [24–27], and CNN-based analyses and ML-based reconstruction are actively being developed for SBN and for DUNE [28, 29].

Motivated by a previous study [6], showing that CNN-based data selection for LArTPC detectors can yield excellent accuracy even when applied solely at raw collection plane data, we have proposed a real-time, 2D CNN-based, frame-by-frame data selection scheme that is found to be a viable solution for the DUNE FD DAQ and trigger system. Leveraging the extensive parallelization and FPGA resources available within the DUNE FD upstream DAQ readout design, in this proposed scheme, 2D image frames streamed at a total rate of 1.25 TB/s are pre-processed and run through hardware-accelerated CNN inference to classify and select interactions of interest on a frame-by-frame basis. The proposed pre-processing and CNN-based selection method yield target signal selection efficiencies that meet the DUNE FD physics requirements, while also providing the needed 10^4 factor of overall data rate reduction.

The FPGA resource utilization for the CNN inference has been optimized with automatized network optimization and with quantization-aware training so as to avoid accuracy loss due to a fixed-point precision implementation in FPGA. The resulting optimized and quantized CNN (**Q-CNN02-DS-OP**) has been shown to fit within available DUNE FD upstream DAQ readout FPGA

resources, and to be executable with sufficiently low latency such that the need for significant buffering resources in the DUNE FD upstream DAQ system can also be relaxed. We note, however, that the pre-processing resource requirements and latency have not been explicitly evaluated, and this will be the subject of future work, as they need to be considered in tandem with the proposed CNN algorithm and implementation.

The findings further motivate future LArTPC readout designs that preserve the physical mapping of readout channels to a contiguous interaction volume as much as possible, in order to minimize pre-processing needs, and preserve spatial correlations that exist within 2D projected views of the interaction volume. Additionally, they motivate the consideration of other image analysis algorithms in the designs of DAQ and trigger systems of future LArTPCs.

Acknowledgments

This work is based upon work supported by the National Science Foundation under Grant No. NSF-1914965.

A Training Details

The Adam optimizer [53] was used with learning rate 0.0029, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and ϵ=1e-8. During training, CNN models were "kept" if the validation accuracy was higher than already-kept highest-accuracy models. The optimized (best) model was found when the validation accuracy stopped improving. The accuracy values quoted in the main text were obtained with the test sample, and they are quoted to sufficiently high precision (as allowed by the statistics used) to assess whether the networks meet DUNE's accuracy requirements.

The training curve showing the training and validation accuracy for the Q-CNN02-DS-OP, as an example, is shown in Fig. 8. The loss curve for the same network is shown in Fig. 9.

References

[1] Alexander Radovic, Mike Williams, David Rousseau, Michael Kagan, Daniele Bonacorsi, Alexander Himmel, Adam Aurisano, Kazuhiro Terao, and Taritree Wongjirad. Machine learning at the energy and intensity frontiers of particle physics. *Nature*, 560(7716):41–48, 2018.

[2] Georgia Karagiorgi, Gregor Kasieczka, Scott Kravitz, Benjamin Nachman, and David Shih. Machine Learning in the Search for New Fundamental Physics. 12 2021.

[3] Farah Fahim et al. hls4ml: An Open-Source Codesign Workflow to Empower Scientific Low-Power Machine Learning Devices. In *tinyML Research Symposium 2021*, 3 2021.

[4] Vladimir Loncar et al. Compressing deep neural networks on FPGAs to binary and ternary precision with HLS4ML. *Mach. Learn. Sci. Tech.*, 2:015001, 2021.

[5] Thea Aarrestad et al. Fast convolutional neural networks on FPGAs with hls4ml. *Mach. Learn. Sci. Tech.*, 2(4):045015, 2021.

[6] Yeon-Jae Jwa, Giuseppe Di Guglielmo, Luca P. Carloni, and Georgia Karagiorgi. Accelerating Deep Neural Networks for Real-time Data Selection for High-resolution Imaging Particle Detectors. In *2019 New York Scientific Data Summit: Data-Driven Discovery in Science and Industry*, 6 2019.

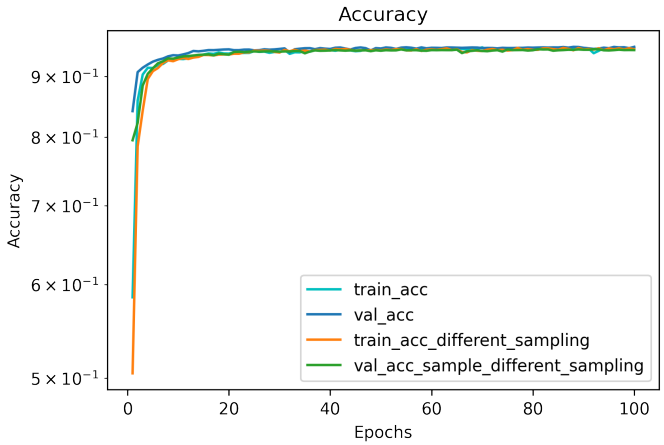


Figure 8. The training and validation accuracy curves of **Q-CNN02-DS-OP**, in cyan and blue, respectively. The best model was found at epoch 88. The training and validation accuracy curves obtained using bootstrapping are overlaid in orange and green, respectively, with the best model found at epoch 92. This comparison further demonstrates that the uncertainty on the accuracy of the network is relatively low.

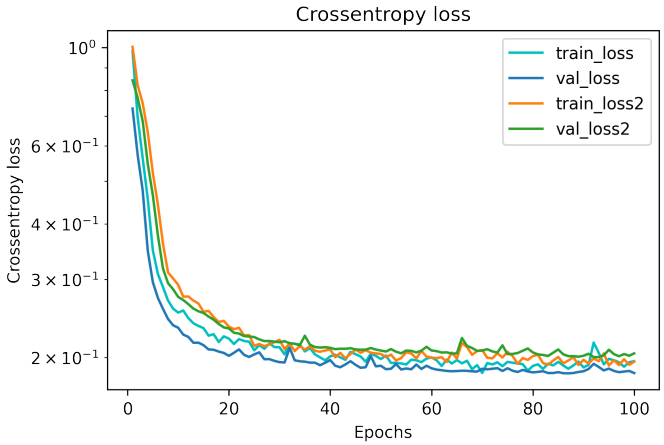


Figure 9. Cross entropy loss of **Q-CNN02-DS-OP** for the training and validation samples, in cyan and blue, respectively. The best model was found at epoch 88. The cross entropy loss curves obtained using bootstrapping are overlaid in orange and green, respectively, with the best model found at epoch 92. This comparison further demonstrates that the uncertainty on the accuracy of the network is relatively low.

[7] Stephen M. Trimberger. Three Ages of FPGAs: A Retrospective on the First Thirty Years of FPGA Technology. *Proceedings of the IEEE*, 103(3):318–331, 2015.

[8] Allison McCarn Deiana et al. Applications and Techniques for Fast Machine Learning in Science. 10 2021.

[9] Tommaso Diotalevi, Marco Lorusso, Riccardo Travaglini, Carlo Battilana, and Daniele Bonacorsi. Deep Learning fast inference on FPGA for CMS Muon Level-1 Trigger studies. *PoS*, ISGC2021:005,

2021.

[10] Abdelrahman Elabd et al. Graph Neural Networks for Charged Particle Tracking on FPGAs. 12 2021.

[11] Georges Aad et al. Artificial Neural Networks on FPGAs for Real-Time Energy Reconstruction of the ATLAS LAr Calorimeters. *Comput. Softw. Big Sci.*, 5(1):19, 2021.

[12] Ekaterina Govorkova et al. Autoencoders on FPGAs for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider. 8 2021.

[13] Aneesh Heintz et al. Accelerated Charged Particle Tracking with Graph Neural Networks on FPGAs. In *34th Conference on Neural Information Processing Systems*, 11 2020.

[14] Yutaro Iiyama et al. Distance-Weighted Graph Neural Networks on FPGAs for Real-Time Particle Reconstruction in High Energy Physics. *Front. Big Data*, 3:598927, 2020.

[15] Vinicius Mikuni, Benjamin Nachman, and David Shih. Online-compatible Unsupervised Non-resonant Anomaly Detection, 11 2021.

[16] Sioni Summers et al. Fast inference of Boosted Decision Trees in FPGAs for particle physics. *JINST*, 15(05):P05026, 2020.

[17] Babak Abi et al. Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume I Introduction to DUNE. *JINST*, 15(08):T08008, 2020.

[18] Babak Abi et al. Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume III: DUNE Far Detector Technical Coordination. *JINST*, 15(08):T08009, 2020.

[19] Babak Abi et al. Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume IV: Far Detector Single-phase Technology. *JINST*, 15(08):T08010, 2020.

[20] Babak Abi et al. Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume II: DUNE Physics. 2 2020.

[21] Dan Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and JÃŒrgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. pages 1237–1242, 07 2011.

[22] Javier Duarte et al. Fast inference of deep neural networks in FPGAs for particle physics. *JINST*, 13(07):P07027, 2018.

[23] Javier Duarte et al. FPGA-accelerated machine learning inference as a service for particle physics computing. *Comput. Softw. Big Sci.*, 3(1):13, 2019.

[24] Roberto Acciarri et al. Convolutional Neural Networks Applied to Neutrino Events in a Liquid Argon Time Projection Chamber. *JINST*, 12(03):P03011, 2017.

[25] Corey Adams et al. Deep neural network for pixel-level electromagnetic particle identification in the MicroBooNE liquid argon time projection chamber. *Phys. Rev. D*, 99(9):092001, 2019.

[26] Paulina Abratenko et al. Convolutional neural network for multiple particle identification in the MicroBooNE liquid argon time projection chamber. *Phys. Rev. D*, 103(9):092003, 2021.

[27] Paulina Abratenko et al. Semantic segmentation with a sparse convolutional neural network for event reconstruction in MicroBooNE. *Phys. Rev. D*, 103(5):052012, 2021.

[28] Roberto Acciarri et al. Cosmic Background Removal with Deep Neural Networks in SBND. 12 2020.

[29] B. Abi et al. Neutrino interaction classification with a convolutional neural network in the DUNE far detector. *Phys. Rev. D*, 102(9):092003, 2020.

[30] Francois Drielsma, Qing Lin, Pierre Côte de Soux, Laura Dominé, Ran Itay, Dae Heun Koh, Bradley J. Nelson, Kazuhiro Terao, Ka Vang Tsang, and Tracy L. Usher. Clustering of electromagnetic showers and particle interactions with graph neural networks in liquid argon time projection chambers. *Phys. Rev. D*, 104(7):072004, 2021.

[31] Dae Heun Koh, Pierre Côte De Soux, Laura Dominé, François Drielsma, Ran Itay, Qing Lin, Kazuhiro Terao, Ka Vang Tsang, and Tracy L. Usher. Scalable, Proposal-free Instance Segmentation Network for 3D Pixel Clustering and Particle Trajectory Reconstruction in Liquid Argon Time Projection Chambers. 7 2020.

[32] Tom O'Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. Keras Tuner. https://github.com/keras-team/keras-tuner, 2019.

[33] KerasTuner. https://keras.io/keras_tuner/. Accessed: December 20, 2021.

[34] Claudionor N. Coelho, Aki Kuusela, Shan Li, Hao Zhuang, Thea Aarrestad, Vladimir Loncar, Jennifer Ngadiuba, Maurizio Pierini, Adrian Alan Pol, and Sioni Summers. Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors. 6 2020.

[35] Benjamin Hawks, Javier Duarte, Nicholas J. Fraser, Alessandro Pappalardo, Nhan Tran, and Yaman Umuroglu. Ps and Qs: Quantization-aware pruning for efficient low latency neural network inference. 2 2021.

[36] Roberto Acciarri et al. Design and Construction of the MicroBooNE Detector. *JINST*, 12(02):P02017, 2017.

[37] Maddalena Antonello et al. A Proposal for a Three Detector Short-Baseline Neutrino Oscillation Program in the Fermilab Booster Neutrino Beam. 3 2015.

[38] Tsuguo Aramaki, Per Hansson Adrian, Georgia Karagiorgi, and Hirokazu Odaka. Dual MeV Gamma-Ray and Dark Matter Observatory - GRAMS Project. *Astropart. Phys.*, 114:107–114, 2020.

[39] Babak Abi et al. Supernova neutrino burst detection with the Deep Underground Neutrino Experiment. *Eur. Phys. J. C*, 81(5):423, 2021.

[40] Babak Abi et al. Prospects for beyond the Standard Model physics searches at the Deep Underground Neutrino Experiment. *Eur. Phys. J. C*, 81(4):322, 2021.

[41] Andrea Borga et al. FELIX based readout of the Single-Phase ProtoDUNE detector. *IEEE Trans. Nucl. Sci.*, 66(7):993–997, 2019.

[42] Eric D. Church. LArSoft: A Software Package for Liquid Argon Time Projection Drift Chambers. 11 2013.

[43] LArSoft. https://larsoft.org/. Accessed: December 20, 2021.

[44] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 9 2014.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. 12 2015.

[46] Keras. https://github.com/keras-team/keras/. Accessed: December 20, 2021.

[47] Ambrose Finnerty and Hervé Ratigner. Reduce power and cost by converting from floating point to fixed point. *Xilinx WP491*, 2017.

[48] Profiling. https://fastmachinelearning.org/hls4ml/api/profiling.html. Accessed: January 2, 2022.

[49] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[50] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. *CoRR*, abs/1502.02551, 2015.

[51] Yunchao Gong, Liu Liu, Ming Yang, and Lubomir D. Bourdev. Compressing deep convolutional networks using vector quantization. *CoRR*, abs/1412.6115, 2014.

[52] QKeras. https://github.com/google/qkeras. Accessed: December 20, 2021.

[53] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. 12 2014.

6.3 Summary

The studies presented in previous sections explored the 2D CNN application as an option for DUNE's real-time data selection. One of the most significant differences between the nominal data selection scheme (using traditional hit reconstruction) and the 2D CNN application scheme is whether to use 1D data (single wire) versus 2D data (wire plane). Handling 1D data may be more straightforward from a data processing point of view, and what is represented in 1D data may be sufficient for a threshold-based selection scheme. The real motivation to pursue 2D representation in data selection comes from the preservation of the physics interactions signatures. DUNE is designed to be a high-resolution imaging particle detector. As seen by raw-digit images in the figures of the papers within this chapter, physics interactions are clearly visualized through raw-digits in 2D data format. 2D CNN-based data selection scheme aims to make use of the full capability of the high-resolution imaging detector along with the rich and powerful developments for ML/DL applications. On the other hand, using the 2D data comes with the cost of extra pre-processing. The formation and preparation of 2D input images for CNN should also be implemented in the target FPGA for DUNE DAQ to fully demonstrate the feasibility of a 2D CNN-based data selection scheme. The authors of "Real-time Inference with 2D Convolutional Neural Networks on Field Programmable Gate Arrays for High-rate Particle Imaging Detectors" are currently working on the full demonstration of end-to-end data selection, including data transfers and pre-processing steps.

Chapter 7: Neutron-antineutron oscillation search in MicroBooNE

In this chapter, the analysis method developed for the first-ever argon-bound neutron-antineutron ($n-\bar{n}$) oscillation search is presented, corresponding to approximately 372 seconds[1] exposure with non-beam MicroBooNE data. Figure 7.1 shows the schematic diagram of the steps for the analysis. The unit for quantifying the data collected in the MicroBooNE active LArTPC is defined with 2.3

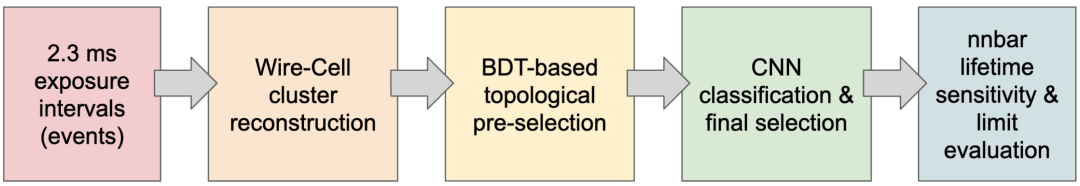


Figure 7.1: The analysis flow of $n-\bar{n}$ search in MicroBooNE.

ms exposure intervals using PMT reconstructed flash time; i.e., "one event" (or one exposure interval) refers to 2.3 ms duration of data collected in MicroBooNE. The length of one exposure interval (2.3 ms) is chosen as the corresponds to the drift time in the active volume of MicroBooNE, with the TPC drift field at 273 V/cm. For the signal simulation, one $n-\bar{n}$ interaction is simulated per one event (one exposure interval) using Monte Carlo (MC) simulation. A data-driven background simulation is used for this analysis. Section 7.1 discusses the signal modeling and background simulation and blind analysis strategy for this analysis. The exposure intervals include reconstructed

[1]The exact exposure may vary from this number.

particle interactions in the liquid argon volume; they are called "clusters" since they are the result of the Wire-Cell cluster reconstruction algorithms using the Wire-Cell 3D imaging and TPC-PMT matching. The clustering algorithms in the Wire-Cell reconstruction paradigm [79] are described in Sec. 7.2. Pre-selection using a Boosted-Decision-Tree (BDT) is developed and applied to the reconstructed signal and background clusters to control the background rate and remove obvious background. The BDT-based pre-selection is described in Sec. 7.3. Final selection using CNN classification is developed using the clusters passing the BDT-based pre-selection. Section 7.4 describes the CNN classification and discusses the final selection efficiency and the background estimates. The systematic uncertainties and the statistical uncertainties of the final selected event prediction are taken into account. The considerations of systematic uncertainties from the simulation and the detector modeling are discussed in Sec. 7.5. The lifetime sensitivity for the $n - \bar{n}$ search is evaluated using a Rolke method; the calculation and the result are presented in Sec. 7.6.

The event selection (BDT-based pre-selection followed by CNN-based final selection) is fully based on image-based analysis, as opposed to full reconstruction-based analysis where individual final state particles and their momenta can be identified. The currently available full reconstruction paradigms for MicroBooNE are developed upon and rely on the topological and kinematic assumptions of the neutrino interactions generated from Fermilab's Boosted Neutrino Beam (BNB), driving the particle flow reconstruction. The image-based analysis is chosen over a full reconstruction approach to avoid introducing these embedded assumptions within current reconstruction paradigms which could limit sensitivity to non-beam related searches such as this one. However, the full reconstruction can be utilized as a verification method for the CNN-based final selection as described in Sec. 7.2.

7.1 Signal and background simulation

7.1.1 Signal simulation

The GENIE v.3.00.04 via uboonecode[2] release v08.00.00.50 is used for the generation of $n-\bar{n}$ oscillation events. The simulation of this process has been developed and was implemented in GENIE as of version v.2.12. The implementation of $n-\bar{n}$ in GENIE is described in Sec. 5.1.1. Nucleon Fermi momentum and binding energy are modeled using the Local Fermi gas model in GENIE as the default setting for MicroBooNE analyses. Alternatively, the Bodek-Ritchie Fermi gas model is configurable in GENIE. Signal $n-\bar{n}$ simulation can vary upon the choice of nuclear modeling (Local Fermi gas model vs. Bodek-Ritchie Fermi gas model), which is further discussed in Sec. 7.5. For simulating the final state interactions (FSI), An empirical data-driven method *hA* with its most recent version *hA*2018 available in GENIE v3 is used for this work. Alternatively, *hN*2018 is configurable, which is a full cascade model. The impact of choosing between *hA*2018 and *hN*2018 on the final signal selection efficiency is discussed in Sec. 7.5.

The signal $n-\bar{n}$ is simulated uniformly across the MicroBooNE active volume (0.08 kton.) In order to account for the full drift of the active volume, a time window of 2.3 ms is set to define the unit of one event. The exposure of the search is further defined as the number of events multiplied by 2.3 ms. The simulated interaction is subsequently processed with GEANT4 and detector simulation before event reconstruction.

7.1.2 $n-\bar{n}$ final states in GENIE

GENIE first simulates $n-\bar{n}$ interaction in ^{40}Ar leaving an intermediate state with unstable particles. Then subsequent FSI are simulated using *hA* INTRANUKE modeling, to determine the GENIE final state particles (what exits the Ar nucleus.) In this section, the generation-level truth for signal simulation is shown at the intermediate GENIE state and the final state. The intermedi-

[2]uboonecode is package in a C++ based framework LArSoft [99]. uboonecode is used for MicroBooNE simulation, reconstruction, and analysis.

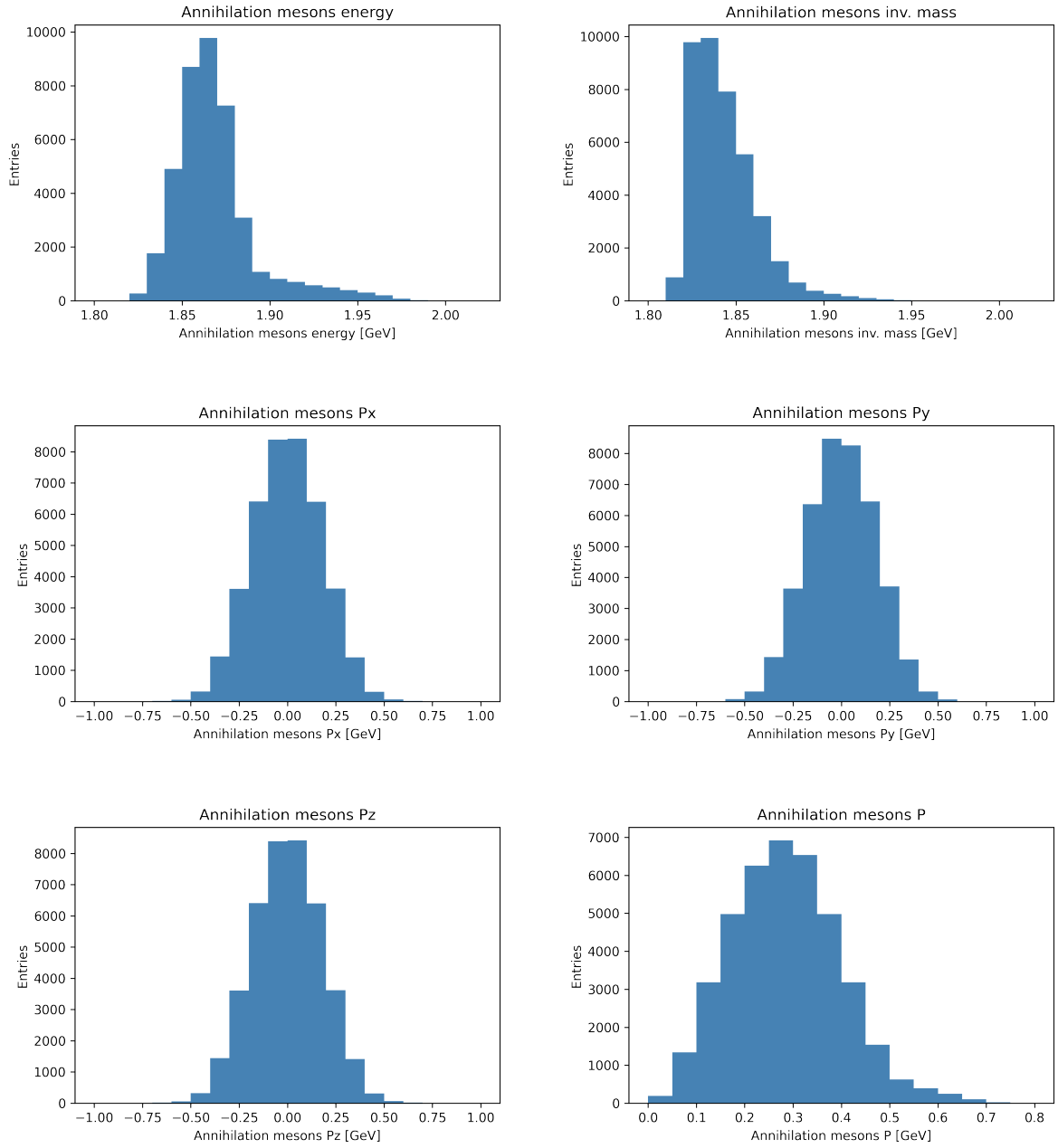


Figure 7.2: Annihilation meson kinematics of $n - \bar{n}$ simulation in the intermediate state of GENIE simulation.

ate GENIE state demonstrates the validation of expected kinematics from the implemented $n - \bar{n}$ oscillation simulation. The final state gives a blueprint of what to expect from the reconstruction of observable activity in the LArTPC.

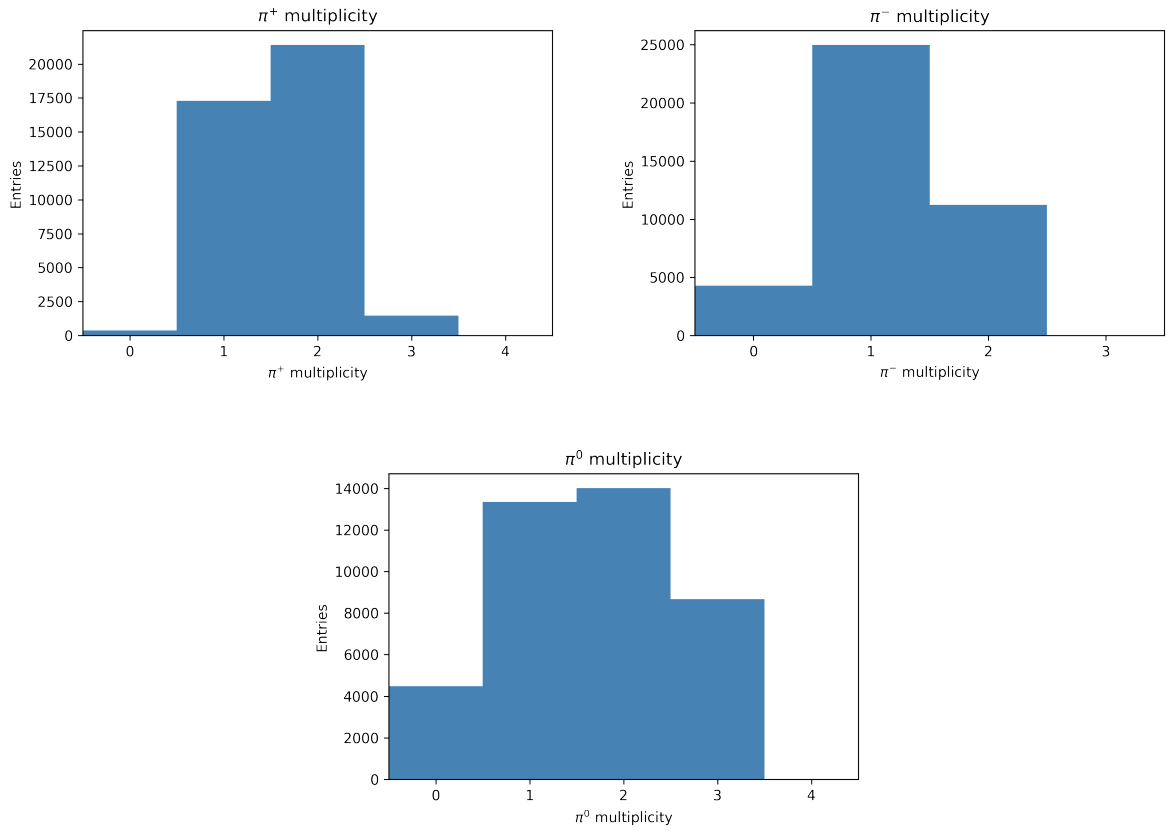


Figure 7.3: π^+, π^-, π^0 multiplicities at the intermediate state of GENIE simulation right after the two-nucleon annihilation following $n - \bar{n}$ oscillation in ^{40}Ar.

Right after the $n - \bar{n}$ oscillation process in ^{40}Ar, $(\bar{n}, p)$ or $(\bar{n}, n)$ annihilation is simulated as described in Sec. 5.1.1 following the branching ratios shown in Tab. 5.1. This annihilation leaves multiple intermediate pions, leaving an unstable nucleus with atomic number 17 or 18, and atomic mass 38. The pions from the pair annihilation have total energy and invariant mass around two nuclei masses (~1.88 GeV) where the Fermi-momentum can spread the distribution. The momentum distributions of these pions have isotropy, where the total momentum peaks around 0.3 GeV. The kinematics of annihilation mesons are shown in Fig. 7.2; the multiplicities of annihilation mesons

are shown in Fig. 7.3

Then these daughter pions from the pair annihilation interact with the remaining unstable nucleus, scattering off protons and neutrons through FSI channels in hA INTRANUKE. The final state, thus, includes multiple protons and neutrons along with pions. Final state neutrons are unlikely to be observed, but protons over 50 MeV kinetic energy threshold are expected to be reconstructed through existing reconstruction paradigms [77]. Protons with kinetic energy over 50 MeV can be considered "visible." The final state particle multiplicities are shown in Fig. 7.4. Note that the pion multiplicities have changed after the FSI of the intermediate state shown in Fig. 7.3. For example, a charge exchange process such as $\pi^+ A \rightarrow p\pi^0 A'$ can remove one π^+ and add one π^0 in the interaction.

The total energy released from the interaction can be defined as the sum of kinetic energies of protons and neutrons in the GENIE final states and the relativistic energies of exiting pions. Assuming 100% reconstruction of charged pions and gamma pairs from neutral pions in the MicroBooNE LArTPC, the visible energy can be defined as the sum of kinetic energies of protons (when over 50 MeV) and the relativistic energies of pions. These variables are shown in Fig. 7.5; the total energy released is close to the expected two-nucleon mass, and the visible energy shows a long tail in the lower energy region due to the missing energy from neutrons and low-energy protons.

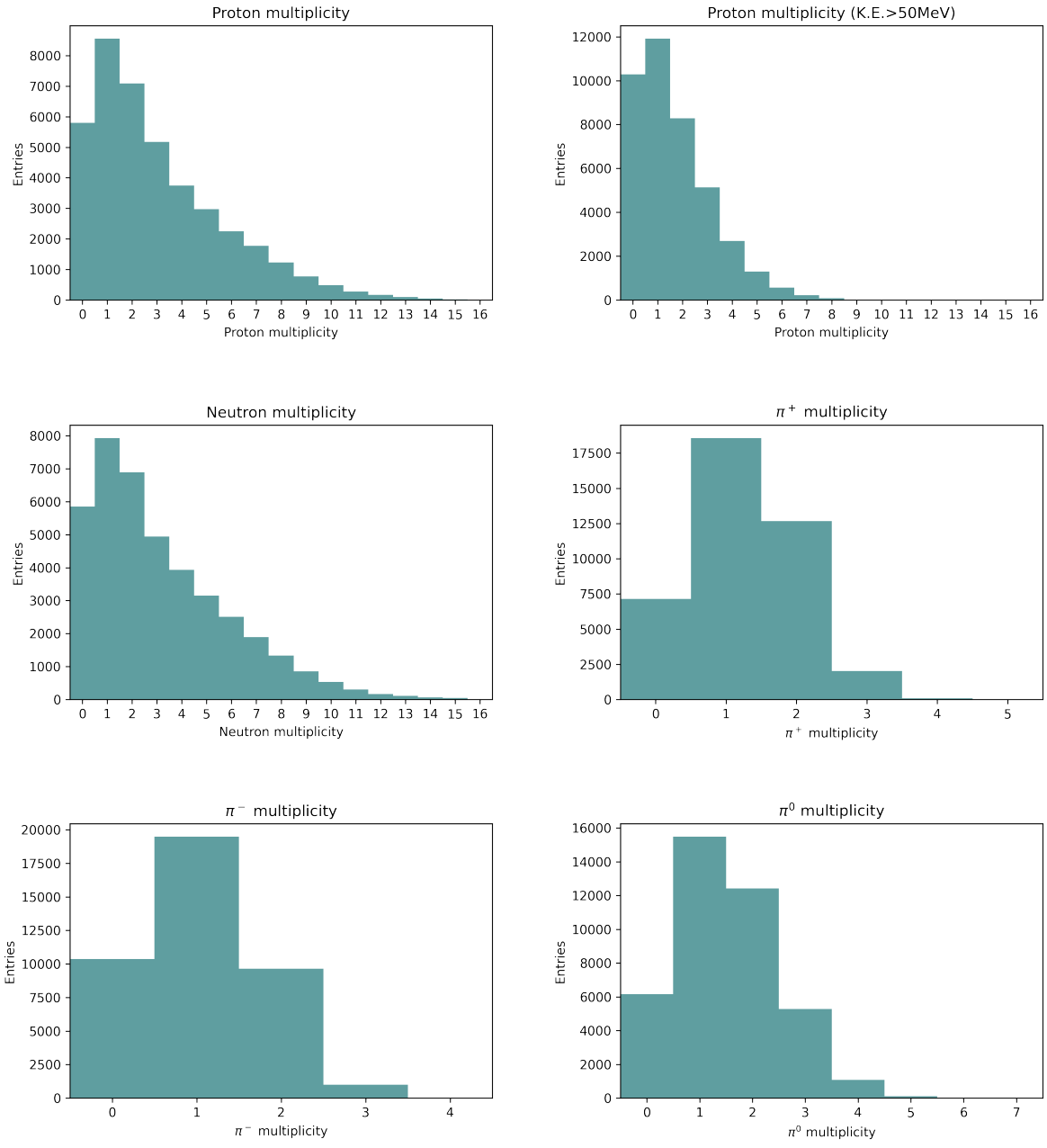


Figure 7.4: GENIE final state particle multiplicities. For the top right figure, protons with their kinetic energy > 50 MeV are considered.

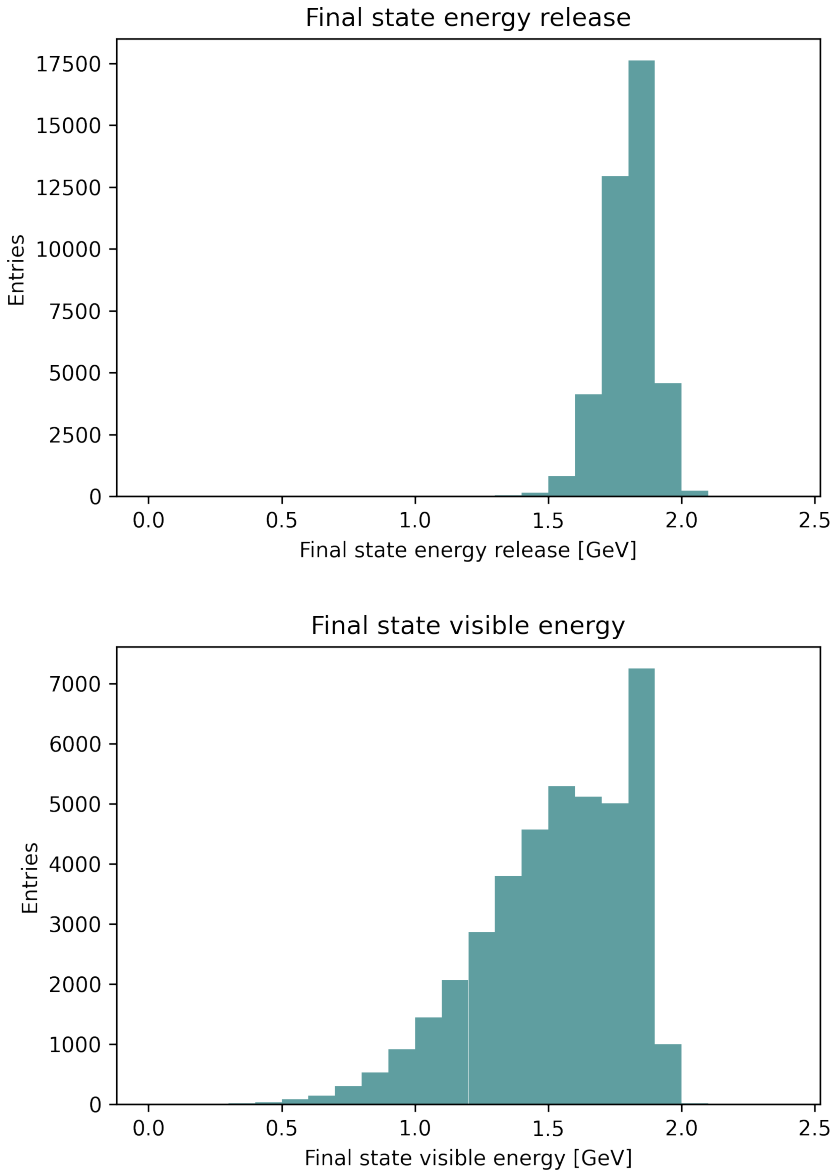


Figure 7.5: (top) Final state energy release. (bottom) Final state visible energy.

7.1.3 Background simulation

Since MicroBooNE is operating just below the earth's surface (about 6 meters underground), without significant overburden, the main background candidate of this analysis in MicroBooNE is cosmogenic interactions, i.e., atmospheric muons. This differentiates the $n - \bar{n}$ search in MicroBooNE from the searches in underground detectors such as Super-Kamiokande and DUNE, where

the main background candidate is atmospheric neutrino interactions. As MicroBooNE receives cosmic rays at 5.5kHz [125], the 2.3 ms window typically includes around a dozen of cosmogenic interactions. While the cosmogenic interactions found in the MicroBooNE detector are predominantly cosmic muon tracks, cosmic events can induce electromagnetic showers and hadronic showers. While the cosmic muon tracks are often identifiable in the TPC readout and easily ruled out as background interactions in the selection, these shower interactions inside the active volume of the detector may contribute as a non-trivial background. The incoming cosmic particles in the MicroBooNE detector are understood through simulation studies; the incoming particle rates and shower backgrounds in the active volume were originally studied and presented in [126]. In the note, electromagnetic (EM) showers created in the active volume were studies using CRY and CORSIKA cosmic simulations and are excerpted here in Tab. 7.1.

Process (in spill)	Source (Primary particle)	CRY PROTON	CORSIKA CMC GHEISHA	CORSIKA PROTON FLUKA	CORSIKA CMC FLUKA
Compton	μ in AV	190±11	297±70	249±20	318±23
Pair production	μ in AV	12220±90	18225±549	13369±150	15917±162
Compton	μ not in AV	4	0	0	0
Pair production	μ not in AV	144±10	33±23	29±7	71±11
Compton	γ	1.3±0.9	0	3.4±2.4	3.3±2.3
Pair production	γ	110±9	264±66	217±19	231±20
Compton	not a μ or γ	12±3	17±17	15±5	28±7
Pair production	not a μ or γ	640±21	1550±160	1222±45	1474±49

Table 7.1: *Excerpted from Table 5 in [126].* Cosmogenic background shower rates for 211 seconds of integrated exposure in MicroBooNE, simulated with CRY [127] and CORSIKA [128] simulations. AV here stands for active volume of the detector. For both cases of Compton scattering and pair production, showers from primary μ in the active volume is the predominant source of interactions. The 211 second is equivalent to 57% of the projected exposure for the $n - \bar{n}$ search in MicroBooNE presented in this thesis.

In Tab. 7.1, it can be found that the pair production from primary μ in the active volume generates the most showers in MicroBooNE. On the other hand, showers from the source "not a μ or γ" includes showers with hadronic primaries. These hadronic showers are of interest in this analysis. The $n - \bar{n}$ signal in MicroBooNE typically generates multiple pions ($\pi^{+/-}, \pi^0$.) The hadronic cosmic showers may contribute as an irreducible background, as the final state and their kinematics

Ancester particle type	No. of cosmogenic shower events
$\bar{p}$	61
π^-	184
e^+	56
e^-	83
π^+	138
p	685
n	295

Table 7.2: *Excerpted from Table 7 in [126].* Non-muon or non-photon primaries using CORSIKA-CMC-FLUKA simulation. Event counts corresponds to a total of 211 seconds.

can be similar to those of the signal $n - \bar{n}$ interactions. The number of showers of these non-muon or non-photon primaries generated by the CORSIKA-CMC-FLUKA simulation are broken down by the ancestor particles and are shown in Tab. 7.2. From Tab. 7.2, showers with $\bar{p}$ ancestor are estimated to be 61 shower events within 211 seconds of MicroBooNE data. These shower events can possibly contribute to the irreducible background, if the proton entering the MicroBooNE active volume has relatively low energy since it can pair-annihilate with a nucleon in the argon nucleus in a similar manner to the $n - \bar{n}$ process, topologically and kinematically. The goals of this analysis are to effectively remove these signatures from cosmogenic interactions (cosmic muon track, cosmogenic shower) during the pre-selection stage, to provide accurate background estimates, and to increase signal efficiency in the final selection.

The former iteration of this $n - \bar{n}$ analysis [129] used fully simulated cosmogenic interactions using CORSIKA-CMC-FLUKA simulation, using a CNN-based classification and Pandora pattern recognition algorithm for the CNN-BDT combination analysis. However, a discrepancy of background estimate between data and MC simulation was present, which motivated the use of a data-driven approach. The current approach described in this thesis uses a data-driven background. This approach simulates the background using the MicroBooNE external data (EXT) collected using random triggers outside the BNB beam window, guaranteeing no on-beam neutrino interactions within.

Using a data-driven approach for background estimation for this search may come along with

the concern of $n-\bar{n}$ interactions (if realized in nature) being contained in the background estimate. The current best 90% C. L. limit on bound $n-\bar{n}$ lifetime is 3.6×10^{32} years (for oxygen bound neutrons [41].) If we assume $n-\bar{n}$ occurs at this lifetime, we can estimate the occurrence in the MicroBooNE detector and show it is negligible: MicroBooNE's active volume is 0.08 kton liquid argon, 1kton liquid argon contains 3.32×10^{32} neutrons. Thus, postulating $n-\bar{n}$ lifetime as 3.6×10^{32} years,

$$occurrence = \frac{3.32\times10^{32}(neutron/kt)\times0.08(kt)}{3.6\times10^{32}(year)} \approx 0.07(neutron/year.) \tag{7.1}$$

Since the search exposure of this analysis is about 200 seconds, possible $n-\bar{n}$ in the background estimate is considered to be negligible,

$$occurrence_{exposure} \approx 0.07(neutron/year)\times200(second) < 10^{-6}neutron. \tag{7.2}$$

Naturally, the short exposure of MicroBooNE is expected to lead to greatly diminished lifetime sensitivity, which is further described in Sec. 7.6. It is worth mentioning here that the objective of the $n-\bar{n}$ search in MicroBooNE is not to set a competitive $n-\bar{n}$ lifetime limit, but to serve as the first-ever demonstration of an $n-\bar{n}$ search in liquid argon and a testbed for future LArTPCs, such as DUNE.[3]

As defined at the beginning of the chapter, one event (or one exposure interval) is 2.3 ms of data in MicroBooNE in this analysis. Thus the expected background rate (b_{rate}) before any event selection is

$$b_{rate} = \frac{3.154\times10^{7}(sec)}{(year)} \times \frac{1(event)}{2.3\times10^{-3}(sec)} = 1.37\times10^{10}(event/year). \tag{7.3}$$

It should be noted that one event (one exposure interval) typically contains multiple cosmic inter-

[3]In principle, the data-driven approach for the background estimate in MicroBooNE does not eliminate possible signal interactions in the background. However a data-driven approach in MicroBooNE is necessary in order to model cosmic interactions for analyses due to the poor MC-data agreement in current cosmic MC simulation.

action clusters, and the final selection is applied on a per cluster basis. After the final selection, the number of events containing at least one cluster is counted. The number of signal and background events containing more than one selected cluster at final selection stage is negligible. The cluster reconstruction and the cluster-based selection are discussed in Sec. 7.2 and Sec. 7.4.

7.1.4 "Overlay" method for signal simulation

MicroBooNE has developed an "overlay" method (a.k.a. MCC9 overlay) technique to account for cosmogenic background in analyses using BNB neutrinos. This "overlay" method is a data-driven simulation, where the simulated interactions of neutrinos (or, in this case $n - \bar{n}$) are placed on top of MicroBooNE EXT data collected outside of the neutrino beam window. This "overlay"

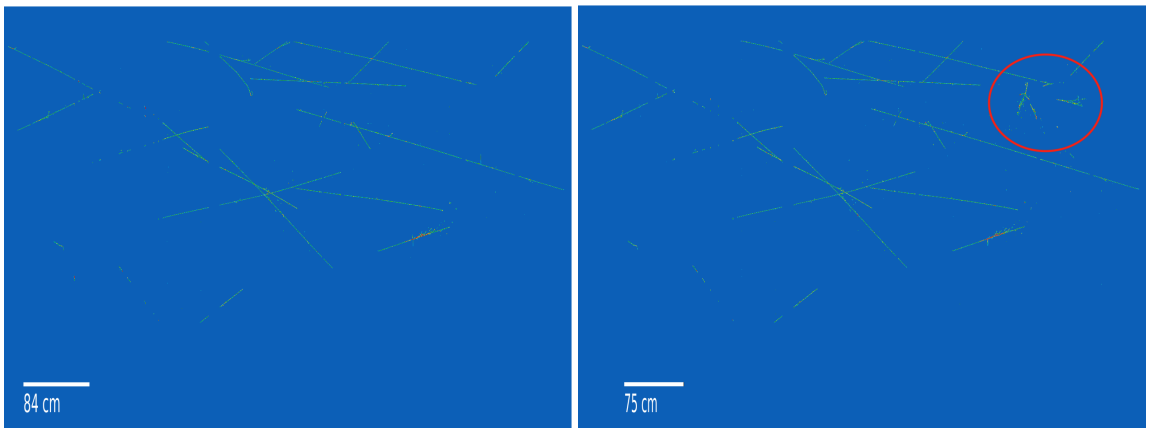


Figure 7.6: (left) Event display of background sample. (right) Event display of $n - \bar{n}$ signal sample. For background simulation, MicroBooNE EXT data is signal processed and reconstructed as discussed in Sec. 3.2.4. For signal simulation, GENIE-simulated interaction is overlaid on MicroBooNE EXT data, then signal processed and reconstructed.

approach has been adopted for other MicroBooNE analyses to resolve the discrepancy between MC simulation and data seen when CORSICA simulation was used. For the signal sample generation, one true $n - \bar{n}$ interaction is simulated per event (2.3 ms exposure interval) uniformly inside the active volume. With this nature of overlay, cosmogenic activities are present in the signal $n - \bar{n}$ sample as well. The clusters reconstructed from the Wire-Cell 3D imaging algorithms are further labeled truly as stemming from the $n - \bar{n}$ interaction or a cosmogenic interaction through the com-

parison of two events from the same EXT background source. This procedure is further described in Sec. 7.2.3.

7.1.5 Sample preparation

For this analysis, the dataset of MicroBooNE EXT taken from December 2017 to August 2018 (Run3b period) is used, which is equivalent to approximately 4.3 million records of 4.8 ms length readout intervals. The total dataset is divided into parts used for (1) analysis development, (2) central value MC ("central value" with reference to systematic variations), (3) fake data analysis, and (4) MicroBooNE search data. The data sample is divided in the way that the central value MC

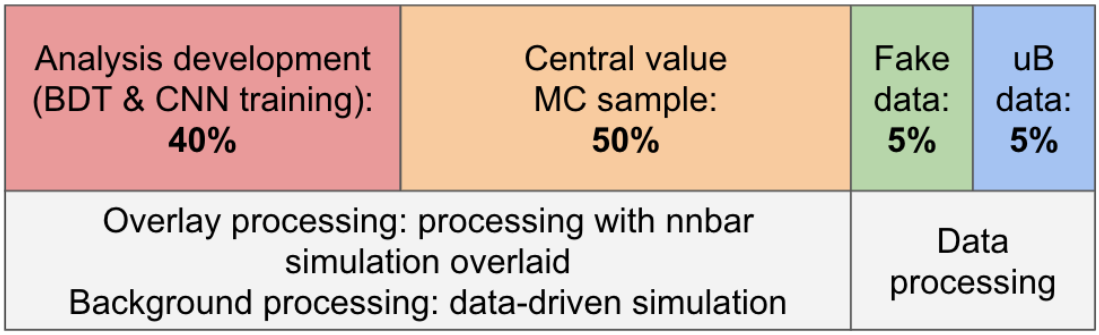


Figure 7.7: Sample division for the blind analysis. The total dataset used corresponds to 4.3 Million events. The development datset (40%) and the central value dataset (50%) are processed twice; once with overlay processing (with signal simulation), once with background processing (with background simulation.) The rest 10% is prepared through data processing.

has ten times larger statistics than the search dataset.[4] At the same time, enough statistics for the development of the selection chain are secured. The data sample division is shown in Fig. 7.7. As described in Sec. 7.1.3, the MicroBooNE EXT can be reconstructed to estimate the data-driven background. Also the same input (EXT) can be processed with $n - \bar{n}$ overlaid as described in Sec. 7.1.4 to simulate signal. The analysis development dataset and central value MC dataset are processed with and without the signal ($n - \bar{n}$) overlay.

[4]The central value statistics is limited in the data-driven approach whereas one can generate full MC as large as one wishes. The central value statistics is recommended to be at least ten times larger than the search region statistics.

The analysis follows a blindness scheme to eliminate the experimenter's bias by opening the data result. The search data may only be analyzed after careful demonstrations of the analysis on the central value MC and the fake datasets were performed. In the process of the final selection development, a small subset of the development sample is set aside to study the performance of CNN classification. Once the BDT classifier and CNN classifier are trained and fixed, the selection efficiency and the sensitivity of $n-\bar{n}$ lifetime in MicroBooNE is evaluated on the central value MC sample, including the uncertainties. Then, the limit of $n-\bar{n}$ lifetime is measured on MicroBooNE search data only.

An additional fake data study can be performed to verify this analysis as well. The fake data is prepared with a given (blind) rate of x% of injected $n-\bar{n}$ signal, where the x% is unknown to the analyzer. When the x% is estimated from the analysis chain with its uncertainty, the true x% is revealed and compared to the estimated value.

The development set (40% of the total dataset) is further broken down for the pre-selection stage and final selection stage. This portion of data is divided again for BDT-based pre-selection (5% of the development dataset, equivalent to 2% of the total dataset) and CNN-based selection (95% of the development dataset, equivalent to 38% of the total dataset) developments, and again separated for BDT (or CNN) training, validation, and testing sets. The breakdown is shown in Tab. 7.3.

Dataset for analysis development (40% of total dataset)					
BDT development			CNN development		
5% of the development set (2% of the total dataset)			95% of the development set (38% of the total dataset)		
Training	Validation	Testing	Training	Validation	Testing
40%	10%	50%	80%	10%	10%

Table 7.3: Dataset division for analysis development. 95% of the development dataset is assigned for the CNN classifier development, 5% of the development dataset is assigned for the BDT classifier development.

It is crucial to use statistically exclusive datasets for the selection development and the sensi-

tivity evaluation, especially for the ML-based selection. Also, careful consideration of allocating training sets for BDT and CNN is absolutely necessary since the procedure of pre-selection and final selection is in sequence, and the overlap between those training sets can tarnish the evaluation of the final efficiency.

7.2 Wire-Cell cluster reconstruction

Equipped with LArTPC's excellent calorimetric and spatial resolutions, multiple event reconstruction paradigms (i.e. Pandora multi-algorithm pattern recognition [78], deep learning with convolutional neural networks [77, 87, 88], and Wire-Cell [79, 130]) have been developed by MicroBooNE, with the focus on BNB neutrino interaction reconstruction. One of the key differences between the Wire-Cell and other reconstruction paradigms is that the Wire-Cell 3D imaging approach capitalizes on the most fundamental detector information (time, charge, and geometry) prior to any pattern recognition. Higher-level full event reconstructions using pattern recognition for the Wire-Cell and other reconstruction paradigms up to the particle flow of neutrino interactions include assumptions around neutrino interactions specifically generated by the Fermilab BNB. This analysis adopted the Wire-Cell approach since this generalized approach of initial interaction reconstruction welcomes the utilization of Wire-Cell reconstruction for non-neutrino interactions such as $n-\bar{n}$ oscillation. And this analysis only makes use of the Wire-Cell reconstruction only up to cluster-level for the pre-selection and final selection, not the Wire-Cell full event reconstruction. Full event reconstruction using the Wire-Cell framework on selected events is possible and shown in Sec. 7.4.2, post-final-selection, only to demonstrate the validation of the final selection of the CNN-based search.

The next subsection briefly discusses the Wire-Cell paradigm's capitalization on time, charge, and geometry of the MicroBooNE TPC and PMTs, and the methodology of 3D imaging, charge-light matching, and clustering algorithms. For a more detailed description, see [79].

7.2.1 Wire-Cell 3D imaging

The readouts of MicroBooNE are signal processed, and the waveforms on the wires are deconvoluted. The charge deposit on the wire is saved as "hit." "Time slice" is defined as 2 μs length of a duration corresponding to 4 sampling ticks in the TPC readout. The three wire planes in the anode plane lie on the same 2D spatial plane. The wires on U, V planes are (60, -60) degrees rotated from the Y plane, and on each plane, wire pitch is defined as the middle boundary of neighboring wires on the same plane. The smallest unit in the Wire-Cell 3D imaging is a "cell," a.k.a. "space point" (spacepoint), defined as a 3D voxel with the smallest possible 2D triangle that can be made from the wire pitches and with the height equal to one time-slice on the third dimension. The consecutive hits on wires generate neighboring hit cells, and these neighboring hit cells form a blob. The charge deposit of the blob is deduced from the sum of hits on the wires defining the blob. The charge of a cell (space point) is deduced by the total charge of the blob that contains the cell, divided by the number of space points within the blob. This 2D plane showing cells, wires, wire pitches, and the blob is shown in Fig. 7.8. This step of finding blobs in a 2D plane is called "tiling." Using time and geometry information, the 2D blobs can be concatenated to form a 3D image of the possible depositions, as shown in Fig. 7.9.

Note that procedures such as "charge solving" and "deghosting" to remove "ghost" blobs (due to MicroBooNE's nonfunctional wires) are not described in this section, since the focus is to summarize the concept of the Wire-Cell reconstruction briefly. However, those steps are in fact crucial to reducing the adverse impact of nonfunctional wires in the MicroBooNE detector. The description of these steps can be found in [79]. After performing the charge solving and deghosting processes on 3D images, the 3D images are called "proto-clusters."

7.2.2 Clustering and "many-to-many charge-light matching"

The proto-clusters in Wire-Cell are solely reconstructed from the TPC readout, making use of the physical proximity of the charge depositions in 3D. When a particle interaction is initiated

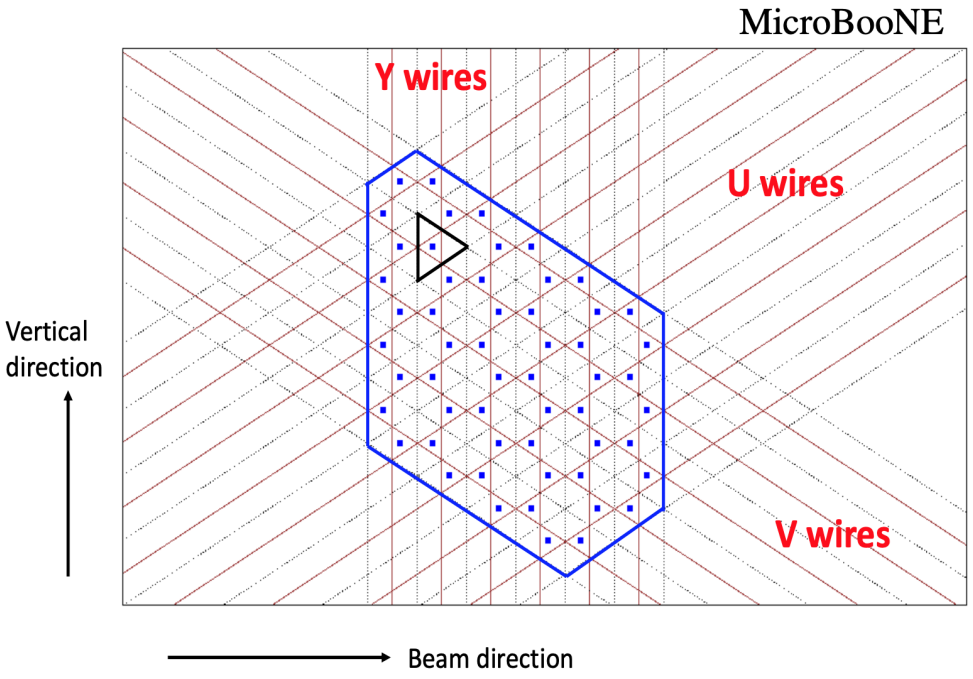


Figure 7.8: The black triangle represents a hit cell defined by wire pitches (dashed black lines) of U, Y, V planes. The blob, the collection of contiguous hit cells, is represented as solid blue lines. Figure 2 of reference [79].

from a primary particle, it can leave disconnected signatures in the LArTPC owing to secondary neutral particles or imperfect signal processing. These proto-clusters are grouped into TPC clusters, where the latter represents one particle physics interaction. The cluster grouping algorithm includes (1) clustering in the presence of gaps, (2) separation of coincidental overlap clusters, (3) further deghosting, and (4) clustering for neutrino events. The details of each step can be found in [79].

After the TPC clusters are grouped to represent candidate physics interactions, TPC cluster and PMT flash matching is performed using a "many-to-many charge-light matching" technique [79]. The matching of PMT flash and TPC cluster allows to find the interaction start time with the time resolution scale of the scintillation light (≈ns), which is much shorter than the TPC drift time (≈ms.) A "flash" is defined as a group of PMT signals in a time window (100ns.) For each recorded event, 20-30 TPC clusters and 40-50 PMT flashes are subjected to be matched. The matching algorithm requires the following rules:

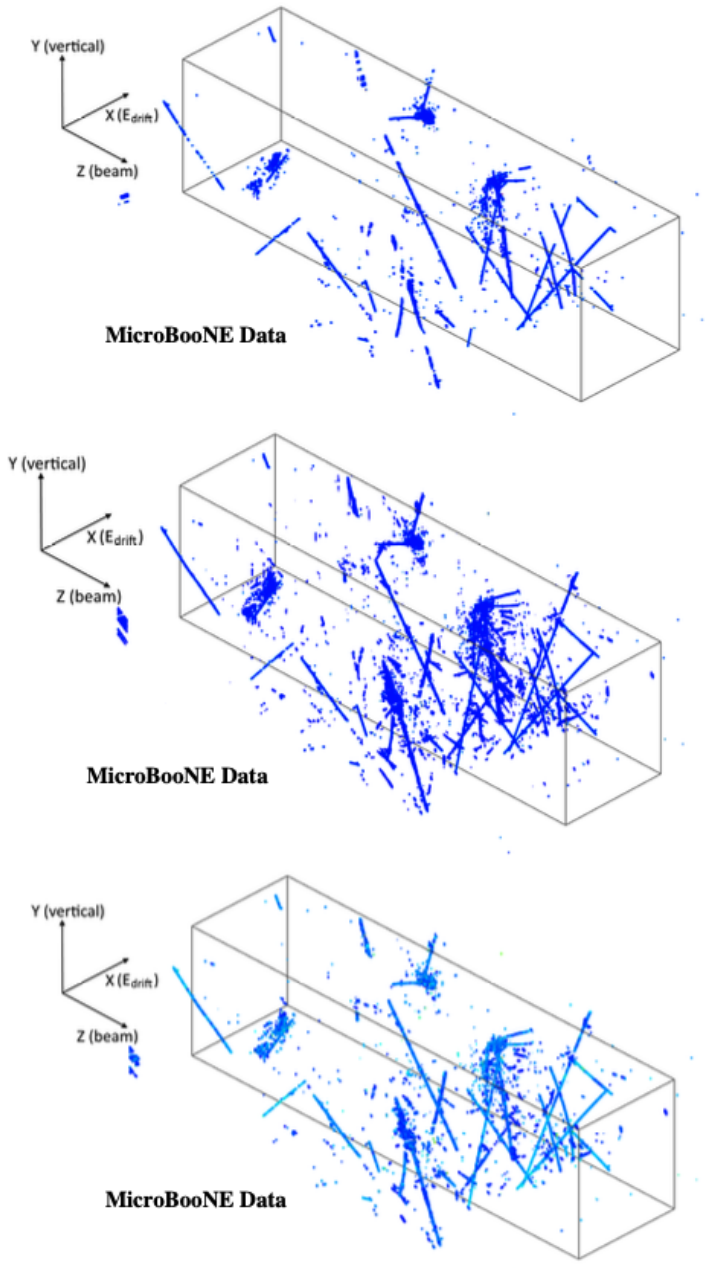


Figure 7.9: Comparison of the hit blobs in the MicroBooNE active volume at different stages. The solid black box represents MicroBooNE active volume. Top: 3-plane tiling with 70% active volume. Middle: 2-plane tiling with 97% active volume. Bottom: 2-plane tiling result after the charge solving. The color scale represents the resulting charge values in the charge solving. Image Credit: [79].

- One TPC cluster can match to zero or at most one PMT flash.
- One PMT flash can match to zero, one, or multiple TPC clusters. These multiple TPC clusters that matching the same PMT flash form a "cluster bundle."

This matching rules ensure that one TPC cluster is associated with at most one PMT flash, which decides the interaction start time. The light signal from the TPC cluster can be predicted and compared to the photoelectron (PE) distribution for 32 PMTs. The "match" is defined as a good agreement between the prediction and the measured flash. Solving this matching problem does not give a unique solution because there are more unknowns than knowns in the system. Introducing l_1-regularization in chi-square optimization can yield a unique solution. The entire matching algorithm in practice includes pre-selection, fitting the l_1-regularized chi-square, and re-examination. Hypothetical TPC-PMT pairs are tested through this algorithm; further details are found in [79]. Generated $n-\bar{n}$ signal and background event samples are reconstructed using this Wire-Cell reconstruction paradigm up to Wire-Cell cluster level, where the outcome is topological clusters consisting of space-points in the TPC with a matched PMT flash.

7.2.3 Low-level Wire-Cell cluster labeling

The Wire-Cell clusters in signal events processed with $n-\bar{n}$ overlaid, as well as those in background events, do not contain MC labels because both include real (unlabeled) data. Labeling reconstructed clusters as originating from the true $n-\bar{n}$ interactions or not is a necessary step prior to the BDT or CNN training. A given signal $n-\bar{n}$ event with 2.3 ms interval typically contains one (or more than one from the reconstruction) cluster from the $n-\bar{n}$ interaction and roughly about ten cosmic clusters from cosmogenic interactions. The $n-\bar{n}$ cluster is defined as a cluster reconstructed and originating from the true $n-\bar{n}$ interaction from the GENIE simulation. True $n-\bar{n}$ clusters are labeled by comparing the pair of two events from the same MicroBooNE EXT input event, where one event is processed with signal ($n-\bar{n}$) simulation overlaid, and the other is without signal simulation. The clusters in the pair events are associated with flashes from the TPC

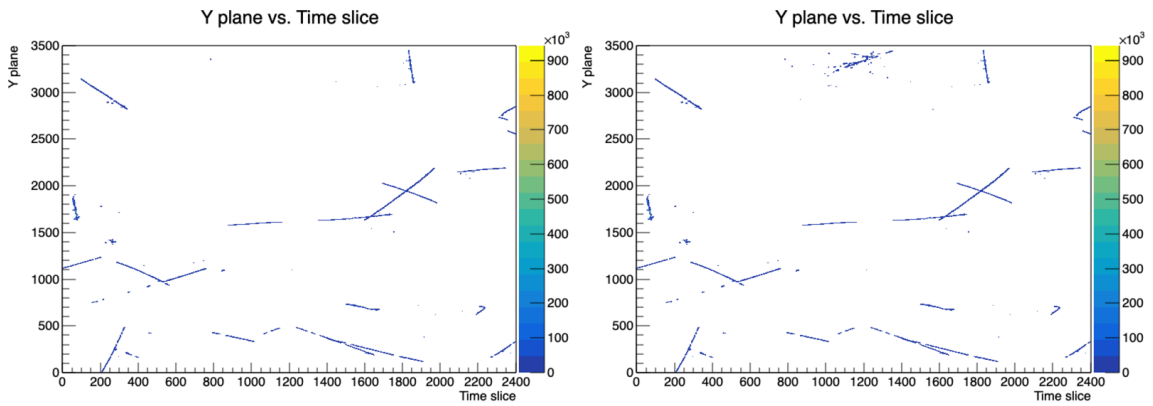


Figure 7.10: 2D projections of reconstructed Wire-Cell clusters on Y plane vs. time slice for (left) background simulation processing and (right) signal simulation processing. The figure on the right includes the signature of $n - \bar{n}$ cluster in the top center, on top of those of multiple cosmic clusters (seen in both images.) These 2D projections are obtained from plotting reconstructed clusters, thus they are further processed through the Wire-Cell clustering described in Sec. 7.2 compared to Fig. 7.6, which are only after low-level signal processing on TPC wires.

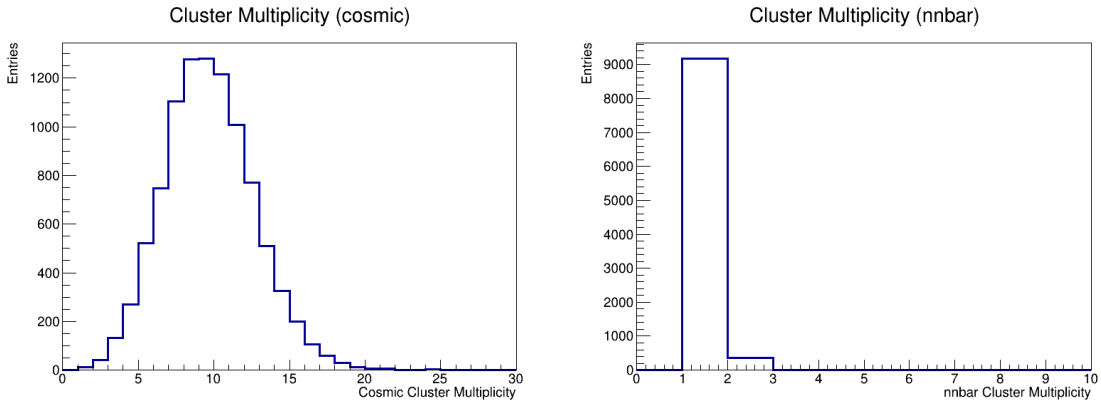


Figure 7.11: The number of clusters reconstructed per event for (left) cosmic sample and (right) nnbar ($n - \bar{n}$) sample.

cluster and PMT flash matching described in Sec. 7.2. Following the matching rule of this TPC-PMT matching, the matched TPC is associated with one flash time with precise timing resolution. When the list of clusters accompanied by the associated flashes are compared in the pair events, the extra flashes originating from the $n - \bar{n}$ signal simulation are easily identified.

After labeling $n - \bar{n}$ clusters and cosmic clusters, each cluster is treated independently in the following event selection, where the event selection is developed in order to select $n-\bar{n}$ clusters and

reject cosmic clusters. For background events, cluster multiplicity in one event is approximately 9. Clusters originated from $n - \bar{n}$ can be one or more in one signal event, as the clustering algorithm can split the interaction into two clusters. The efficiency of the selection is established on cluster selection, then converted into event selection which is further discussed in Sec. 7.4.

7.3 Pre-selection

A cluster, after the Wire-Cell cluster-level reconstruction, is a collection of space points, where each of space point is associated with an index position on the U, V, Y plane, time slice (from TPC sampling), and the charge deposition. Using 2D projections of these clusters, the extent of the signature of the reconstructed interaction along with the U, V, Y plane, and time-axis is found, as well as the number of space points associated with the cluster.

Figure 7.12 shows the distributions of these variables for $n - \bar{n}$ and cosmic clusters. The pre-selection of the reconstructed clusters is developed using Boosted Decision Tree (BDT) classification based on the following five topological variables:

- Space points: the number of cells (i.e. space points) in the cluster
- Time-tick extent: (max time slice value - min time slice value) in the cluster
- Y-plane extent: (max Y-plane index - min Y-plane index) in the cluster
- U-plane extent: (max U-plane index - min U-plane index) in the cluster
- V-plane extent: (max V-plane index - min U-plane index) in the cluster,

where the index equals to the wire number. These BDT input variables are possibly correlated / anti-correlated. For example, a muon track could be long along one wire plane and short along others because the three wire planes are tilted by 60° with each other on the vertical plane (See Sec. 3.2.1.) A BDT classifier can perform whether the input variables are correlated or not.

The objectives of this BDT pre-selection are: 1) to remove obvious cosmic prior to the subsequent CNN classification analysis stage; 2) to suppress the cosmic cluster rate to backgounrds

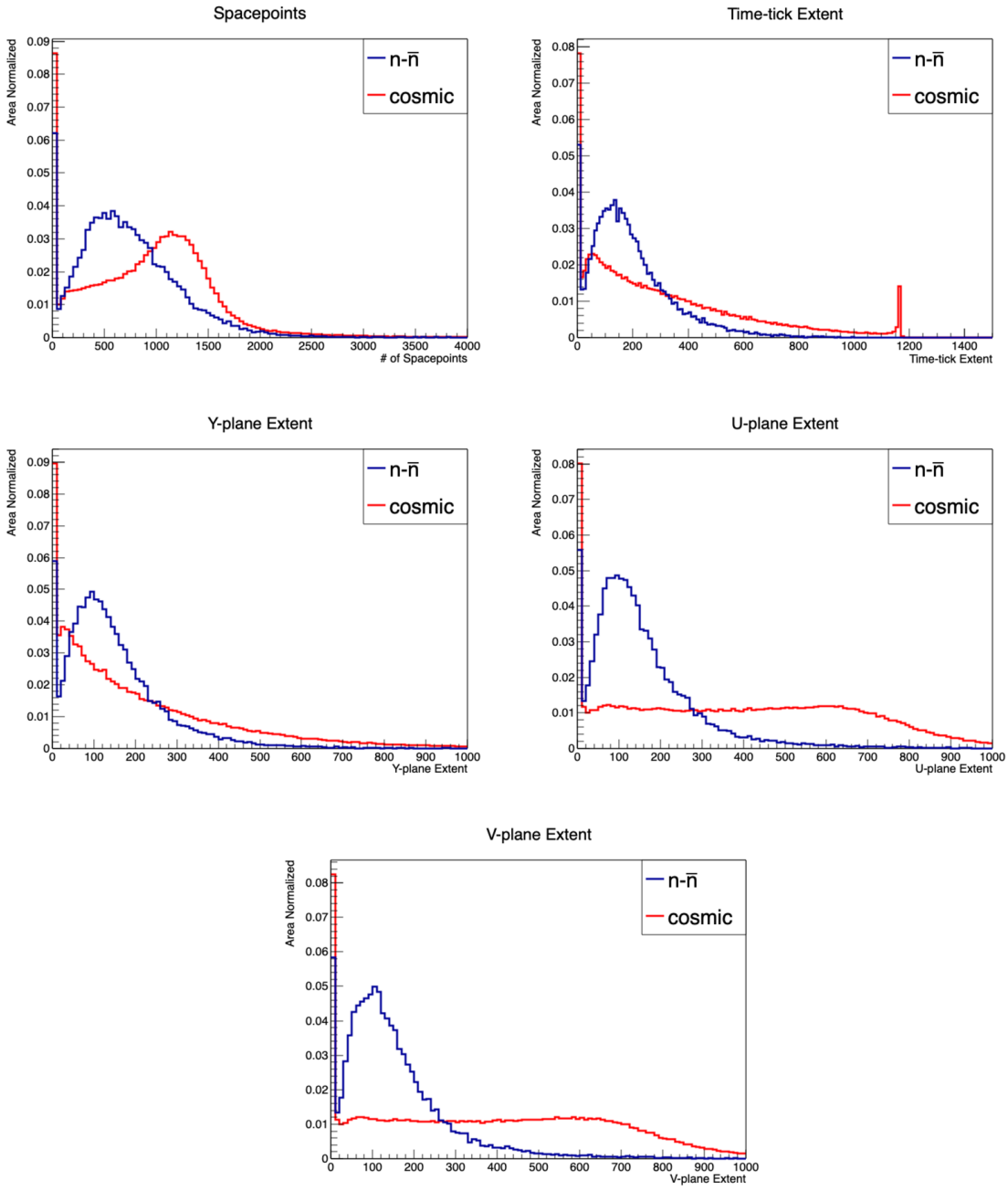


Figure 7.12: Distributions for topological variables from 2D cluster projections for $n-\bar{n}$ and cosmic clusters. These variables are used for the BDT pre-selection.

under 10%. The goal of pre-selection is to remove obvious cosmic clusters effectively, thus allowing CNN to learn more intricate features off more obscure background and signal data. Since roughly a dozen of cosmic clusters are found per event, the BDT pre-selection aimed to suppress the cosmic cluster rate to be under 10%, without losing significant $n - \bar{n}$ clusters.

The XGBoost framework [131] is used for BDT training for 300 iterations. The BDT training and validation during the training is performed using ≈23,000 $n - \bar{n}$ clusters and ≈200,000 cosmic clusters. By selecting clusters with higher BDT score than 0.1, the efficiency for $n - \bar{n}$ clusters and cosmic clusters is found to be 86.48% and 8.58%, using 26,525 and 237,062 clusters each, respectively. This test sample corresponds to the BDT development testing dataset in Tab. 7.3. Figure 7.13 shows the BDT score distributions for $n - \bar{n}$ and cosmic clusters. Signal $n - \bar{n}$ clusters

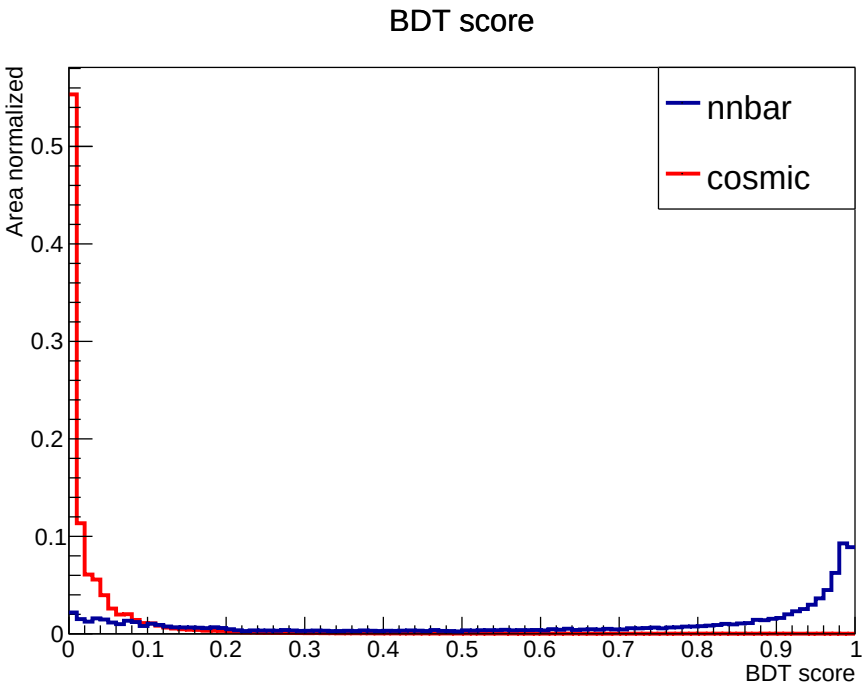


Figure 7.13: Classification performance of pre-selection BDT for $n - \bar{n}$ and cosmic clusters. The entries are area-normalized to one. The BDT cut is placed at BDT score>0.1. This shows the selection efficiency 86.48% for $n - \bar{n}$ (nnbar) clusters and 8.58% for cosmic clusters.

with different BDT scores are shown as 2D projections in Fig. 7.14. The left-hand-side figure shows the projections of an $n - \bar{n}$ cluster with a high BDT score (0.94.) The right-hand-side figure

shows the projections of an $n - \bar{n}$ cluster with a low BDT score (0.00145.) The right-hand-side cluster looks like a straight line track in terms of topology.

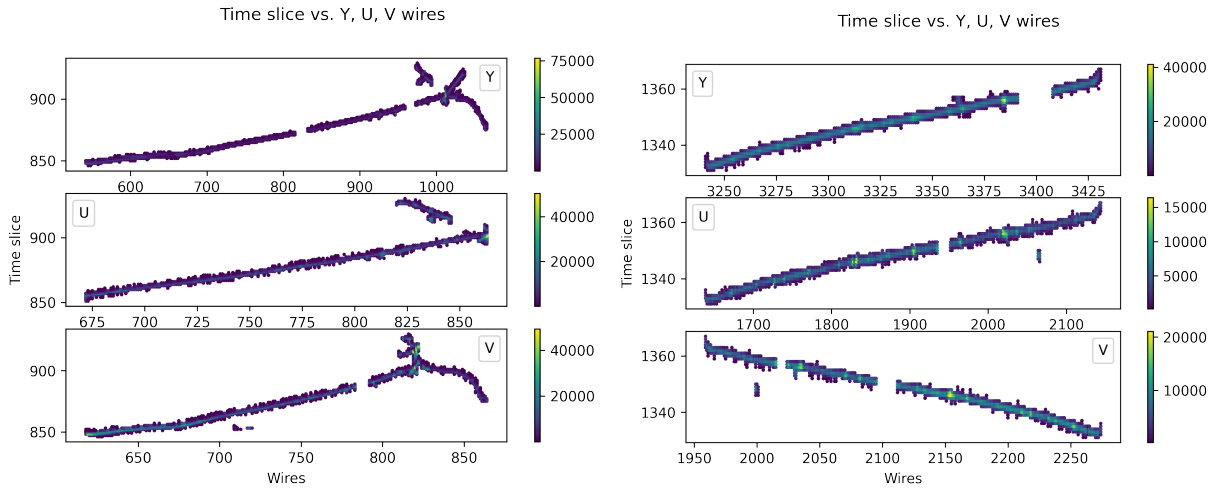


Figure 7.14: (left) $n - \bar{n}$ cluster passing the BDT pre-selection with BDT score 0.94. (right) $n - \bar{n}$ cluster failing the BDT pre-selection with BDT score 0.00145.

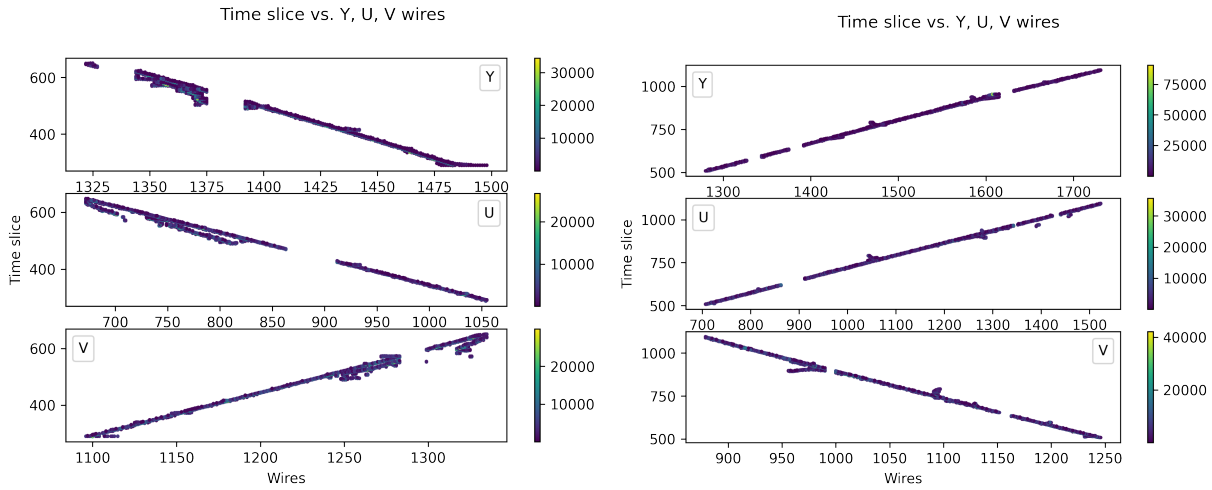


Figure 7.15: (left) cosmic cluster passing the BDT pre-selection with a BDT score of 0.35. (right) cosmic cluster failing the BDT pre-selection with BDT score 0.0063.

The topology-driven pre-selection BDT, since it does not take into account the charge depo-

sition associated with the space point, can not differentiate a cosmic muon from such a straight line signal topology. However, since we expect more than two daughter particles, typically in pair-annihilation, such a topology is not frequent in the $n - \bar{n}$ clusters. Cosmic clusters with different BDT scores are shown in Fig. 7.15. The figure on the left-hand side shows the projections of a cosmic cluster with a relatively high BDT score (0.35); the figure on the right-hand side shows the projections of a cosmic cluster with a low BDT score (0.0063). The right-hand side cluster seems to be more sizable than the left-hand side cluster in all planes and on the time-tick axis. The right-hand side cluster could be a cosmogenic muon track. The left-hand-side cluster, while it has a straight-line topology, also has signatures of shorter tracks or showers, forming a vertex-like kink. The clusters passing BDT pre-selection (with a higher BDT score than 0.1) are used for training and validating the CNN classifier, described in the next section.

7.4 Final selection

The CNN classification is used to develop the final selection of this analysis. The pre-selection BDT developed in the previous section is applied to the dataset used for CNN development, which corresponds to 95% of the total development set (see Tab. 7.3.) Therefore, only clusters with a higher BDT score than 0.1 are kept for CNN-based selection development. The 2D projections of the clusters on MicroBooNE's three planes (U, V, Y) are used as the CNN input. This data format naturally resembles 2D pixeled RGB images, which have been shown to be effectively classified by CNN image classification in numerous ML applications. Here the 2D projections are used to utilize the resemblance of those with 2D pixeled RGB images. However, one can alternatively format the same input in 4D space points (U, V, Y, T) and construct a CNN classifier working on 4D input. Representations of data can be chosen in various formats, and finding better representation for certain data types is an active area of research (e.g. graph representation of particle interactions in graph neural networks [132, 133].)

However, these 2D projections from the Wire-Cell clusters significantly differ in density from

the pixel RGB 2D images. The ratio of important pixels in these 2D projections is sub-percent of the full 2D image from wire readouts (the number of space points in a cluster is on the order of thousands, while the number of 2D pixels from wire × time-tick is the order of millions.) This input data type is called "sparse," and when they are treated the same way as a dense input, issues can arise as discussed in Sec. 4.3.1. Instead, the 2D projections are formatted as a columnar dataset with (N, 3) shape for each projection, where N refers to the number of the Wire-Cell space points, and the three entries in a row hold wire position (x position), time position (y position), and hit value per space point, respectively.

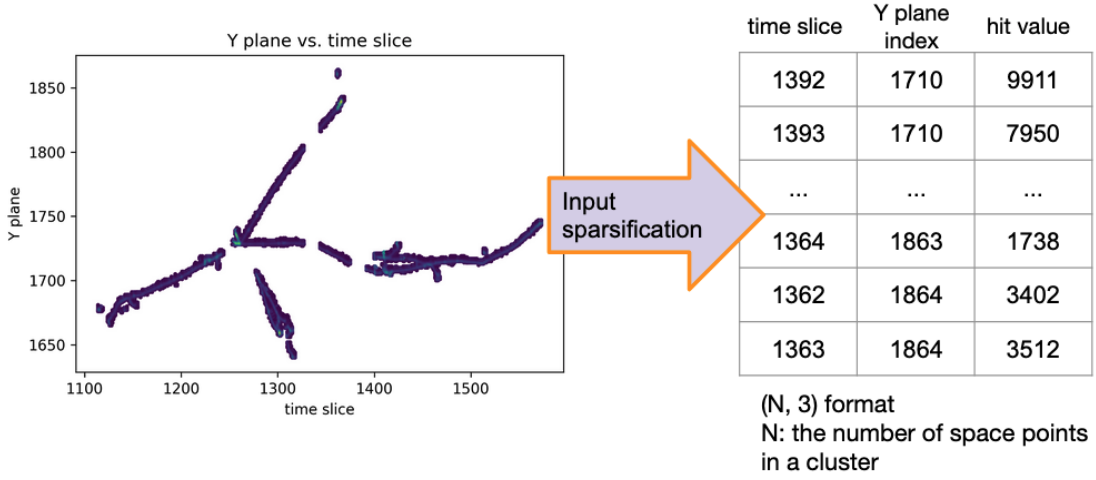


Figure 7.16: (left) 2D projection of signal cluster on conventional representation, (right) same cluster with on sparsified columnar representation.

This way of formatting the CNN input has some major advantages. This can store the localized cluster rather than saving the entire readout; thus, it is highly memory-efficient. Also, there is no need to assign a frame with fixed dimensions for the input (often 2D CNN inputs need to be fixed square, such as 224×224 or 64×64.) Thus, the size of these 2D projections can be as small or large as they are; there is no need for lossy down-sampling or up-scaling. The columnar (sparse) representation of the 2D projections is shown in Fig. 7.16 schematically. The generalized convolution used in *Minkowski Engin* [89]] can accommodate this sparse representation for 2D convolution as

described in Sec. 4.3.1. All three planes are used for CNN classifier training. 80% of the CNN development dataset is used for training, 10% is used for validation, and 10% is used for testing, as shown in Tab. 7.3 (the testing set is separately reserved for CNN inference prior to the application on the central value MC.) The total training size, including both $n - \bar{n}$ and cosmic clusters, are 1.5 M clusters. As a general rule of thumb, a large training set in CNN can add more quality to training by avoiding the possibility of overfitting by repeatedly sampling on a small dataset. Training loss and validation accuracy are monitored during the CNN training as they are the metrics of training quality. At the end of each iteration, the training loss is calculated and the weight of CNN is updated through backpropagation. Training loss, as discussed in Sec. 4.1, is the classification error in the training batch and is minimized over iterations. The training loss is monitored every 20 iterations, and the training batch size is set at 256. The validation accuracy is the binary classification accuracy on the validation sample set. The validation accuracy is monitored every 200 iterations. Figure 7.17 shows the training loss and validation accuracy during the training.

For CNN-based classification, 20,000 iterations were chosen by monitoring training loss and validation accuracy. 20,000 iterations using batch size 256 means sampling of ($20000 \times 256 = 5.12 \times 10^6$) out of the training set with the size of 1.5×10^6. Thus the entire training sample is used with the frequency of $\frac{5.12 \times 10^6}{1.5 \times 10^6} = 3.4$; this number is also called an epoch. If the number of epochs is too small (<1), the training set is not fully used; if the number of epochs is too high, the risk of overfitting increases.

The CNN classifier after 20,000 iterations of training is again validated on an exclusively reserved 10% testing dataset. The result is shown in Fig. 7.18. The accuracy was evaluated at 97.7% on this testing dataset, where 100,906 $n - \bar{n}$ clusters and 90,499 cosmic clusters are tested. When a cluster (either $n - \bar{n}$ or cosmic) is tested through the trained CNN classifier, the 'softmax' scores are returned for each categorical class at the final layer of the CNN classifier. These softmax scores are a vector with the length of the number of categorical classes (two for binary classification as this CNN classifier) and are normalized so that they are summed to one. The signal score and the background score here mean the softmax scores for $n - \bar{n}$ class and cosmic class, respectively.

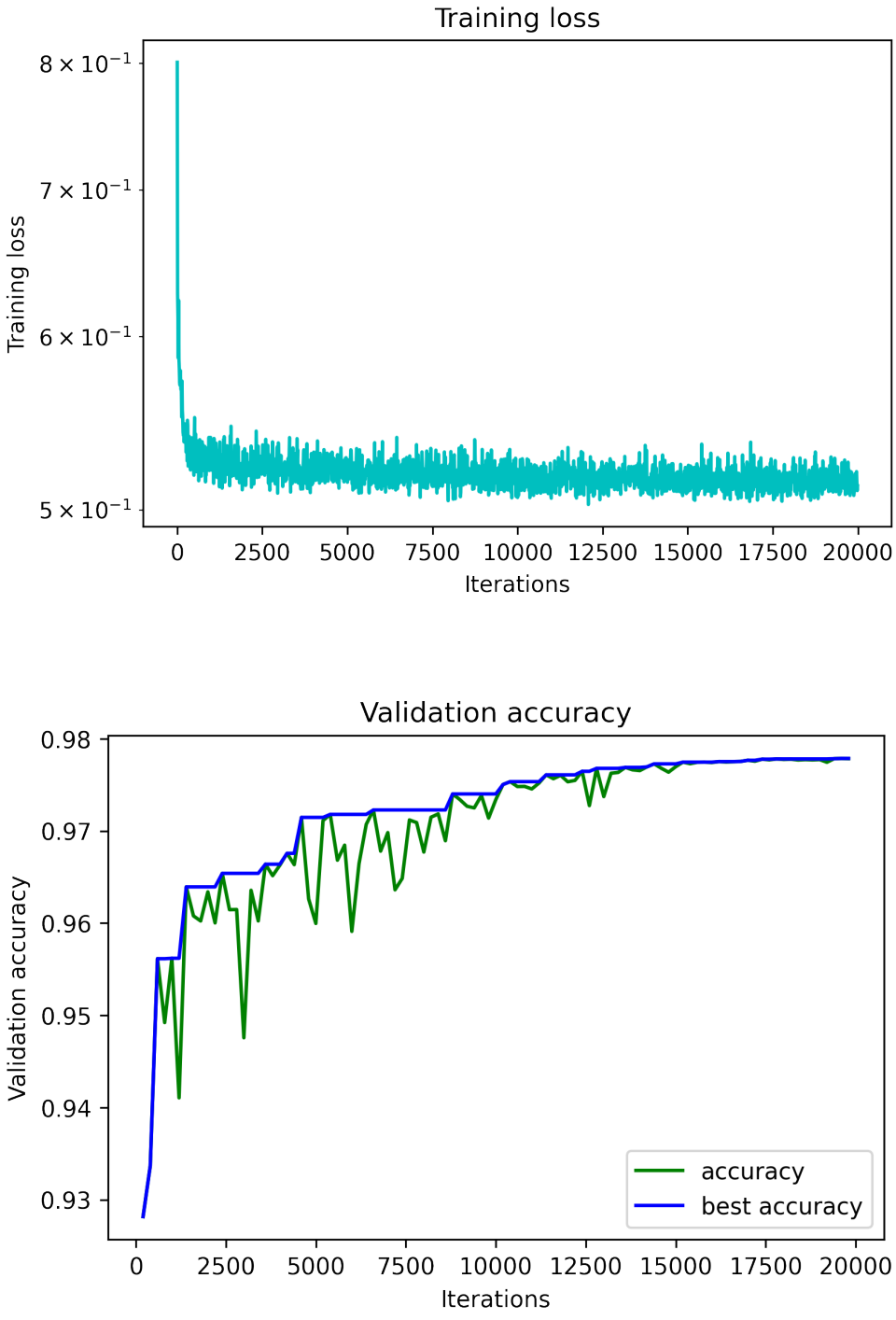


Figure 7.17: (top) Monitored training loss over iterations. (bottom) Monitored validation accuracy over iterations. The current accuracy is shown as green solid line, the best accuracy is shown as blue solid line.

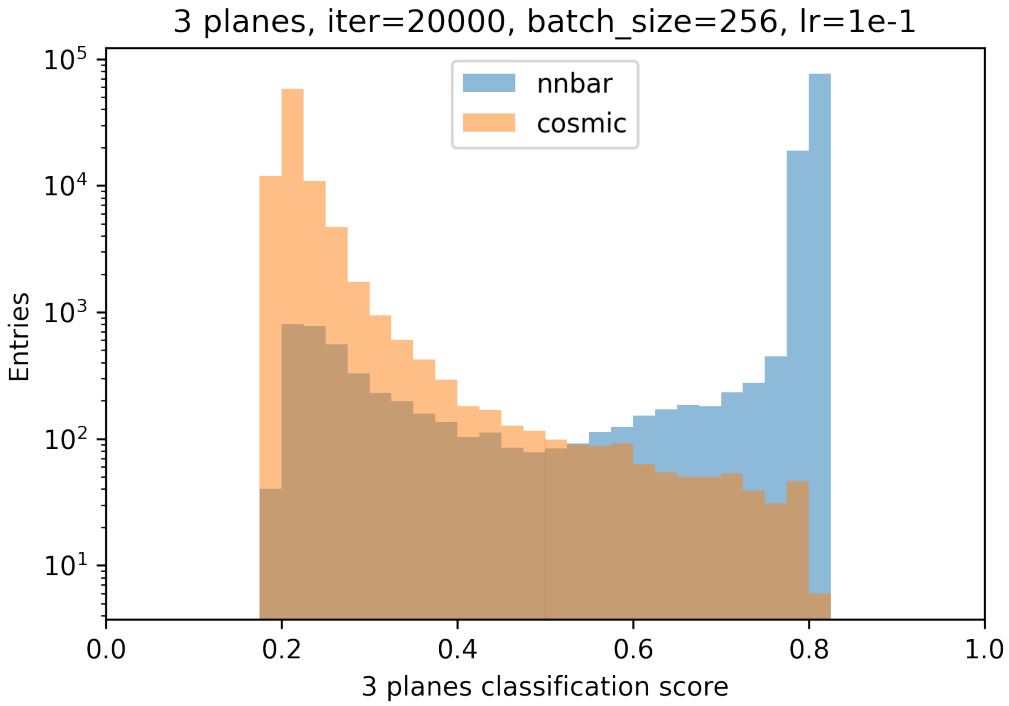


Figure 7.18: CNN signal score distribution after 20000 iterations of training. A dedicated testing set (100,906 $n-\bar{n}$ clusters and 90,499 cosmic clusters) is used for the preliminary inference.

The classification prediction is made by choosing the class with the highest softmax score. The classification accuracy shows the rate of correct prediction compared to the truth label.

On the other hand, the softmax score can be utilized to place a selection cut. Fig. 7.18 shows the $n-\bar{n}$ signal softmax scores for both $n-\bar{n}$ clusters and cosmic clusters in the testing dataset (subset in development dataset.) It is clearly shown that the CNN classifier can differentiate $n-\bar{n}$ clusters from cosmic clusters.

7.4.1 Final selection optimization

The CNN score cut for the final selection is optimized with respect to the projected sensitivity using the CV sample. As the search region statistics corresponds to a size of 10% of the CV sample, the sensitivity calculation follows this assumption. Figure 7.19 shows the CNN score distributions for $n-\bar{n}$ and cosmic clusters in the CV sample using the CNN classifier trained on the development

dataset.

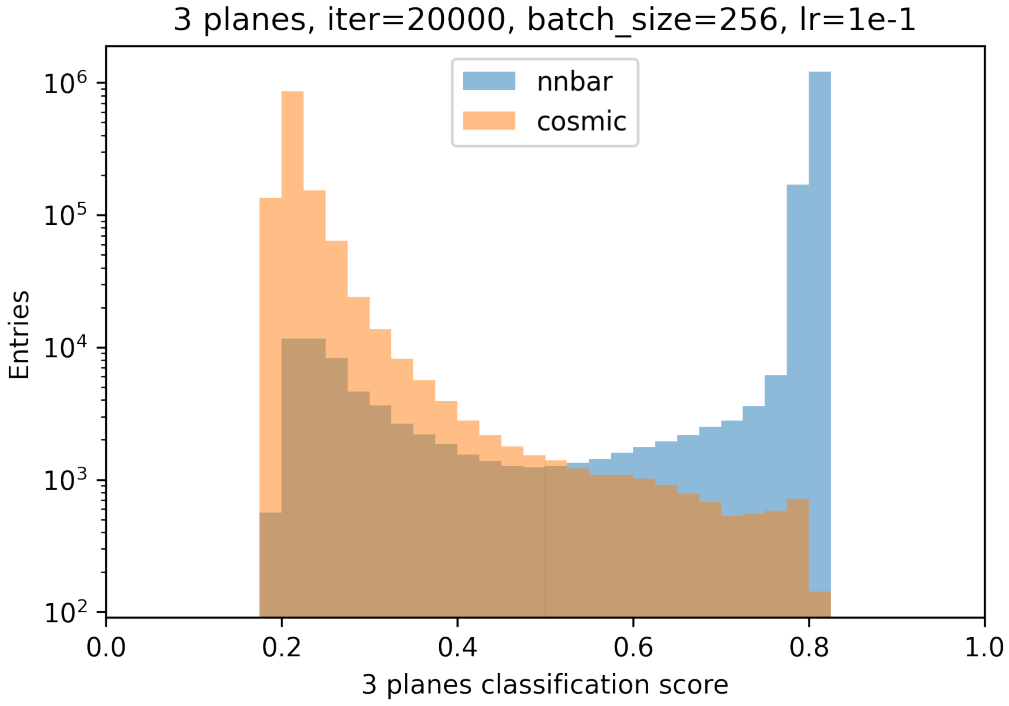


Figure 7.19: CNN signal score distribution for $n - \bar{n}$ and cosmic clusters using CV sample. 1,633,525 $n - \bar{n}$ clusters and 1,618,827 cosmic clusters are classified by the trained CNN classifier.

After a CNN score cut, the final selected events are the events that include the clusters in the final selection (for example, if more than one cluster from one event passes the final CNN cut, the number of events is counted instead of the number of clusters[5].) The final efficiency for $n - \bar{n}$ events and background events are shown in Fig. 7.20 on the curve with respect to varying CNN score cuts.

For these selection efficiencies, preliminary sensitivity values are calculated using the TRolke package [134] in ROOT. Below are the assumptions for this calculation.

- The search region statistics are estimated to be 10% of the CV statistics.

[5]This is found to be negligible for background events and sub-percent for signal events. Table 7.5 shows the exact numbers of entities for the CV sample.

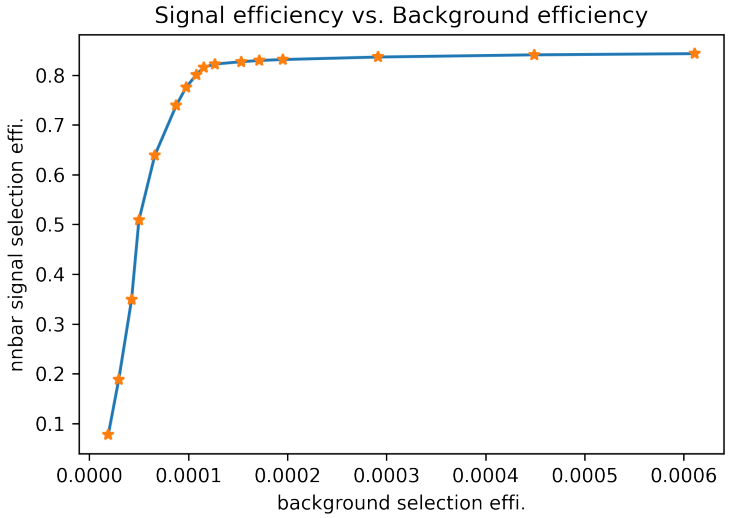


Figure 7.20: Signal selection efficiency versus background selection efficiency after final selection, varying the CNN score cut, where the cut values are [0.77, 0.78, 0.79, 0.795, 0.796, 0.797, 0.798, 0.7985, 0.799, 0.7995, 0.8, 0.801, 0.802, 0.803, 0.804, 0.805]. The stars in the figure represent the cut values.

- The uncertainty in signal selection efficiency is assumed to be 15%, which includes the statistical and systematic uncertainties.
- Only statistical uncertainty in the background selection is assumed.

The statistical uncertainty in signal selection is almost negligible due to the high selection efficiency on the high statistics in the CV sample. Thus, the 15% total uncertainty in signal selection assumes 15% systematic uncertainty. On the other hand, the statistical uncertainty in background selection is significant due to the very low selection efficiency. The optimal CNN cut is found at 0.8, achieving 73.6±0.28 % signal efficiency and 8.77e-3±2.33e-3 % background efficiency. The background statistical error assumes the 10% of the CV sample statistics. The statistical uncertainties for signal and background selection are shown in Tab. 7.4 along with the background estimate and the calculated sensitivity around the optimal CNN cut.

The events (2.3 ms intervals) in the signal and background CV sample at the production stage have similar statistics (1.6 million each.) At the cluster reconstruction stage, the number of clusters

CNN cut	Signal effi.	Bkg. effi.	Bkg. estimate	Sensitivity
0.797	0.827 ± 0.0003	1.53e-4 ± 9.7e-6	24.8 ± 1.6	2.62e+25 yrs
0.798	0.822 ± 0.0003	1.27e-4 ± 8.8e-6	20.5 ± 1.4	2.83e+25 yrs
0.799	0.801 ± 0.00031	1.08e-4 ± 8.2e-6	17.5 ± 1.3	2.98e+25 yrs
0.8	0.736 ± 0.00034	8.77e-5 ± 7.4e-6	14.2 ± 1.2	2.99e+25 yrs
0.801	0.639 ± 0.00038	6.61e-5 ± 6.4e-6	10.7 ± 1.0	2.95e+25 yrs
0.802	0.508 ± 0.00039	5.0e-5 ± 5.6e-6	8.1 ± 0.9	2.65e+25 yrs
0.803	0.349 ± 0.00037	4.26e-5 ± 5.1e-6	6.9 ± 0.8	1.95e+25 yrs

Table 7.4: Preliminary sensitivity around the optimized CNN cut 0.8. The signal and background efficiencies, and the background estimate are shown. The errors shown indicate finite-MC statistical uncertainty.

can outnumber the input events, especially for the cosmic sample, approximately by a factor of 9. This is consistent with the cluster multiplicity shown in Fig. 7.11.

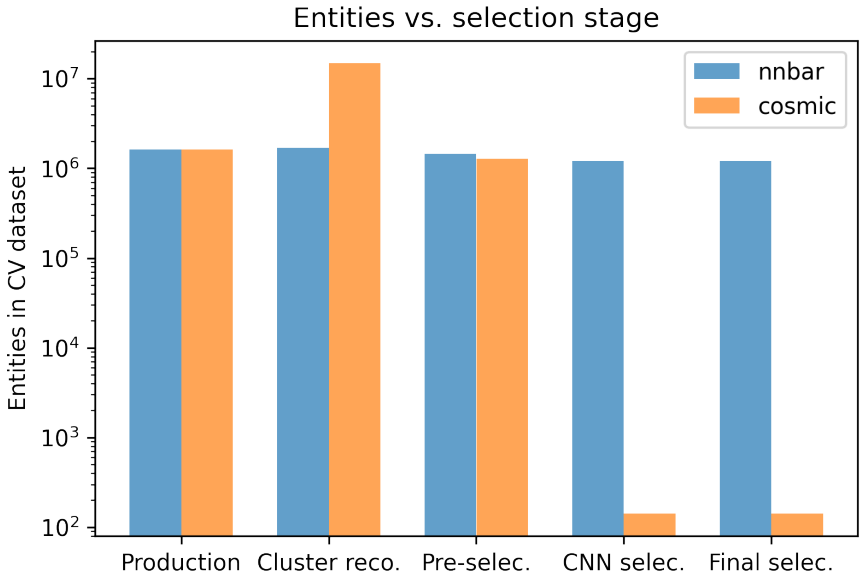


Figure 7.21: Entities during the selection stages using CV sample. The blue (orange) histograms represent the number of entities for the $n - \bar{n}$ (cosmic) sample at each stage.

During the pre-selection, the number of cosmic clusters is reduced, and the number of $n - \bar{n}$ clusters and cosmic clusters become the same order. Then, the number of cosmic clusters is strongly suppressed after the final selection, while the number of $n - \bar{n}$ clusters remains high.

Figure 7.21 shows the number of entities at stages of reconstruction and selection. The exact numbers of these entities and the selection efficiency are shown in Tab. 7.5.

Entities	$n - \bar{n}$	cosmic
Events	1,633,525	1,618,827
Reconstructed clusters	1,684,516	14,857,224
Clusters (after pre-selection)	1,455,214	1,283,074
Clusters (after final-selection)	1,207,153	142
Events (after final-selection)	1,202,281	142
Selection efficiency	0.736	8.77e-5

Table 7.5: The number of entities for each reconstruction and selection stage. The numbers are evaluated using the central value samples. The selection efficiency indicates the ratio between the 'Events' and 'Events (after final-selection)'.

7.4.2 Selected events: topologies & kinematics

In this section, 2D projections of the clusters that pass the final selection are shown for well-classified cases and otherwise. Signal and background events passing the final selection are further reconstructed using the Wire-Cell (WC) full event reconstruction to investigate the kinematics of these clusters, as post-analysis validation. Reconstructed variables are compared with truth variables.

7.4.2.1 Selected event 2D projections

Figure 7.22 shows example 2D projections of $n - \bar{n}$ clusters that passed the final selection with a high CNN score. Notice that the U-plane projection of the left-hand-side cluster is very small, thus looks discrete due to binning. The Y-plane and V-plane projections are well representing the signal topology; thus, this event scores highly on CNN classification with all-planes inference.

The right-hand-side cluster demonstrates the inefficacy of U plane and V plane wires. However, the cluster is still highly scored on the CNN classification (as signal-like).

Figure 7.23 shows the example 2D projections of cosmic clusters failing the final selection after

passing the BDT pre-selection. The left-hand-side cluster shows a muon track; the right-hand-side cluster seems to have two prongs, but this could be crossing two cosmic muon tracks reconstructed as one cluster.

Some signal clusters failed to pass the final selection as shown in Fig. 7.24. The left-hand-side cluster shows no reconstructed space-points on the Y-plane, and insignificant reconstruction on the U-plane. Both U-plane and V-plane signatures are small in time and wire axes. The right-hand-side cluster shows small extents on both axes as well. These clusters do not manifest the signature "star-like" topology and are classified with a low CNN score.

Cosmic clusters passing the final selection are shown in Fig. 7.25. The left-hand-side cluster shows a highly energetic cosmogenic interaction, including shower activities. As discussed in Sec. 7.1.3, highly energetic cosmogenic interactions accompanied by showers can be tricky background candidates to distinguish from the signal. The right-hand-side cluster seems to be a cluster reconstructed with multiple energetic cosmogenic interactions.

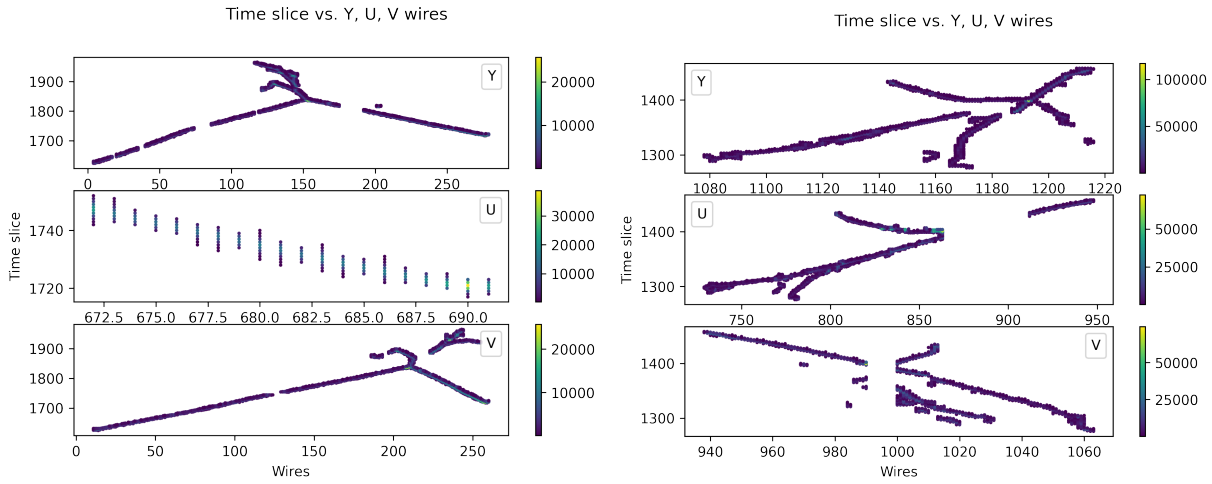


Figure 7.22: (left) $n - \bar{n}$ cluster that passed the BDT pre-selection with BDT score 0.981 and the CNN selection with CNN score 0.8036. (right) $n - \bar{n}$ cluster that passed the BDT pre-selection with BDT score 0.947 and the CNN selection with CNN score 0.804.

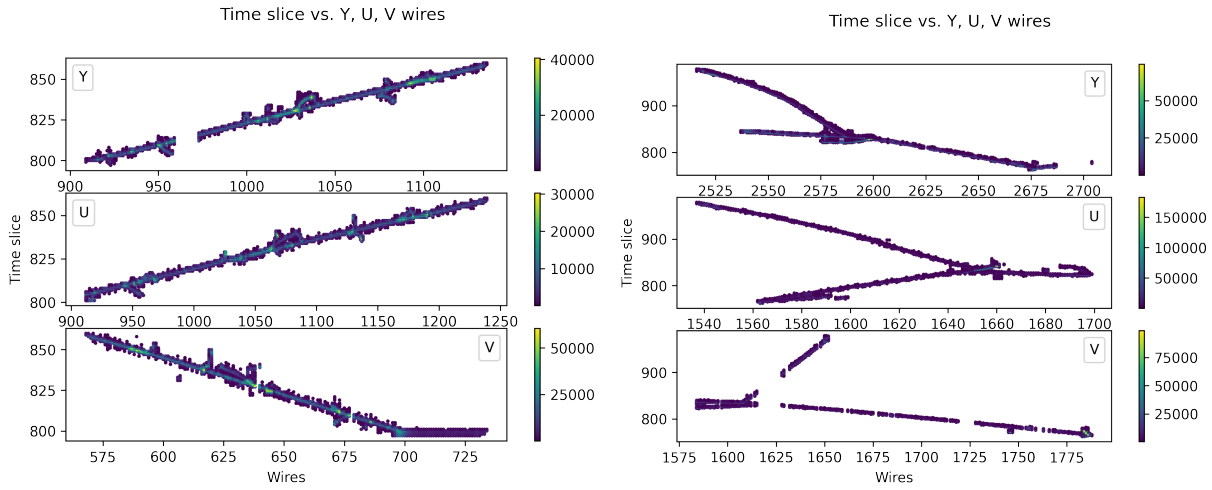


Figure 7.23: (left) cosmic cluster that passed the BDT pre-selection with BDT score 0.178 and failed the CNN selection with CNN score 0.204. (right) cosmic cluster passing the BDT pre-selection with BDT score 0.977 and failed the CNN selection with CNN score 0.205.

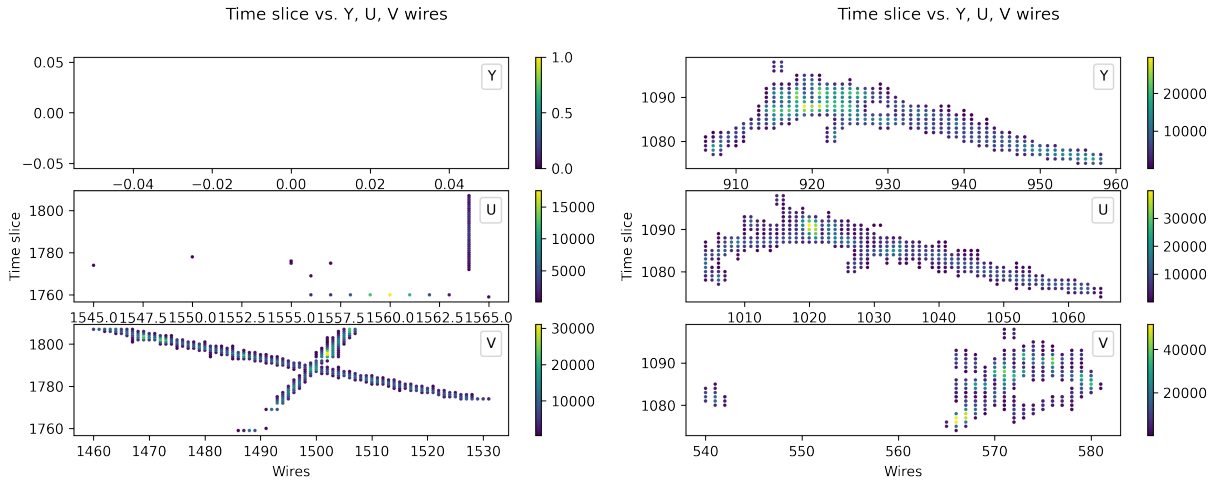


Figure 7.24: (left) $n-\bar{n}$ cluster that passed the BDT pre-selection with BDT score 0.177 and failed the CNN selection with CNN score 0.737. (right) $n-\bar{n}$ cluster that passed the BDT pre-selection with BDT score 0.221 and failed the CNN selection with CNN score 0.662.

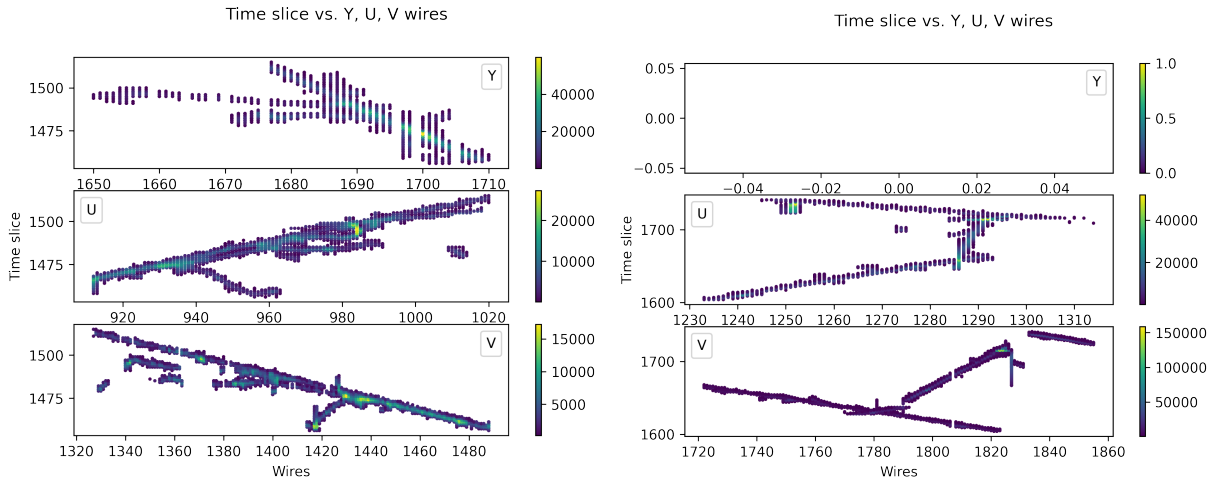


Figure 7.25: (left) cosmic cluster passing the BDT pre-selection with BDT score 0.663 and passing CNN selection with CNN score 0.805. (right) cosmic cluster passing the BDT pre-selection with BDT score 0.927 and passing the CNN selection with CNN score 0.801.

7.4.2.2 Full event reconstruction on selected events

After the final selection, 142 background events from the CV sample (corresponding to 3723 seconds) remain. These background events, along with signal events that pass the final selection, are processed for full-event reconstruction using the Wire-Cell (WC) reconstruction. The corresponding histograms from the GENIE final state of the signal simulation are shown alongside the reconstruction.

The reconstructed energy distributions for signal and background are shown in Fig. 7.26, as well as the true visible energy calculated at the final state of the GENIE. The visible energy is defined as the sum of kinetic energies of protons over 50 MeV and the relativistic energies of pions. Compared to the visible energy, the reconstructed energy deviates to the lower energy region by a few hundred MeVs. The agreement is reasonable considering that the GENIE final state does not include any reinteraction in argon or detector effect. The two reconstructed energy distributions for the signal and the background show a clear difference in the energy range. It is suggested that we can further reject background events using kinematic variables from the current final selection,

which already shows the 8.77e-5 efficiency for background events.

The distributions of momentum scalar sum[6] over reconstructed particles for signal and background are shown in Fig. 7.27, as well as the final state momentum scalar sum over visible particles at the GENIE final state. The visible particles include protons with kinetic energy greater than 50

[6]Reconstructed momentum vector sum can be also a useful variable to look at. At the moment, the Wire-Cell reconstruction does not save momentum vectors of particles after the reconstruction. Developed for neutrino event analyses, the WC reconstruction saves limited kinematics. Modifications can be applied in the near future to enable a more comprehensive study on reconstructed kinematics.

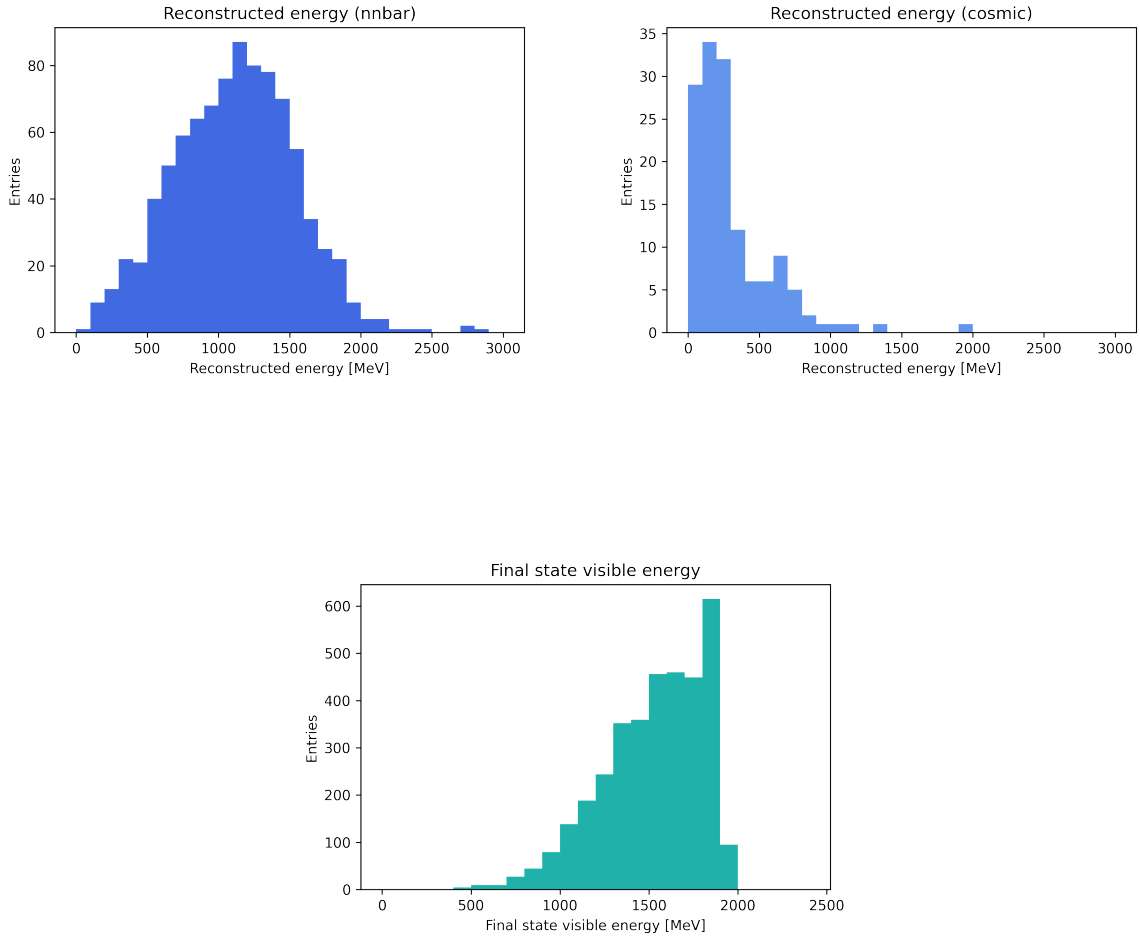


Figure 7.26: Reconstructed energy distributions of the signal ('nnbar') clusters (top left), background ('cosmic') clusters (top right) after final selection. The true visible energy from the final state of GENIE is shown (bottom).

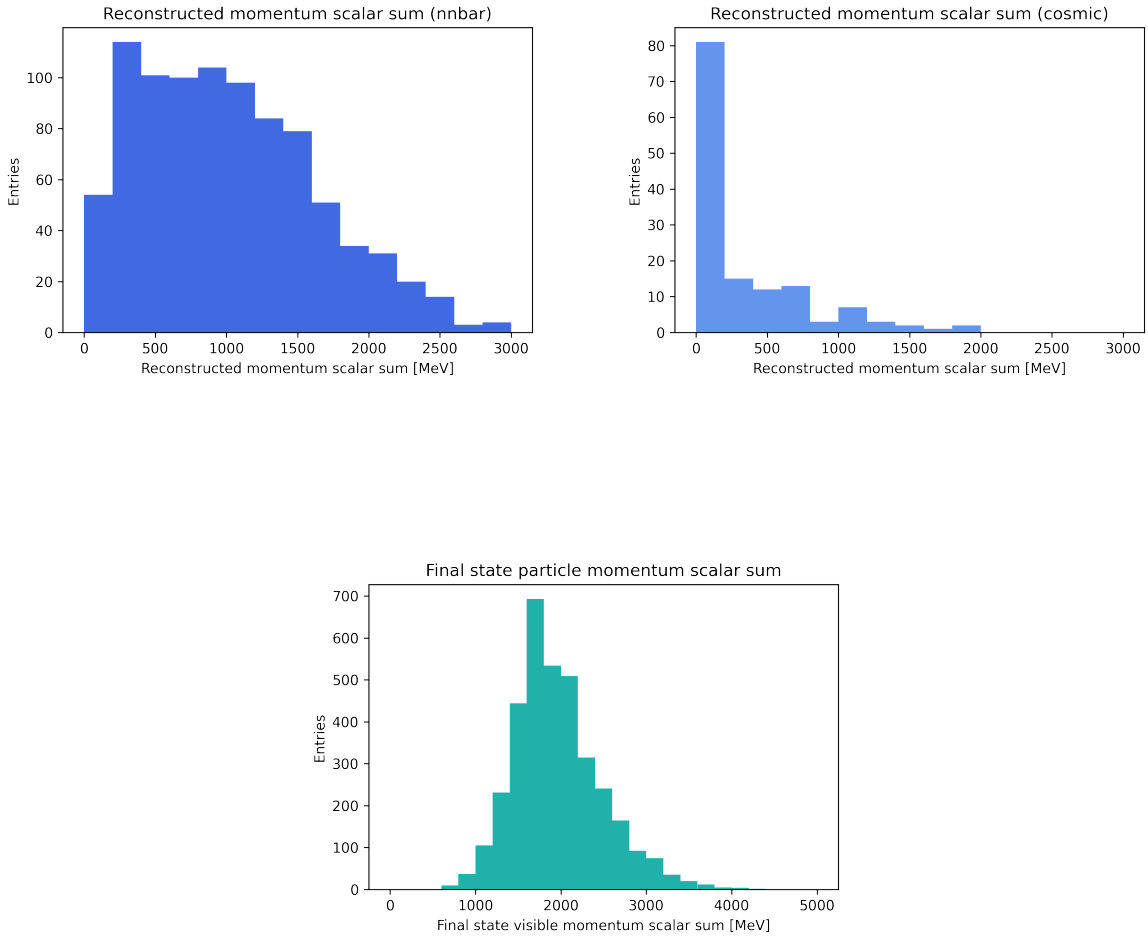


Figure 7.27: Reconstructed momentum scalar sum distributions of the signal ('nnbar') clusters (top left), background ('cosmic') clusters (top right) after final selection. The momentum scalar sum from GENIE final states is shown (bottom.)

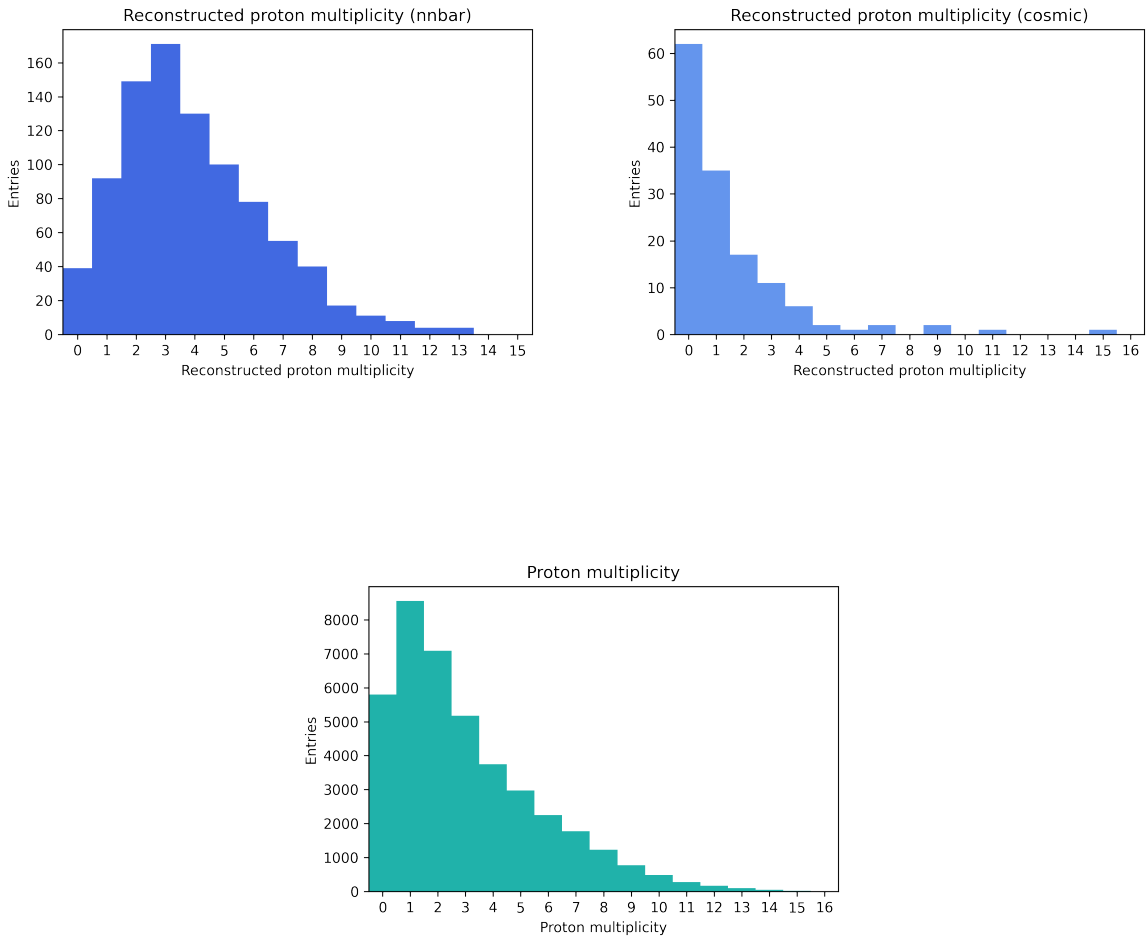


Figure 7.28: Reconstructed proton multiplicity distributions of the signal (‘nnbar’) clusters (top left), background (‘cosmic’) clusters (top right) after final selection. The true proton multiplicity from the GENIE final states is shown (bottom).

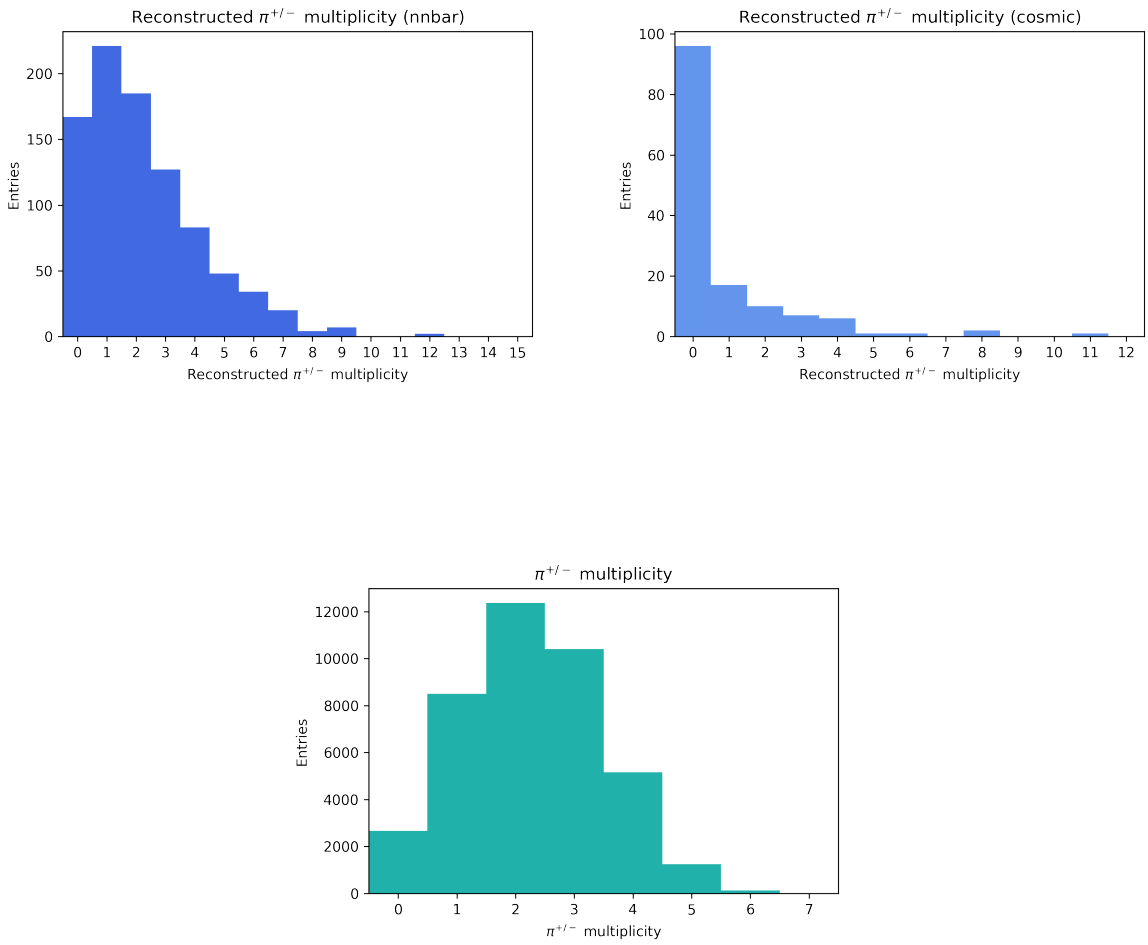


Figure 7.29: Reconstructed charged pion multiplicity distributions of the signal (‘nnbar’) clusters (top left), background (‘cosmic’) clusters (top right) after final selection. The true charged pion multiplicity from the GENIE final state is shown (bottom).

MeV, charged pions, and neutral pions. Here, similar observations can be said from the reconstructed energy. The final state and the signal reconstruction are different by a few hundred MeVs, and the signal reconstruction and the background reconstruction show a significant difference.

The full-event reconstruction of the WC is equipped with particle identification for protons, charged pions, muons, and electrons. The multiplicity of protons for signal and background reconstruction is shown in Fig. 7.28, compared to the true number of protons in the final state of GENIE.final state of GENIE. The multiplicity of charged pions (inclusively for π^+ and π^-) for signal and background reconstruction is shown in Fig. 7.28, with the true number of charged pions in the GENIE final state.

7.5 Systematic uncertainties

Uncertainties from the generator (GENIE) modeling, hadron reinteraction (via GEANT4), and detector modeling can have a significant impact on $n-\bar{n}$ signal selection efficiency. The uncertainty of signal selection for variations in the generator and detector modeling is studied with exclusive samples that are generated with different models. Inelastic reinteractions of hadrons (π^+, π^-, p) in the liquid argon volume are simulated by GEANT4; the cross-sections of these hadronic reinteractions are varied to account for the corresponding systematic uncertainty. In this section, evaluations of these systematic uncertainties on $n - \bar{n}$ signal selection efficiency are discussed.

7.5.1 Uncertainties from the $n - \bar{n}$ generator modeling

GENIE via uboonecode supports Bodek-Ritchie (BR) and Local Fermi gas (LFG) models for nuclear modeling. For the final state interaction (FSI), the empirical modeling *hA* and the full-cascade modeling *hN* are available. Combining these variations, four modeling variations in GENIE are possible (hA_LFG, hN_LFG, hA_BR, hN_BR) in generating $n - \bar{n}$ signal simulation. The final state π^0 multiplicity in $n - \bar{n}$ interaction with respect to those combinations are shown in Fig. 7.30, as an example of how these uncertainties affect the signal. Final-state kinematics is

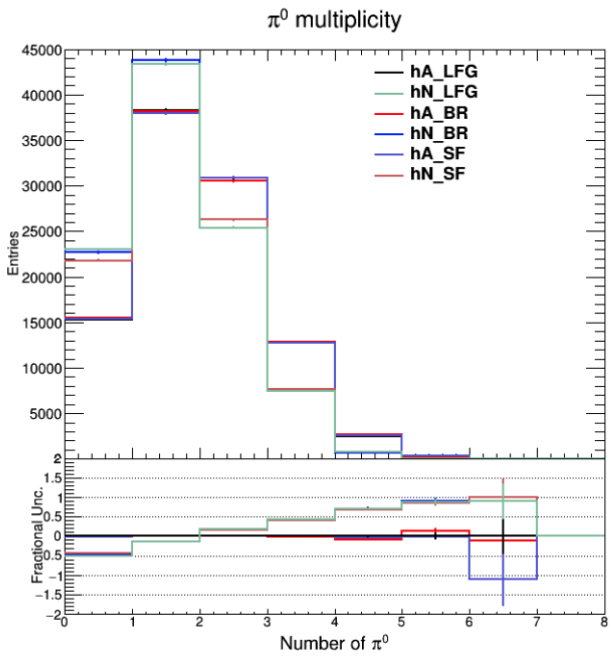


Figure 7.30: π^0 multiplicity in GENIE final state for different FSI modeling and nuclear modelings. hA and hN are two FSI modelings, LFG stands for Local Fermi gas model, BR stands for Bodek-Ritchie model, and the SF stands for spectral function model.

also expected to be affected. Such variations also affect the multiplicities of $\pi^{+/-}$ and protons. The default GENIE via uboonecode uses the hA_LFG setting. The signal efficiency using samples with other possible variations is being evaluated.

7.5.2 Uncertainties from reinteraction of hadronic final states

Charged hadrons can interact with external argon nuclei while traveling inside the liquid argon volume. This reinteraction of charged hadrons through elastic and inelastic scatterings is simulated using GEANT4 [112]. The uncertainty of these scatterings of protons and charged pions can be significant, especially when there are many charged hadrons in the final state, such as in $n-\bar{n}$ interactions. Currently, the uncertainty of the hadron reinteraction is evaluated using an event reweighting scheme [135]. In the event reweight scheme, the reinteraction cross-sections are varied around 30%, the final state particles are weighted accordingly by the change of cross-section from the nominal cross-section, and the events are weighted as the compounded effect from the collec-

tive particle reweight. The impact of hadron reinteraction uncertainty on $n - \bar{n}$ signal efficiency is being evaluated using this method.

7.5.3 Uncertainties from the detector modeling

The evaluation of the systematic uncertainty from detector modeling follows MicroBooNE's approach; the details on the approach can be found in [136]. There are four major categories of detector systematic uncertainties: 1) variations in the electron-ion recombination modeling, 2) variations related to the light yield (LY) and propagation simulation, 3) variation in the space-charge effect in the liquid argon continuum, and 4) variation in the TPC waveform. The light response variation includes light yield reduction, change in light attenuation length, and change of Rayleigh length. Variations in the TPC waveform include waveform modifications using parameterized functions obtained by data/MC comparison of TPC hit reconstruction. x, y, z are the drift direction, the vertical direction, the beam direction positions, and $\Delta x, \Delta y, \Delta z$ are the size of the reconstructed hit on each axis. The list of detector variations studied is as follows.

- Recombination (Recomb2) : Electronion recombination simulation using an alternative recombination modeling "Recomb2"
- Light yield down (LYDown) : Applying 25% uniform reduction in light yield in the PMTs
- Light yield attenuation (LYAtt) : Applying a 10 meter light attenuation length to account for distance-dependent mismodeling
- Light yield Rayleigh (LYRayleigh) : Modifying the Rayleigh scattering length to 120 cm (from the default 60 cm)
- Space charge effect (SCE) : Simulation using an alternative data-driven space charge map
- Wire modification x (WModX) : Waveform modification by applying a parameterized function of x

- Wire modification y/z (WModYZ) : Waveform modification by applying a parameterized function of y/z
- Wire modification θ_{xz} (WModThetaXZ) : Waveform modification by applying a parameterized function of θ_{xz} (where $\theta_{xz} = arctan(\Delta x/\Delta z)$)
- Wire modification θ_{yz} (WModThetaYZ) : Waveform modification by applying a parameterized function of θ_{yz} (where $\theta_{yz} = arctan(\Delta y/\Delta z)$)

For each variation, a new independent signal MC sample is generated. The final selection is applied to these variation samples, and signal efficiency for each variation MC sample is obtained. Table 7.6 shows the signal efficiency difference for each variation compared to the CV sample.

Detector systematics		signal effi. diff. (%)	signal effi. diff. + stat. unc. (%)
Recombination	Recomb2	0.479 ± 0.0023	0.481
Light yield	LYDown	1.06 ± 0.005	1.1
	LYAtt	0.175 ± 0.0008	0.18
	LYRayleigh	0.142 ± 0.0007	0.14
Space charge effect	SCE	0.0847 ± 0.0004	0.085
Waveform modeling	WModX	0.564 ± 0.0024	0.57
	WModYZ	3.63 ± 0.015	3.6
	WModThetaXZ	2.4 ± 0.011	2.4
	WModThetaYZ	4.88 ± 0.023	4.9
Total			6.7

Table 7.6: The signal efficiency differences between detector variations from CV sample is shown in 'signal effi. diff. (%)' column with the statistical uncertainties. The statistical uncertainties are added to the signal efficiency differences in 'signal effi. diff. + stat. unc. (%)' column to represent a conservative estimate of the detector variation systematic error. The square sum this column shows 6.7%, estimating the total detector uncertainty.

The signal efficiency difference and the statistical uncertainty from the finite size of the MC variation sample are added to obtain a conservative approximation of each variation systematic uncertainty. A total of 6.7% uncertainty in the detecter effect is obtained from the square sum of those conservative uncertainties.

7.6 Sensitivity evaluation

The sensitivity for MicroBooNE's $n-\bar{n}$ search is calculated assuming 372s of exposure, corresponding to 3.13 neutron·years. The Rolke method [134, 137] with Gaussian background is used to evaluate the sensitivity. This method, in comparison to the multidimensional integral approach followed in Sec. 5.5, has been favored by more recent $n-\bar{n}$ searches in large detectors [41, 51]. The Rolke method profiles the likelihood for signal events for the given signal selection efficiency and the background estimate, rather than performing a numerical multidimensional integral, as described in Sec. 5.5.

Signal selection unc.	Sensitivity
Stat. only (0.047%)	3.09 e+25 yrs
Stat. + detector (6.7%)	3.07 e+25 yrs
Total uncertainty (15%)	2.99 e+25 yrs

Table 7.7: 90% C.L. sensitivity for $n-\bar{n}$ oscillation lifetime assuming 372 second exposure in MicroBooNE. For the background uncertainty, finite-MC statistical uncertainty is used.

The estimation of total uncertainty for signal selection is not yet complete; the sensitivities considering only statistical uncertainty, detector effect, and 15% assumption as the total uncertainty are shown in Tab. 7.7 using the Rolke method.

Conclusion

The work presented in this thesis ultimately aims to help establish the future $n - \bar{n}$ oscillation search in DUNE. In Ch. 2, the theoretical motivation for searching for baryon number violation (BNV) in $|\Delta\mathcal{B}| = 2$ is discussed, and the experimental $n - \bar{n}$ searches until today are summarized. In Ch. 5, the projected sensitivity after DUNE's 10 years of data-taking is presented using an ML-based event selection scheme. In Ch. 6, novel trigger schemes are explored to enable data selection for such rare interactions over continuous readout in a large-scale LArTPC using CNN-based triggering algorithms. In Ch. 7, the search for $n - \bar{n}$ in MicroBooNE, developed to demonstrate the ML-based approach of such a search analysis in a LArTPC, is presented. The remaining pieces of MicroBooNE analysis include the complete systematic uncertainty evaluation and the fake data analysis for final validation. Evaluation of the signal uncertainties in generator modelings and hadronic re-interaction modeling have been finalized and are entering MicroBooNE collaboration review, prior to unblinding, at the time of submission of this thesis. Once this search is performed in MicroBooNE, it will represent the first-ever search for $n - \bar{n}$ in a LArTPC. Given the small exposure of available data in MicroBooNE, this search does not aim to set a competitive limit to the current existing results. However, this work will serve as an important reference for the future search in DUNE, which shares the same LArTPC technology.

Some of the possible ways to improve the DUNE analysis are already discussed in Sec. 5.7, but there is more to learn from the development of the MicroBooNE analysis. First of all, the generalized sparsed CNN, which is used in MicroBooNE analysis, can be adopted for the DUNE analysis. This is expected to enable training on much larger data statistics than the current analysis

(millions versus tens of thousands). This will improve the quality of the CNN classifier. Secondly, different reconstruction paradigms can be explored to maximize the sensitivity. A DL-based high-energy event reconstruction is being developed by DUNE collaborators [138]; in the future, various options as alternatives to Projection Matching Algorithm (PMA) reconstruction will also be available [121, 138–140]. Especially, reconstruction with a strong particle identification power can significantly improve the analysis. Lastly, more realistic atmospheric simulations can be used. There have been improvements implemented in atmospheric MC simulation in DUNE, including a new method of neutrino oscillation calculation [105]. Eventually, a data-driven atmospheric background constraint can be an option once DUNE data is available.

References

[1] F. Abe *et al.* (CDF Collaboration). "Observation of Top Quark Production in $p\bar{p}$ Collisions with the Collider Detector at Fermilab". In: *Physical Review Letters* 74.14 (1995), pp. 26262631. URL: http://dx.doi.org/10.1103/PhysRevLett.74.2626.

[2] S. Abachi *et al.* (D0 Collaboration). "Observation of the Top Quark". In: *Physical Review Letters* 74.14 (1995), pp. 26322637. URL: http://dx.doi.org/10.1103/PhysRevLett.74.2632.

[3] K. Kodama *et al.* (DONUT Collaboration). "Observation of tau neutrino interactions". In: *Physics Letters B* 504.3 (2001), pp. 218224. URL: http://dx.doi.org/10.1016/S0370-2693(01)00307-0.

[4] G. Aad *et al.* (ATLAS Collaboration). "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 129. URL: http://dx.doi.org/10.1016/j.physletb.2012.08.020.

[5] S. Chatrchyan *et al.* (CMS Collaboration). "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 3061. URL: http://dx.doi.org/10.1016/j.physletb.2012.08.021.

[6] P. A. R. Ade et al. "Planck 2015 results. XIII. Cosmological parameters". In: *Astron. Astrophys.* 594 (2016), A13. arXiv: 1502.01589 [astro-ph.CO].

[7] P. A. Zyla *et al.* (Particle Data Group). "Big Bang Nucleosynthesis". In: *Prog. Theor. Exp. Phys.* (2020), p. 083C01.

[8] A. D. Dolgov. "NonGUT baryogenesis". In: *Phys. Rept.* 222 (1992), pp. 309–386.

[9] A. D. Dolgov. "Baryogenesis, 30 years after". In: *25th ITEP Winter School of Physics*. July 1997. arXiv: hep-ph/9707419.

[10] A. D. Dolgov and Ya. B. Zeldovich. "Cosmology and elementary particles". In: *Rev. Mod. Phys.* 53 (1 1981), pp. 1–41. URL: https://link.aps.org/doi/10.1103/RevModPhys.53.1.

[11] A D Sakharov. "VIOLATION OF CP INVARIANCE, C ASYMMETRY, AND BARYON ASYMMETRY OF THE UNIVERSE." In: *JETP Lett.* (Jan. 1967). URL: https://www.osti.gov/biblio/4449128.

[12] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to quantum field theory*. Reading, USA: Addison-Wesley, 1995. ISBN: 978-0-201-50397-5.

[13] C. S. Wu et al. "Experimental Test of Parity Conservation in Beta Decay". In: *Phys. Rev.* 105 (4 1957), pp. 1413–1415. URL: https://link.aps.org/doi/10.1103/PhysRev.105.1413.

[14] J. H. Christenson et al. "Evidence for the 2π Decay of the K_2^0 Meson". In: *Phys. Rev. Lett.* 13 (4 1964), pp. 138–140. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.138.

[15] A. Alavi-Harati et al. "Observation of Direct CP Violation in $K_{S,L} \to \pi\pi$ Decays". In: *Phys. Rev. Lett.* 83 (1 1999), pp. 22–27. URL: https://link.aps.org/doi/10.1103/PhysRevLett.83.22.

[16] V. Fanti et al. "A new measurement of direct CP violation in two pion decays of the neutral kaon". In: *Physics Letters B* 465.1 (1999), pp. 335–348. URL: https://www.sciencedirect.com/science/article/pii/S0370269399010308.

[17] G. 't Hooft. "Symmetry Breaking through Bell-Jackiw Anomalies". In: *Phys. Rev. Lett.* 37 (1 1976), pp. 8–11. URL: https://link.aps.org/doi/10.1103/PhysRevLett.37.8.

[18] Gerard 't Hooft. "Computation of the Quantum Effects Due to a Four-Dimensional Pseudoparticle". In: *Phys. Rev. D* 14 (1976). Ed. by Mikhail A. Shifman. [Erratum: Phys.Rev.D 18, 2199 (1978)], pp. 3432–3450.

[19] V.A. Kuzmin, V.A. Rubakov, and M.E. Shaposhnikov. "On anomalous electroweak baryon-number non-conservation in the early universe". In: *Physics Letters B* 155.1 (1985), pp. 36–42. URL: https://www.sciencedirect.com/science/article/pii/0370269385910287.

[20] David E Morrissey and Michael J Ramsey-Musolf. "Electroweak baryogenesis". In: *New Journal of Physics* 14.12 (2012), p. 125003. URL: https://doi.org/10.1088/1367-2630/14/12/125003.

[21] M.E. Shaposhnikov. "Baryon asymmetry of the universe in standard electroweak theory". In: *Nuclear Physics B* 287 (1987), pp. 757–775. URL: https://www.sciencedirect.com/science/article/pii/0550321387901271.

[22] Jogesh C. Pati and Abdus Salam. "Is Baryon Number Conserved?" In: *Phys. Rev. Lett.* 31 (10 1973), pp. 661–664. URL: https://link.aps.org/doi/10.1103/PhysRevLett.31.661.

[23] Howard Georgi and S. L. Glashow. "Unity of All Elementary-Particle Forces". In: *Phys. Rev. Lett.* 32 (8 1974), pp. 438–441. URL: https://link.aps.org/doi/10.1103/PhysRevLett.32.438.

[24] R. Barbier *et al.* "R-Parity-violating supersymmetry". In: *Physics Reports* 420.1 (2005), pp. 1–195. URL: https://www.sciencedirect.com/science/article/pii/S0370157305003327.

[25] Rabindra N Mohapatra. "Supersymmetry and R-parity: an overview". In: *Physica Scripta* 90.8 (2015), p. 088004. URL: https://doi.org/10.1088/0031-8949/90/8/088004.

[26] R. N. Mohapatra and R. E. Marshak. "Local $\mathcal{B} - \mathcal{L}$ Symmetry of Electroweak Interactions, Majorana Neutrinos, and Neutron Oscillations". In: *Phys. Rev. Lett.* 44 (20 1980), pp. 1316–1319. URL: https://link.aps.org/doi/10.1103/PhysRevLett.44.1316.

[27] A. Takenaka et al. "Search for proton decay via $p \to e^+\pi^0$ and $p \to \mu^+\pi^0$ with an enlarged fiducial volume in Super-Kamiokande I-IV". In: *Physical Review D* 102.11 (2020). URL: http://dx.doi.org/10.1103/PhysRevD.102.112011.

[28] S.L. Glashow. "The Future of Elementary Particle Physics". In: *NATO Sci. Ser. B* 61 (1980), p. 687.

[29] D.G. Phillips et al. "Neutron-antineutron oscillations: Theoretical status and experimental prospects". In: *Physics Reports* 612 (2016), pp. 145. URL: http://dx.doi.org/10.1016/j.physrep.2015.11.001.

[30] K.S Babu and R.N Mohapatra. "Observable neutron-antineutron oscillations in seesaw models of neutrino mass". In: *Physics Letters B* 518.3 (2001), pp. 269–275. URL: https://www.sciencedirect.com/science/article/pii/S0370269301010772.

[31] Shmuel Nussinov and Robert Shrock. "n- $\overline{n}$ Oscillations in Models with Large Extra Dimensions". In: *Phys. Rev. Lett.* 88 (17 2002), p. 171601. URL: https://link.aps.org/doi/10.1103/PhysRevLett.88.171601.

[32] Bhaskar Dutta, Yukihiro Mimura, and R. N. Mohapatra. "Observable $N-\overline{N}$ Oscillation in High Scale Seesaw Models". In: *Phys. Rev. Lett.* 96 (6 2006), p. 061801. URL: https://link.aps.org/doi/10.1103/PhysRevLett.96.061801.

[33] K. S. Babu et al. "Post-sphaleron baryogenesis and an upper limit on the neutron-antineutron oscillation time". In: *Phys. Rev. D* 87 (11 2013), p. 115019. URL: https://link.aps.org/doi/10.1103/PhysRevD.87.115019.

[34] Howard Georgi and S. L. Glashow. "Unity of All Elementary-Particle Forces". In: *Phys. Rev. Lett.* 32 (8 1974), pp. 438–441. URL: https://link.aps.org/doi/10.1103/PhysRevLett.32.438.

[35] S. Raby et al. "DUSEL Theory White Paper". In: (2008). arXiv: 0810.4551 [hep-ph].

[36] P. A. Zyla *et al.* (Particle Data Group). "Neutrino Masses, Mixing, and Oscillations". In: *Prog. Theor. Exp. Phys.* (2020), p. 083C01.

[37] P. A. Zyla *et al.* (Particle Data Group). "CP Violation in the Quark Sector". In: *Prog. Theor. Exp. Phys.* (2020), p. 083C01.

[38] E. Friedman and A. Gal. "Realistic calculations of nuclear disappearance lifetimes induced by $n\bar{n}$ oscillations". In: *Phys. Rev. D* 78 (1 2008), p. 016002. URL: https://link.aps.org/doi/10.1103/PhysRevD.78.016002.

[39] M. Baldo-Ceolin et al. "A new experimental limit on neutron-antineutron oscillations". In: *Zeitschrift für Physik C Particles and Fields* 63.3 (1994), pp. 409–416. URL: https://doi.org/10.1007/BF01580321.

[40] Joshua Barrow et al. "Progress and simulations for intranuclear neutron-antineutron transformations in Ar 18 40". In: *Physical Review D* 101 (Feb. 2020). URL: https://link.aps.org/doi/10.1103/PhysRevD.101.036008.

[41] K. Abe and et al. (SK Collaboration). "Neutron-antineutron oscillation search using a 0.37 megaton-years exposure of Super-Kamiokande". In: *Physical Review D* 103.1 (2021). URL: http://dx.doi.org/10.1103/PhysRevD.103.012008.

[42] M. J. Frost. "The NNbar Experiment at the European Spallation Source". In: (2016). arXiv: 1607.07271 [hep-ph].

[43] M. L. Cherry et al. "Experimental Test of Baryon Conservation: A New Limit on Neutron-Antineutron Oscillations in Oxygen". In: *Phys. Rev. Lett.* 50 (18 1983), pp. 1354–1356. URL: https://link.aps.org/doi/10.1103/PhysRevLett.50.1354.

[44] T. W. Jones and et al. (Irvine-Michigan-Brookhaven Collaboration). "Search for $n - \bar{n}$ Oscillation in Oxygen". In: *Phys. Rev. Lett.* 52 (9 1984), pp. 720–723. URL: https://link.aps.org/doi/10.1103/PhysRevLett.52.720.

[45] G. Fidecaro and et al. "EXPERIMENTAL SEARCH FOR NEUTRON ANTI-NEUTRON TRANSITIONS WITH FREE NEUTRONS". In: *Phys. Lett. B* 156 (1985), pp. 122–128.

[46] M. Takita and et al. (Kamiokande Collaboration). "Search for neutron-antineutron oscillation in ^{16}O nuclei". In: *Phys. Rev. D* 34 (3 1986), pp. 902–904. URL: https://link.aps.org/doi/10.1103/PhysRevD.34.902.

[47] G. Bressi and et al. “Final results of a search for free neutron anti-neutron oscillations”. In: *Nuovo Cim. A* 103 (1990), pp. 731–750.

[48] Ch. Berger and et al. “Search for neutron-antineutron oscillations in the Frèjus detector”. In: *Physics Letters B* 240.1 (1990), pp. 237–242. URL: https://www.sciencedirect.com/science/article/pii/0370269390904418.

[49] J. Chung and et al. “Search for neutron-antineutron oscillations using multiprong events in Soudan 2”. In: *Physical Review D* 66.3 (2002). URL: http://dx.doi.org/10.1103/PhysRevD.66.032004.

[50] K. Abe et al. “Search for $n - \overline{n}$ oscillation in Super-Kamiokande”. In: *Phys. Rev. D* 91 (7 2015), p. 072006. URL: https://link.aps.org/doi/10.1103/PhysRevD.91.072006.

[51] B. et al. Aharmim. “Search for neutron-antineutron oscillations at the Sudbury Neutrino Observatory”. In: *Phys. Rev. D* 96 (9 2017), p. 092005. URL: https://link.aps.org/doi/10.1103/PhysRevD.96.092005.

[52] Alessandro Bettini et al. “The ICARUS liquid argon TPC: a complete imaging device for particle physics”. In: *Nuclear Instruments & Methods in Physics Research Section A* 315 (1992), pp. 223–228.

[53] Roxanne Guenette. “The ArgoNeuT experiment”. In: (2011). arXiv: 1110.0443 [physics.ins-det].

[54] J. Paley et al. “LArIAT: Liquid Argon In A Testbeam”. In: (2014). arXiv: 1406.5560 [physics.ins-det].

[55] Roberto Acciarri et al. “Design and Construction of the MicroBooNE Detector”. In: *JINST* 12.02 (2017), P02017. arXiv: 1612.05824 [physics.ins-det].

[56] A.A. Aguilar-Arevalo et al. “Significant Excess of Electronlike Events in the MiniBooNE Short-Baseline Neutrino Experiment”. In: *Physical Review Letters* 121.22 (2018). URL: http://dx.doi.org/10.1103/PhysRevLett.121.221801.

[57] MicroBooNE collaboration et al. “First Measurement of Inclusive Electron-Neutrino and Antineutrino Charged Current Differential Cross Sections in Charged Lepton Energy on Argon in MicroBooNE”. In: (2021). arXiv: 2109.06832 [hep-ex].

[58] MicroBooNE Collaboration et al. “First Measurement of Energy-dependent Inclusive Muon Neutrino Charged-Current Cross Sections on Argon with the MicroBooNE Detector”. In: (2021). arXiv: 2110.14023 [hep-ex].

[59] P. Abratenko et al. “Measurement of the flux-averaged inclusive charged-current electron neutrino and antineutrino cross section on argon using the NuMI beam and the MicroBooNE detector”. In: *Phys. Rev. D* 104 (5 2021), p. 052002. URL: https://link.aps.org/doi/10.1103/PhysRevD.104.052002.

[60] MicroBooNE collaboration et al. “Search for an Excess of Electron Neutrino Interactions in MicroBooNE Using Multiple Final State Topologies”. In: (2021). arXiv: 2110.14054 [hep-ex].

[61] MicroBooNE collaboration et al. “Search for an anomalous excess of inclusive charged-current ν_e interactions in the MicroBooNE experiment using Wire-Cell reconstruction”. In: (2021). arXiv: 2110.13978 [hep-ex].

[62] MicroBooNE collaboration et al. “Search for an anomalous excess of charged-current ν_e interactions without pions in the final state with the MicroBooNE experiment”. In: (2021). arXiv: 2110.14065 [hep-ex].

[63] MicroBooNE collaboration et al. “Search for an anomalous excess of charged-current quasi-elastic ν_e interactions with the MicroBooNE experiment using Deep-Learning-based reconstruction”. In: (2021). arXiv: 2110.14080 [hep-ex].

[64] MicroBooNE collaboration et al. “Search for Neutrino-Induced Neutral Current Δ Radiative Decay in MicroBooNE and a First Test of the MiniBooNE Low Energy Excess Under a Single-Photon Hypothesis”. In: (2021). arXiv: 2110.00409 [hep-ex].

[65] P. Abratenko et al. “Search for heavy neutral leptons decaying into muon-pion pairs in the MicroBooNE detector”. In: *Phys. Rev. D* 101 (5 2020), p. 052001. URL: https://link.aps.org/doi/10.1103/PhysRevD.101.052001.

[66] P. Abratenko et al. “Search for a Higgs Portal Scalar Decaying to Electron-Positron Pairs in the MicroBooNE Detector”. In: *Phys. Rev. Lett.* 127 (15 2021), p. 151803. URL: https://link.aps.org/doi/10.1103/PhysRevLett.127.151803.

[67] P. Abratenko et al. “Semantic segmentation with a sparse convolutional neural network for event reconstruction in MicroBooNE”. In: *Phys. Rev. D* 103 (5 2021), p. 052012. URL: https://link.aps.org/doi/10.1103/PhysRevD.103.052012.

[68] MicroBooNE collaboration et al. “Wire-Cell 3D Pattern Recognition Techniques for Neutrino Event Reconstruction in Large LArTPCs: Algorithm Description and Quantitative Evaluation with MicroBooNE Simulation”. In: (2021). arXiv: 2110.13961 [physics.ins-det].

[69] B. Abi et al. “Volume IV. The DUNE far detector single-phase technology”. In: *Journal of Instrumentation* 15 (Aug. 2020), T08010–T08010.

[70] Jyoti Joshi and Xin Qian. "Signal Processing in the MicroBooNE LArTPC". In: (2015). arXiv: 1511.00317 [physics.ins-det].

[71] "Software Trigger". In: *MicroBOONE-Internal-DocDB-5205* (2016). URL: https://microboone-docdb.fnal.gov/cgi-bin/sso/ShowDocument?docid=5205.

[72] Andy Furmanski. "Triggering in microboone". In: *MicroBOONE-Internal-DocDB-17143* (2018). URL: https://microboone-docdb.fnal.gov/cgi-bin/sso/ShowDocument?docid=17143.

[73] Afroditi Papadopoulou et al. "Technical Note: MCC9 Overlay GENIE simulated BNB and Cosmic data". In: *MicroBOONE-Internal-DocDB-20737* (2019). URL: https://microboone-docdb.fnal.gov/cgi-bin/sso/ShowDocument?docid=20737&version=1.

[74] P. Abratenko et al. "Neutrino event selection in the MicroBooNE liquid argon time projection chamber using Wire-Cell 3D imaging, clustering, and charge-light matching". In: *Journal of Instrumentation* 16.06 (2021), P06043. URL: http://dx.doi.org/10.1088/1748-0221/16/06/P06043.

[75] C. Adams et al. "Ionization electron signal processing in single phase LArTPCs. Part I. Algorithm Description and quantitative evaluation with MicroBooNE simulation". In: *Journal of Instrumentation* 13.07 (2018), P07006–P07006. URL: https://doi.org/10.1088/1748-0221/13/07/p07006.

[76] C. Adams et al. "Ionization electron signal processing in single phase LArTPCs. Part II. Data/simulation comparison and performance in MicroBooNE". In: *Journal of Instrumentation* 13.07 (2018), P07007–P07007. URL: https://doi.org/10.1088/1748-0221/13/07/p07007.

[77] P. Abratenko *et al.* (MicroBooNE Collaboration). "Electromagnetic Shower Reconstruction and Energy Validation with Michel Electrons and π^0 Samples for the Deep-Learning-Based Analyses in MicroBooNE". In: (2021). arXiv: 2110.11874 [hep-ex].

[78] R. Acciarri *et al.* "The Pandora multi-algorithm approach to automated pattern recognition of cosmic-ray muon and neutrino events in the MicroBooNE detector". In: *The European Physical Journal C* 78.1 (2018). URL: http://dx.doi.org/10.1140/epjc/s10052-017-5481-6.

[79] P. Abratenko *et al.* (MicroBooNE Collaboration). "Neutrino Event Selection in the MicroBooNE Liquid Argon Time Projection Chamber using Wire-Cell 3D Imaging, Clustering, and Charge-Light Matching". In: (2021). arXiv: 2011.01375 [physics.ins-det].

[80] Babak Abi et al. “Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume I: Introduction to DUNE”. In: (2020). arXiv: 2002.02967 [physics.ins-det].

[81] Babak Abi et al. “Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume II DUNE Physics”. In: (Feb. 2020). arXiv: 2002.03005 [hep-ex].

[82] ImageNET project. *ImageNET*. https://www.image-net.org/. Jan. 2022.

[83] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Commun. ACM* 60.6 (May 2017), pp. 8490. URL: https://doi.org/10.1145/3065386.

[84] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2014. arXiv: 1409.1556 [cs.CV].

[85] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[86] P. A. Zyla *et al.* (Particle Data Group). “Machine Learning”. In: *Prog. Theor. Exp. Phys.* (2021), p. 083C01.

[87] P. Abratenko *et al.* (MicroBooNE Collaboration). “Vertex-finding and reconstruction of contained two-track neutrino events in the MicroBooNE detector”. In: *Journal of Instrumentation* 16.02 (2021), P02017P02017. URL: http://dx.doi.org/10.1088/1748-0221/16/02/P02017.

[88] Jarrett Moon. “Using Deep Learning Techniques to Search for the MiniBooNE Low Energy Excess in MicroBooNE with > 3σ Sensitivity”. In: (2020). arXiv: 2010.14505 [physics.data-an].

[89] Christopher Choy, JunYoung Gwak, and Silvio Savarese. *4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks*. 2019. arXiv: 1904.08755 [cs.CV].

[90] Farah Fahim et al. *hls4ml: An Open-Source Codesign Workflow to Empower Scientific Low-Power Machine Learning Devices*. 2021. arXiv: 2103.05579 [cs.LG].

[91] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[92] Herbert E. Robbins. “A Stochastic Approximation Method”. In: *Annals of Mathematical Statistics* 22 (2007), pp. 400–407.

[93] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[94] Benjamin Graham and Laurens van der Maaten. *Submanifold Sparse Convolutional Networks*. 2017. arXiv: 1706.01307 [cs.NE].

[95] Chris Choy. *Chris Choy*. https://chrischoy.github.io/. Jan. 2022.

[96] Parker Allen Tew. “An investigation of sparse tensor formats for tensor libraries”. In: 2016.

[97] François Chollet et al. *Keras*. https://keras.io. 2015.

[98] Xilinx. *Vivado*. URL: https://www.xilinx.com/support/documentation-navigation/design-hubs/dh0012-vivado-high-level-synthesis-hub.html.

[99] E. L. Snider and G. Petrillo. “LArSoft: Toolkit for Simulation, Reconstruction and Analysis of Liquid Argon TPC Neutrino Detectors”. In: *J. Phys. Conf. Ser.* 898.4 (2017). Ed. by Richard Mount and Craig Tull, p. 042057.

[100] Jeremy E.T. Hewes. “Searches for Bound Neutron-Antineutron Oscillation in Liquid Argon Time Projection Chambers”. PhD thesis. Manchester U., 2017.

[101] Roger D. Woods and David S. Saxon. “Diffuse Surface Optical Model for Nucleon-Nuclei Scattering”. In: *Phys. Rev.* 95 (2 1954), pp. 577–578. URL: https://link.aps.org/doi/10.1103/PhysRev.95.577.

[102] A. Bodek and J. L. Ritchie. “Further studies of Fermi-motion effects in lepton scattering from nuclear targets”. In: *Phys. Rev. D* 24 (5 1981), pp. 1400–1402. URL: https://link.aps.org/doi/10.1103/PhysRevD.24.1400.

[103] Costas Andreopoulos et al. “The GENIE Neutrino Monte Carlo Generator: Physics and User Manual”. In: (Oct. 2015). arXiv: 1510.05494 [hep-ph].

[104] G. D. Barr et al. “Three-dimensional calculation of atmospheric neutrinos”. In: *Phys. Rev. D* 70 (2 2004), p. 023006. URL: https://link.aps.org/doi/10.1103/PhysRevD.70.023006.

[105] Joshua L. Barrow. “Towards Neutron Transformation Searches”. PhD thesis. U. Tennessee, Knoxville, 2021.

[106] E. Church et al. “Nucleon Decay, Atmospheric Neutrinos, and Cosmic Rays at DUNE: September 2016 Progress Report”. In: (2016). URL: https://digitalcommons.imsa.edu/sir_progress_reports/1.

[107] Aaron Higuera, Yeon jae Jwa, and Gerogia Karagiorgi (DUNE Collaboration). “Proton Decay $p \to \bar{\nu}K^+$”. In: *DUNE-Internal-DocDB-12136* (2018). URL: https://docs.dunescience.org/cgi-bin/private/ShowDocument?docid=12136.

[108] “Prob3plusplus”. In: *GitHub repository* (2018). URL: https://github.com/rogerwendell/Prob3plusplus.git.

[109] V. Barger et al. “Matter effects on three-neutrino oscillations”. In: *Phys. Rev. D* 22 (11 1980), pp. 2718–2726. URL: https://link.aps.org/doi/10.1103/PhysRevD.22.2718.

[110] J. A. Formaggio and G. P. Zeller. “From eV to EeV: Neutrino cross sections across energy scales”. In: *Reviews of Modern Physics* 84.3 (2012), pp. 13071341. URL: http://dx.doi.org/10.1103/RevModPhys.84.1307.

[111] R. Acciarri et al. “Measurement of and ŕ neutral current 0 production in the ArgoNeuT detector”. In: *Physical Review D* 96.1 (2017). URL: http://dx.doi.org/10.1103/PhysRevD.96.012006.

[112] D. H. Wright and M. H. Kelsey. “The Geant4 Bertini Cascade”. In: *Nucl. Instrum. Meth. A* 804 (2015), pp. 175–188.

[113] LArSoft. *Projection Matching Algorithm*. URL: https://larsoft.org/single-record/?pdb=102.

[114] David Caratelli et al. “MicroBooNE TPC Pre-commissioning Electronics Checks And Calibration Tests”. In: *MicroBooNE-Internal-DocDB-3785* (2014). URL: https://microboone-docdb.fnal.gov/cgi-bin/sso/ShowDocument?docid=3785.

[115] Yangqing Jia et al. “Caffe: Convolutional Architecture for Fast Feature Embedding”. In: *arXiv preprint arXiv:1408.5093* (2014).

[116] A. Hoecker et al. *TMVA - Toolkit for Multivariate Data Analysis*. 2007. arXiv: physics/0703039 [physics.data-an].

[117] Robert E Schapire. “Explaining adaboost”. In: *Empirical inference*. Springer, 2013, pp. 37–52.

[118] E. Friedman and A. Gal. “Realistic calculations of nuclear disappearance lifetimes induced by $n\overline{n}$ oscillations”. In: *Phys. Rev. D* 78 (1 2008), p. 016002. URL: https://link.aps.org/doi/10.1103/PhysRevD.78.016002.

[119] Steven G. Johnson. *Multi-dimensional adaptive integration (cubature) in C*. URL: https://github.com/stevengj/cubature.

[120] B. Abi et al. "Prospects for beyond the Standard Model physics searches at the Deep Underground Neutrino Experiment". In: *The European Physical Journal C* 81.4 (2021). URL: http://dx.doi.org/10.1140/epjc/s10052-021-09007-w.

[121] J. S. Marshall and M. A. Thomson. "The Pandora software development kit for pattern recognition". In: *The European Physical Journal C* 75.9 (2015). URL: http://dx.doi.org/10.1140/epjc/s10052-015-3659-3.

[122] E.S. Golubeva, J.L. Barrow, and C.G. Ladd. "A model of antineutron annihilation in experimental searches for neutron-antineutron transformations". In: *Physical Review D* 99.3 (2019). URL: http://dx.doi.org/10.1103/PhysRevD.99.035002.

[123] Yeon-Jae Jwa et al. "Accelerating Deep Neural Networks for Real-time Data Selection for High-resolution Imaging Particle Detectors". In: *2019 New York Scientific Data Summit (NYSDS)*. 2019, pp. 1–10.

[124] Yeon-jae Jwa et al. "Real-time Inference with 2D Convolutional Neural Networks on Field Programmable Gate Arrays for High-rate Particle Imaging Detectors". In: (Jan. 2022). arXiv: 2201.05638 [physics.ins-det].

[125] R. Acciarri et al. "Measurement of cosmic-ray reconstruction efficiencies in the MicroBooNE LArTPC using a small external cosmic-ray counter". In: *Journal of Instrumentation* 12.12 (2017), P12030–P12030. URL: https://doi.org/10.1088/1748-0221/12/12/p12030.

[126] "Cosmic Shielding Studies at MicroBooNE". In: (May 2016).

[127] Christian Hagmann, David Lange, and David M. Wright. "Cosmic-ray shower generator (CRY) for Monte Carlo transport codes". In: *2007 IEEE Nuclear Science Symposium Conference Record* 2 (2007), pp. 1143–1146.

[128] D. Heck et al. "CORSIKA: A Monte Carlo code to simulate extensive air showers". In: (Feb. 1998).

[129] MicroBooNE Collaboration. "Progress Toward the First Search for Bound Neutron Oscillation into Antineutron in a Liquid Argon TPC". In: *MICROBOONE-NOTE-1093-PUB* (2020).

[130] X. Qian et al. "Three-dimensional imaging for large LArTPCs". In: *Journal of Instrumentation* 13.05 (2018), P05032P05032. URL: http://dx.doi.org/10.1088/1748-0221/13/05/P05032.

[131] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. URL: http://doi.acm.org/10.1145/2939672.2939785.

[132] Jie Zhou et al. “Graph Neural Networks: A Review of Methods and Applications”. In: (2021). arXiv: 1812.08434 [cs.LG].

[133] Saúl Alonso-Monsalve et al. “Graph neural network for 3D classification of ambiguities and optical crosstalk in scintillator-based neutrino detectors”. In: *Physical Review D* 103.3 (2021). URL: http://dx.doi.org/10.1103/PhysRevD.103.032005.

[134] J. Lundberg et al. “Limits, discovery and cut optimization for a Poisson process with uncertainty in background and signal efficiency: TRolke 2.0”. In: *Computer Physics Communications* 181.3 (2010), pp. 683686. URL: http://dx.doi.org/10.1016/j.cpc.2009.11.001.

[135] Kirsty Duffy. “Geant4 Reweight”. In: *MicroBOONE-Internal-DocDB-28376* (). URL: https://microboone-docdb.fnal.gov/cgi-bin/sso/ShowDocument?docid=28376.

[136] "A. Ashkenazi et al. (MicroBooNE Collaboration)". “Detector systematics internal note”. In: *MicroBOONE-Internal-DocDB-27009* (). URL: https://microboone-docdb.fnal.gov/cgi-bin/sso/ShowDocument?docid=20737&version=1.

[137] Wolfgang A. Rolke and Angel M. López. “Confidence intervals and upper bounds for small signals in the presence of background noise”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 458.3 (2001), pp. 745758. URL: http://dx.doi.org/10.1016/S0168-9002(00)00935-9.

[138] Junze Liu et al. “Deep-Learning-Based Kinematic Reconstruction for DUNE”. In: (Dec. 2020). arXiv: 2012.06181 [physics.ins-det].

[139] B. Abi et al. “Long-baseline neutrino oscillation physics potential of the DUNE experiment”. In: *The European Physical Journal C* 80.10 (2020). URL: http://dx.doi.org/10.1140/epjc/s10052-020-08456-z.

[140] Chao Zhang, Brett Viren, and Milind Diwan. “Three-dimensional imaging for large LArTPCs”. In: *Journal of Instrumentation* 13 (Mar. 2018).