Selection of $\bar{\nu}_\mu$ charged-current single-pion events with boosted decision tree particle identification in the ND280 detector

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Gabriel Charles Penn

February 2023

Abstract

Tokai to Kamioka (T2K) is a long-baseline neutrino oscillation experiment situated in Japan, studying electron-neutrino appearance in a muon-neutrino beam and thus probing multiple parameters of the Pontecorvo-Maki-Nakagawa-Sakata matrix. A beam of muon-(anti)neutrinos is produced at the J-PARC accelerator complex and sampled by near and far detectors to compare the beam content before and after oscillation. The off-axis near detector, ND280, serves both to characterise signals and backgrounds observed at the far detector, and to measure a variety of neutrino interaction cross-sections on carbon and oxygen nuclear targets. These measurements contribute to our understanding of both neutrino oscillations and the interactions between neutrinos and atomic nuclei.

Selecting specific interaction types from the ND280 data requires accurate identification of charged particle tracks, but the conventional cut-based methods perform poorly in many areas. To make full use of the wide array of information available from all of the ND280 subdetectors and thus develop a high-performing particle identification (PID) algorithm, a multivariate analysis approach is necessary, but such a tool has yet to be developed and deployed at T2K. This thesis presents the development of a general-purpose PID tool for charged particle tracks in ND280 using boosted decision trees, and demonstrates its superior classification power compared to the existing PID algorithms currently in use. The tool has been designed to be broadly applicable and outperforms purpose-built conventional PID at identifying each particle type in almost all kinematic regions tested. Furthermore, it has been integrated into a selection algorithm for muon-antineutrino charged-current single-pion events, and found to substantially improve the selection performance. These results demonstrate the power of multivariate PID methods for charged particle tracks in ND280, and strongly motivate their use in future T2K analyses.

"The first principle is that you must not fool yourself — and you are the easiest person to fool."

— *Richard Feynman*

Acknowledgements

First I must thank my supervisor, Professor Neil McCauley, for his support, guidance, feedback, and encouragement over the course of my PhD and particularly in the difficult final year; and in matters both academic and personal. I am immensely grateful for the support he's given me throughout this process, and especially for his patience, understanding and robust but always helpful advice when things have not gone as planned. It has all come together in the end, and I truly appreciate how well he's helped me navigate this challenging process amid difficult circumstances.

It's been a great pleasure to be a member of the T2K collaboration, which has provided both an exciting research environment and a very enjoyable social experience, and I'd like to thank everyone I've had the pleasure to work with during my time here. Special thanks go to Georgios Christodolou, Ka Ming Tsui and Matthew Lawe, for the help they've each given me at various stages of my PhD and with various elements of the physics and software problems I've had to get my head around. I must also thank the staff at the University of Liverpool, particularly the diligent HEP computing team for providing and maintaining the computing resources without which my analysis would've been impossible.

Thank you to all the friends I've made over the course of my PhD. I was fortunate enough to spend my LTA in Tōkai-mura in the excellent company of Sam, Jordan, Tristan, Joe, Charlie and Yue, known amongst ourselves as the 'ECal RMM 3 Appreciation Society' for infamous reasons. Thank you all for making it such a fun stay — I will fondly remember those late-night 'friendly' poker games and the implausible levels of yakiniku consumption, and you have my apologies for filling the karaoke queue with David Bowie. My thanks also go to my fellow T2K Liverpool members Adrian, Pratiksha, Lauren, Francis and Pruthvi, who made for great mentors and/or travel buddies over the years, and who could always be relied upon for good company at either end of Eurasia; and to Andrew, Harry, Ron, Matt, and my other peers at Liverpool who, between lockdowns and LTAs, I haven't been able to spend as much time with as I'd have liked. I wish you all the best as we each go from this chapter of our lives to the next!

And of course, thank you to all the friends and family who've been there for me these past five years (and occasionally made valiant efforts to understand what on earth I've been working on). There have been many trials and tribulations, and it's thanks to you all that I've weathered them and completed this process that means so much to me. Thank you to Tyler, Alex, Max, Nick, Eli, and all the other friends who've helped keep me sane

via cyberspace amid the isolation of the Covid era. Thank you Midori, Josh, Ryouji-san and Kazue-san, for your great kindness and trust. Thank you Andrew and Kathryn, for all your help in difficult times. Thank you to my brother Joe, for always being there and for the vicarious joy of seeing things coming together for you much as they have for me.

To Mum and Dad, thank you so much for everything you've done for me: for all your love and care these past five years and the twenty-two before that. You've always encouraged me to follow my curiosity and passions and to make the most of my potential, and given me the support I needed to do so. I can't thank you enough for that, and I hope I've made you proud.

And to Mari, thank you for your love, your support, your unwavering care, your extraordinary patience, and the joy you bring to my life. I couldn't have done this without you.

Declaration

The work presented in this thesis is the work of the author, except as noted here and where the work of others has been cited.

In Chapter 3, the ECal bar-to-bar calibration process was developed by previous ND280 collaborators in the role, and the scripts used to generate the plots shown were only modified slightly to ensure compatibility with current software. As the analyser responsible for the bar-to-bar calibration over the latter four years of the PhD, the author performed the bar-to-bar constant generation, validation and uploading for T2K Runs 8 and 9. The author also worked on Long-Term Attachment as part of a team responsible for the day-to-day running of the ECal during T2K Run 10, which included daily calibrations and monitoring, but did not develop any of the tools or processes used.

In Chapter 4, the existing $\bar{\nu}_\mu$ CC1π^- event selection was developed by other ND280 collaborators, including the standard pre-selection cuts which were used by the author for other selections as described in Chapters 6 and 7. The author developed the modifications that lead to the 'improved' selection, namely the antimuon candidate PID loop and the ECal PID cuts.

In Chapter 6, the input variables used for the BDT were developed by other ND280 collaborators.

Gabriel Charles Penn

Contents

Abstract **i**

Acknowledgements **v**

Declaration **vii**

Contents **xii**

Glossary **xiv**

Acronyms **xvii**

Introduction **1**

1 Neutrino Physics **3**

1.1 The Standard Model 4

1.2 A history of the neutrino 6

1.3 Neutrino oscillations 9

1.3.1 Solar neutrinos 13

1.3.2 Atmospheric neutrinos 16

1.3.3 Reactor neutrinos 19

1.3.4 Accelerator neutrinos 21

1.4 Neutrino-nucleus interactions 23

1.4.1 Neutrino-nucleon interactions 24

1.4.2 Nuclear effects 24

1.4.3 Pion production 26

2 The T2K Experiment **29**

2.1 Motivations 30

2.2 Neutrino beam 30

2.3 Super-Kamiokande 33

2.4 INGRID 35

2.5 ND280 36
2.5.1 Fine-Grained Detectors 38
2.5.2 Time Projection Chambers 38
2.5.3 Pi-Zero Detector 41
2.5.4 Electromagnetic Calorimeter 43
2.5.5 Magnet and Side Muon Range Detector 43
2.5.6 Software and data processing 45

3 ND280 Electromagnetic Calorimeter 47
3.1 Design of the Electromagnetic Calorimeter 48
3.1.1 Scintillator bars 49
3.1.2 ECal modules 50
3.1.3 MPPC photosensors 52
3.1.4 Readout electronics 52
3.2 ECal charge calibration 54
3.2.1 Pedestal and gain calibrations 55
3.2.2 Bar equalisation calibrations 56
3.2.3 Fibre scaling correction 60
3.2.4 Scintillator ageing 61
3.3 ECal reconstruction and performance 62
3.3.1 Clustering 62
3.3.2 Energy and shape fits 65
3.3.3 Global reconstruction 67

4 Muon-antineutrino CC1pi selection development in ND280 70
4.1 Data and Monte Carlo samples 71
4.2 The current ND280 $\bar{\nu}_\mu$ CC1π^- event selection 72
4.2.1 Particle identification 73
4.2.2 Limitations 78
4.3 Improved $\bar{\nu}_\mu$ CC1π^- event selection 80
4.3.1 Removing leading track requirements 81
4.3.2 ECal particle ID cuts 81
4.3.3 Improved selection performance 86

5 Multivariate analysis techniques 92
5.1 Treatment of multivariate data 93
5.2 Machine learning 94
5.3 Bayesian statistics 95
5.4 Multivariate analysis methods 96
5.4.1 Grid searches 96
5.4.2 Linear methods 97

5.4.3 Naïve Bayes classifier . . . 99
5.4.4 Kernel-based methods . . . 100
5.4.5 Neural networks . . . 101
5.4.6 Decision trees . . . 105
5.5 Ensemble learning . . . 106
5.5.1 Boosting . . . 107
5.5.2 Bagging . . . 108
5.5.3 Random forests . . . 109
5.6 Performance metrics . . . 109
5.7 MVA implementation . . . 111
5.7.1 The TMVA package . . . 112
5.7.2 BDT configuration . . . 114
5.7.3 Highland integration . . . 115

6 Particle ID development with boosted decision trees 116
6.1 Charged particle ID in ND280 event selections . . . 117
6.1.1 ECal PID . . . 117
6.1.2 Current usage . . . 123
6.2 Towards a global PID with boosted decision trees . . . 125
6.3 Training and testing samples . . . 126
6.3.1 Event weighting and momentum sensitivity . . . 129
6.4 Candidate BDT input variables . . . 131
6.4.1 Kinematic variables . . . 132
6.4.2 TPC particle ID . . . 135
6.4.3 FGD particle ID . . . 141
6.4.4 ECal particle ID . . . 142
6.4.5 SMRD particle ID . . . 145
6.5 Performance analysis and optimisation . . . 146
6.5.1 BDT training parameters . . . 147
6.5.2 Comparing representations . . . 151
6.5.3 Variable removal . . . 156
6.6 Final configuration performance . . . 160
6.6.1 Comparisons to existing PID . . . 167
6.6.2 Overtraining . . . 178
6.7 Conclusions . . . 181

7 Muon-antineutrino CC1pi selection with BDT PID 182
7.1 Track identification by BDT preference . . . 183
7.2 Optimised cuts . . . 187
7.3 Comparisons to real data . . . 192
7.4 Conclusions and outlook . . . 202

Bibliography **204**

Glossary

bar-to-bar calibration Correction to equalise the response of ND280 ECal scintillator bars. 56

baseline The distance between source and (far) detector in a neutrino oscillation experiment. 11

BDT selection $\bar{\nu}_\mu$ CC1π^- event selections using the BDT PID tool presented in this thesis. 182

event selection A computer algorithm designed to select events of a particular category from ND280 data. 1

existing selection The standard $\bar{\nu_\mu}$ CC1π^- event selection currently implemented in the Highland framework. 78

Geant4 An open-source software toolkit for simulating the passage of particles through matter. 45

GENIE A software package for generating simulated neutrino-nucleus interaction events. 46

global (regarding particle identification) using information from each subdetector a track crosses. 125

Highland Software framework for selecting and analysing events from ND280 data and MC. 46

improved selection $\bar{\nu}_\mu$ CC1pi event selection consisting of the modified selection plus cuts on ECal variables. 86

MipPion ECal PID variable designed to distinguish MIP-like from showering pion-like objects. 82

modified selection $\bar{\nu_\mu}$ CC1π^- event selection with the kinematic restrictions of the existing selection removed. 81

multiclassification Classification in which the number of classes is greater than 2. 112

NEUT A software package for generating simulated neutrino-nucleus interaction events. 46

particle gun A generator for Monte Carlo simulations that creates particles with user-defined kinematics as opposed to simulating an initial interaction. 126

pre-selection Cuts applied in a selection before PID to remove poor quality, unsuitable or background events. 72

preference The largest of a set of PID variables corresponding to different particle hypotheses, which may be taken as the PID decision for a track. 147

pull PID variable describing the difference between expected and measured energy loss. 74

ROOT An open-source software framework for data analysis. 45

segment Reconstruction object specific to a particular ND280 subdetector. 76

selection See ‘event selection’. 1

significance Figure of merit defined as $S/\sqrt{S+B}$ for a data sample containing S signal and B background events. 84

TMVA ROOT software package for multivariate analysis. 112

topology The set of particles leaving the atomic nucleus following a neutrino-nucleus interaction. 71

Trip-T Application-specific integrated circuits used to integrate MPPC signals in the ND280 ECal. 52

wrong-sign background The background for $\bar{\nu_\mu}$ CC1π^- of events originating from ν_μ. 71

Acronyms

ANN Artificial neural network. 101

BDT Boosted decision tree. 107

BSM Beyond Standard Model. 6

CC Charged current. 8

CC1pi Charged-current single-pion. 26

CP Charge-parity. 9

DIS Deep inelastic scattering. 24

DS-ECal Downstream Electromagnetic Calorimeter. 47

DT Decision tree. 105

ECal Electromagnetic Calorimeter. 37

EM Electromagnetic. 83

ES Elastic scattering. 24

FGD Fine-Grained Detector. 37

FHC Forward Horn Current. 30

FSI Final state interaction. 25

FV Fiducial volume. 72

GQ Good quality. 135

HIP Highly ionising particle. 123

HMPT Highest-momentum positive track. 73

HMT Highest-momentum track. 73

K-S Kolmogorov-Smirnov. 178

LLR Log-likelihood ratio. 122

MC Monte Carlo. 23

MIP Minimum ionising particle. 48

MLP Multilayer perceptron. 102

MPPC Multi-pixel photon counter. 36

MPV Most probable value. 74

MSW Mikheyev-Smirnov-Wolfenstein. 17

MVA Multivariate Analysis. 92

NC Neutral current. 8

ND280 Near Detector at 280 m. 29

NN Neural network. 101

nTPCs Number of TPC segments. 135

PDF Probability density function. 123

PG Particle gun. 126

PID Particle identification. 71

PMNS Pontecorvo-Maki-Nakagawa-Sakata. 9

PMT Photomultiplier tube. 13

POT Protons on target. 31

PØD Pi-Zero Detector. 37, 41–43, 46

QE Quasi-elastic scattering. 24

RES Resonant scattering. 24

RHC Reverse Horn Current. 31

ROC Receiver operating characteristic. 166

SK Super-Kamiokande. 8

SM Standard Model. 4

SMRD Side Muon Range Detector. 37

SSM Standard Solar Model. 8

T2K Tokai to Kamioka. 21

TPC Time Projection Chamber. 37

WLS Wavelength-shifting. 36

Introduction

Neutrino physics is one of the main frontiers of modern particle physics research. The discovery of neutrino oscillations required a modification of the Standard Model of Particle Physics, and this remains a highly active area of study comprising several large international experiments. One such experiment is Tokai to Kamioka (T2K), which studies electron-neutrino appearance in a muon-neutrino beam. The off-axis near detector ND280 analyses the T2K beam before oscillation and thus helps characterise signals at the far detector, and also measures rates of various interaction types, improving our understanding of neutrino-nucleus interactions.

Muon-antineutrino ($\bar{\nu}_\mu$) charged-current single-pion (CC1pi) interactions are a neutrino-nucleus event type of interest to T2K, as they are both a major background for charged-current quasi-elastic measurements and a potentially valuable channel for the oscillation analysis. However, the signature for $\bar{\nu}_\mu$ CC1π^- is a single antimuon and a single negative pion, which the existing selection algorithm used in ND280 fails to distinguish from the common 'wrong-sign' signature of a single muon and a single positive pion. This happens due to the similar behaviour of muons and pions, particularly in the ND280 time projection chambers which are used as the standard source of PID information. Thus the existing $\bar{\nu}_\mu$ CC1π^- event selection suffers from poor background rejection and cannot support precise cross-section measurements. This thesis presents efforts to improve the $\bar{\nu}_\mu$ CC1π^- event selection algorithm, initially by adding conventional rectangular cuts using information from the ND280 electromagnetic calorimeter (ECal), and then by developing a multivariate and 'global' PID tool using boosted decision trees (BDTs). Although the original motivation was to improve the $\bar{\nu}_\mu$ CC1π^- selection specifically, the BDT has been trained in a 'selection-agnostic' manner, favouring no particular event or particle type and with flat prior distributions of momentum and direction, with the intention of developing a general-purpose tool that can be used a wide variety of event selection contexts.

Chapter 1 covers the history and theory of neutrino physics, with a particular focus on the phenomenon of neutrino oscillation and the interactions between neutrinos and atomic nuclei. Chapter 2 outlines the T2K experiment and in particular the ND280 near detector, and Chapter 3 covers the ND280 ECal which has been an important part of the analysis presented in this thesis. Chapter 4 describes the initial $\bar{\nu}_\mu$ CC1π^- selection development work using conventional PID methods, which motivated a move towards multivariate analysis (MVA) methods. Chapter 5 provides an overview of the principles of multivariate analysis and machine learning and outlines some of the methods available, as well as the software tools that enable MVA implementation in ND280 data analysis. Chapter 6 details the development of the BDT PID tool and demonstrates its effectiveness at identifying particle tracks compared to conventional PID methods. Chapter 7 presents the results of applying the BDT PID tool within a $\bar{\nu}_\mu$ CC1π^- selection with full neutrino event MC, and discusses the potential for further development of this tool and the contribution it can make to T2K physics analyses.

Chapter 1

Neutrino Physics

Neutrinos are the most abundant massive particles in the universe, yet many of their properties remain unknown or poorly understood. These properties could offer insights into many of the problems currently facing high-energy physics (HEP) and cosmology, including baryon asymmetry and dark matter. Probing these properties is difficult, as neutrinos are highly elusive particles: they interact only via the weak nuclear force, leading to very small interaction cross-sections that make them challenging to detect.

In the decades since the discovery of the neutrino, theoretical understanding and detector technology have both progressed rapidly. In particular, the theory of neutrino flavour oscillation was a major extension of the Standard Model of particle physics. Today many experiments around the world are probing the parameters of neutrino oscillation theory, using a wide variety of sources and detection techniques. However, almost all such experiments rely on interactions between neutrinos and atomic nuclei. The cross-sections of these interactions are difficult to model, since they typically occur deep within the nucleus and are therefore subject to a variety of nuclear effects. As the precision of neutrino experiments increases, interaction models are becoming a major source of systematic uncertainties, making neutrino-nucleus interactions a crucial area of research in neutrino physics.

This chapter gives an overview of the neutrino's place within the Standard Model of particle physics, the history of its discovery and study, the phenomenon of flavour oscillation, and the interactions between neutrinos and atomic nuclei.

1.1 The Standard Model

The Standard Model (SM) of particle physics is a unified theory of matter at the fundamental level. It attempts to fully describe the elementary particles of nature and the forces that govern their interactions. Particles in the SM are classified into two groups according to their spin[1]: fermions with half-integer spin, and bosons with integer spin. The SM describes three fundamental forces which are mediated or 'carried' by spin-1 bosons: the electromagnetic force, mediated by the photon; the strong force, mediated by the gluon; and the weak force, mediated by the massive $W^{\pm}$ and Z^0 bosons. The final SM particle is the Higgs boson; the excitation of the Higgs field gives rise to the non-zero masses of other particles. The SM particles and their properties are summarised in Figure 1.1.

Fermions are divided into quarks and leptons, each of which comprise three distinct generations of paired particles. All stable matter in the universe is made up of first-generation particles, which are the lightest; heavier particles of higher generation decay to lighter, more stable ones. Additionally, for each fermion there is a corresponding antiparticle related to it by charge conjugation: particles and their antiparticles have identical mass and spin, but opposite charge and other internal quantum numbers.

The quarks interact with all three forces, and exist in six varieties or 'flavours': up, down, charm, strange, top, and bottom. These can be grouped by their charge into up-type quarks of charge $+\frac{2}{3}e$ (where e is the absolute value of the electron charge) and down-type quarks of charge $-\frac{1}{3}e$. In addition to electric charge, quarks possess one of three 'colour' charges associated with the strong force; all three colours together (or one colour and anti-colour) cancel. Quarks are not observed in free space: instead they bond via the strong force to form hadrons, which in nature consist of baryons (three quarks) and mesons (quark-antiquark pairs), though the possibility of structures of four or more quarks has recently been confirmed by high-energy experiments [2]. Free particles always have non-zero overall colour and individual quarks cannot be isolated; this is known as 'colour confinement'.

The particles of the lepton sector do not interact by the strong force. Three generations of charged lepton exist: the electron (e), the muon (μ) and the tau particle (τ), all with charge $-1e$. Each of these possesses a corresponding lepton flavour. Similarly, there exist three generations of neutral leptons: the neutrinos. These have no electric charge and so do not interact with the electromagnetic force (leaving only the weak interaction), but share

[1] Spin takes values of integer or half-integer multiples of $\hbar$, which will here be expressed in natural units such that $\hbar = 1$ for convenience.

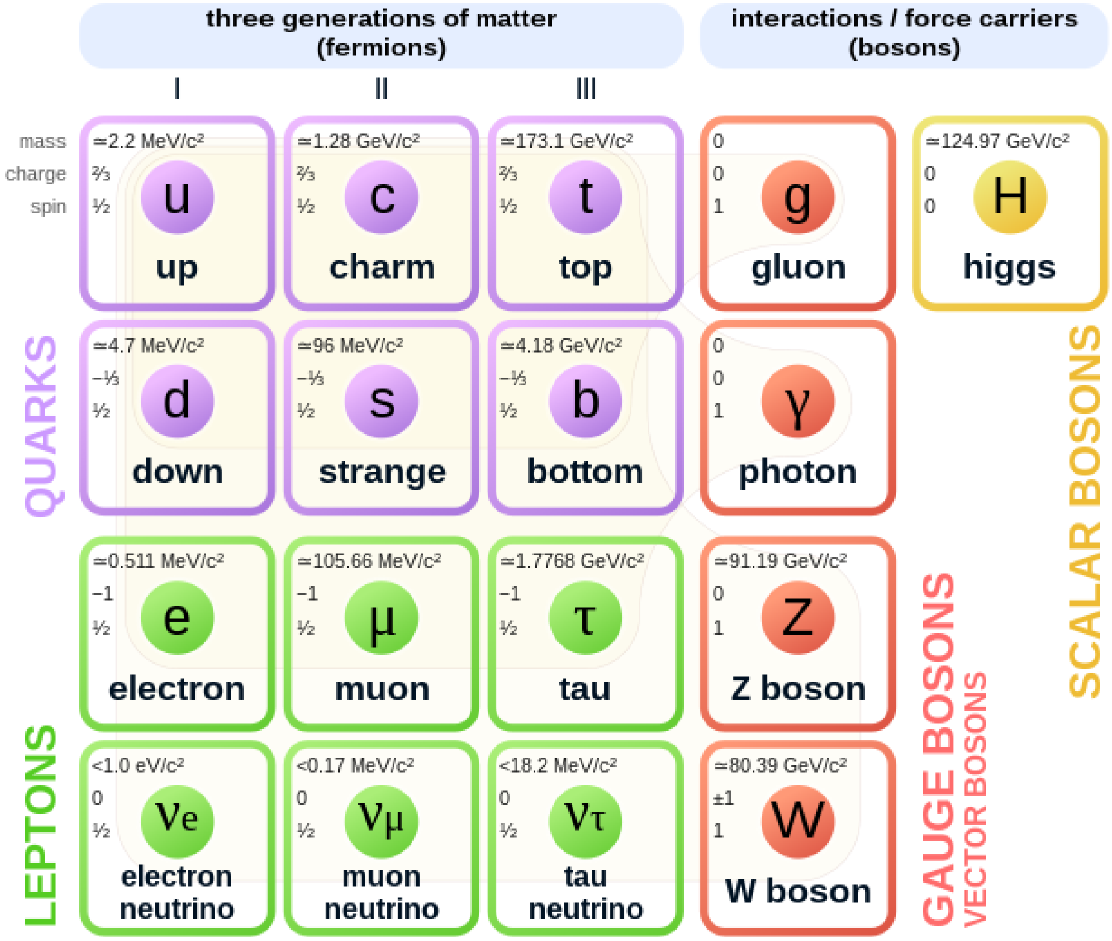


Figure 1.1: Elementary particles described by the Standard Model [1].

the same three flavours, and so are known as the electron-neutrino (ν_e), muon-neutrino (ν_μ) and tau-neutrino (ν_τ). The SM originally predicted that neutrinos should be massless, but this has been disproved by the phenomenon of neutrino oscillation (see Section 1.3 below), which also violates the conservation of lepton flavour that is otherwise observed.

The Standard Model has been extremely successful in predicting most particle interactions, but is not a complete theory of fundamental physics: it still has theoretical issues, and fails to account for several observed phenomena. Although the SM has now been modified to account for neutrino oscillations, it cannot address a number of important cosmological observations: the gravitational force and general relativity (which cannot yet be unified with quantum field theories), the existence of dark matter and dark energy, or the baryon asymmetry of the Universe. As a result, one of the primary goals of modern particle physics research is to search for physics beyond the Standard Model (BSM) and so work towards a full ‘theory of everything’. Neutrino physics relates to several of these issues: for example, charge-parity violation in neutrino oscillation may lead to leptogenesis [3] and thus an explanation of baryon asymmetry; and several parameters of neutrinos, such as their masses, remain unknown or poorly understood. Neutrino physics is consequently one of the broadest and most active areas of study within particle physics.

1.2 A history of the neutrino

Neutrinos were first discovered as a result of the study of radioactive decays in the late 19th and early 20th centuries. Following the discovery of nuclear decay by Henri Becquerel in the late 1890s [4], the existence of three distinct types of nuclear radiation was established by Ernest Rutherford and Paul Villard; Rutherford named these alpha, beta and gamma rays [5], each arising from a corresponding mode of nuclear decay. In 1914, James Chadwick demonstrated that the energy spectrum of electrons emitted in beta decay is continuous [6], in contrast to the narrow energy distributions of alpha and gamma particles. The beta particle was the only observed emission of beta decay at the time, so this appeared to violate conservation of energy. Additionally, it was found that beta decay always resulted in an integer change in nuclear spin, but with the electron spin being $\frac{1}{2}$, conservation of angular momentum was also seemingly being violated.

An explanation for these discrepancies was posited by Wolfgang Pauli in 1930, in his famous ‘letter to the radioactive ladies and gentlemen’ [7]. Pauli proposed that another, unseen particle was being emitted alongside the electron, carrying away the apparently missing energy. For consistency with known conservation laws and the observed beta decay spectrum, this particle would have to be electrically neutral, spin $\frac{1}{2}$, and have very small (or zero) mass. Pauli named this new particle the ‘neutron’. The particle we now know as the neutron would be discovered by Chadwick two years later [8], but its mass was too large to be Pauli’s particle, which was subsequently renamed the ‘neutrino’ (‘little neutral one’) by Enrico Fermi. Fermi would go on to formulate a theory of beta decay [9] using Pauli’s neutrino, describing it as the decay of a neutron to produce a proton, an electron and a neutrino:

$$n \to p + e^- + \bar{\nu}. \tag{1.1}$$

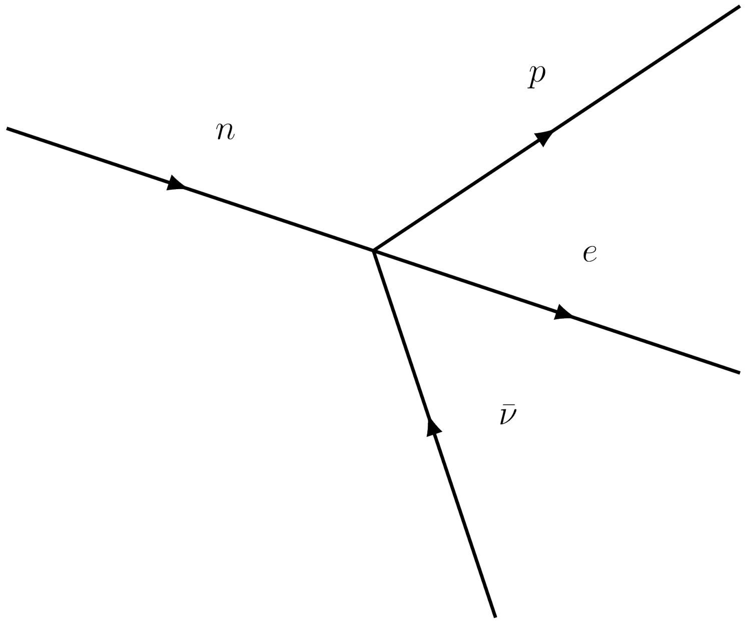


Figure 1.2: Feynman diagram (time left to right) of Fermi's model of beta decay, with a neutron decaying directly into a proton, an electron and an (anti)neutrino.

Fermi's model was incomplete as it had the four particles coupling directly as shown in Figure 1.2, rather than via a W boson as would later be understood, but it was nevertheless the first model to correctly describe beta decay as $n \to p + e^- + \bar{\nu}$. Crucially, it also predicted other interactions such as inverse beta decay:

$$\bar{\nu} + p \to e^+ + n \tag{1.2}$$

which would lead to the first detection of the neutrino itself two decades later.

In 1956, the Cowan-Reines experiment reported the first direct detection of neutrinos [10] — specifically electron-antineutrinos from beta decay, exploiting the huge neutrino flux in the vicinity of a nuclear reactor. Water tanks providing the target protons were sandwiched between tanks of liquid scintillator for gamma detection. Light produced by gamma rays passing through the scintillator was detected by photomultiplier tubes. The water target was doped with cadmium chloride ($CdCl_2$), which emits a gamma ray when it absorbs a neutron. This resulted in a characteristic signal of two gamma rays from annihilation of the positron, followed by a third gamma from neutron capture several microseconds later. This setup yielded a neutrino detection rate of approximately three events per hour, which was confirmed by shutting down the reactor and observing a reduction in the rate of detected events. This result confirmed the existence of the neutrino, and established the usefulness of large detectors to compensate for the very low cross-sections of neutrino interactions.

The existence of the muon-neutrino was later discovered by the first accelerator neutrino experiment [11], using the Alternating Gradient Synchroton (AGS) at Brookhaven National Laboratory. This established the existence of neutrino flavours corresponding to those of the charged leptons. The discovery of the tau particle in 1975 [12] led to the expectation that a corresponding tau-neutrino should exist, and this was confirmed in 2000 by the DONUT experiment [13].

The first hints of the phenomenon of neutrino oscillation were seen in the 1960s in the form of the solar neutrino problem. Models of nuclear fusion processes in the Sun using SM physics predicted large fluxes of neutrinos [14], but measurements made by the Homestake experiment [15] using capture of electron-neutrinos on chlorine atoms showed only about a third of the expected detection rate. This discrepancy persisted despite further improvement of both the solar models and the Homestake detector. The solution, proposed by Bruno Pontecorvo [16], was that neutrinos had non-zero mass and could oscillate between flavours, particularly over astronomical scales such as the distance between the Sun and the Earth. Pontecorvo showed that neutrino oscillation between ν_e and ν_μ states could result in a deficit like that observed by Homestake: solar fusion processes produce only ν_e, but a large proportion of these had oscillated to ν_μ, to which the Homestake detector was not sensitive. This would eventually be confirmed in 2002 [17] by the Sudbury Neutrino Observatory (SNO), which could detect neutral-current (NC) neutrino interactions and was thus sensitive to all three neutrino flavours, as well as being able to measure the ν_e rate in isolation via charged-current (CC) interactions. The SNO results showed that while there was a deficit in ν_e consistent with that of Homestake and other experiments such as Super-Kamiokande (SK) [18], the overall neutrino flux was consistent with the Standard Solar Model (SSM) predictions.

Meanwhile, other evidence for neutrino oscillations had been growing. A similar anomaly was observed in atmospheric neutrinos: in 1988, the Kamiokande experiment reported a deficit in ν_μ flux compared to predictions, but no such deficit in the ν_e flux [19]. This was reinforced by results from other experiments, including Kamiokande's successor experiment Super-Kamiokande [20], which showed that the ν_μ deficit was dependent on the direction of (and therefore the distance travelled by) the neutrino. Furthermore, disappearance of reactor antineutrinos was observed by the KamLAND experiment in 2003 [21]. Together, these observations of flavour change in solar, atmospheric and reactor neutrinos confirmed the theory of neutrino oscillations.

Since then, many experiments have been undertaken to measure the parameters of the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix that describes neutrino flavour mixing. These and the PMNS matrix itself are summarised in the following section. The mixing angles θ_{12}, θ_{23} and θ_{13} have all been measured, along with the mass square differences Δm^2_{12} and $|\Delta m^2_{23}|$, but other parameters remain unknown or poorly understood. The phase δ_{CP} which describes charge-parity (CP) violation in the neutrino sector is an important parameter with possible implications for baryon asymmetry. The T2K experiment in 2020 [22] published results which exclude $\delta_{\text{CP}} = 0$ at the 90% confidence level, strongly suggesting that such CP violation does exist in the neutrino sector, but its precise value remains unknown. The absolute values and ordering of the masses also remain unknown, as does the possibility that neutrinos may be Majorana particles — that is, particles which are their own antiparticles. Additionally, evidence such as the reactor anomaly [23] suggests the possibility of other neutrino flavours, which may be 'sterile' i.e. non-interacting with the weak force, and could have very large masses and thus provide a candidate for dark matter (although recent results from the MicroBooNE experiment do not support this [24]). Numerous experiments are currently working to probe these and other open questions in neutrino oscillation theory.

1.3 Neutrino oscillations

Neutrino oscillations are a quantum mechanical phenomenon in which neutrinos of one flavour appear to change to another after propagating through space. This happens because the three neutrino flavour eigenstates are not eigenstates of the Hamiltonian. Instead they are superpositions of three mass eigenstates:

$$|\nu_\alpha\rangle = \sum_{i=1}^{3} U_{\alpha i} |\nu_i\rangle \tag{1.3}$$

where α are the neutrino flavours e, μ, τ; i are the neutrino masses $1, 2, 3$; and U is the PMNS matrix. The PMNS matrix is typically parameterised by the three mixing angles θ_{12}, θ_{23}, θ_{13} and the CP-violating phase δ_{CP}, with which it can be written as the product of three rotation matrices:

$$U = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{pmatrix} \begin{pmatrix} c_{13} & 0 & s_{13}e^{-i\delta_{\mathrm{CP}}} \\ 0 & 1 & 0 \\ -s_{13}e^{i\delta_{\mathrm{CP}}} & 0 & c_{13} \end{pmatrix} \begin{pmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{1.4}$$

$$U = \begin{pmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta_{\mathrm{CP}}} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{i\delta_{\mathrm{CP}}} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta_{\mathrm{CP}}} & s_{23}c_{13} \\ s_{12}s_{23} - c_{12}c_{23}s_{13}e^{i\delta_{\mathrm{CP}}} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{i\delta_{\mathrm{CP}}} & c_{23}c_{13} \end{pmatrix} \tag{1.5}$$

where s_{ij} and c_{ij} are $\sin\theta_{ij}$ and $\cos\theta_{ij}$ respectively.

The PMNS matrix can be used to predict neutrino oscillations as follows. Assuming propagation in vacuum, we start with some initial neutrino flavour eigenstate at $t = 0$:

$$|\nu(t=0)\rangle = |\nu_\alpha\rangle = \sum_{k=1}^{3} U^*_{\alpha k} |\nu_k\rangle . \tag{1.6}$$

Temporal evolution of the mass eigenstates is governed by the Schrödinger equation:

$$i\frac{d}{dt}|\nu_k(t)\rangle = H|\nu_k(t)\rangle \tag{1.7}$$

where H is the Hamiltonian. Using a plane wave approximation $|\nu_k(t)\rangle = e^{-iE_k t}|\nu_k\rangle$, for the time evolution of the state in Equation 1.6 we arrive at:

$$|\nu(x,t)\rangle = \sum_{k}^{3} U^*_{\alpha i} e^{-i(E_k t - p_k x)} |\nu_i\rangle \tag{1.8}$$

where E_k and p_k are the energy and momentum of the kth mass eigenstate. This leads to a probability that a neutrino with flavour α will be later observed as flavour β after time t of

$$P(\nu_\alpha \to \nu_\beta, t) = |A(\nu_\alpha \to \nu_\beta, t)|^2 \tag{1.9}$$

where A is the probability amplitude given by

$$A(\nu_a \to \nu_\beta, t) = \langle \nu_\beta | \nu(t)\rangle = \sum_{k}^{3} U^*_{\alpha k} e^{-iE_k t} \langle \nu_\beta | \nu_k \rangle = \sum_{i}^{3}\sum_{k}^{3} U_{\beta i} U^*_{\alpha k} e^{-iE_k t} \langle \nu_i | \nu_k \rangle = \sum_{k}^{3} U_{\beta k} e^{-iE_k t} U^*_{\alpha k}. \tag{1.10}$$

Given that the neutrino masses are very small, we can assume they must be highly relativistic, so we can approximate the distance travelled (in natural units such that $c = 1$) as $L \approx t$ and the momentum of mass state i as

$$p_i = \sqrt{E_i^2 - m_i^2} \approx E_i - \frac{m_i^2}{2E_i} \tag{1.11}$$

using which we can express the mass eigenstate after travelling a distance L as

$$|\nu_i(L)\rangle = e^{-i\frac{m_i^2 L}{2E}} |\nu_i(0)\rangle \,. \tag{1.12}$$

Hence we can find the probability for flavour change as a function of distance:

$$P(\nu_\alpha \to \nu_\beta, L) = |\langle \nu_\beta(L)|\nu_\alpha\rangle|^2 = \left| \sum_i U_{\alpha i}^* U_{\beta i} e^{-i\frac{m_i^2 L}{2E}} \right|^2 \tag{1.13}$$

which can be written as

$$\begin{aligned} P(\nu_\alpha \to \nu_\beta, L) = \delta_{\alpha\beta} - 4\sum_{i>j} \mathrm{Re}\left(U_{\alpha i}^* U_{\beta i} U_{\alpha j} U_{\beta j}^*\right) \sin^2\left(\frac{\Delta m_{ij}^2 L}{4E}\right) \\ + 2\sum_{i>j} \mathrm{Im}\left(U_{\alpha i}^* U_{\beta i} U_{\alpha j} U_{\beta j}^*\right) \sin\left(\frac{\Delta m_{ij}^2 L}{2E}\right) \end{aligned} \tag{1.14}$$

where $\Delta m_{ij}^2 \equiv m_i^2 - m_j^2$ are known as the mass-squared differences. Thus it can be seen that neutrino oscillation probabilities depend on the energy, the distance propagated (the 'baseline' of an experiment), the parameters of the PMNS matrix, and two independent parameters arising from the mass squared differences, Δm_{23}^2 and Δm_{21}^2. Consequently neutrino oscillation experiments cannot measure the absolute values of the neutrino masses, only the differences of their squares, and their ordering currently remains unknown. That is, the measured Δm_{23}^2 and Δm_{21}^2 permit two different interpretations, illustrated in Figure 1.3.

It is extremely cumbersome to write Equation 1.14 in terms of the mixing angles of the PMNS matrix. However, since θ_{13} is small and $|\Delta m_{23}^2| \gg |\Delta m_{21}^2|$, there are many situations where only two neutrinos participate significantly in mixing. In these cases, only one mixing angle θ is required and so we can use a two-neutrino model with mixing matrix

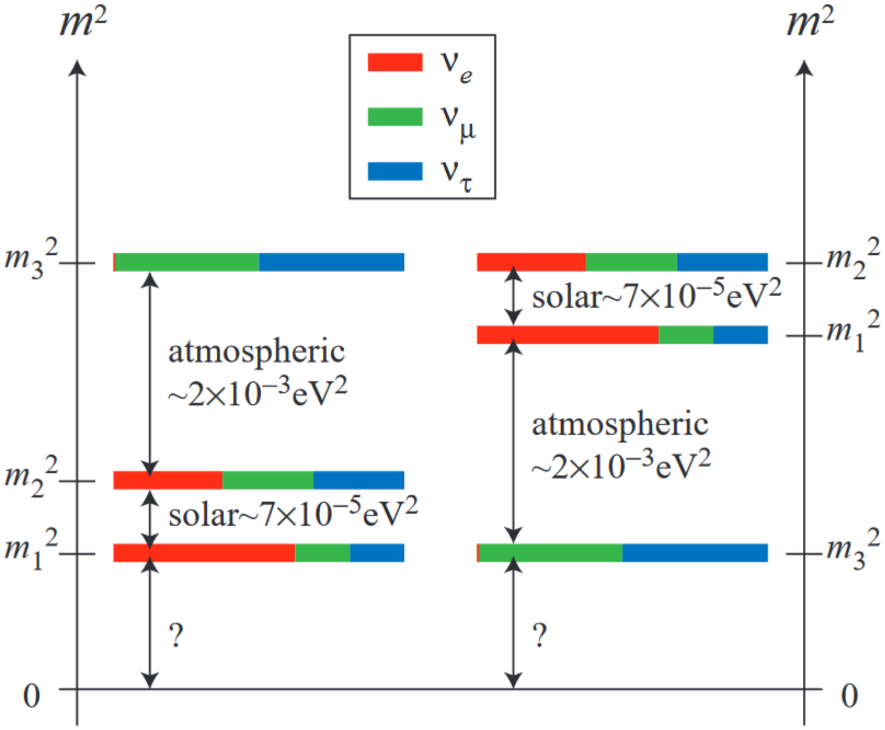


Figure 1.3: Possible neutrino mass orderings consistent with current measurements of Δm^2_{23} and Δm^2_{21}. The left-hand pattern is known as the 'normal hierarchy', in which the highest mass is much larger than the smaller two, as seen in the generations of other SM particles. The right-hand pattern is known as the 'inverted hierarchy'. The sizes of the coloured bands in each mass state represent the probabilities of finding a neutrino of each flavour from that mass eigenstate [25].

$$U = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \tag{1.15}$$

for which Equation 1.14 simplifies to

$$P(\nu_\alpha \to \nu_\beta, L) = \sin^2(2\theta)\sin^2\left(\frac{\Delta m^2 L}{4E}\right). \tag{1.16}$$

The maxima of P are known as 'oscillation maxima'. By controlling the energy E and the baseline L in an experiment, the detector can be placed at an oscillation maximum. This maximises the oscillation probability and therefore the sensitivity to oscillation parameters.

Neutrino oscillations can be measured using neutrinos from a variety of sources. Some of these are naturally-occurring, such as the fusion processes in the Sun or cosmic ray interactions in Earth's atmosphere. Others are man-made, such as nuclear fission reactors and particle accelerators. Different sources can produce neutrinos of different energy distributions, as well as different feasible baselines for experiment. Thus the oscillation parameters an experiment can probe depend in large part on what neutrino source it uses, and as a result, our understanding of the various neutrino oscillation parameters comes from the combination of various different experiments and sources.

1.3.1 Solar neutrinos

The Sun is a major source of neutrinos in our immediate environment. Nuclear fusion processes in its core produce ν_e in enormous quantities, which escape into space resulting in a flux at Earth's surface of order 10^{10} s^{-1}cm^{-2}. A number of different fusion reactions occur in the Sun, producing neutrinos of different energy spectra, some continuous and some discrete, as shown in Figure 1.4.

As described in Section 1.2 above, solar neutrinos provided the first evidence for neutrino oscillations. Solar neutrino experiments are predominantly sensitive to the mixing angle θ_{12} (which for this reason is sometimes known as the 'solar mixing angle') and Δm^2_{12}, and have very slight sensitivity to θ_{13}. The currently allowed regions for these parameters are the combination of results from multiple experiments, summarised in Figure 1.9.

Following their initial detection by the Homestake experiment, various solar neutrino experiments have been conducted. The deficit in ν_e observed by Homestake was confirmed by the Kamioka Nucleon Decay Experiment (Kamiokande) [28]. Kamiokande was originally intended to search for proton decay, but was also suitable for detecting solar neutrinos from the ^{8}B reaction via elastic scattering:

$$\nu_e + e^- \rightarrow \nu_e + e^-. \tag{1.17}$$

Kamiokande used a large cylindrical tank containing 3000 tons of water and instrumented with 1000 photomultiplier tubes (PMTs) to detect Cherenkov radiation, placed 1km underground in the disused Kamioka zinc mine to provide shielding from cosmic rays. Electrons scattered by neutrino interactions produced Cherenkov light detected only by the PMTs in their direction of travel, enabling reconstruction of the neutrino direction. Thus, in addition to measuring a $> 2\sigma$ discrepancy in the neutrino flux compared to SSM

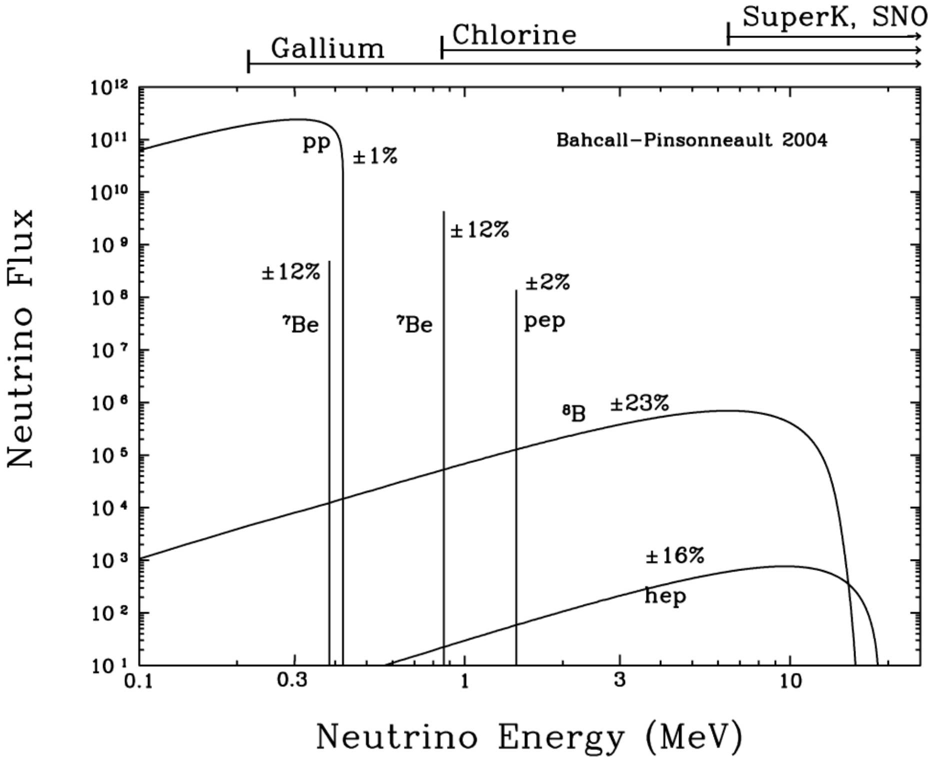


Figure 1.4: Energy spectrum of solar neutrinos separated by fusion reaction type, as predicted by the BP04 solar model by J. Bahcall and M. Pinsonneault [26]. The flux is given in number of neutrinos $\mathrm{cm}^{-2}\mathrm{s}^{-1}\mathrm{MeV}^{-1}$ for continuous sources, and in number of neutrinos $\mathrm{cm}^{-2}\mathrm{s}^{-1}$ for line sources. The total theoretical uncertainty is shown for each source, and the regions of sensitivity for different experiments along the top of the plot [27].

predictions, Kamiokande was able to use the directional information to isolate the neutrino signal from the Sun.

As the solar neutrino problem persisted, further experiments attempted to verify the predictions of the SSM. An important step in this was to probe the flux of neutrinos produced by the pp reaction, which make up the bulk of solar neutrinos but have low energies (0–0.42 MeV) to which Homestake and Kamiokande were not sensitive. This was addressed by use of inverse beta decay on gallium:

$$\nu_e + {}^{71}Ga \rightarrow {}^{71}Ge + e^- \tag{1.18}$$

with the resulting germanium being extracted by chemical methods. A number of gallium experiments made use of this method: the GALLium EXperiment (GALLEX) [29], the Gallium Neutrino Observatory (GNO) [30], and the Soviet-American Gallium Experiment (SAGE) [31]. Together, they measured a neutrino flux deficit similar to that observed by Homestake and Kamiokande.

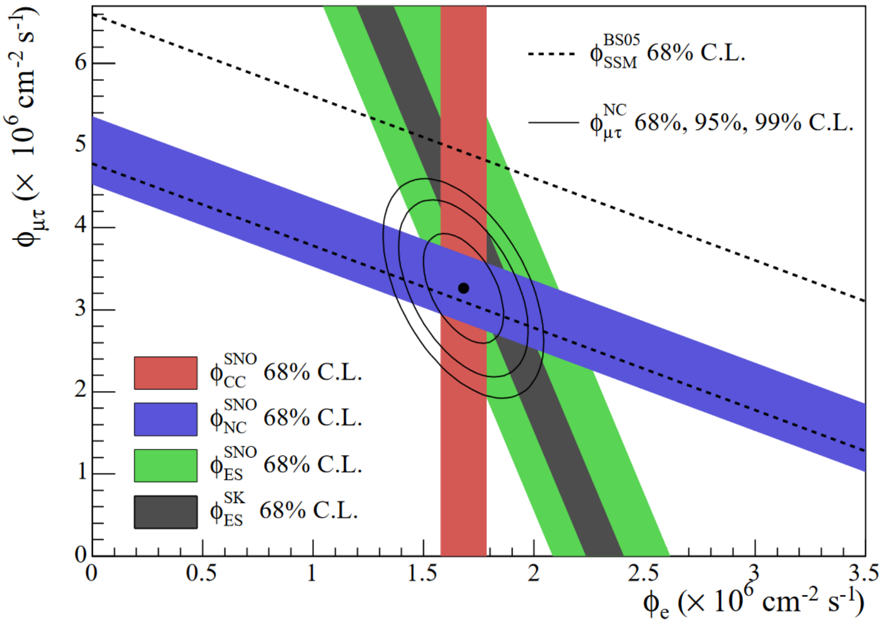


Figure 1.5: Plot of the $\nu_\mu + \nu_\tau$ flux vs the ν_e flux measured by the SNO experiment. The coloured bands indicate the 1σ confidence level for the different interaction types, with the black band representing the equivalent elastic scattering results from Super-Kamiokande. The dashed lines indicate the SSM flux prediction. The black point represents the flux of ν_e from CC, and $\nu_\mu + \nu_\tau$ from NC-CC difference, and is encircled by the 68%, 95% and 99% confidence level contours [32].

The Cherenkov detector experiments Super-Kamiokande [18] and SNO [17] were able to make high-precision measurements of ^{8}B solar neutrinos. Super-Kamiokande was the successor to Kamiokande, using a much larger tank containing 50 kton of water and about 11,200 PMTs (please see Section 2.3 for a full description of the SK detector). SNO used a 1000-tonne heavy-water (D_2O) sensitive volume and 9,522 PMTs mounted on a geodesic sphere. The use of deuterium enabled SNO to detect not only charged-current ν_e reactions:

$$\nu_e + D \rightarrow p + p + e^- \tag{1.19}$$

but also neutral-current (NC) and elastic scattering (ES) reactions:

$$\nu_\alpha + D \rightarrow p + n + \nu_\alpha \tag{1.20}$$

$$\nu_\alpha + e^- \rightarrow \nu_\alpha + e^- \tag{1.21}$$

and thus measure the overall neutrino flux across all flavours $\alpha = e, \mu, \tau$ directly. The SNO results, summarised in Figure 1.5, confirmed that the deficit of solar ν_e seen by previous experiments was due to neutrino flavour change to ν_μ and ν_τ, and that the overall flux was in line with SSM predictions. Results from SK and SNO continue to dominate the solar neutrino constraints on the oscillation parameters as shown in Figure 1.9 below.

1.3.2 Atmospheric neutrinos

The interactions of cosmic rays with Earth's atmosphere provide another major natural source of neutrinos. Cosmic ray protons and heavier nuclei interact with atoms in the atmosphere to produce air showers. These showers contain large numbers of energetic mesons (mainly charged pions and kaons) which decay, producing neutrinos. These neutrinos typically have much higher energy than solar neutrinos, with an energy spectrum that peaks around 1 GeV; atmospheric neutrino energies have been observed from the 100s of MeVs to the 100s of TeVs. Additionally, since neutrinos can pass through the entirety of the Earth, atmospheric neutrinos can be detected over a wide range of baselines. These properties give atmospheric neutrinos a wide range of L/E, making them a very useful probe of neutrino oscillations.

Similarly to solar neutrinos with θ_{12}, atmospheric neutrinos are mainly sensitive to θ_{23} (known as the 'atmospheric mixing angle') as well as Δm^2_{23}. This results in a signal of ν_μ disappearance in upward-going muon-like interactions (and appearance of ν_τ, which are not directly produced in the atmosphere in significant amounts, although these are difficult to reconstruct). 'Upward-going' refers to neutrinos that pass through the Earth before being detected and thus have sufficient L/E for significant oscillation, whereas 'downward-going' neutrinos have travelled a much shorter distance so do not. The difference between the two is known as the 'up-down asymmetry', and is also affected by the Mikheyev-Smirnov-

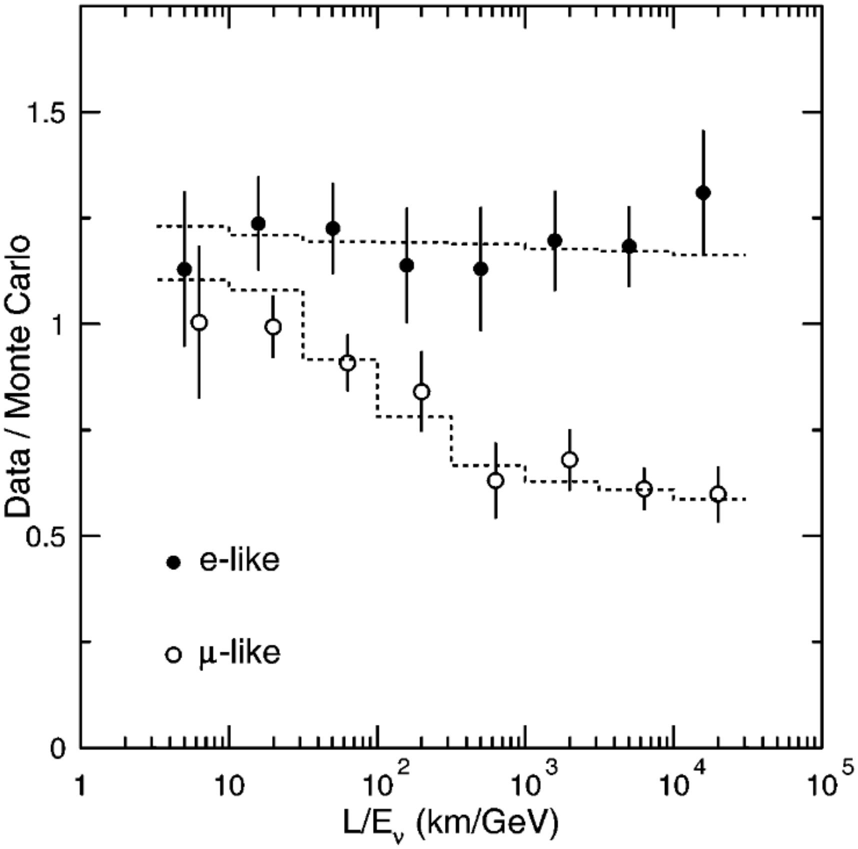


Figure 1.6: Comparison of SK atmospheric neutrino data to SM predictions in the absence of oscillations. The expected shape for $\Delta m^2_{23} = 2.2 \times 10^{-3}\text{eV}^2$ and $\sin^2 2\theta = 1$ is shown as dashed lines [20].

Wolfenstein (MSW) effect, whereby neutrino oscillations in matter are modified by the presence of electrons [33].

Atmospheric neutrinos were first detected in 1965 by experiments at the Kolar Gold Field mines in India [34] and the East Rand Proprietary mine in South Africa [35], which used extremely deep underground laboratories for shielding purposes. Later, Kamiokande was used to examine atmospheric neutrinos and found a deficit in ν_μ flux compared to predictions, whereas the ν_e flux was as expected [19]. This became known as the 'atmospheric neutrino anomaly'.

Super-Kamiokande published a measurement of the atmospheric neutrino flux as a

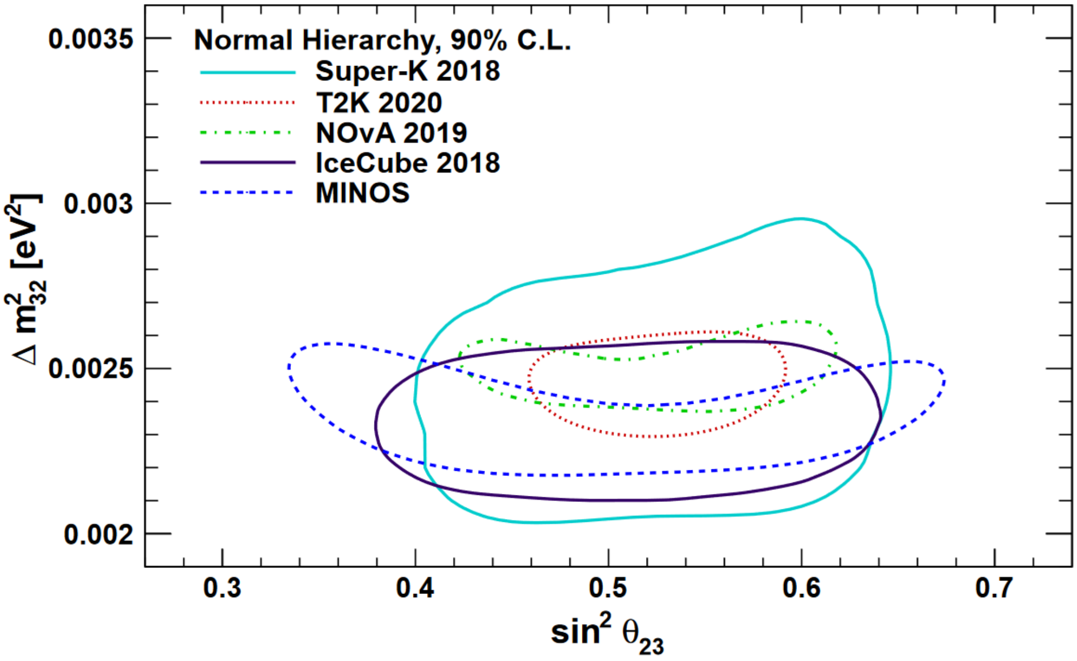


Figure 1.7: Allowed regions for Δm^2_{32} and θ_{23} measured by the atmospheric neutrino experiments Super-Kamiokande and IceCube, as well as the accelerator neutrino experiments T2K, NOνA and MINOS [38].

function of direction in 1998 [20], demonstrating the up-down asymmetry for the first time and providing strong evidence for atmospheric neutrino oscillations as shown in Figure 1.6. SK continues to record atmospheric neutrino data, and contributes to current constraints of atmospheric neutrino mixing parameters alongside the neutrino telescopes ANTARES [36] and IceCube [37]. The current measurements of SK and IceCube (as well as accelerator experiments sensitive to the same parameters) are summarised in Figure 1.7. Atmospheric neutrinos also offer methods of probing the neutrino mass ordering through the MSW effect, since neutrinos and antineutrinos will be affected differently depending on the ordering, though current experiments have yet to resolve this due to the difficulty of distinguishing ν_e from $\bar{\nu}_e$ with water Cherenkov detectors [38].

1.3.3 Reactor neutrinos

Nuclear reactors produce large quantities of electron-antineutrinos from beta decay processes. These neutrinos have relatively low energies, potentially allowing for long baselines, but the neutrino flux decreases with distance according to the inverse square law. The choice of baseline for a reactor neutrino experiment is therefore a tradeoff between oscillation amplitude and detection rate. Such experiments can be divided into short-baseline (SBL) with L of order 1 km or less, and long-baseline (LBL). Reactor antineutrinos are typically detected by inverse beta decay on atomic nuclei, but they have low energy so this only occurs for $\bar{\nu}_e$; $\bar{\nu}_\mu$ and $\bar{\nu}_\tau$ cannot be detected. Consequently reactor neutrino oscillations can only be measured by $\bar{\nu}_e$ disappearance, so they cannot probe δ_{CP} as this requires the appearance channel. This can be seen as beneficial however, as the absence of this unknown parameter improves the precision with which other parameters can be measured [38].

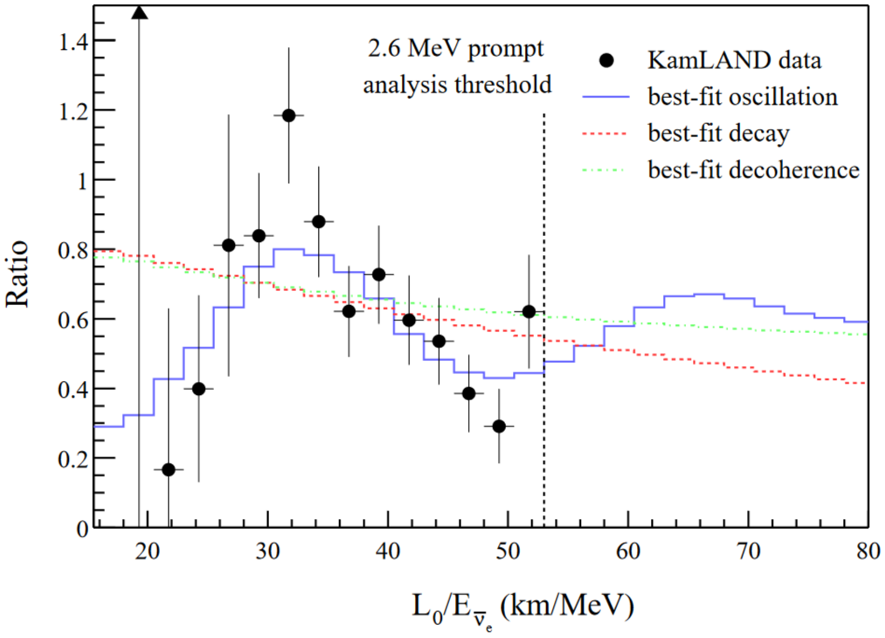


Figure 1.8: Ratio of the $\bar{\nu}_e$ spectrum observed by KamLAND to the expectation for no oscillation. The prediction of the oscillation model is shown in blue, in good agreement with the data. All data points and models were plotted with baseline $L_0 = 180$ km, as if all antineutrinos detected were due to a single reactor at the average distance [39].

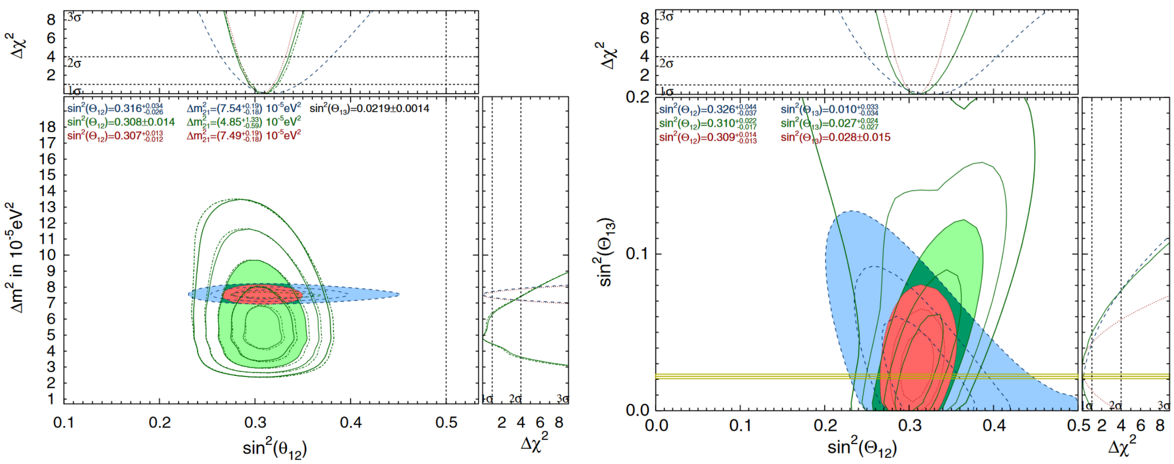


Figure 1.9: Allowed regions for Δm^2_{21} vs θ_{12} (left) and θ_{13} vs θ_{12} (right) from all solar neutrino data plus the KamLAND reactor antineutrino experiment. Filled regions represent the 3σ confidence level from solar neutrino data (green), KamLAND (blue), and the combined result (red). The $\sin^2\theta_{13}$ measurement from reactor neutrino data is shown as a yellow band on the right-hand plot [40].

The Kamioka Liquid Scintillator AntiNeutrino Detector (KamLAND) was the first experiment to observe neutrino oscillations in reactor antineutrinos [21]. Situated in the same mine cavern that previously housed Kamiokande, KamLAND uses 1000 tons of liquid scintillator surrounded by 1,879 PMTs to detect $\bar{\nu}_e$ from 53 nuclear reactors, with an average baseline of approximately 180 km. This enabled the first significant observation of $\bar{\nu}_e$ disappearance as shown in Figure 1.8, and a high-precision measurement of Δm^2_{21} [39].

Reactor neutrino experiments have particularly good sensitivity to θ_{12} and Δm^2_{21}. This makes them complementary to solar neutrino experiments, as shown with the inclusion of KamLAND results in Figure 1.9. They also have good sensitivity to θ_{13}, and in 2012 the discovery of non-zero θ_{13} was made by the Daya Bay [41] and Reactor Experiment for Neutrino Oscillation (RENO) [42] reactor experiments. Future reactor neutrino experiments such as the Jiangmen Underground Neutrino Observatory (JUNO) [43] will greatly improve the precision of oscillation parameter measurements, and are also expected to be capable of determining the neutrino mass ordering.

1.3.4 Accelerator neutrinos

Intense beams of neutrinos can be produced using particle accelerators. The conventional method used by most experiments is to generate a beam of protons which is fired at a nuclear target, producing charged pions and kaons which produce (anti)neutrinos in their decay. Pion decays mainly produce muon-(anti)neutrinos so these will dominate the beam, while a small proportion will be electron-(anti)neutrinos from kaon and muon decay. The accompanying $\mu^{\pm}$ and any remaining mesons are then 'ranged out' in solid matter, leaving a neutrino beam. Positive or negative pions can be selected to produce a neutrino or antineutrino beam respectively. The beam can then be sampled by detector(s) after it has travelled some distance to measure how the flavour composition has evolved. This is often done using two (or more) detectors: a far detector placed at an oscillation maximum, and a near detector at a location close to the beam target where oscillation remains minimal, to better control systematic uncertainties.

Neutrino beams give experiments valuable control over E/L. They can be produced with very high intensities, so long baselines are feasible. The neutrino energy can also be controlled: although placing the detector directly on the beam axis provides the greatest integrated flux, a much narrower energy spectrum can be obtained with an off-axis configuration as a result of the Jacobian peak (see Figure 2.3 below).

Since conventional neutrino beams are primarily made up of muon-neutrinos, accelerator neutrinos are sensitive to the 'atmospheric' parameters θ_{23} and Δm^2_{32} through the $\nu_\mu/\bar{\nu}_\mu$ disappearance channel; current measurements of these parameters from accelerator neutrino experiments are shown in Figure 1.7 alongside atmospheric experiments. Additionally, the $\nu_e/\bar{\nu}_e$ appearance channel provides sensitivity to θ_{13} and δ_{CP}.

Today, the experiments most sensitive to the atmospheric parameters are the accelerator experiments Tokai to Kamioka (T2K) and NuMI Off-Axis ν_e Appearance (NOvA) [44], both using proton accelerator neutrino beams with off-axis detectors. The T2K experiment, described in Chapter 2, made the first observations of ν_e appearance in a ν_μ beam [45] and the first significant constraint on δ_{CP} [22]. The NOvA detector samples the NuMI beam at an angle of 0.84°, resulting in an energy spectrum peaked at approximately 2 GeV, with a 810 km baseline. Both T2K and NOvA continue to increase the precision of measurements of θ_{23}, Δm^2_{32}, θ_{13} and δ_{CP} as they record more data and improvements are made to their respective beams. The current constraints on δ_{CP} and the mixing angles from these experiments are shown in Figure 1.10. Planned future accelerator experiments include the Deep Underground Neutrino Experiment (DUNE) [46] and Tokai to Hyper-Kamiokande [47], which are expected to have excellent sensitivity to δ_{CP} and the mass ordering respectively.

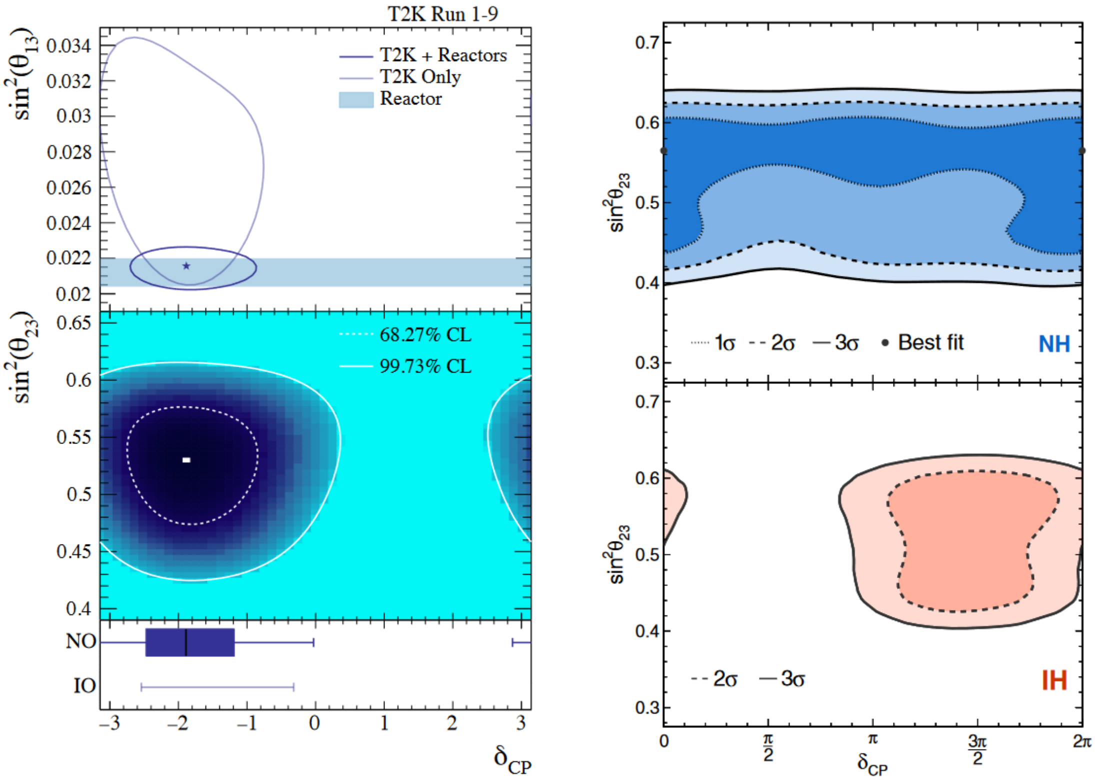


Figure 1.10: Constrained regions in δ_{CP} and the mixing angles by T2K (left [22]) and NOvA (right [48]). The top T2K plot shows the 68.27% confidence level assuming normal hierarchy, comparing and combining T2K results with reactor experiments. The middle T2K plot shows the 68.27% and 99.73% confidence intervals (dashed and solid white lines respectively) for T2K data combined with the reactor θ_{13} constraint, again assuming normal hierarchy. The bottom T2K plot shows the 68.27% (shaded region) and 99.73% (error bar) confidence intervals for δ_{CP} for each ordering/hierarchy, again combined with the reactor θ_{13} constraint. The NOvA plots show the allowed regions for δ_{CP} and θ_{23} from NOvA data assuming the normal (top) and inverted (bottom) hierarchies.

1.4 Neutrino-nucleus interactions

Most neutrino oscillation experiments rely on interactions between neutrinos and atomic nuclei for neutrino detection, but these interactions are far from trivial to model. While there is some data from hydrogen bubble chambers [49], hydrogen targets are uncommon due to the difficulty of safely containing liquid hydrogen; most experiments use heavier target materials such as organic scintillator (CH), water (H_2O), iron (Fe) or noble liquids (Ar). As a result the neutrino interacts not with a free nucleon, but with one bound in a nucleus (or indeed with the nucleus as a whole), so both the initial and final states of the interaction are subject to a variety of nuclear effects. These effects are very difficult to probe because we cannot observe the initial neutrino-nucleon interaction, only the particles that leave the nucleus, so initial and final state nuclear effects cannot be measured separately. Neutrino-nucleus interactions must be well understood in order to measure the neutrino energy and interaction rate accurately, but current models require refinement with both theoretical developments and more experimental data. Systematic uncertainties arising from neutrino-nucleus interaction models are becoming a major limiting factor on precision in modern oscillation experiments, so this is a crucial area of research for neutrino physics [50].

Modelling neutrino-nucleus interactions begins with an understanding of the interactions between a neutrino and an individual free nucleon (proton or neutron). These interactions are modified by factors arising from the nuclear environment, collectively referred to as ‘nuclear effects’: these include short-range interactions, long-range screening effects, and final state interactions. Nuclear effects also give rise to entirely new modes of interaction such as coherent scattering, in which the neutrino interacts with the nucleus as a whole.

Nuclear models have been implemented in a number of Monte Carlo (MC) event generators, which are able to simulate a wide variety of neutrino interactions and nuclear effects. These have been greatly refined over recent decades as models improve, and include NEUT [51], GENIE [52], NuWro [53], and GiBUU [54]. The development of these generators and their underlying nuclear models has been supported by data from a number of cross-section experiments, such as ArgoNeuT [55], MiniBooNE [56], MINERνA [57], NOvA [44], and T2K. However, discrepancies remain between different generators, and between generators and experimental data [58], so further work is needed in order to continue building more accurate neutrino-nucleus interaction models.

1.4.1 Neutrino-nucleon interactions

Neutrinos can interact with individual nucleons in a number of ways. These are categorised broadly into charged-current (exchange of a $W^{\pm}$) and neutral-current (exchange of a Z^0), and more specifically into scattering types:

- Elastic scattering (ES) is an exclusively NC process in which the particle content remains the same, with only the four-vectors changing:

  $$\nu_l(k_1) + a(k_2) \to \nu_l(k_1') + a(k_2') \tag{1.22}$$

  where $a = n, p$ is the nucleon, l is the lepton flavour, and k are the four-vectors.

- Quasi-elastic scattering (QE) is the CC analogue of ES, in which the particle content changes but remains a two-particle final state:

  $$\nu_l(k_1) + n(k_2) \to l^-(k_1') + p(k_2') \tag{1.23}$$

  $$\bar{\nu}_l(k_1) + p(k_2) \to l^+(k_1') + n(k_2'). \tag{1.24}$$

- In resonance scattering (RES), one or more mesons are produced via a resonant state. The simplest example is single pion production, e.g.:

  $$\nu_l + p \to l^- + \Delta^{++} \to l^- + p + \pi^+ \tag{1.25}$$

  though RES processes also include the production of multiple pions or the heavier meson species.

- In deep inelastic scattering (DIS), the neutrino has sufficient energy to resolve the individual quarks of the nucleon, resulting in a jet of hadrons rather than a single nucleon in the final state.

1.4.2 Nuclear effects

A variety of nuclear effects are known to influence interactions between neutrinos and atomic nuclei, affecting both the initial and final states of the interaction. These can have a substantial effect on the energy and composition of the observed outgoing particles compared to an interaction on a free nucleon.

Estimates of neutrino energy depend either on a sum of the total energy exiting the nucleus, in which case one needs to detect as many of the outgoing particles as possible and correct for any missed with a good physics model; or in the QE case, the neutrino energy can be inferred from the final state lepton kinematics. This assumes a stationary target, but a nucleon bound in a nucleus is not at rest: it is subject to Fermi motion. This is often modelled with a Relativistic Fermi Gas (RFG) treatment, based on the Smith-Moniz model [59], in which the nucleus is modelled as a simple potential well populated by neutrons and protons. This results in a smearing of the neutrino energy measured in the QE method; most other nuclear effects lead to an underestimate [38].

It is possible for a neutrino to interact with multiple correlated nucleons. These are known as multinucleon processes, or npnh (‘n particles-n holes’), since they result in some number n nucleons in the final state. There is growing evidence that the $n = 2$ case (2p2h) makes up a significant proportion of neutrino-nucleus interactions [38], and can have a substantial impact on measurements of neutrino oscillation parameters [50]. Alternatively, a neutrino may scatter off the nucleus as a whole, leaving it intact with little energy transfer: this is known as ‘coherent scattering’. Coherent scattering may be quasi-elastic, or may result in the production of mesons.

Following a neutrino-nucleon interaction, the final state particles must traverse the nuclear medium before exiting the nucleus. Consequently they may undergo interactions with nucleons and/or each other before they can be observed in any detector. These are known as final state interactions (FSIs), and include:

- Rescattering: the final state particles may undergo further elastic or inelastic scattering interactions, which change their energy and/or content (e.g. charge exchange),
- Production: additional particles may be created,
- Reabsorption: final state particles may be absorbed by the nucleus, so they cannot be observed among the outgoing particles.

Since we can only detect the particles that leave the nucleus, FSIs can change the apparent reaction products. Figure 1.11 shows an example of this: although the initial interaction is resonant pion production, the pion is reabsorbed so cannot be detected.

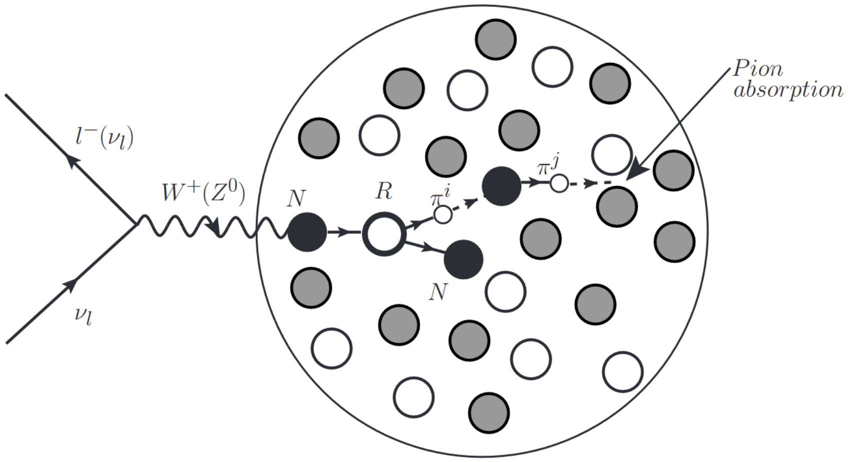


Figure 1.11: Illustration of an example of a final state interaction. Resonant pion production has occurred in the initial interaction, but the pion has been reabsorbed [38].

Only the outgoing lepton can be observed, and thus the reaction signature appears to be quasi-elastic. In NEUT, FSIs are simulated using a cascade model [51]: each hadron produced in the initial interaction is propagated step-by-step until it leaves the nucleus, with any interactions occurring at each step decided according to the mean free paths of the modelled interaction channels. Models such as these enable MC simulation of the outgoing particles that may be observed as a result of any initial interaction. Since we cannot observe the initial interaction, neutrino-nucleus interaction events can only be categorised by the topology: the set of particles that leave the nucleus after any FSIs.

1.4.3 Pion production

At the energies of accelerator neutrino experiments such as T2K, many of the above processes can result in the presence of one or more pions in the outgoing particles. Pions can be produced in the initial interaction by resonant processes or coherent scattering; or they can be the product of FSIs [60]. As a result, topologies containing pions are common. Using topologies as defined by T2K, the simplest of these is neutrino-induced charged-current single-pion production (CC1pi): charged-current interactions in which the particles that leave the nucleus contain a single charged pion (interactions producing a π^0 are considered separately) and no other mesons. In the initial interaction, CC1pi can occur through two RES channels (illustrated in Figure 1.12):

$$\nu_l(\bar{\nu}_l) + p \rightarrow l^-(l^+) + \pi^+(\pi^-) + p \tag{1.26}$$

$$\nu_l(\bar{\nu}_l) + n \rightarrow l^-(l^+) + \pi^+(\pi^-) + n \tag{1.27}$$

and coherent scattering as follows:

$$\nu_l(\bar{\nu}_l) + A \rightarrow l^-(l^+) + A + \pi^+ \tag{1.28}$$

where A is the target nucleus. If the number of outgoing mesons remains unchanged, these will result in a CC1pi topology. Alternatively, single-pion topologies can arise as a result of FSIs: a QE final state may undergo pion production as it traverses the nucleus, or a multiple-pion final state may have all but one pion be reabsorbed in the nucleus. Pions may also undergo charge exchange before leaving the nucleus, so the observed pion species may not be the same as that produced in the initial interaction.

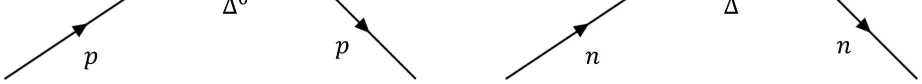


Figure 1.12: Feynman diagrams of resonant single pion production in scattering between an antineutrino and a proton (left) and neutron (right) respectively. Each interaction proceeds via a Δ resonance.

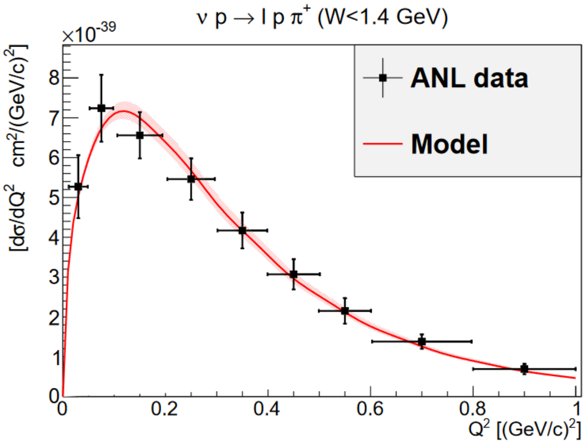


Figure 1.13: Fit of a neutrino-nucleon single-pion production model to cross-section data from the ANL experiment, showing the cross-section in relation to the squared 4-momentum transfer $Q^2 = (p_l - p_\nu)^2$, where p_l and p_ν are the 4-momenta of the lepton and neutrino respectively. The Q^2-differential cross-section for the interaction $\nu_l + p \to p + \pi^+$ is plotted as a function of Q^2, with a cut on the invariant mass of $W < 1.4$ GeV. The shaded area shows the variation in the model prediction due to the uncertainty of a form factor parameter [60].

Topologies containing pions make up a significant fraction of charged-current events at T2K, and of these, CC1pi events are the most common subtype. They are therefore a major background for CCQE measurements and also a potentially valuable channel for the T2K oscillation analysis, but require precise measurements to control systematic uncertainties on measurements made at the T2K far detector. Moreover, measurements of CC1pi cross-sections are needed in order to improve neutrino-nucleus interaction models (an example of a CC1pi cross-section model fit to experimental data is shown in Figure 1.13). This thesis presents work undertaken as part of the wider efforts at T2K to better select and understand CC1pi events in the ND280 detector.

Chapter 2

The T2K Experiment

Tokai to Kamioka (T2K) is a long-baseline neutrino oscillation experiment based in Japan, designed to measure the mixing angle θ_{13} via ν_e appearance in a ν_μ beam. As illustrated in Figure 2.1, a beam of muon-(anti)neutrinos is produced from a high-intensity proton beam at the Japan Proton Accelerator Research Complex (J-PARC) in Tokai and sampled by near and far detectors. Properties of the beam before oscillation are determined by a suite of near detectors situated in Tokai. The far detector Super-Kamiokande (SK) is situated 295 km away and samples the beam following oscillation. The beam is oriented 2.5° off-axis with respect to SK and the magnetised near detector, ND280, to provide a narrow band of neutrino energies with a peak at ~ 0.6 GeV which maximises oscillation at the far detector.

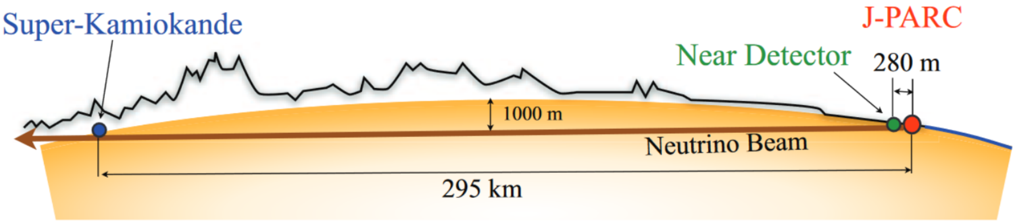


Figure 2.1: Schematic of the T2K experiment baseline showing the neutrino beam axis and the locations and relative distances of the J-PARC accelerator where the beam is produced, the near detector complex in Tokai, and the far detector Super-Kamiokande in Kamioka. The beam travels 295 km through the Earth's surface beneath the main island of Japan before reaching SK. Sea level is shown in yellow and ground level in black [61].

This chapter gives an overview of the T2K experiment, describing its physics goals, detector design and software. A particular focus will be placed on the ND280 detector, for which the event selections and particle identification tools presented in this thesis were developed.

2.1 Motivations

The T2K experiment was designed primarily to measure $\nu_\mu \to \nu_e$ oscillation, giving it sensitivity to θ_{13}, and also to measure $\sin^2\theta_{23}$ and the mass difference Δm^2_{23} through ν_μ disappearance, as well as the CP-violating phase δ_{CP} by comparing oscillation of neutrinos and antineutrinos. These design requirements motivated the use of both near and far detectors, as well as the choice of an off-axis beam, which reduces the flux but offers a much narrower neutrino energy distribution than an on-axis beam, and hence enables maximisation of the ν_e appearance rate at the far detector. The beam can be run in either neutrino (ν_μ) or antineutrino ($\bar{\nu}_\mu$) mode and thus oscillation rates can be compared between the two.

The near detectors fulfil a variety of functions. Sampling the unoscillated beam, they provide measurements of the neutrino energy spectrum, flavour content and interaction rates which are essential to characterise the signals recorded at the far detector. In addition, they are designed to support a wide programme of inclusive and exclusive neutrino-nucleus interaction cross-section measurements.

2.2 Neutrino beam

The T2K muon-(anti)neutrino beam is produced at the J-PARC accelerator complex. A proton beam is accelerated to 30 GeV kinetic energy and supplied to the T2K neutrino beamline in spills of eight bunches at a time. The beamline, shown in Figure 2.2, consists of primary and secondary sections which transport the proton beam and convert it into a neutrino beam respectively. In the primary beamline, the proton beam extracted from the accelerator main ring is bent by 80.7° to point towards the direction of Kamioka. It then enters the secondary beamline and impinges on a graphite target to produce charged pions, which are focused by magnetic horns before reaching the decay volume in which they decay to produce the beam neutrinos. The magnetic horns can be run in two different modes by switching the polarity of the current: in ‘forward horn current’ (FHC) mode, π^+ are

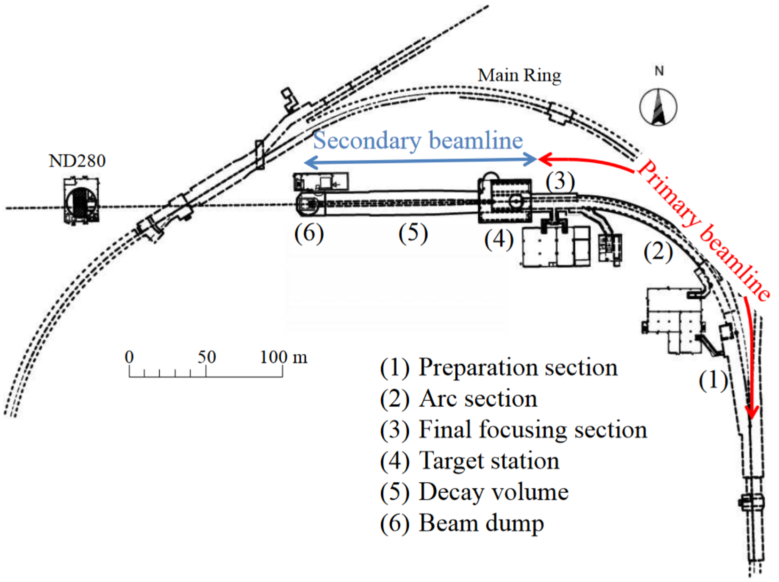


Figure 2.2: Diagram of the T2K neutrino beamline as viewed from above. The primary beamline, labelled in red, transports the proton beam to point towards Kamioka. The secondary beamline, labelled in blue, converts the proton beam into a muon-(anti)neutrino beam [61].

selected to yield a neutrino beam; in 'reverse horn current' (RHC) mode, π^- are selected to yield an antineutrino beam. The pions decay mainly into muons and muon-neutrinos:

$$\pi^+ \rightarrow \mu^+ + \nu_\mu \tag{2.1}$$

$$\pi^- \rightarrow \mu^- + \bar{\nu}_\mu . \tag{2.2}$$

A small contamination of electron-(anti)neutrinos is present as a result of decays of muons and kaons. Particles leaving the decay volume then enter the beam dump, which absorbs any remaining hadrons and muons below ~ 5 GeV/c, leaving only the neutrino beam and muons above ~ 5 GeV/c. A muon monitor is located behind the beam dump to detect these high-energy muons, which enables indirect monitoring of the intensity and direction of the neutrino beam. Data statistics for T2K beam runs are quantified by protons on target (POT), that is, the number of beam protons incident on the target constituting a particular run.

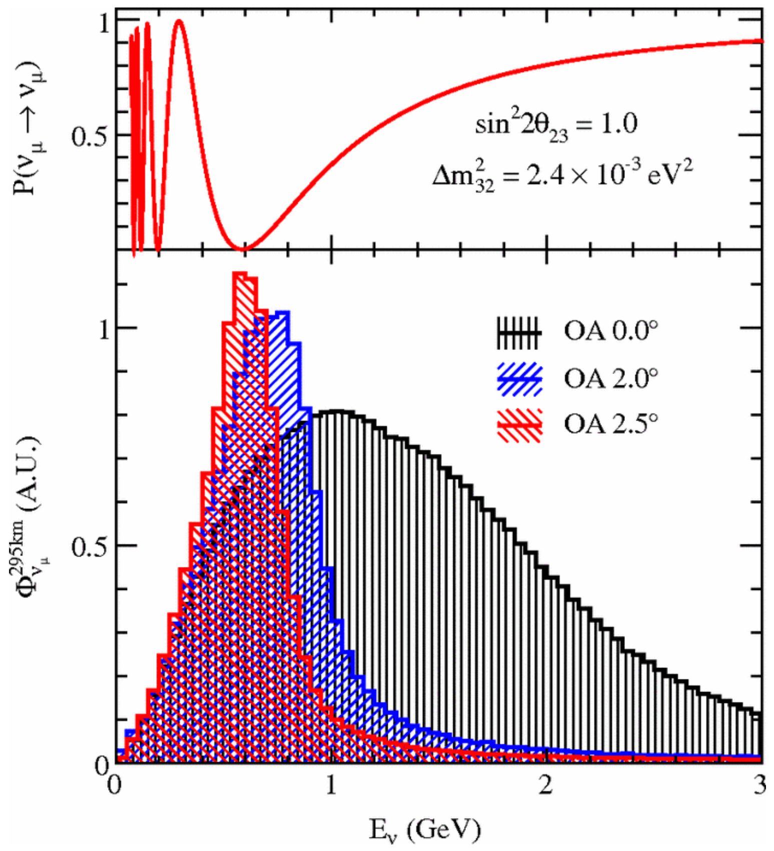


Figure 2.3: Neutrino fluxes (bottom) and predicted muon-neutrino survival probability at 295 km (top) for the T2K beam. The broad on-axis energy spectrum is shown in black, and the sharply-peaked spectrum at 2.5° in red, chosen to position the peak at the oscillation maximum. The minimum configurable off-axis angle of 2.0° is shown in blue. The assumed oscillation parameters are displayed alongside the prediction [62].

The beam is oriented at an angle of 2.5° off-axis with respect to the SK and ND280 detectors, which can be reduced to a minimum of 2.0° to tune the energy spectrum at SK. Figure 2.3 shows the neutrino energy spectra on-axis and at angles of 2.0° and 2.5°, together with the predicted ν_μ survival probability at SK. Although the on-axis beam has greater integrated flux, its energy spectrum is very broad. The 2.5° angle is chosen for a well-defined energy peak centred at the oscillation maximum for the 295 km baseline. This off-axis angle is a key feature of T2K as it enables greater precision in oscillation measurements than an on-axis beam: the energy of the beam is well-defined so any higher-energy neutrinos can be discarded as backgrounds.

2.3 Super-Kamiokande

The pre-existing Super-Kamiokande (SK) detector is used as the far detector for T2K, sampling the neutrino beam at a distance of 295 km from the beam source to measure rates of ν_e appearance and ν_μ disappearance. Super-Kamiokande is a water Cherenkov detector situated in the disused Mozumi mine approximately 1 km underground beneath Mt. Ikenoyama, and was a groundbreaking neutrino oscillation experiment in its own right before its use as part of T2K (as touched on in Chapter 1). Its underground location provides shielding from cosmic rays, and its directional sensitivity enables the separation of T2K beam neutrinos from those from other sources. As well as acting as the T2K far detector, Super-Kamiokande continues to contribute substantially to other areas of research, such as nucleon decay searches [63] and supernova detection [64].

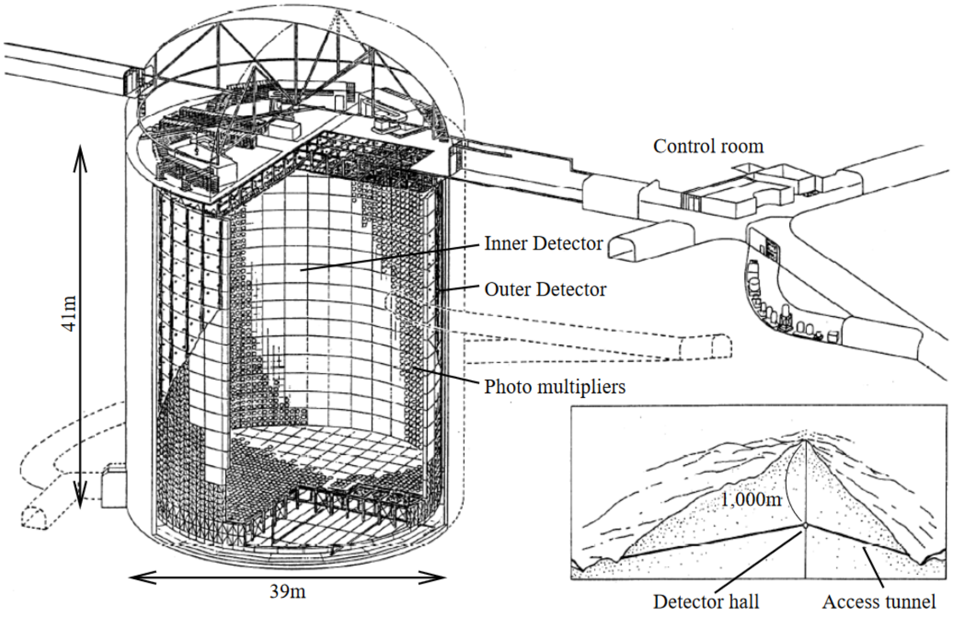
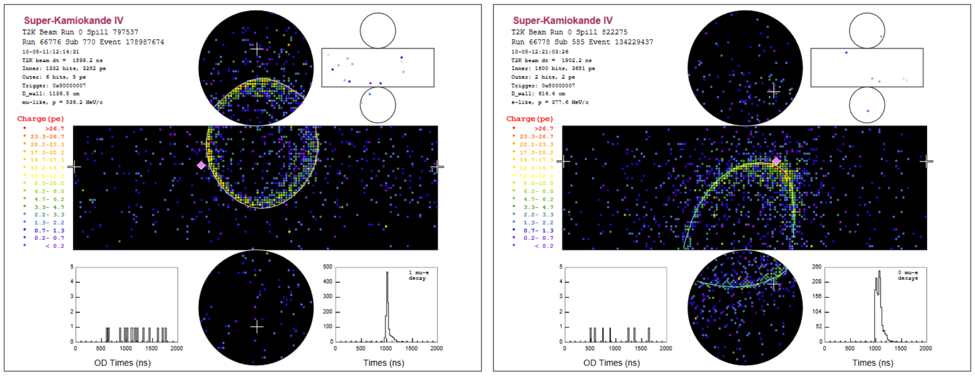


Figure 2.4: Diagram of the Super-Kamiokanda detector (left) and access tunnels [65].

The Super-Kamiokande tank is a cylindrical volume containing 50 kton of pure water

39 m in diameter and 42 m in height, instrumented with 13,014 photomultiplier tubes (PMTs). It comprises two volumes, separated by a cylindrical stainless steel scaffold structure and optically isolated from each other; a schematic is shown in Figure 2.4. The inner detector (ID) is the inner of the two volumes, a cylindrical space 33.8 m in diameter and 36.2 m in height, containing 11,129 PMTs mounted on the scaffold and facing inwards. This is enclosed in the outer detector (OD), a hollowed cylindrical space which is approximately 2 m thick radially and at either end, containing 1,885 PMTs mounted on the scaffold and facing outwards. The ID is the part of the detector designed to reconstruct neutrino interactions: it is well instrumented and thus has sufficient spatial resolution to measure a number of physical quantities from the detected Cherenkov light. The OD is sparsely instrumented by comparison, lacking the resolution and geometry needed for detailed event reconstruction, and acts instead as an active veto of cosmic ray muons and other backgrounds.



(a) muon-like event

(b) electron-like event

Figure 2.5: Examples of reconstructed T2K events in Super-Kamiokande, showing a muon-like (a) and an electron-like (b) ring. Each pixel in the event display represents a single photomultiplier tube, coloured according to the amount of charge recorded. The reconstructed cone is shown as a white line, and the location of the reconstructed vertex as a white cross [61].

Super-Kamiokande detects charged particles produced in neutrino interactions by their Cherenkov light cones. A charged particle traversing the water medium with sufficient energy produces a cone of photons, which form a ring-shaped hit pattern on the PMTs when they reach the ID walls. This information can be used to extract information about neutrino interactions such as the vertex position and the directions, energies and identities of the outgoing particles. In particular, muons and electrons can be distinguished by characteristic differences in the ring shapes they produce on the detector wall. Muons have larger mass and are therefore resilient to changes in their momentum, and consequently produce a well-defined cone of Cherenkov radiation. This results in a clear, sharp ring. Electrons, on the other hand, scatter more easily and tend to induce electromagnetic showers at the energies relevant to SK. Thus their light cones are less well-defined, effectively the sum of many overlapping cones with slight differences in orientation, resulting in a 'fuzzy' ring. Examples of muon- and electron-like rings are shown in Figure 2.5. This ability to distinguish between electrons and muons is essential for identifying ν_e and ν_μ events respectively, and thus measuring the corresponding appearance and disappearance rates in the T2K beam.

2.4 INGRID

The Interactive Neutrino GRID (INGRID) detector [61] is one of the T2K near detectors. It is located in the near detector pit 280 m from the target and is placed on-axis with respect to the neutrino beam. INGRID is designed to monitor the direction and intensity of the beam via neutrino interactions in iron, and is able to measure the beam centre with better than 10 cm precision, corresponding to 0.4 mrad. It comprises 16 identical modules, 14 of which are arranged in two arrays (horizontal and vertical) forming a cross centred on the neutrino beam axis (defined as 0° with respect to the proton beamline direction). The other two are placed separately at off-axis locations (see Figure 2.6). Additionally, a 'proton module' of similar but modified design is also placed at the centre of the cross between the two arrays. The INGRID cross modules sample the beam with a transverse section of 10 m × 10 m, measuring the location of the beam centre, while the separate off-axis modules check the axial symmetry. The proton module is designed to detect muons and protons produced by the neutrino beam and thus identify the QE neutrino interaction channel for comparison with MC simulations.

The 14 cross and 2 off-axis modules share the same design: they consist of a sandwich structure of alternating layers of iron and tracking scintillator. Each module contains 11 scintillator planes and 9 iron sheets. The iron sheets are 6.5 cm thick with dimensions 124 cm × 124 cm perpendicular to the beam axis, and provide a neutrino target mass of 7.1 tons per module. The tracking scintillator planes each comprise 24 horizontal and 24 vertical scintillator bars, each of dimensions 1.0 cm × 5.0 cm × 120.3 cm; the scintillation light

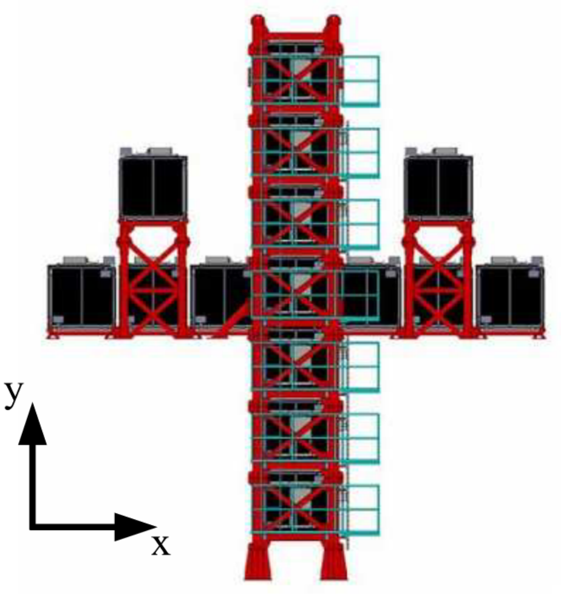


Figure 2.6: Diagram of the INGRID on-axis near detector [61].

from each bar is collected and transported by a wavelength-shifting (WLS) fibre attached to a multi-pixel photon counter (MPPC). Each module is surrounded by veto scintillator planes to reject charged particles coming from outside the modules. The proton module lacks the iron layers of the other modules, consisting instead of 34 tracking scintillator planes each comprising 32 bars in alternating orientations, giving it higher resolution than the other modules. It is similarly surrounded by veto planes.

The INGRID modules identify neutrino events by detecting tracks from muons. After a muon-neutrino interacts in an iron layer, the resulting muon passes downstream through the scintillator layers and deposits energy via scintillation. Scintillation light is collected by the WLS fibres and recorded by the MPPCs, with the alternating bar orientations providing position information. Hence the path of the muon can be reconstructed from MPPC hits.

2.5 ND280

The Near Detector at 280 m (ND280) [61] is a large magnetised off-axis detector that measures the flux, energy spectrum and flavour content of the T2K neutrino beam, as well as the rates of various neutrino interaction types. As the main off-axis near detector for T2K, it is placed at 280 m from the beam target at the same 2.5° off-axis angle as Super-Kamiokande, and serves both to characterise signals and backgrounds observed at the far detector and to measure neutrino-nucleus interaction cross-sections on carbon and oxygen.

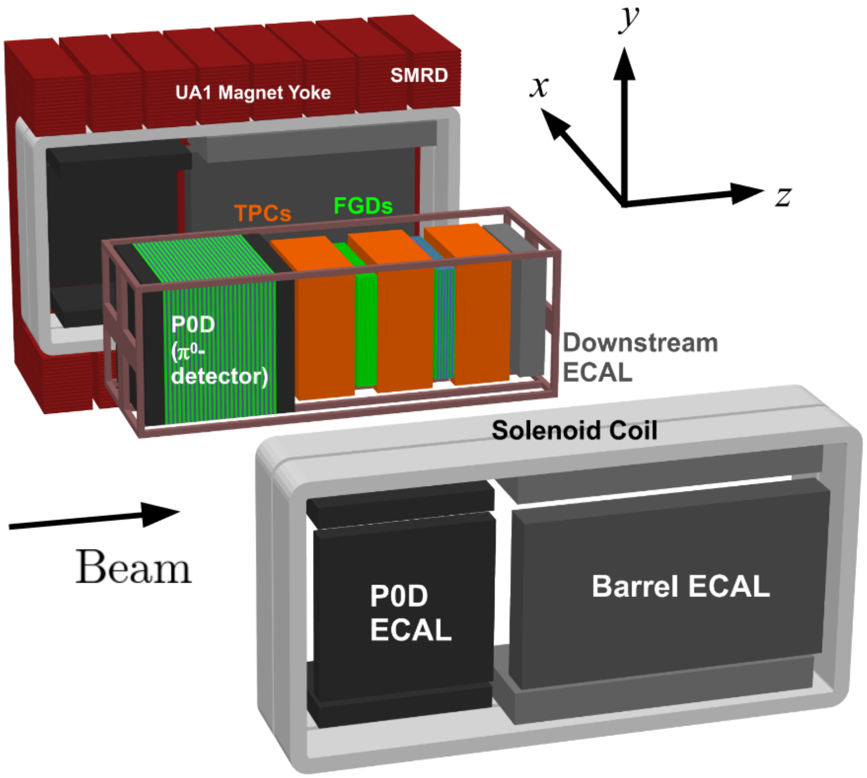


Figure 2.7: Exploded view of the ND280 detector [61].

ND280 consists of multiple subdetectors: the ‘tracker’, a sandwich of fine-grained detectors (FGDs) and time projection chambers (TPCs), is placed downstream of a pi-zero detector (PØD); these are together enclosed in an electromagnetic calorimeter (ECal) and placed within the recycled UA1 magnet which provides a 0.2 T magnetic field. The magnet yoke is also instrumented with scintillator to serve as a side muon range detector (SMRD).

2.5.1 Fine-Grained Detectors

The purpose of the two fine-grained detectors (FGDs) [61] is to provide target mass for neutrino interactions and to track charged particles coming from the interaction vertex. They are constructed from layers of plastic scintillator bars oriented perpendicular to the beam, as illustrated in Figure 2.8. These form an active target, both providing carbon nuclei for neutrino interactions and tracking the outgoing charged particles via scintillation light, which is collected by WLS fibres that transport it to MPPCs for detection. The scintillator bars have dimensions 9.61 mm × 9.61 mm × 1864.3 mm and are arranged in layers of 192 bars, with alternating orientation in the horizontal (x) and vertical (y) directions. Pairs of adjacent layers are attached to form structural units referred to as 'XY modules'. Each FGD contains 1.1 tons of target material and has outer dimensions 2300 mm × 2400 mm × 365 mm ($x \times y \times z$, where z is the beam direction), and is enclosed in a dark box for optical isolation from the rest of the detector.

The two FGDs have different designs. The more upstream of the two, FGD1, contains only scintillator as target mass, and is constructed from 15 XY modules comprising 5,760 bars. The more downstream, FGD2, also incorporates water sections to facilitate comparisons with interactions at SK, where water is the target mass. FGD2 consists of seven XY modules comprising 2,688 bars, alternating with six water layers. The water layers are 2.5 cm thick and are built from sheets of hollow corrugated polycarbonate, which are filled with water and kept below atmospheric pressure to prevent water leakage into the FGD interior.

Charged particles produced in and traversing the FGDs deposit energy in the form of scintillation light, which is recorded as hits by the electronics. These hits are reconstructed into tracks, which are used to locate the neutrino interaction vertex and may be matched to adjacent tracks in the TPCs and/or ECals if the particle leaves the FGD. The FGDs are generally not used for particle identification except for tracks that stop before leaving the FGD, in which case the FGD hits are the only available information.

2.5.2 Time Projection Chambers

The three time projection chambers (TPCs) [67] sandwich the FGDs such that there is a TPC immediately upstream and downstream of each FGD, and provide high-resolution 3D imaging of charged particle paths via ionisation in a low-pressure gas medium. This enables them to perform three key functions. Firstly, the number and orientations of charged

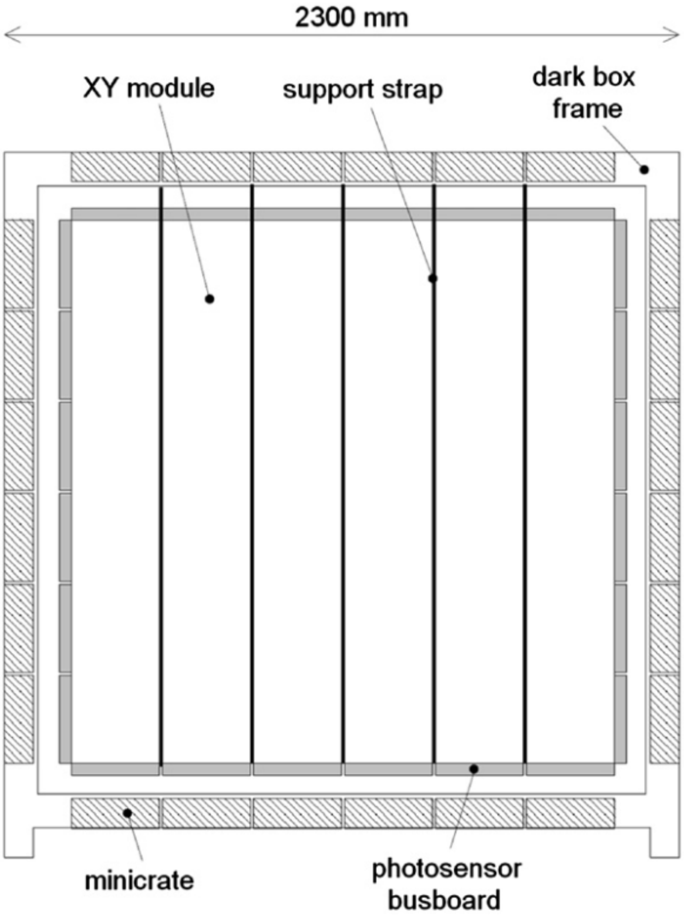


Figure 2.8: Cross-sectional view of an FGD, seen from the beam direction. The locations of the scintillator modules, photosensors, electronics minicrates and dark box are shown, as well as the structural support straps [66].

particles can be determined. Secondly, the curvature of tracks due to the magnetic field can be used to calculate their momenta, enabling reconstruction of the energy of interacting neutrinos with high resolution (as shown by Figure 2.10). Thirdly, the amount of ionisation energy deposited along the track can be used together with the momentum for particle identification, which is particularly useful for distinguishing muons from electrons and thus ν_μ events from ν_e events. Methods of particle identification with the TPCs will be discussed in Section 4.2.1.

Each TPC consists of an inner box (1808 mm × 2230 mm × 854 mm) placed within

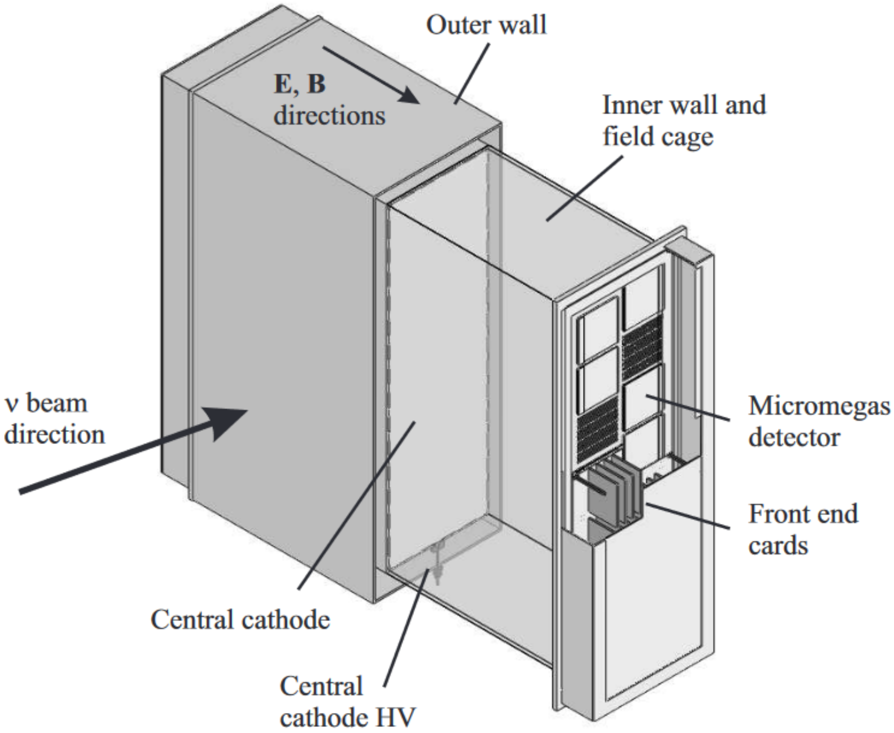


Figure 2.9: Simplified cutaway schematic of the ND280 TPC design [61].

an outer box, containing the drift gas and the insulating gas respectively. The drift gas is an argon-based mixture (Ar:CF_4:iC_4H_{10} in a 95:3:2 ratio) chosen for its high drift speed, low diffusion, and good performance with micro mesh gas ('micromegas') detectors, and is kept at 0.4 mbar pressure. The insulating gas is CO_2, providing electrical insulation between the inner box and ground and excluding atmospheric oxygen from entering the inner box. A cathode is placed at the midpoint of the inner box, and readout planes at either end containing micromegas detectors with 7.0 mm × 9.8 mm anode pad segmentation, generating an electric field of approximately 275 V/cm. Charged particles traversing the TPCs deposit energy in the drift gas by ionisation, producing ionisation electrons that drift under the effect of the electric field towards the readout planes, where they are detected. The 2D pattern recorded in the pad plane is combined with the timing information of the hits using the known electron drift velocity to yield a 3D image of the path of the particle.

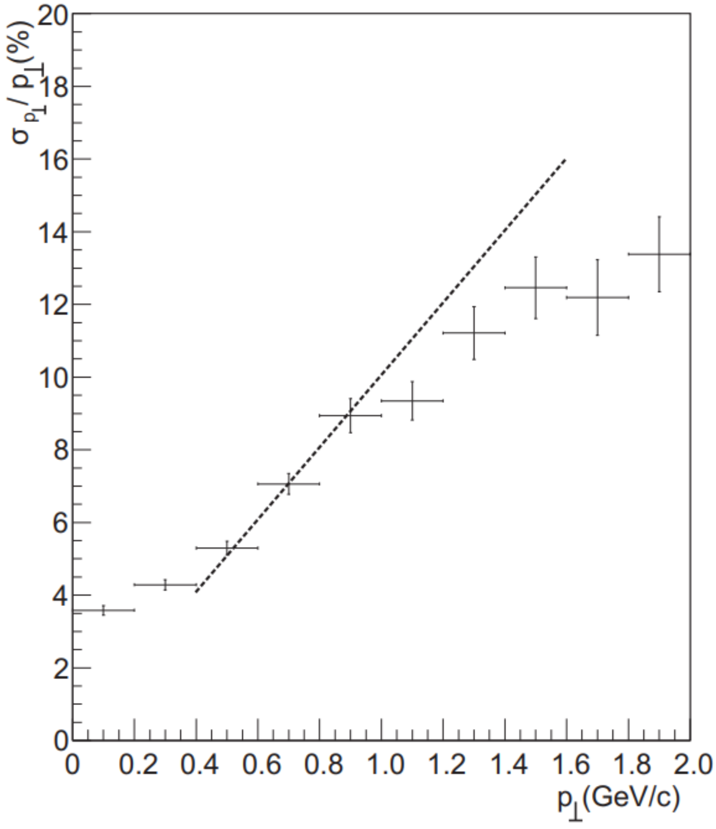


Figure 2.10: The momentum resolution of an ND280 time projection chamber, shown as a function of momentum perpendicular to the magnetic field. These values are based on predictions from Monte Carlo simulation of muons, using only tracks that cross at least 50 out of the 72 pad columns of the TPC volume. The design goal for the momentum resolution is shown as a dashed line [67].

2.5.3 Pi-Zero Detector

The pi-zero detector (PØD) [68] is a water-target plastic-scintillator-based detector designed to detect neutral pions. Its primary goal is to measure the neutral current process

$$\nu_\mu + N \rightarrow \nu_\mu + N + \pi^0 + X \tag{2.3}$$

which is a major background for ν_e appearance at SK. The PØD is composed of plastic scintillator layers, metal sheets (brass and lead), and water bags which can be filled or left empty, enabling a subtraction method to determine cross sections on water. The scintillator layers have sufficient resolution to reconstruct both charged particle tracks (muons and pions) and EM showers induced by the metal layers (electrons and photons from π^0).

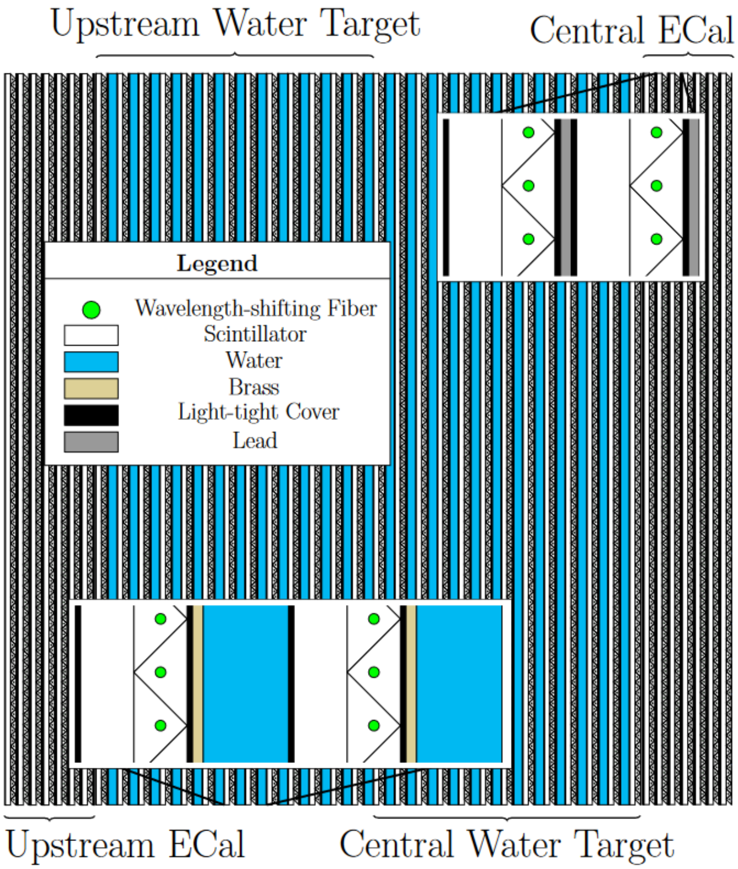


Figure 2.11: Schematic of the ND280 pi-zero detector viewed from the side, with the beam going from left to right. Details of the different PØDule types are shown in insets [61].

Whereas the scintillator bars used in other ND280 subdetectors have square or rectangular cross-sections, those of the PØD are triangular (see Figure 2.11). They are arranged in 40 modules, referred to as ‘PØDules’, each consisting of 134 vertical (2200 mm long) and 126 horizontal (2340 mm) bars. As elsewhere, scintillation light is collected by WLS fibers and detected by MPPCs. The 40 PØDules are grouped into four ‘super-PØDules’ which are stacked one after another in the beam direction. The inner two are water target super-PØDules; the upstream (central) water target is a sandwich of 13 PØDules alternating with 13 (12) 28 mm-thick water layers and 13 (12) 1.5 mm-thick brass sheets. The most upstream and downstream are ECal super-PØDules, which are a sandwich of seven PØDules and 4 mm-thick lead sheets and contain no water, being designed instead to ensure any escaping photons are detected. The water target fiducial region is designed to contain approximately 1.9 tons of water, and the PØD as a whole comprises 10,400 scintillator bar channels and has dimensions 2103 mm × 2239 mm × 2400 mm ($x \times y \times z$, where z is the beam direction).

While the PØD has provided the target for several cross-section analyses, its upstream location means it is generally not relevant to event selections using the FGDs as their target volume, such as the ones presented in this thesis.

2.5.4 Electromagnetic Calorimeter

The ND280 ECal is a sampling electromagnetic calorimeter built from plastic scintillator layers and lead absorber sheets. It surrounds the inner detectors (FGDs, TPCs and PØD) and provides near-hermetic coverage for particles exiting them. The ECal is designed to detect photons and measure their energy and direction, and also to detect and provide PID information for charged particles. The ability to detect photons is critical to reconstructing any π^0 that may be produced in neutrino-nucleus interactions. The design and functionality of the ECal will be described in detail in Chapter 3.

2.5.5 Magnet and Side Muon Range Detector

The tracker, PØD and ECal are placed inside a magnet reused from the UA1 experiment, which provides a horizontally-oriented dipole magnetic field of 0.2 T. This enables the detector to determine the sign of charged particles and measure their momentum with good resolution, as showin in Figure 2.10. The magnet consists of water-cooled aluminium coils and a steel flux return yoke. The return yoke also contains the Side Muon Range Detector (SMRD), an array of plastic scintillator modules inserted into the 1.7 cm air gaps between the 4.8 cm-thick steel plates.

The SMRD serves three main functions. Firstly, it is able to detect muons that escape the inner detectors with high-angle trajectories and measure their momenta; most other particles are absorbed in the inner detectors or the magnet so do not reach the SMRD. Secondly, it provides a trigger for cosmic ray muons entering ND280. Thirdly, it helps

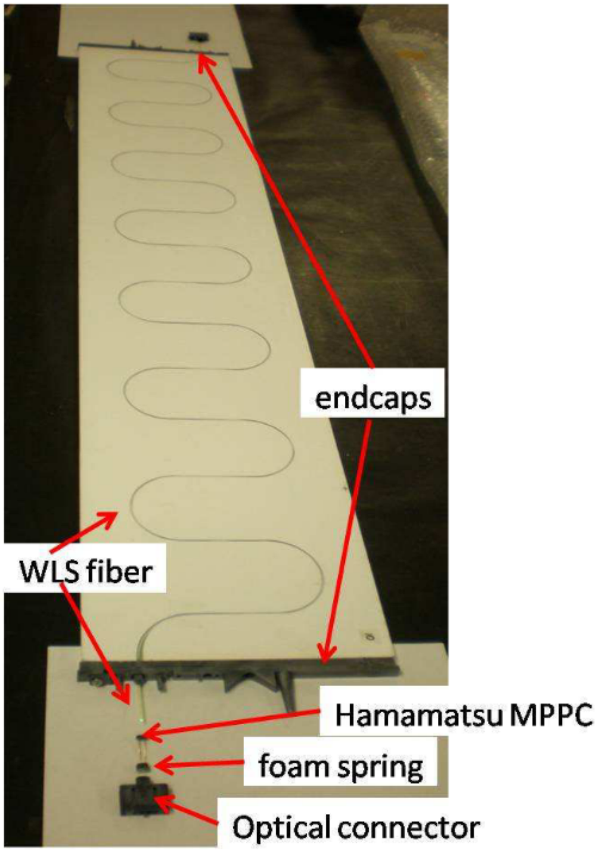


Figure 2.12: View of a single SMRD scintillation counter with components labelled [61].

identify beam neutrino interactions that do not originate from the inner detector (such as those that occur in the surrounding cavity walls). It is constructed from 440 scintillator modules arranged in layers of three to six (depending on the yoke section, since the air gap geometry varies) placed in the innermost gaps of the return yoke. The modules are composed of scintillation counters, which are 7 mm-thick polystyrene panels 875 mm long; horizontal modules contain four counters of 187 mm width and vertical modules contain five of 175 mm width, to maximise the active area in each air gap. The counters are coated with a white diffuse layer to reflect light, and are machined with an S-shaped groove to accommodate a WLS fiber (see Figure 2.12) to collect scintillation light more efficiently than a straight channel would. The WLS fiber transports light to an MPPC. The SMRD contains 192 horizontal modules and 248 vertical for a total of 4,016 scintillation counter channels. The SMRD's large scintillator panel dimensions give it poor resolution compared to the ECal or FGDs; it is intended for particle tagging and energy measurement rather than detailed tracking.

2.5.6 Software and data processing

A large number of software packages make up the ND280 offline software suite. They have a modular structure and use standard particle physics software libraries as their foundation, specifically ROOT [69] as the underlying framework and data storage model, and Geant4 [70] as the basic simulation library; consequently they are mostly written in C++.

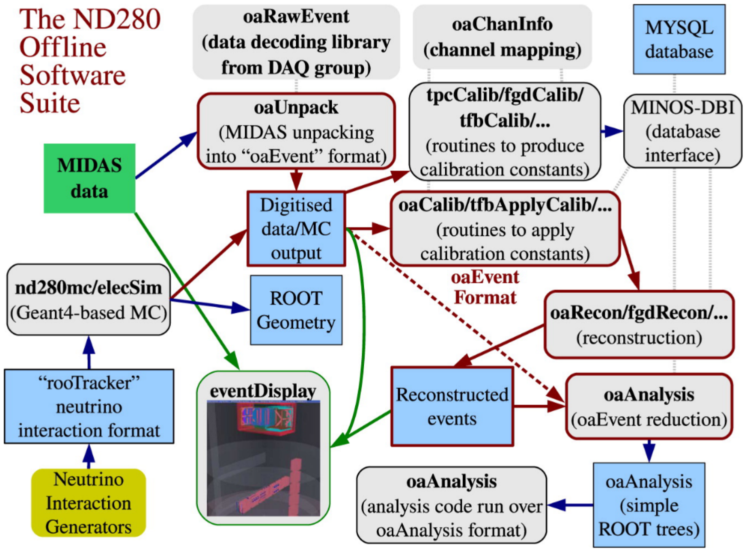

The ND280 Offline Software Suite
oaRawEvent (data decoding library from DAQ group)
oaChanInfo (channel mapping)
MYSQL database
oaUnpack (MIDAS unpacking into "oaEvent" format)
tpcCalib/fgdCalib/ tfbCalib/... (routines to produce calibration constants)
MINOS-DBI (database interface)
MIDAS data
Digitised data/MC output
oaCalib/tfbApplyCalib/... (routines to apply calibration constants)
oaEvent Format
nd280mc/elecSim (Geant4-based MC)
ROOT Geometry
oaRecon/fgdRecon/... (reconstruction)
"rooTracker" neutrino interaction format
eventDisplay
Reconstructed events
oaAnalysis (oaEvent reduction)
Neutrino Interaction Generators
oaAnalysis (analysis code run over oaAnalysis format)
oaAnalysis (simple ROOT trees)

Figure 2.13: Schematic of the package structure of the ND280 offline software suite [61].

Figure 2.13 shows the general structure of the software suite. Raw data files in the MIDAS [71] format are converted to the ND280 format defined by the *oaEvent* library for offline use, and then processed in three stages. In the calibration stage, controlled by the *oaCalib* package, calibration constants are applied from a centralised MySQL database. In the reconstruction stage, controlled by *oaRecon*, objects (such as tracks and showers) are reconstructed from the data from each subdetector and then matched and combined to form global reconstruction objects. In the reduction stage, controlled by *oaAnalysis*, the large reconstruction files are reduced into smaller files built from ROOT trees which are lightweight enough to be used for analysis.

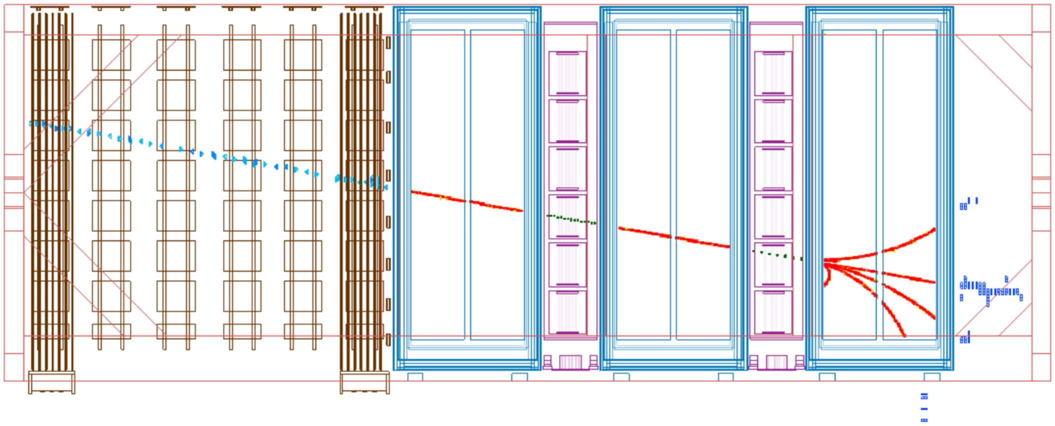


Figure 2.14: ND280 event display showing hits in each subdetector from a muon track. The muon enters from the front face of the PØD and passes through the tracker, producing secondary tracks in TPC3. The secondary particles stop in the downstream and barrel ECals [61].

To produce Monte Carlo simulated data, a simulation of the neutrino beam interfaces with a neutrino interaction event generator package (T2K primarily uses NEUT [51], with GENIE [52] as a cross-check). The ND280 software includes a representation of the detector geometry which is used first to determine the nuclear target used in the event generator, and then for the propagation of final state particles with Geant4. The energy deposits simulated by Geant4 are then inputted into the *elecSim* package which simulates the response of the active detectors and electronics. This results in MC files that can then be processed in the same way as real data.

Event selections and cross-section analyses can be performed using the highLevelAnalysis (‘Highland’) framework. Taking *oaAnalysis* files as input, Highland provides numerous classes that enable users to develop event selection algorithms, extract cross-sections, plot results, and evaluate the impact of systematic errors. The event selections presented in this thesis were built using version `2.68` of the Highland framework.

Chapter 3

ND280 Electromagnetic Calorimeter

The ND280 ECal is a sampling electromagnetic calorimeter which surrounds the inner detectors (FGDs, TPCs and PØD). It consists of layers of plastic scintillator bars interleaved with lead absorber sheets, and is divided into 13 independent modules in three main sections: the downstream ECal (DS-ECal), which is placed immediately downstream of TPC3; the barrel ECal, which surrounds the tracker detectors; and the PØD-ECal, which surrounds the pi-zero detector. The barrel and PØD ECal each comprise six modules which surround their respective inner detectors on four sides parallel to the z (beam) axis, as can be seen from Figure 3.1.

The main purpose of the ECal is to detect photons and thus reconstruct neutral pions produced in neutrino interactions, and to measure the energy of electromagnetic showers. The ECal can also provide valuable information for particle ID, since different particle types exhibit different characteristic behaviours as they traverse or stop within it. These capabilities make the ECal a key component of ND280, and ECal information makes a crucial contribution to the event selections presented in this thesis.

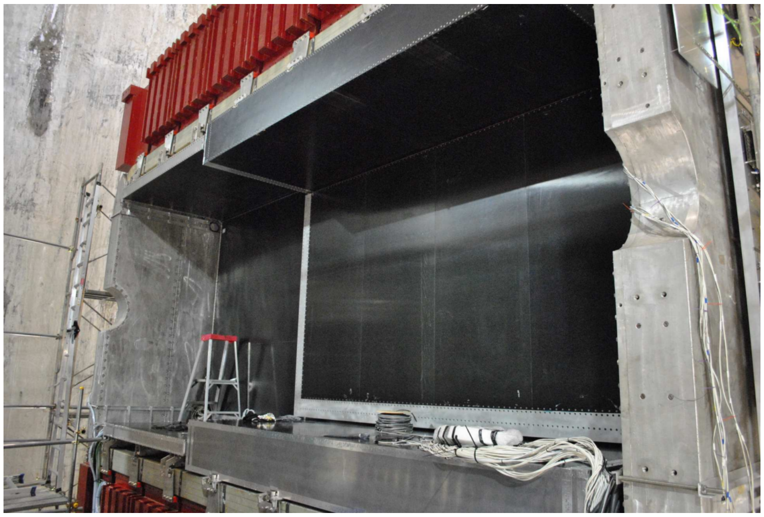


Figure 3.1: One side of the ECal installed within the ND280 magnet. Three of the six barrel-ECal modules can be seen on the right, with three of the thinner PØD-ECal modules on the left. The basket, which contains the DS-ECal, is not present [72].

This chapter describes the design of the tracker (barrel and downstream) ECal, as well as its charge calibration procedures and reconstruction methods. Its particle identification capabilities will be discussed as part of Chapters 4 and 6. The PØD-ECal, which has a different design and more limited capabilities, will not be discussed as it is not relevant to the work presented in this thesis.

3.1 Design of the Electromagnetic Calorimeter

The ECal is designed to induce and measure electromagnetic showers. Each ECal module contains layers of lead converter: dense material which causes photons and most other particles to stop and produce showers (and can also provide a target for neutrino interactions). These showers are detected by layers of plastic scintillator bars with alternating perpendicular orientations. These two orientations per module provide two 2D views which can be reconstructed into a full 3D view as illustrated by Figure 3.2. Scintillation light produced in each bar is collected and transported by a WLS fibre, located at the centre of the bar, to a MPPC photosensor. Some particles such as muons tend not to shower in the lead layers and instead behave as minimum ionising particles (MIPs), but can still be detected by the

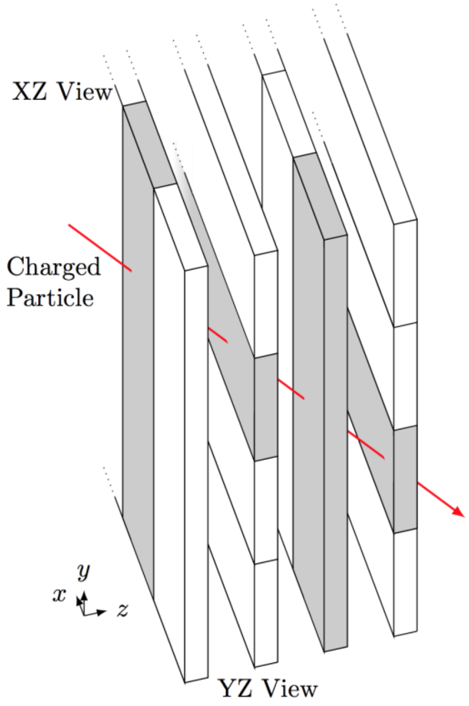


Figure 3.2: Schematic view of the passage of a MIP-like particle through a side barrel ECal module. Vertical bars together provide a view in the XZ plane, and horizontal bars a view in the YZ plane, which are combined into a 3D track by the reconstruction [73].

energy they deposit along their path in the scintillator.

3.1.1 Scintillator bars

Scintillator bars are the basic active element of the ECal. They are made from extruded polystyrene doped with organic fluors: 1% polyphenylene oxide (PPO) and 0.03% 1,4-bis(5-phenyloxazol-2-yl) benzene (POPOP). All bars have a cross-section of 40 mm × 10 mm: the 10 mm thickness was chosen to minimise the overall depth of the ECal while still producing sufficient light yield, while the 40 mm width is a compromise between reconstruction efficiency (since widths greater than 50 mm were found to seriously compromise π^0 reconstruction) and channel cost (since thinner bars would be more numerous and thus require more channels to read out). The lengths of the bars vary according to module and

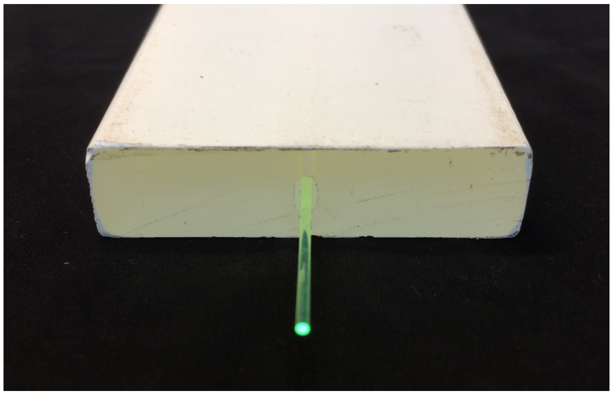


Figure 3.3: End view of an ECal scintillator bar and WLS fibre [73].

orientation. Each bar has a 0.25 mm-thick reflective coating containing TiO_2 to provide optical isolation and increase light yield collected by the WLS fibre, which is inserted into a 2 mm-diameter hole running longitudinally through the centre of the bar (see Figure 3.3).

The WLS fibres collect scintillation light from the bars and transport it to the MPPC photosensors for readout. All ECal modules use 1 mm-diameter Kuraray Y-11(200)M CS-35J [74], which are polystyrene optical fibres with 200 ppm wavelength shifting dye and double cladding for high trapping efficiency. Depending on the module and orientation of the bar, the WLS fibre may be read out at both ends (double-ended readout), or only one end (single-ended readout) in which case the other end is mirrored with a vacuum deposition of aluminium. The DS-ECal contains only double-ended bars, while the longitudinal and perpendicular bars of the barrel ECal are double- and single-ended respectively. At the readout end(s) the fibre is optically coupled to the corresponding MPPC entrance window via a transparent Teflon ferrule glued to the fibre end.

3.1.2 ECal modules

The ECal modules each comprise a number of layers, each of which consists of scintillator bars bonded to a lead sheet, with the bars of each layer being perpendicularly oriented with respect to its neighbours. The DS-ECal is the region of the ECal placed downstream of the inner detectors, occupying the last 50 cm of the structural basket that contains the FGDs, TPCs and PØD. It is a single module consisting of 34 lead-scintillator layers oriented in the x and y directions (perpendicular to the beam axis), each of which comprises 50 scintillator bars. Its other attributes are summarised in Table 3.1. Due to conservation of momentum, the kinematics of typical neutrino interaction events are such that a large proportion of particle tracks will follow closely to the beam direction and thus pass through the DS-ECal rather than the barrel modules. For this reason, the DS-ECal has more layers than the barrel modules.

	Downstream ECal
Dimensions (mm)	2300 × 2300 × 500
Weight (kg)	6500
No. of layers	34
No. of bars per layer	50
Total no. of bars	1700
Bar length (mm)	2000
Bar orientation	x/y
Lead thickness (mm)	1.75

Table 3.1: Properties of the Ds-ECal design, summarising the dimensions and layer makeup.

	Barrel-ECal top/bottom	Barrel-ECal sides
Dimensions (mm)	4140 × 1676 × 462	4140 × 2500 × 462
Weight (kg)	8000	10000
No. of layers	31	31
No. of bars per layer	38 long., 96 perp.	57 long., 96 perp.
Total no. of bars	2280 long., 6144 perp.	1710 long., 3072 perp.
Bar length (mm)	3840 long., 1520 perp.	3840 long., 2280 perp.
Lead thickness (mm)	1.75	1.75

Table 3.2: Properties of the barrel-ECal design, summarising the dimensions and layer makeup (divided into longitudinally- and perpendicularly-oriented bars where appropriate).

The barrel-ECal surrounds the tracker on its four outward-facing sides, and comprises 6 modules (2 top, 2 bottom, 1 left, 1 right) which are affixed to the inside surface of the magnet return yoke. The top and bottom sections are divided into two along the beam direction to permit opening of the magnet and access to the tracker for maintenance. The barrel-ECal modules are made up of 31 lead-scintillator layers each with alternating orientation similarly to those of the DS-ECal. The dimensions and other attributes of the barrel ECal modules are summarised in Table 3.2 and a photograph of one module is shown in Figure 3.4.

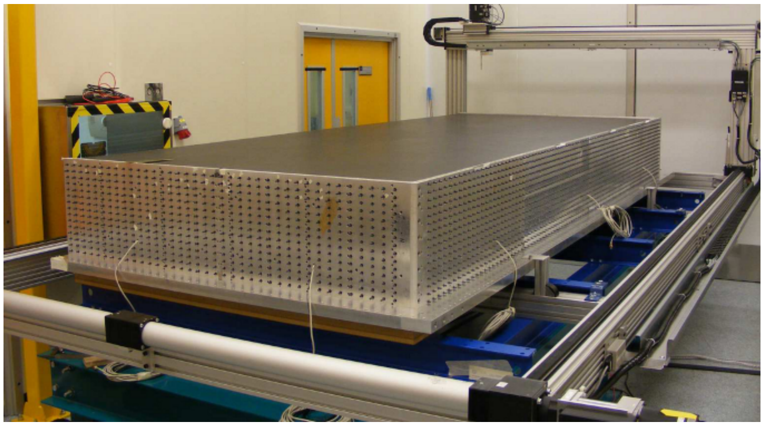


Figure 3.4: One of the top barrel-ECal modules lying horizontally during construction at the University of Liverpool. The ends of the WLS fibres, which are encased in Teflon ferrules, can be seen protruding from the aluminium bulkheads [72].

3.1.3 MPPC photosensors

Like the other ND280 scintillator detectors, the ECal uses multi-pixel photon counters (MPPCs) to read out light from the WLS fibres. Traditionally PMTs would be used for this purpose, but they are relatively bulky and would not function within the 0.2 T magnetic field, so they would have to be placed outside the detector and the light signal transported several metres to reach them. MPPCs on the other hand are both compact enough to fit inside the ECal modules and able to function inside the magnetic field, and also have a higher detection efficiency for photons of the wavelength produced by the WLS fibres. They are solid-state photosensors consisting of many independent sensitive pixels, each of which is a Geiger-mode avalanche photodiode [75]. The output of the MPPC is the analogue sum of the fired pixels; this is usually expressed in 'pixel energy units' (PEU), 1 PEU being the charge seen when a single pixel fires. A specialised MPPC with a sensitive area of 1.3×1.3 mm^2 containing 667 pixels was developed for ND280 by Hamamatsu Photonics K.K. [76]. Its parameters are listed in Table 3.3. Each scintillator bar is connected to either one (single-ended) or two (double-ended) MPPCs, with a total of 22336 across all modules.

3.1.4 Readout electronics

The ND280 back-end electronics consists of several different boards: Trip-T front-end boards (TFBs) which receive the MPPC output directly; readout merger modules (RMMs) which

Parameter	
Number of pixels	667
Active area	1.3 × 1.3 mm^2
Pixel size	50 × 50 μm^2
Operational voltage	68–71 V
Gain	$\approx 10^6$
Photon detection efficiency at 525 nm	26–30%
Dark rate above 0.5 PEU at 25°C	$\leq$ 1.35 MHz

Table 3.3: Parameters of the T2K MPPCs. The dark noise rate is given for a threshold equivalent to half the charge of a single pixel firing.

control and read out the TFBs; a master clock module (MCM) providing central control via slave clock modules (SCMs) for each subdetector; and two cosmic trigger modules (CTMs) which support calibration and monitoring by providing a selection of cosmic-ray muon triggered events.

The ECal MPPCs are connected to Trip-T front-end boards (TFBs) which form the front-end of the readout electronics. A TFB has 64 channels, each of which can read out a single MPPC and also records MPPC monitoring data such as voltage and temperature, the latter of which is recorded via a built-in temperature sensor and a port to connect to an external one. Each TFB contains 4 Trip-T integrated circuits, which were originally designed for the D0 experiment at FNAL. A Trip-T integrates the charge detected by connected MPPCs in a preset time interval (for T2K usage, this is programmed to synchronise with the timing of the neutrino beam) followed by a similarly programmable reset time of at least 50 ns. Once 23 readout cycles [1] have been completed, the stored data are digitized by analogue-digital converters, and sent by the TFB (along with timestamp information) to a RMM back-end board. 12 RMMs are used in the ECal (2 for the Ds-ECal, 8 for the barrel-ECal, 2 for the PØD ECal), each of which controls and receives signals from up to 48 TFBs. Data from the RMMs is sent to commercial PCs for collection and processing.

The ND280 MCM receives signals directly from the accelerator indicating when the neutrino beam will be active, and uses this information to distribute trigger and clock signals across the whole detector. Control is fanned out to the subdetectors via SCMs; the ECal SCM passes trigger and clock signals to the ECal RMMs, and can also be set as the master controller in order to run the ECal autonomously from the rest of ND280 for the

[1] This is the maximum number of cycles that can be read out from the Trip-T chip pipeline, capturing each bunch of the spill and also after-spill data [77].

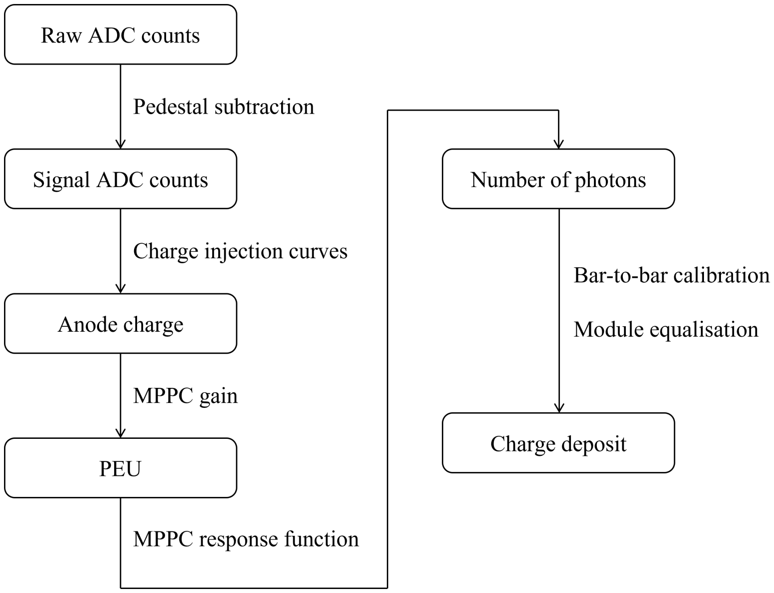


Figure 3.5: Flowchart illustrating the series of steps used to calculate the charge deposited in the ECal scintillator from the recorded ADC values.

purposes of routine calibration, debugging etc.

3.2 ECal charge calibration

ECal hits must be calibrated precisely in order to account for instrumental effects: differences in scintillation yield, attenuation along WLS fibres, time lags in the electronics, and so on. The calibration procedures developed for this purpose can be divided into two main categories: charge calibrations, which together convert the ADC value registered by the electronics to the charge deposited in the scintillator; and timing calibrations, which correct for delays in the hit time recorded by the electronics. Only the charge calibrations will be discussed here; see [78] for an overview of the timing calibrations.

The ECal charge calibrations consist of a series of steps as illustrated in Figure 3.5. These calibrations convert each ECal channel's response from an ADC value to a charge deposit in a multi-step calculation, working back along the readout chain from the electronics, to the MPPC, to the WLS fibre, and finally to the scintillator bar. Corrections are applied to account for factors such as differences in sensitivity between channels and diurnal temperature changes. The charge distribution is later used to compute an energy deposit in the reconstruction stage, and also forms the basis of ECal particle identification.

3.2.1 Pedestal and gain calibrations

The first step of converting ADC counts to anode charge is to calculate and subtract the electronics pedestal: the ADC value registered in the absence of any physics signal. Each channel has its own pedestal, which varies with each readout cycle. Diurnal temperature variations at the electronics can result in pedestal changes of a few ADC counts, so pedestals are calculated every three hours using noise spectra recorded by the DAQ in normal running. Dedicated pedestal calibration runs are also taken during weekly beam down-time. The calculated values are subtracted from any ADC counts recorded by the DAQ.

Converting the pedestal-subtracted ADC into an anode charge is achieved by mapping the electronics response, which is not perfectly linear and includes a transition from low- to high-gain ADC channels. The TFB is used to inject a known charge onto each channel, enabling the channel's low- and high-gain responses to be mapped. The resulting charge injection curves are parametrised for low- and high-gain channels using two cubic polynomials, with the transition being modelled with a sigmoid function for a smooth transition from low to high gain. The parameters for this calibration are updated around once per running period since they are typically stable over time. An example of the high and low gain responses of a typical channel is shown in Figure 3.6.

Next, the calculated anode charge is converted to an estimated number of photons incident on the MPPC. This is done by first converting the charge into a number of PEU, i.e. dividing by the MPPC gain. Similarly to pedestals, the gains are different for each channel and are affected by diurnal temperature variations, so they are calculated every three hours using the same noise spectra. They are measured by fitting the position of the first non-pedestal peak, which corresponds to a single MPPC pixel firing. For each channel, the gain is taken to be the difference between the fitted positions of the pedestal and single-pixel peaks.

Having divided by the gain, the number of pixels is converted to a number of photons by dividing by the MPPC response function. The response function is stable with time, but is complicated and cannot easily be calculated analytically, so is modelled on testbench measurements instead. A single response function is used for all channels, the parameters of which are functions of the MPPC gain. This enables the conversion to account for changes in PDE and other parameters with the overvoltage supplied to the MPPC.

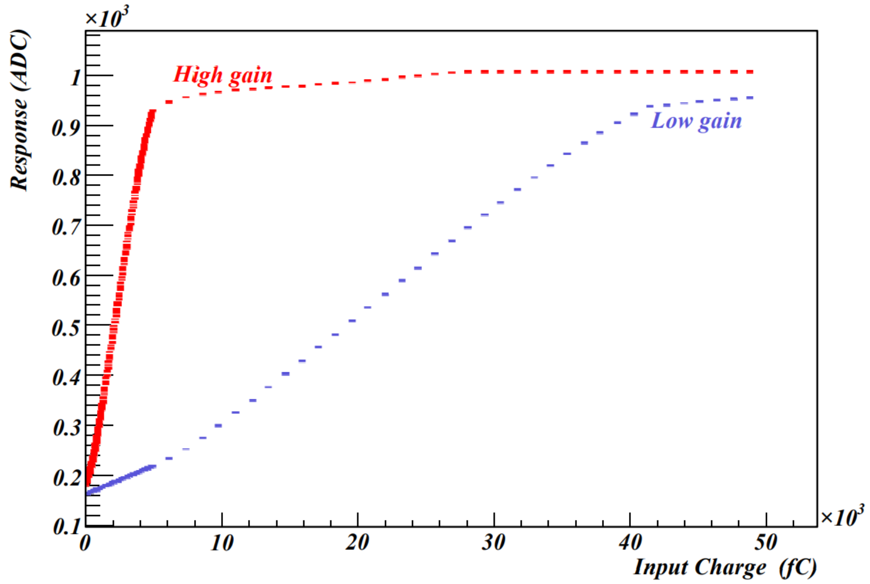


Figure 3.6: High and low gain ADC response of a typical TripT channel as a function of input charge. The input charge is calculated from the digital-to-analogue converter controlling the capacitor used to perform the charge injection [72].

3.2.2 Bar equalisation calibrations

The next steps of the charge calibrations are designed to equalise the response of all bars. This accounts for differences in sensitivity between different bars of the same type (the 'bar-to-bar' correction) and differences between bar types (the 'module equalisation' correction) [79].

The bar-to-bar calibration constants scale the response of each bar to the average for its type (module, orientation, and endedness) and are somewhat stable with time, so they are calculated once per T2K run. This is done using an empirical data sample of cosmic ray muons, which are a plentiful and well-understood control sample. They behave as MIPs so their energy loss in the scintillator is simple and well understood. Only cosmic muon tracks that pass through (i.e. do not stop in) the ECal are selected for the calibration sample.

The sample is first processed with all calibrations except the bar-to-bar and module equalisation corrections applied, including a correction for differences in track path length through the bar. The charge spectrum of all hits in each bar is fitted with a Landau-Gaussian function: a Landau distribution (modelling the energy deposit of the particle)

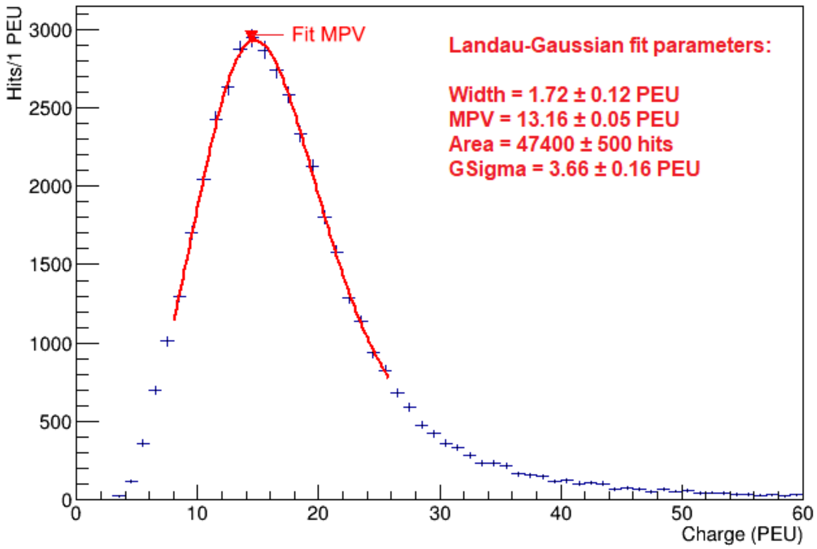


Figure 3.7: Example of a cosmic muon hit charge spectrum for one ECal channel. The Landau-Gaussian function fitted to the data is shown in red.

convolved with a Gaussian (modelling the smearing due to detector resolution), as shown in Figure 3.7. The fit is attempted across multiple truncations of each channel spectrum to find a range that will yield a successful fit; however if the histogram contains insufficent hits as in Figure 3.8, the fit may fail on all attempts. The most probable value (MPV) of the Landau-Gaussian fit is taken to represent the response of that bar. These MPVs are then used to generate a set of constants such that the MPV of each bar is scaled to the 6σ truncated mean for its type. These are the bar-to-bar constants, which when applied are used as divisors to correct each hit.

Channels are only included in the calculation of the mean for their bar type if they meet two criteria: first, the fit must have been successful, and second, the MPV must meet the truncation requirement i.e. it must fall within 6σ of the raw mean, as calculated by a Gaussian fit. Channels that do not meet these criteria will not receive a bar-to-bar correction (their constants will be set to 1). In the T2K Run 9 calibrations from which the plots in this chapter are taken, 20313 channels were successfully calibrated, and the remaining 95 were excluded due to failed fitting or truncation.

These calibration constants are subject to two stages of validation. The first stage takes

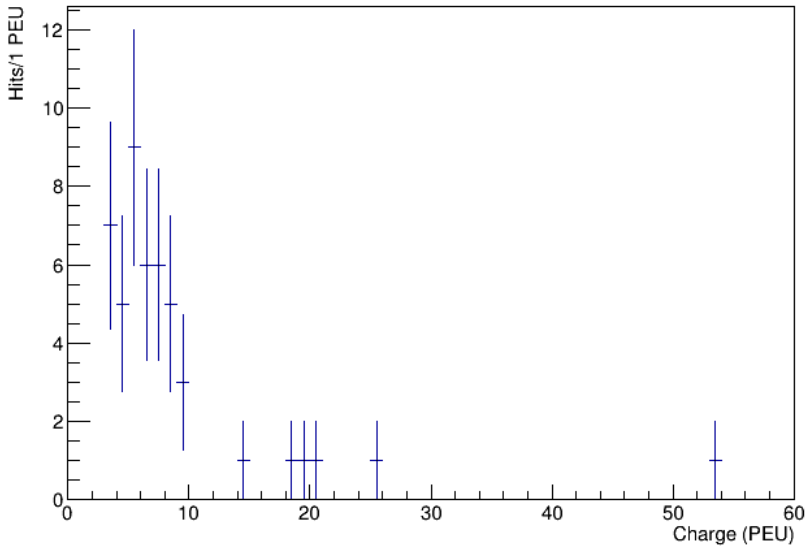


Figure 3.8: Example of a cosmic muon hit charge spectrum for which the Landau-Gaussian fit fails. This happens because not enough hits have been recorded for a successful fit in this channel.

place before the constants are uploaded to the T2K database, and serves to check that the constants have been calculated correctly and that the table file is valid. The constants are applied to cosmic muon data, and the Landau-Gaussian fit is performed once again for each channel. The distributions of the MPVs with and without the correction are compared, as shown in Figure 3.9 for the same data and Figure 3.10 for an independent sample.

The second stage takes place after the constants are uploaded to the database, with the purpose of checking that the constants have been uploaded successfully and can be applied correctly. The processing of the raw cosmic muon data is repeated, with the only change being that the bar-to-bar corrections are now applied. The Landau-Gaussian fitting is then performed for the two versions (with and without bar-to-bar and module equalisation calibrations) of the cosmic muon data set, and the distributions of the resulting MPVs are compared; if the constants have been uploaded and applied correctly, the corrected distribution will have a narrower spread.

Bar-to-bar (and other) calibration constants are stored in the T2K database. The database contains all sets of bar-to-bar constants for each T2K run, along with validity

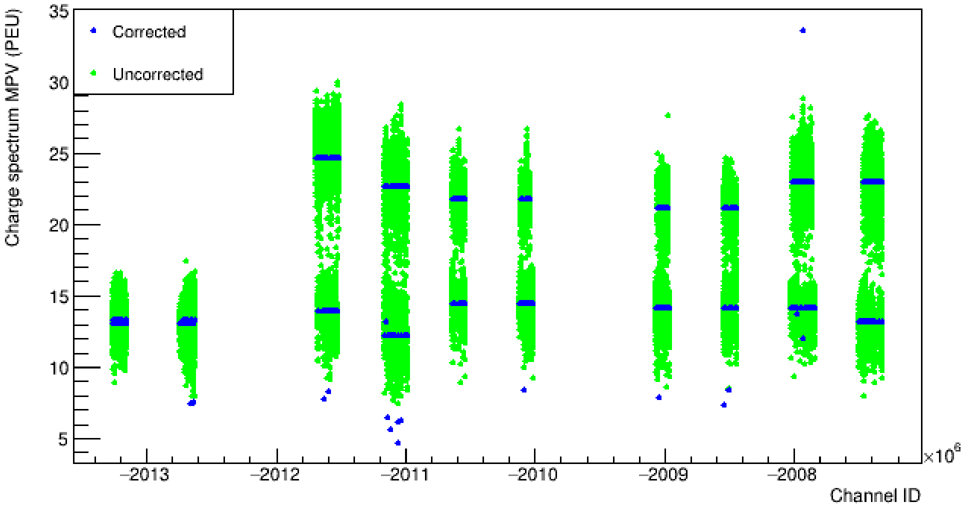


Figure 3.9: A pre-upload validation plot for ECal bar-to-bar constants for T2K Run 9, comparing cosmic muon hit MPVs before (green) and after (blue) application of the constants, generated from the same data sample used to calculate the constants. For each channel, the charge spectrum MPV is shown by the y-axis and the channel ID number by the x-axis, which groups them according to their module. Consequently, two bands can be seen for modules which contain two different bar types. The uncorrected points show a spread of MPV values for each bar type, whereas the calibrated points have been corrected to a single average value for each bar type, indicating that the calibration is working as intended. Outliers in the corrected points correspond to channels where the fit has failed or been truncated; a small number of these is acceptable.

information that determines which constants will be used for data taken during any given period. This ensures that constants will only be applied to the data for which they are valid (i.e. data from the same run as the cosmic muon data from which the constants were generated) and that any sets of constants found to have been incorrectly calculated or uploaded can be overridden by newly-calculated ones.

Following the application of the bar-to-bar constants, the module equalisation correction is applied to equalise the response across all modules and bar types. This is done by scaling the average response for each bar type (module, orientation and endedness) to that of the DS-ECal bars. The relative responses of different bar types are expected to be stable with time, so these constants are not recalculated alongside the bar-to-bar constants.

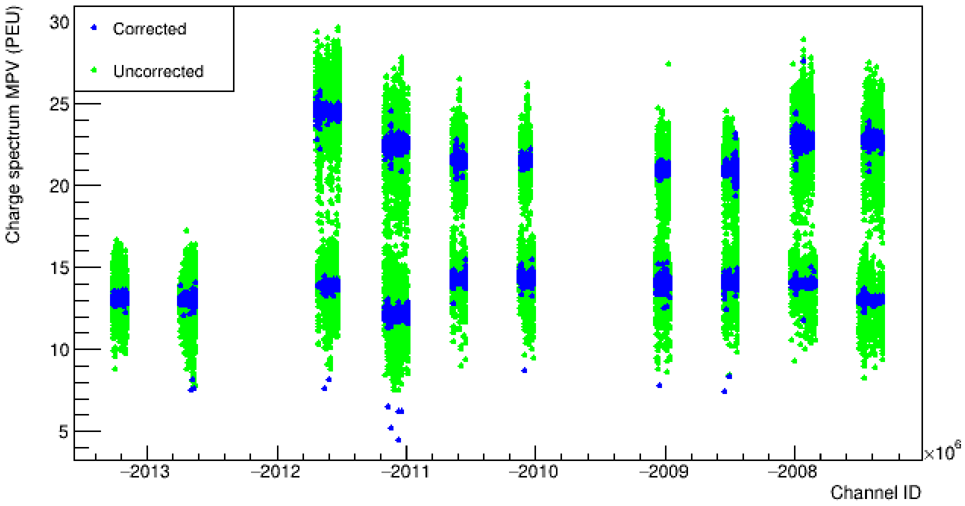


Figure 3.10: A pre-upload validation plot for ECal bar-to-bar constants as in Figure 3.9, but generated from a statistically independent cosmic muon data sample from the same T2K run. Some spread can be seen in the corrected points about the mean values as a result of statistical noise, but this is greatly reduced compared to the uncorrected points.

3.2.3 Fibre scaling correction

There is some attenuation of photons as they traverse the WLS fibres, so the light yield recorded in an ECal hit will depend on how far along the bar's length the hit originated. To account for this, the 'fibre scaling' correction is applied. Since this requires longitudinal positional information that a single hit does not provide, it cannot be applied until the reconstruction stage, when hits are matched between the two views (i.e. layer orientations) of the ECal module. Then the component of the hit position along the length of the bar can be estimated using the adjacent hit(s) or, where two hits are detected in the same bar, the timing information.

The attenuation profile of the fibre is modelled as a sum of two exponential functions, as shown in Figure 3.11. In the case of single-ended bars, an additional correction is applied to account for light reflected from the mirrored end. The constants governing these functions were calculated separately for each bar type using cosmic muon tracks similarly to the bar-to-bar correction, but are stable with time so are not recalculated each run.

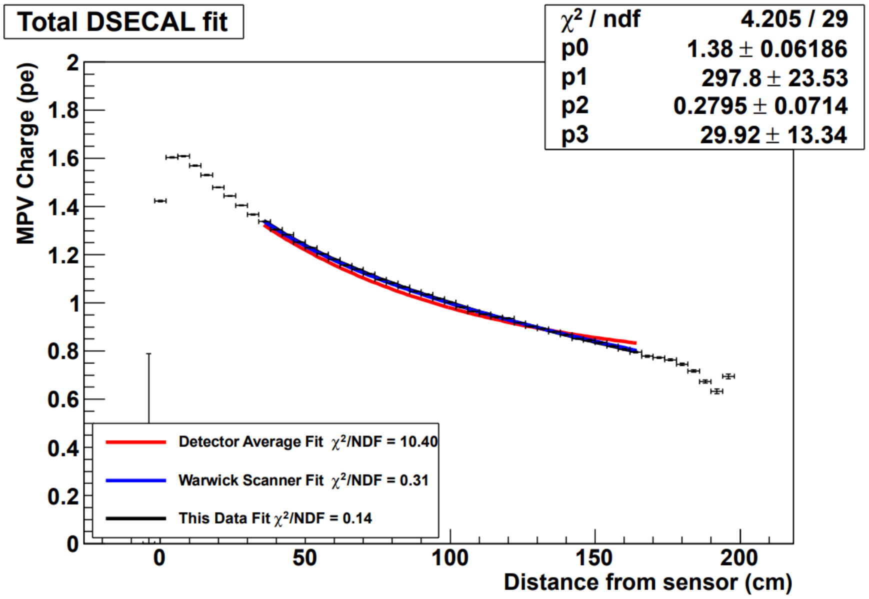


Figure 3.11: The double exponential attenuation curve fitted to data (black) used for the fibre scaling correction in the DS-ECal. Previous fits are shown in red and blue [78].

3.2.4 Scintillator ageing

The light yield of plastic scintillator detectors tends to degrade over time; this is referred to as ‘scintillator ageing’. The response of the scintillator-based T2K subsystems has been observed to reduce by 0.9–2.2% per year [73], and it has been confirmed that this is due to the scintillator plastic rather than that of the WLS fibres. The exact cause remains unknown, though it is hypothesised to be some combination of crazing or shearing due to mechanical stress, fogging due to water penetrating into the material and condensing, and oxidation of the scintillator through photochemical processes [73]. This is corrected for by fitting an exponential function to the measured light yield degradation over time, which is then used to scale the MPVs for each time bin such that they remain constant with time [80]. It has been projected that the ECal scintillator light yield will drop by ~50–60% for all bar types by 2040, and although this remains above the minimum charge threshold of 5.5 PEU required by the current reconstruction algorithms, there is a risk that information will be lost for particle interactions that deposit energy below the MIP MPV if improvements are not made [73].

3.3 ECal reconstruction and performance

Hits recorded in the tracker ECals are reconstructed into track and shower objects by the *ecalRecon* software package [72]. Clusters of hits are identified in each 2D view and matched to form 3D cluster objects, and Michel electron candidates are identified using timing information. The energy and dimensions of each cluster are computed by fits of the hits associated with it. The resulting track and shower objects are passed to the global reconstruction to be matched with objects from other subdetectors.

First, hits in each view are grouped by time to form ‘hit selections’, such that the time between successive hits in a selection is less than 50 ns. Each hit selection is treated separately by the reconstruction algorithms. Where two hits are recorded in the same double-ended bar corresponding to the same physical energy deposit, they need to be recombined, and in this case the effective speed of light can be used to estimate the position along the bar. This in turn is used to apply the fibre scaling correction described previously.

3.3.1 Clustering

Hits in a selection are ‘clustered’ to group the hits arising from a single incident particle into one object in each 2D view. This is done in three stages by three separate algorithms: Basic Clustering, Combine Clusters and Expand Clusters. In Basic Clustering, the hit with the highest charge is used as a seed to cluster nearby hits. A hit is clustered with the seed if it satisfies all of the following criteria (illustrated in Figure 3.12):

- The hit lies in the same, neighbouring, or next-to-neighbouring layer as the seed
- Along the layer, the hit lies within one bar either side of the seed bar
- The hit time is within 15 ns of that of the seed.

Once this has been completed, each clustered hit becomes the seed, continuing recursively until no more hits can be grouped with any of the hits in the cluster. The process is then repeated with any hits that are not associated with the first cluster, continuing until no more clusters can be formed.

The Combine Clusters algorithm aims to merge basic clusters for events that have regions with sparse hits, as is often seen for hadronic showers. The cluster with the most hits is used as a seed, and is merged with nearby smaller clusters if their average hit times are within 40 ns and their average positions and widths (determined by principle component

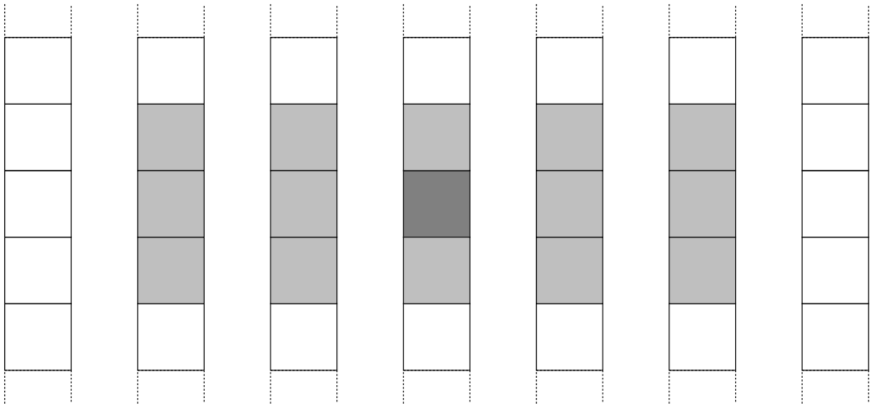


Figure 3.12: Schematic illustrating the spatial criteria of the basic clustering algorithm, where the boxes represent bars of one view with the layers shown vertically; the gaps between these layers are the layers of the other view. The seed hit shown in dark grey will be clustered with hits in the bars shown in lighter grey [81].

analysis) are found to match geometrically. Clusters must have at least three hits to be valid for Combine Clusters. The Expand Clusters algorithm is similar, and attempts to match unclustered hits into the existing clusters. Each unclustered hit is considered in turn and the algorithm attempts to match it to the existing clusters. A hit is matched and added to a cluster if its time is within 40 ns of the cluster average and its position matches the cluster geometrically.

Next, clusters in each 2D view are matched to create 3D cluster objects. A likelihood variable which takes into account the ratio of their charges and the difference in starting layer (the closest hit layer to the centre of ND280) determines whether a pair of clusters should be matched. All possible cluster pairs are considered, and those with the best likelihood values are combined provided the likelihood passes a quality cut. Each 2D cluster may only be matched with one cluster from the other view, as illustrated in Figure 3.13. Any clusters that fail to match may be 'rematched' with unclustered hits from the other view, and are otherwise stored separately as 2D cluster objects. Once 2D clusters have been matched, the hit positions along the bar can be recalculated using information from the other view, providing a more accurate result than the recombination estimate. This is then used to re-calibrate the fibre scaling correction.

The overall reconstruction efficiencies (that is, the probabilities of an ECal cluster being reconstructed for an incident particle) of the downstream and barrel ECals are summarised

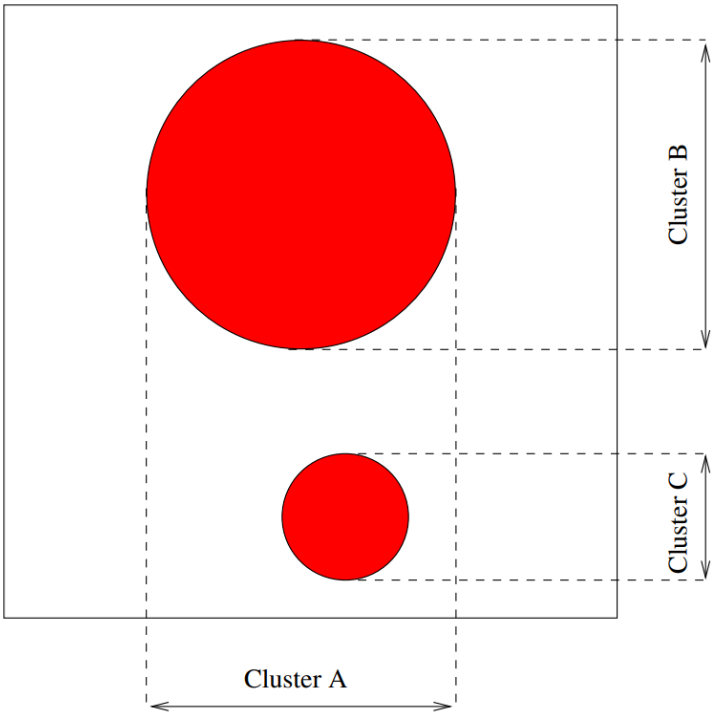


Figure 3.13: Illustration of how overlapping clusters are handled by the 2D-3D matching. Two separate clusters (B and C) are seen in one view, while in the other view they are superimposed as a single cluster (A). Due to their similarity, cluster A would be matched to cluster B to form a 3D cluster object, leaving cluster C as an unmatched 2D cluster [81].

in Figure 3.14, and their momentum and angle dependencies are shown in Figures 3.15 and 3.16. It can be seen that the efficiency of the downstream ECal is higher than that of the barrel, and the efficiency for tracks is higher than for showers. There are also strong dependencies on the kinematics, with higher-momentum particles generally having better reconstruction efficiency, and perpendicular impact angles resulting in better efficiency than shallower ones. Further information on the reconstruction efficiency of the ECal can be found in [82].

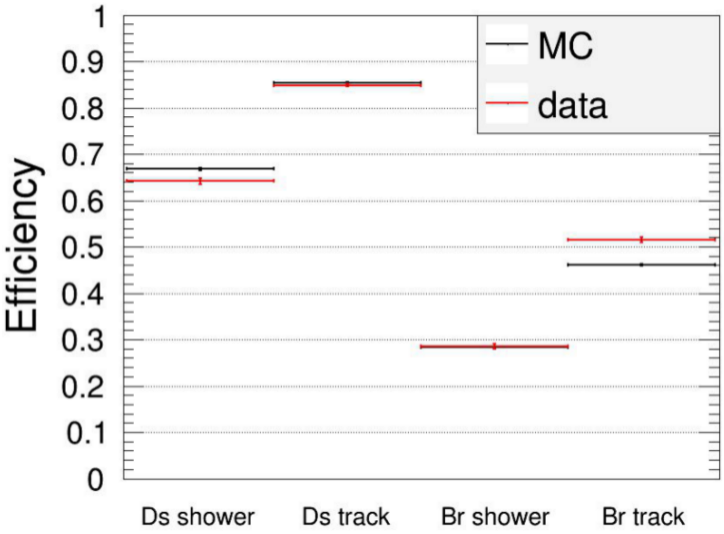


Figure 3.14: Overall efficiencies for shower and track reconstruction in the downstream and barrel ECals, calculated using enhanced $e^{\pm}$ (shower-like) and enhanced μ^{-} (track-like) control samples of both MC and real data. These were selected by TPC PID from samples of TPC tracks with trajectories approaching the ECals. The efficiencies here are defined as the proportion of shower-like/track-like TPC candidates for which a shower-like/track-like object is seen in the expected ECal. It can be seen that the reconstruction efficiency is generally higher for tracks than for showers (though this may be due to muon and pion impurities in the $e^{\pm}$ control sample), and higher for the downstream than the barrel ECals (thought to be due to all DS-ECal scintillator bars being double-ended, while the barrel ECals contain both double- and single-ended bars, the latter of which perform more poorly). The data-MC discrepancy for track-like objects in the barrel ECal is due to a detector geometry issue which is corrected for in real data but not MC [82].

3.3.2 Energy and shape fits

An energy fitting algorithm is applied to all reconstructed 3D clusters to calculate an energy deposit from the calibrated hit charges. The fit assumes an electromagnetic shower and is a likelihood fit of parameters of the cluster charge distribution: the total charge, the RMS divided by mean, and the skew (though the result is mainly dependent on the total charge). The fit is tuned on a Monte Carlo particle gun photon sample with energies ranging from 75 MeV to 25 GeV to ensure validity across the expected energy range. The resolution of this energy measurement has been calculated from test beam measurements to be around

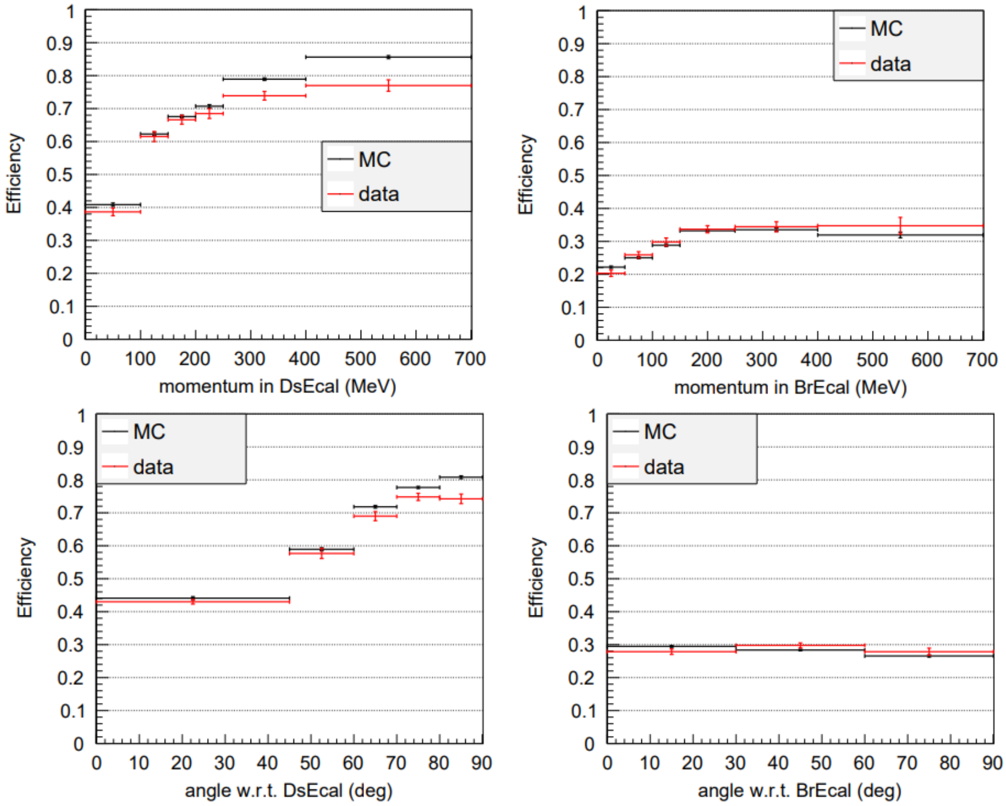


Figure 3.15: Shower reconstruction efficiency in the downstream (left) and barrel (right) ECals, plotted as a function of momentum (top) and impact angle with respect to the ECal surface (bottom). The efficiencies are defined and calculated as in Figure 3.14 from the same control samples [82].

8–20% for EM showers [72][83].

Each cluster is then fitted as both track and shower objects: the track fit consists of 2D linear fits in each view which are combined into a 3D linear fit, and the shower fit is a 3D principle component analysis. These fits assign position, direction and shape information to the cluster, and particle identification is later used to determine which hypothesis (shower or track) is used.

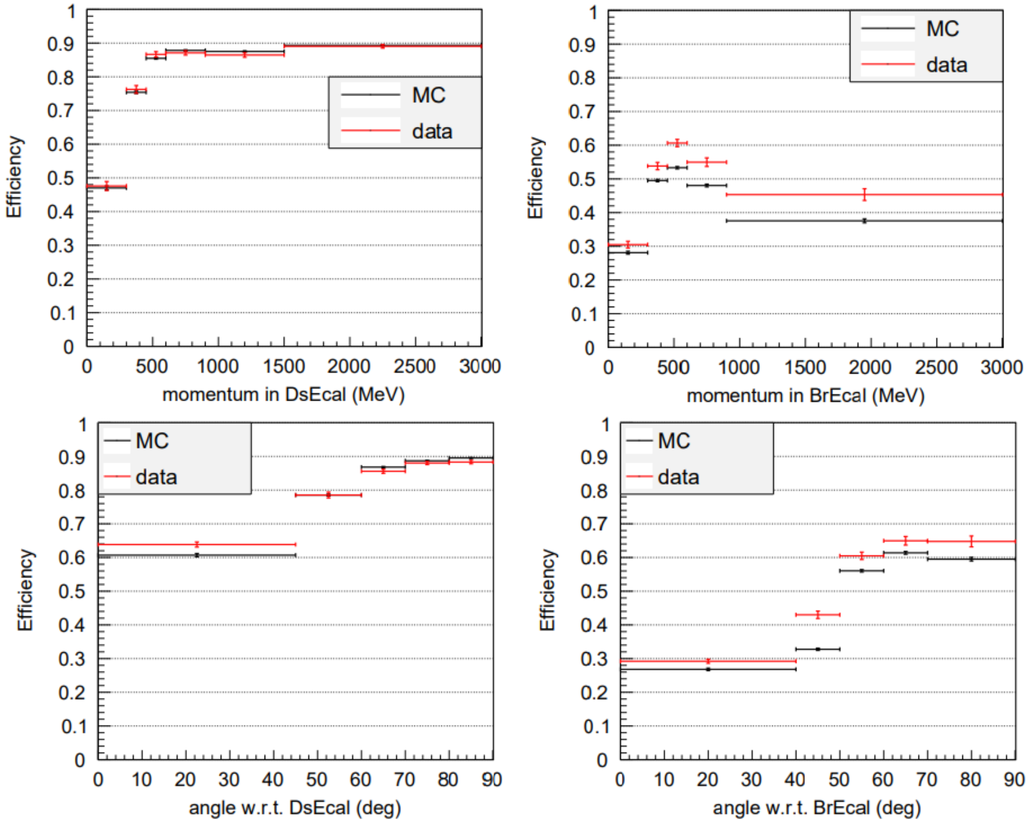


Figure 3.16: Track reconstruction efficiency in the downstream (left) and barrel (right) ECals, plotted as a function of momentum (top) and impact angle with respect to the ECal surface (bottom). The efficiencies are defined and calculated as in Figure 3.14 from the same control samples [82].

3.3.3 Global reconstruction

Reconstructed ECal objects are passed to the global reconstruction, which matches and combines them with objects from the other subdetectors to form global objects. A Kalman filter [84] is used to refit the individual objects (provided they are track-like, not shower-like) and to refit combinations of matched objects. The reconstruction attempts to iteratively match pairs of objects in adjacent subdetectors by extrapolating one to a matching plane with the other; they are matched if the calculated χ^2 is less than 100 (200 when the objects include the PØD or SMRD) and the time difference is within 300 ns. Only two objects are matched at a time, and the resulting combined object is used in the next iteration. This begins with objects in the tracker subdetectors, then proceeds to unmatched objects in the PØD and ECal respectively until no more matches can be found [81]. Global tracks are the main reconstruction objects used at the analysis level, but the local objects remain accessible as well.

ND280 event selections typically require a track with at least one TPC segment, so the efficiency for matching tracks between the TPC and ECal is an important quantity to consider. Measurements of this efficiency have been made by selecting tracks that appear to enter the downstream or barrel ECals; the matching efficiency is then the proportion of such tracks for which an ECal segment is reconstructed [82]. These efficiencies have been measured for e-like, μ-like and proton tracks; an example is shown in Figure 3.17 and the full set of plots can be found in [82]. It has been found that the matching efficiency generally increases with momentum and is higher for shallower track angles, as well as differing between particle types, with μ-like tracks generally showing the highest efficiencies [82].

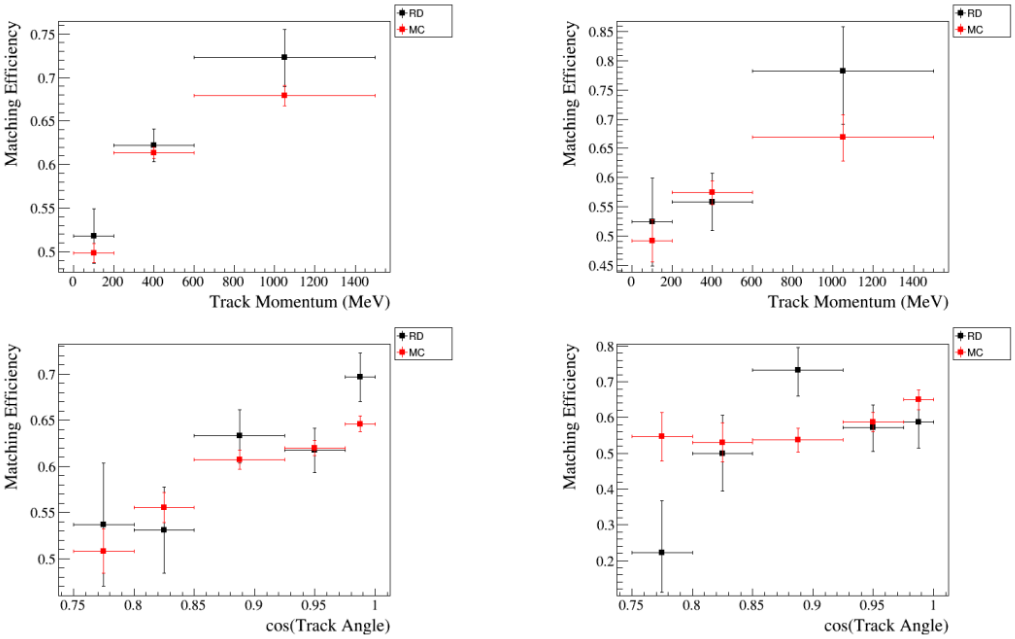


Figure 3.17: TPC-ECal matching efficiency for e-like tracks entering the DS-ECal, shown as a function of momentum and impact angle with respect to the ECal surface. The efficiencies are shown for both real data (black) and Monte Carlo (red), and from ν mode (left) and $\bar{\nu}$ mode (right) samples [82].

Chapter 4

Muon-antineutrino CC1pi selection development in ND280

As described in Section 1.4 above, neutrino-nucleus interactions are an essential and highly active topic of research in neutrino physics. They are difficult to describe theoretically, since the target nucleon is not free but bound within a nucleus. The interaction is subject to nuclear effects which are difficult to model, and significant discrepancies still exist between current models and experimental data. As neutrino oscillation experiments move towards higher and higher precision, neutrino interaction models have become a critical source of systematic uncertainties. Cross-section measurements are needed in order to constrain and improve models to increase the precision of experiments that rely on neutrino-nucleus interactions.

It is impossible to directly observe the products of the initial interaction, since it occurs inside the nucleus; only the particles that leave the nucleus can be detected. For this reason, T2K classifies events based on the topological information: the set of particles that leave the nucleus. One such category is known as $\nu_\mu(\bar{\nu}_\mu)$ CC1pi: events in which a muon-(anti)neutrino interacts with a nucleus via the charged current weak interaction, producing an (anti)muon, a charged pion and no other particles besides protons. In FHC mode, this is known as ν_μ CC1π^+:

$$\nu_\mu + A \rightarrow A' + \mu^- + \pi^+ + Np \tag{4.1}$$

and $\bar{\nu}_\mu$ CC1π^- in RHC mode:

$$\bar{\nu}_\mu + A \rightarrow A' + \mu^+ + \pi^- + Np \tag{4.2}$$

where A is the nuclear target, and $N \geq 0$ since the topology definition allows any number of protons. The interaction channels that can give rise to CC1pi topology are discussed in Section 1.4.3 above. The work in this thesis only considers events originating in FGD1, for which the nuclear target is primarily carbon with a small proportion of hydrogen.

One-pion topologies make up a significant proportion of charged-current events at T2K. As a result they form a major background for CCQE samples which are essential to the oscillation analysis, and can themselves be used as a sample for oscillation fits. This requires precise cross-section measurements, which can also contribute towards the testing and development of neutrino-nucleus interaction models. In turn, cross-section measurements require the development of high-performing event selection algorithms, here referred to simply as 'selections'.

A selection is a sequence of cuts on the reconstructed event information, designed to select events of a particular signal topology and reject its backgrounds. As part of this process, the selection designates reconstructed objects (e.g. global tracks) as candidates for each of the particles in the event topology. A $\bar{\nu}_\mu$ CC1π^- selection should therefore identify candidate tracks for both a μ^+ and a π^- (and any number of protons), and reject events otherwise. This is challenging, since the RHC beam has a significant contribution from ν_μ, here termed the 'wrong-sign background'. The wrong-sign background can for example give rise to ν_μ CC1π^+ events, for which the signature is a μ^- and a π^+. This mimics the signal μ^+ and π^- since muons and pions can behave very similarly. For this reason, particle identification (PID) is a crucial element of $\bar{\nu}_\mu$ CC1π^- selections in ND280. However, previous attempts have been derivative of the $\bar{\nu}_\mu$ CC inclusive selection and have not used PID suited to the particular challenges of selecting the CC1π^- topology. As a result they perform poorly. This thesis presents efforts to improve on the existing ND280 FGD1 $\bar{\nu}_\mu$ CC1π^- event selection by developing high-performing PID algorithms, with a focus on muon-pion discrimination to address the wrong-sign background issue.

4.1 Data and Monte Carlo samples

The data and MC samples used for the work presented in this chapter and Chapter 7 were taken from the T2K Production 6T RHC samples for runs 5, 6, 7 and 9. The POT statistics of these samples are summarised in Table 4.1. The MC sample is T2K Production 6T simulated data, which was generated using ND280 software version `nd280v11r31p43` and NEUT [51] version 5.4.0.1, with the neutrino beam flux predicted by the JNUBEAM [62] software. The relative quantities of each interaction type are as predicted by NEUT and have not been modified. Further information concerning Production 6T can be found in [85].

ND280 Run	MC POT $\times 10^{20}$	Data POT $\times 10^{20}$
Run 5	21.9125	0.445156
Run 6	26.0132	3.42020
Run 7	32.3090	2.43498
Run 9	5.10102	2.30244
Total	85.3358	8.60279

Table 4.1: Summary of real data and Monte Carlo production POT used for the work presented in Chapters 4 and 7.

4.2 The current ND280 $\bar{\nu}_\mu$ CC1π^- event selection

The selection development described in this thesis takes the pre-existing ND280 event selection as its starting point. This selection is part of a branched selection for multiple pion analyses [86]. This begins with a number of pre-selection cuts designed to select events of suitable data quality and remove certain backgrounds before applying PID:

- **Event quality:** Only events compatible with timing information from the beam are selected. The T2K beam is produced by eight proton 'bunches' at a time, each of width 15 ns. To pass this cut, events must be associated with the beam trigger and compatible with one of the bunches; that is, they must fall within 4σ of the centre of a bunch.

- **Total multiplicity:** Events must have at least one reconstructed track crossing TPC2 or TPC3.

- **Track quality and fiducialisation:** The reconstructed vertex of the event must fall within the fiducial volume (FV) of FGD1. Additionally, the TPC track must have more than 18 clusters (sets of contiguous pads in a row or column) in order to be selected. This requirement is imposed because the TPC reconstruction is less reliable for shorter tracks.

- **Leading track:** Of all tracks originating in the FGD1 FV, the highest-momentum positively-charged track (HMPT) is considered the antimuon candidate. The leading track cut requires that this also be the highest-momentum track (HMT) overall, in order to reduce π^+ contamination.

- **Upstream background veto:** Events in which the second-highest-momentum track starts > 150 mm upstream of the muon candidate are rejected. This cut is applied in order to remove misreconstructed events in which the true muon originated upstream of the FGD1 FV but the reconstructed track starts inside it.

- **Broken track:** If the muon candidate track starts in the last (most downstream) two layers of the FGD, and a fully FGD-contained track is also present, the event is rejected. This is to remove misreconstructed events in which the muon track is 'broken' into two tracks.

Events that pass the pre-selection are then subject to a 'primary' PID cut to identify an antimuon and thus select an inclusive sample of $\bar{\nu}_\mu$ charged-current events, and additional tracks (such as pions and protons) are identified by applying 'secondary' PID to each of them. The primary and secondary PID are described in Sections 4.2.1 and 4.2.1 respectively. At this point the sample is split into three selection branches according to the number of reconstructed mesons: CC0pi (zero mesons), CC1π^- (one negative pion only, no other pions), and CC-Other (any other topologies). The CC1π^- branch is selected by requiring one reconstructed π^- candidate and no other pion signatures. The π^- candidate must be in the same time bunch and start in the same FGD FV as the muon candidate. If it enters one of the TPCs downstream of the starting FGD, then it must also pass the TPC quality cut. Events tagged as containing π^+, π^0 or additional π^- are rejected.

4.2.1 Particle identification

The existing $\bar{\nu}_\mu$ CC1π^- selection relies on TPC PID (or where it is unavailable, FGD PID). This section describes how the TPC and FGD PID variables are constructed, and how they are applied in the primary and secondary PID.

TPC PID variables

Particle ID in the TPCs is based on the rate of energy loss per unit distance travelled (‘dE/dx’) in the gas. This is measured from the ionisation charge clusters reconstructed along the track. Since the processes of energy loss in the gas are subject to large stochastic fluctuations, the mean dE/dx given by the Bethe-Bloch function is not useful in this case; instead, the ‘straggling function’ is used, which describes the distribution of energy loss for a thin gas layer [87]. The most probable value (MPV) of this function depends only on the mass and momentum of the particle, and can therefore be used for particle ID in conjunction with the reconstructed momentum. Since the distribution of dE/dx has a long tail, the mean of the measured energy loss is a poor approximation to the MPV of the straggling function. Instead a truncated mean is used, in order to measure a quantity more closely related to the peak of the distribution by discarding measurements with very large energy deposition. This is defined:

$$C_T = \frac{1}{\alpha N} \sum_{i}^{\alpha N} C_C(i) \tag{4.3}$$

where $C_C(i)$ is the energy in cluster i (with the clusters ordered in increasing energy), N is the numbers of cluster energy measurements in the TPC, and α is the truncation fraction which has been set at 70% to optimise energy resolution [87]. C_T is further calibrated to account for differences in the direction and number of clusters of the track [87][88] to a calibrated truncated mean $\bar{C}_T$.

The high-level particle ID for the ND280 TPCs compares the measured $\bar{C}_T$ to the expected MPV C_E for different particle types (μ, π, proton, electron), considering the probability that a particle of mass m and with the measured momentum yields the observed $\bar{C}_T$. For each particle type, a pull δ is defined which describes the distance between the expected and measured values. For the i-th particle hypothesis the pull is defined:

$$\delta_i = \frac{\bar{C}_T - C_E(i)}{\sigma_o(i)} \tag{4.4}$$

where $\sigma_o(i)$ is the total width, defined as

$$\sigma_o(i) = \sigma_T(i) \oplus (dC_E/dp)\sigma_p \tag{4.5}$$

where σ_T is the standard deviation of the distribution of $\bar{C}_T$ whose mean is C_E, (dC_E/dp)

the derivative of C_E with respect to momentum, and σ_p the uncertainty of the momentum measurement [87]. Furthermore for a given track, the pulls can be used to construct a likelihood variable L for each hypothesis:

$$L_i = \frac{e^{-\delta_i^2}}{\sum_l e^{-\delta_l}} \tag{4.6}$$

where l are the full set of particle hypotheses. Hence the likelihoods L_i represent the estimated probability that the identity of the particle producing the track is the same as the hypothesis i, based on energy loss in the TPC.

FGD PID variables

If a track is contained to the FGD, that is, if a particle stops before leaving the FGD, no TPC information is available so PID can only be performed via FGD information. Like the TPC PID, the FGD PID is based on the energy loss of the particle due to ionisation in the medium, also known as the 'stopping power' $S(E)$. This is described by the Bethe-Bloch equation [89]:

$$S(E) = -\frac{dE}{dx} = 4\pi N_A r_e^2 m_e c^2 z^2 \frac{Z}{A}\frac{1}{\beta^2}\left[\frac{1}{2}\ln\frac{2m_e c^2\beta^2\gamma^2 T_{max}}{I^2} - \beta^2 - \frac{\delta}{2}\right] \tag{4.7}$$

where z is the charge of the incident particle as a multiple of the electron charge; Z and A are the atomic number and atomic mass number of the medium respectively; N_A is Avogadro's number; r_e is the classical electron radius; β and γ are the usual relativistic kinematic variables; $m_e c^2$ is the electron rest mass; I is the mean excitation energy; T_{max} is the maximum possible kinetic energy that can be imparted to a free electron; and δ is a parameter that corrects for density effects. The stopping power can be used to calculate the expected range R of a particle traversing a medium:

$$R = \int_0^R dx = \int_E^0 \frac{dx}{dE} dE = \int_0^E \frac{dE}{S(E)}. \tag{4.8}$$

At low energies ($\beta \ll 1$), the stopping power of particles with the same kinetic energy differs according to their mass; hence stopping particles of the same kinetic energy and different mass will have different track lengths in the FGD, and we can use the measured energy and track length to construct PID variables. The track length is defined as the length of the straight line between the initial and final 3D positions, which are obtained by fitting a straight line in each 2D projection. The total energy deposited is obtained from the sum of the energy recorded for each hit making up the track. Using these and Equation 4.8, a set of pulls δ_i are constructed for the muon, pion and proton hypotheses:

$$\delta_i = \frac{E - E_i(x)}{\sigma_i(x)} \tag{4.9}$$

where $i = \mu, \pi, p$ is the particle hypothesis, E is the measured energy deposited, x is the measured track length, and E_i and σ_i are the expected energy deposited and resolution respectively for the ith hypothesis. These pulls assume a stopping particle, so are only valid when the track is found to be contained to the FGD [90].

Primary PID

In ND280 $\bar{\nu}_\mu$ CC selections the antimuon candidate track is referred to as the 'primary track', and all other tracks as 'secondary tracks'. The primary track is required by the pre-selection to have a TPC segment, so TPC PID is used to identify it. This is implemented as follows:

- **Antimuon TPC PID:** Particle identification is performed using the TPC likelihood variables of the designated antimuon candidate, considering the muon, pion, electron and proton hypotheses as defined in Section 4.2.1. The following cuts are applied, accepting events for which:

  $$L_\mu > 0.1 \tag{4.10}$$

  $$L_{\text{MIP}} = \frac{L_\mu + L_\pi}{1 - L_p} > 0.9 \text{ if } p < 500 \text{ MeV/c} \tag{4.11}$$

  where L_μ, L_π, L_p and L_{MIP} are the muon, charged pion, proton and minimum ionising particle (MIP) likelihoods respectively; and p is the reconstructed momentum of the track.

In the case of the pre-existing $\bar{\nu}_\mu$ CC selection, the primary track is the highest-momentum positive track in the event (and also required to be the highest-momentum track overall). This PID is applied to select an inclusive selection of $\bar{\nu}_\mu$ charged-current events, accepting approximately 93% of signal (all $\bar{\nu}_\mu$ charged-current topologies) and rejecting 77% of backgrounds, yielding a 'CC-inclusive' sample that is 82% pure[1].

Secondary PID

Following the primary PID cut, the sample is split into three selection branches according to the number of detected mesons: CC0pi (zero mesons), CC1π^- (one negative pion only, no other pions), and CC-Other (any other topologies). This requires identification of all secondary tracks in the event. To be considered as a secondary track, a track must be in the same time bunch and start in the same FGD FV as the antimuon candidate. If a track passes through the TPC, then it must pass the TPC quality cut and PID is performed using the TPC likelihoods; if it is contained to the FGD, then the FGD pulls are used instead. For TPC PID of secondary tracks, only the pion, proton and electron hypotheses are considered, since muons and pions cannot be distinguished in the TPC (see Section 4.2.2) and it is assumed that the antimuon candidate is the only μ in the event.

The secondary PID process begins by identifying pion candidates. Positive TPC tracks are identified as π^+ if L_π is greater than L_p and L_e. The proton hypothesis is not needed for negative TPC tracks, which are identified as π^- if $L_\pi > 4L_e$. FGD-contained tracks are identified as charged pions if $-2.0 < \delta_\pi < 2.5$ where δ_π is the FGD pion pull. Charged pions may also be tagged in the FGD via Michel electrons which are produced in the decay of muons, themselves produced in pion decay. Michel electrons are identified by the presence of time-delayed hits in the same FGD as the interaction vertex. A Michel electron tag is more likely to indicate a π^+ than a π^- since the latter are more likely to be absorbed, so for the purposes of this selection a Michel electron is considered evidence of a π^+.

Neutral pions do not themselves leave tracks in ND280, but they can be identified by positrons and electrons pair-produced by their decay photons. Positive tracks in the TPCs are identified as positrons if L_e is greater than L_p and L_π (provided that the momentum is below 900 MeV/c, as otherwise the track is more likely to be a proton). Negative tracks are identified as electrons if $L_\pi < 4L_e$. Tracks contained to the FGD are identified as $e^\pm$ if $\delta_\pi < -2.0$, or $\delta_\pi < -3.0$ if one or more Michel electrons have been tagged. Since $e^\pm$ TPC tracks are unlikely to be produced by any other source[2], any reconstructed $e^\pm$ are

[1] These values were obtained from testing with the MC sample used in this thesis.

[2] While $e^\pm$ can also be produced in muon decay or $\nu_e/\bar{\nu}_e$ interactions, these do not generally result in TPC tracks in $\nu_\mu/\bar{\nu}_\mu$ selections. Muon decay typically occurs after the muon stops in a solid medium such as the FGDs or TPCs, and the resulting Michel $e^\pm$ are produced following a delay so are reconstructed

considered evidence of π^0 decay.

Following the identification of pion candidates, proton candidates are selected from any remaining FGD-contained or positive TPC tracks. Positive TPC tracks are identified as protons if $L_p > 0.5$, and FGD-contained tracks if $\delta_p > -4$ where δ_p is the FGD proton pull.

Following this, all tracks that have been identified as μ^+, $\pi^\pm$, $e^\pm$, or protons are considered to assign the event topology. If a track has not been assigned a PID identity, then it is disregarded. The CC1π^- branch is selected by requiring one and only one reconstructed π^- candidate (either a negative TPC pion or an FGD-contained pion) and no other pion signatures (negative TPC pions, Michel electrons, or π^0 electrons). This is referred to as the **one pion cut**. Optionally, an 'ECal π^0 veto' may be applied in order to further remove events with neutral pions, but this has been found to reject significant numbers of signal events [91] so is not generally used. The one pion cut accepts 30% of $\bar{\nu}_\mu$ CC1π^- signal and rejects 96% of backgrounds, but since the backgrounds are here much more plentiful than the signal, this yields a sample with only 48% purity (values again obtained from testing with the MC sample used in this thesis).

4.2.2 Limitations

The above set of cuts will here be referred to as the 'existing selection'. Previous testing of this selection on MC simulated events [86] found that the selected CC1π^- sample was only 45.4% pure i.e. contained only 45.4% true $\bar{\nu}_\mu$ CC1π^- events. Non-$\bar{\nu}_\mu$ CC processes made up 36.7% of the sample, of which 77.9% were ν_μ interactions. The ability to better reject wrong-sign events is therefore a crucial target for efforts to improve this selection. Furthermore, the selection testing undertaken for this analysis also shows that the one pion cut rejects a large proportion of signal ($\sim$ 70%), so increasing selection efficiency is also desirable.

The leading track cut described above was originally introduced to mitigate the wrong-sign background, but it brings with it its own issues. Only the highest-momentum track is considered as a μ^+ candidate, so any signal events in which the π^- and/or a proton has higher reconstructed momentum than the μ^+ will be rejected. This makes the selection highly dependent on the event kinematics, and therefore on the neutrino interaction model. Ideally this should be avoided, so it is desirable to remove this cut, but doing so results in

separately from other tracks. A $\nu_e/\bar{\nu}_e$ event typically will not contain muon-like tracks, which are already required by the primary PID.

even greater wrong-sign contamination of the sample (demonstrated below). This further adds to the importance of accurately rejecting wrong-sign backgrounds.

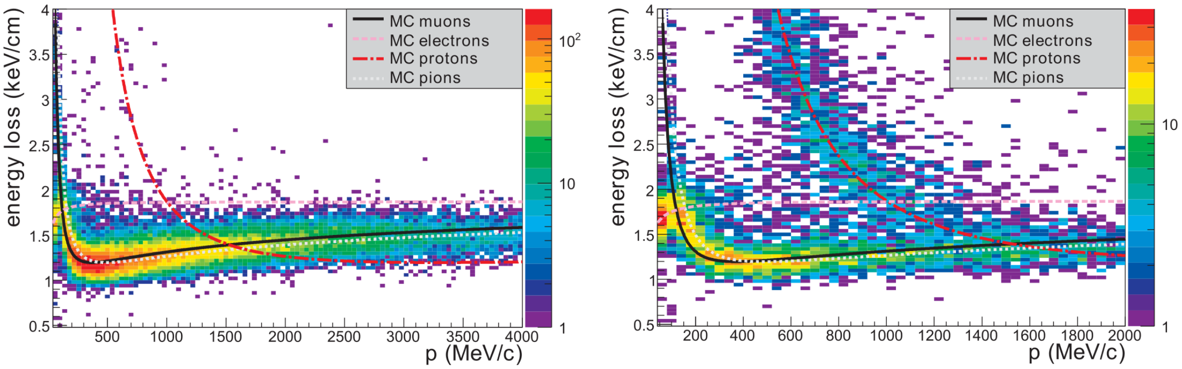


Figure 4.1: Measured distributions of energy loss as a function of momentum for negatively charged (left) and positively charged (right) particles in the ND280 TPC. Energy loss curves predicted by MC simulations are shown for muons, electrons, protons and pions [67].

The misidentification of wrong-sign events as signal arises as a result of incorrect track PID. This can be seen by considering the track purity: the proportion of tracks that have been assigned the correct identity by the PID. Of the μ^+ candidate tracks in the selected MC sample, only 64.2% are associated with true μ^+. The dominant backgrounds are true π^+ and protons, which make up 19.7% and 13.6% respectively. Similarly, of selected π^- candidates with TPC segments, 67.7% are true π^- and 30.1% are μ^-. This high misidentification rate is a result of the PID relying entirely on TPC energy loss (for tracks that cross the TPC). Figure 4.1 shows the measured TPC dE/dx as a function of momentum, compared with the expected dE/dx curves for different particle types. The curves for muons and pions are very similar across the momentum spectrum, and are thus very difficult to distinguish at the resolution of the TPC. Additionally, for positive tracks, the proton dE/dx curve crosses the muon curve at around 1500 GeV/c, so protons will be indistinguishable from muons in this momentum region. The consequence of this can be seen in Figure 4.2: the proton contamination of the μ^+ candidate track peaks at around 1500 MeV/c, causing a corresponding drop in the μ^+ purity. The π^+ contamination is approximately constant across the momentum spectrum.

Although μ^+ candidate PID using only TPC energy loss information performs well at

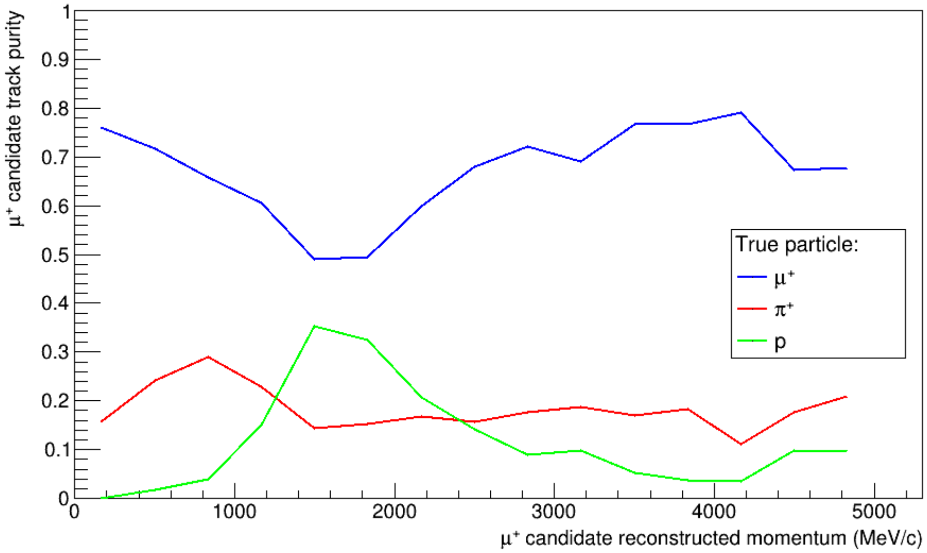


Figure 4.2: MC predicted purities of the μ^+ candidate track in the existing $\bar{\nu}_\mu$ CC1π^- selection, plotted as a function of reconstructed momentum. Only the dominant three particle types (antimuons, positive pions and protons) are shown.

rejecting e^+/e^- since they have a distinct dE/dx curve, it is clear that it performs poorly for muon-pion discrimination, and for muon-proton discrimination at around 1500 MeV/c. To better identify tracks and reject wrong-sign backgrounds, additional sources of PID information are required. One such source is the ECal, which can induce pions to shower and thus distinguish them from muons, making ECal PID an obvious choice to supplement TPC PID. The following section describes an attempt to improve the performance of this selection by developing additional PID cuts using ECal information.

4.3 Improved $\bar{\nu}_\mu$ CC1π^- event selection

The goal of this work was to improve the performance of the $\bar{\nu}_\mu$ CC1π^- selection, with a focus on improving the wrong-sign background rejection to the extent that the leading track cut would no longer be necessary. This was pursued by developing cuts on ECal variables to be added to the PID of the μ^+ and π^- candidate tracks.

4.3.1 Removing leading track requirements

As described above, the existing selection only considers the highest-momentum positive track as a μ^+ candidate, and rejects events for which the HMPT is not the highest-momentum track overall. This means the selection rejects any signal events for which the π^- or a proton track is reconstructed with higher momentum than the μ^+, introducing a model dependency. Removing this dependency involved two steps:

- **No leading track cut:** the requirement that the HMPT be the HMT was removed. This allows π^- tracks to have higher momentum than the μ^+.
- **Antimuon candidate selection loop:** the selection of the μ^+ candidate track was modified to consider all good-quality positive TPC tracks. The TPC antimuon PID cut is applied to each such track. If exactly one track passes the cut, this is chosen as the μ^+ candidate. Otherwise, if zero or more than one μ^+-like tracks are found, the event is rejected. This allows proton tracks to have higher momentum than the μ^+.

With these changes made, the selection is referred to as the 'modified selection'. Testing the effect of these changes on the MC sample showed a severe impact on the purity of the selection, since large numbers of wrong-sign events are no longer being screened by the kinematic restrictions. The purity of $\bar{\nu}_\mu$ CC1π^- events fell from 47.8% to 30.6%, with the non-$\bar{\nu}_\mu$ CC backgrounds rising from 33.2% to 57.1%. The purity of the μ^+ and π^- candidate tracks fell considerably (see Table 4.3 below). This demonstrates the scale of the wrong-sign background issue, and the extent to which the leading track requirements mitigate it in the existing selection. The failure of the current PID to distinguish pions from muons is clear in the large contamination of the μ^+ and π^- candidates by π^+ and μ^- respectively. To make a $\bar{\nu}_\mu$ CC1π^- selection viable with the above changes, a new high-performance particle ID must be developed.

4.3.2 ECal particle ID cuts

Particles behave very differently in the ND280 ECal compared to the TPCs and FGDs, since the lead layers of the ECal can induce the production of hadronic and electromagnetic showers. The behaviour of different particle types in the ECal results in different distributions of the deposited charge, which can be used to identify them. Muons tend to behave as MIPs and so pass through the ECal without showering, leaving a single track. Protons

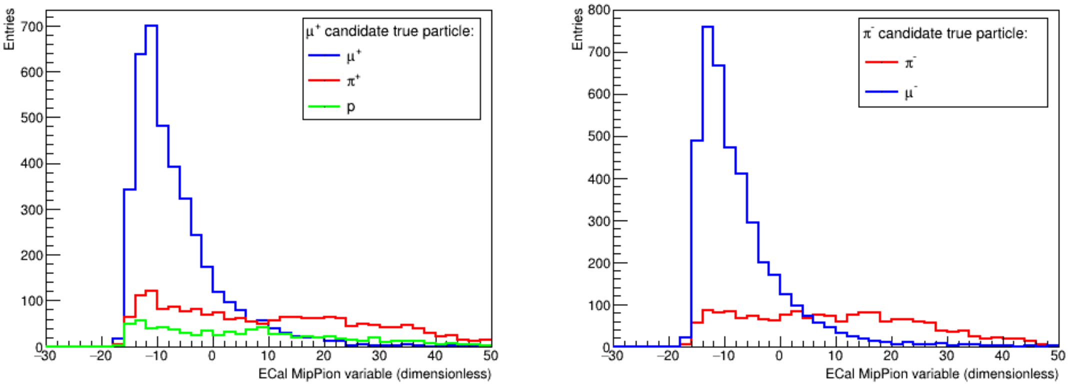


Figure 4.3: MC predicted MipPion distributions for μ^+ (left) and π^- (right) candidate tracks in the $\bar{\nu}_\mu$ CC1π^- selection with leading track requirements removed. Only the signal particle and the dominant backgrounds are shown.

tend to immediately produce hadronic showers upon entering the ECal. Charged pions can behave as MIPs similarly to muons, or they may produce showers — though unlike protons, they often pass some distance into the ECal before showering. Quantifying these different charge distributions yields PID variables that are independent of and complementary to the dE/dx-based TPC PID. This is discussed in greater detail in Section 6.1.1.

The aforementioned differences in behaviour between muons, pions and protons made the addition of ECal PID an obvious approach to reducing the wrong-sign background contamination. Approximately 60% of TPC μ^+ and π^- candidate tracks in the modified selection have an ECal segment associated with them, and 86% of events have an ECal segment associated with at least one of the two, so ECal information is usually available for PID. High-level ECal PID variables have already been developed using neural network techniques (see Section 6.1.1 and references [92] and [93]); of these, the ‘MipPion’ variable was chosen for this application because it is designed to distinguish between MIP-like and showering-pion-like behaviour. The distributions of MipPion for the μ^+ and π^- candidate tracks are shown in Figure 4.3, from which its separating power can be seen: true μ tend to have lower values, and true π/p tend to have higher values. However, there is significant overlap between the particle types, so a second ECal PID variable was constructed to provide further separating power.

The energy loss behaviour of MIPs is different to that of showering particles: all MIPs deposit approximately the same energy per unit length. We can therefore divide the total

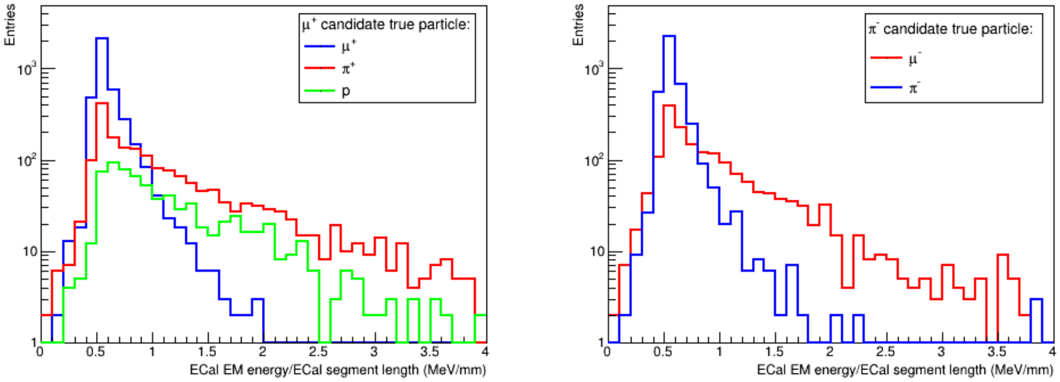


Figure 4.4: MC predicted E/L distributions for μ^+ (left) and π^- (right) candidate tracks in the $\bar{\nu}_\mu$ CC1π^- selection with leading track requirements removed. Only the signal particle and the dominant backgrounds are shown. A logarithmic scale is used to make the long tail of the pion and proton distributions more clearly visible.

EM energy[3] of an ECal track by its length, and expect this variable (here called 'E/L') to have a sharply peaked distribution for MIP-like tracks. For showers, on the other hand, the EM energy will be dependent on the total energy of the stopping particle, and the 'length' of the shower cluster is defined differently. Hence we can expect a wide range of E/L values for showering tracks. This can be seen in Figure 4.4: a sharp MIP-like peak can be seen for muons and part of the pion distribution, whereas a shower-like long tail can be seen for pions and protons. Thus E/L too can be used as a PID variable for μ/π and μ/p discrimination. The correlation factor between MipPion and E/L is 0.68 for μ^+ candidates and 0.62 for π^- candidates in the modified selection (calculated from 2D histograms using the `TH2::GetCorrelationFactor` function in ROOT [69]), so these variables are moderately correlated. Some degree of correlation is to be expected since both separate track-like and shower-like behaviour, but since these factors are not close to 1, the variables may be sufficiently independent to be complementary for PID.

Cuts on MipPion and E/L for the μ^+ and π^- candidate tracks of the modified selection were optimised by maximising the efficiency multiplied by purity for the track. Only tracks with an ECal segment associated were considered; if no such segment exists, ECal

[3] The ECal EM energy is computed under the assumption of an EM shower, which does not hold for MIPs, but the fit depends mainly on the total charge so it still behaves as needed for the E/L PID variable. The total charge itself would be a more appropriate choice, but was not used due to a bug in its implementation in the ND280 software.

PID cannot be performed so the ECal cut will be waived. The purity was defined as the proportion of μ^+ (π^-) candidates associated with true μ^+ (π^-), and the efficiency as the proportion of true μ^+ (π^-) tracks passing the modified selection that go on to pass the ECal cuts. The significance $S/\sqrt{S+B}$, where S is the number of μ^+ (π^-) candidate tracks associated with true μ^+ (π^-) and B is the number not associated with true μ^+ (π^-), was also optimised and found to yield the same results as efficiency*purity.

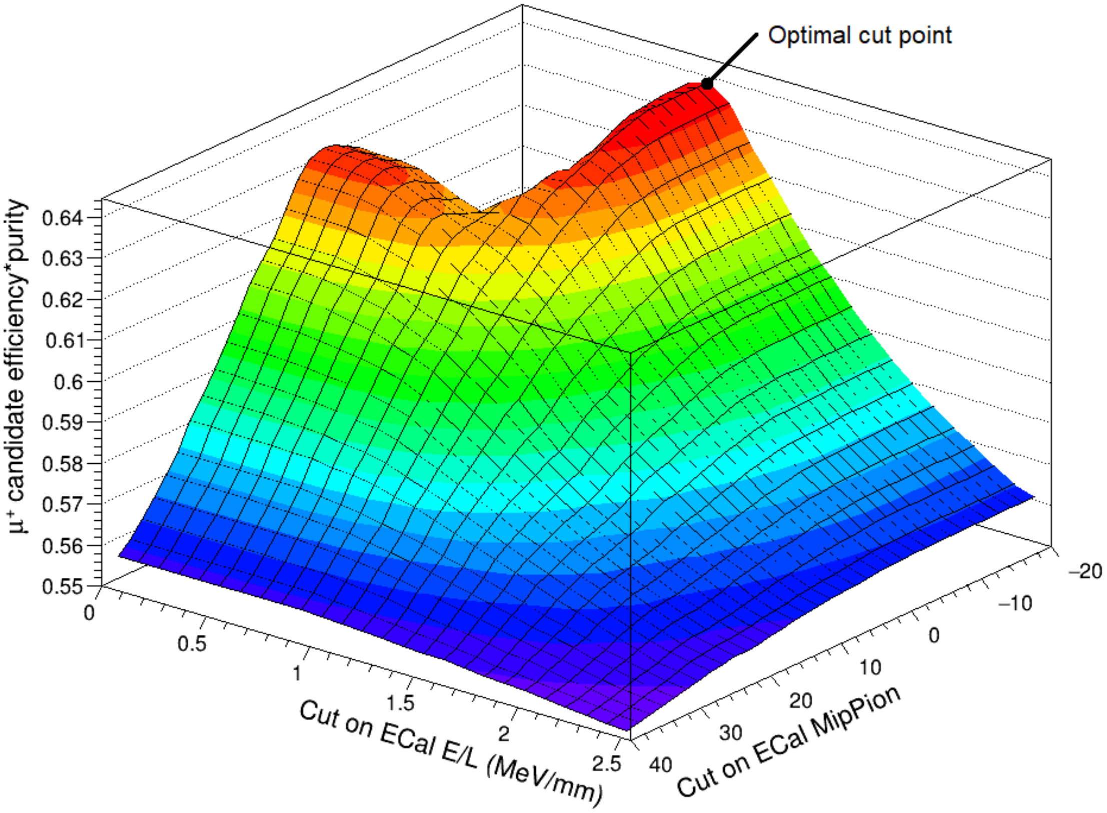


Figure 4.5: Surface plot for optimisation of ECal PID cuts on the μ^+ candidate track. The product of efficiency and purity of the track is shown as a function of cut values on the two variables considered. The purity is defined as the proportion of μ^+ candidates associated with true μ^+, and the efficiency as the proportion of true μ^+ tracks passing the modified selection that go on to pass the ECal cuts. The maximal efficiency*purity is 0.64, for a cut accepting tracks of E/L $\leq$ 0.88 MeV/mm and no cut on MipPion.

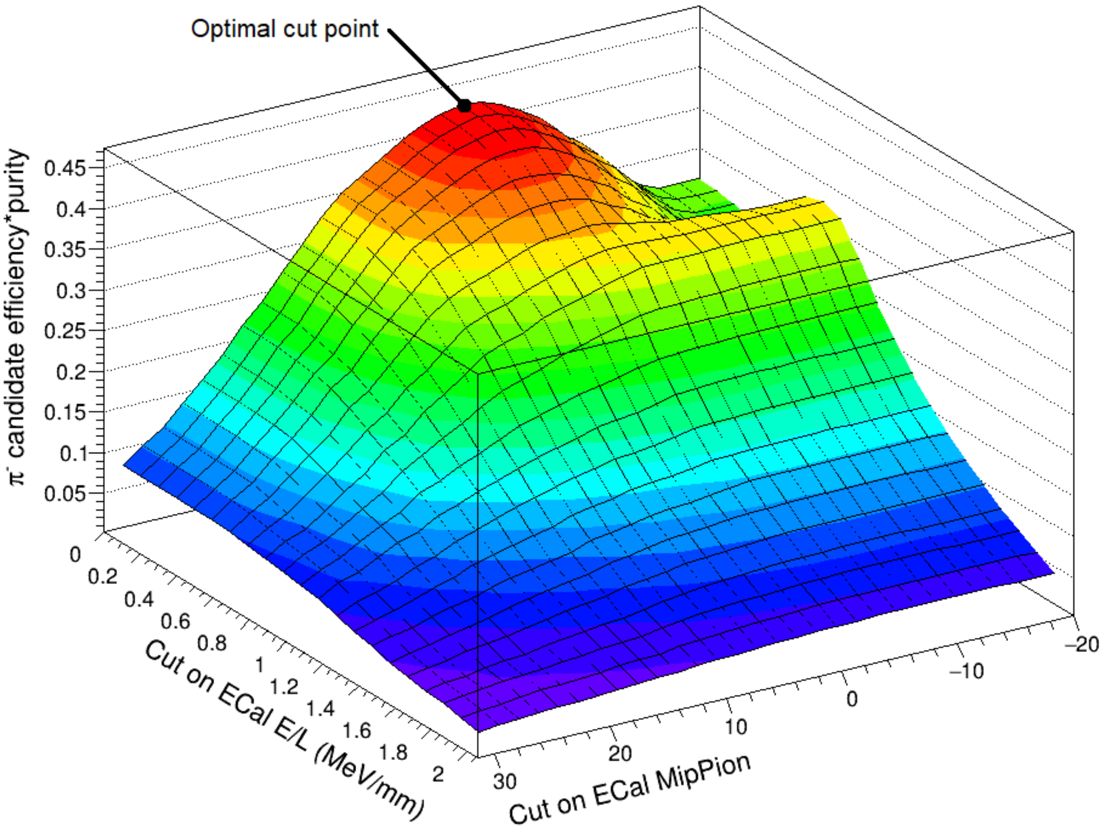


Figure 4.6: Surface plot for optimisation of ECal PID cuts on the π^- candidate track. The product of efficiency and purity of the track is shown as a function of cut values on the two variables considered. The purity is defined as the proportion of π^- candidates associated with true π^-, and the efficiency as the proportion of true π^- tracks passing the modified selection that go on to pass the ECal cuts. The maximal efficiency*purity is 0.43, for a cut accepting tracks of MipPion ≥ 1.0 and no cut on E/L.

The optimisation was performed by a regular grid search (see Section 5.4.1) of cuts on MipPion and E/L. A total of 2500 pairs of cuts (50 values of each variable) were tested, and the efficiency*purity was computed for each point in this grid, as shown in Figures 4.5 and 4.6. The pair of cuts yielding the highest efficiency*purity was deemed to be the optimised PID for use in the selection. Despite the expected complementarity of the two variables, the optimisation favoured a cut on only one variable in each case: E/L for the

μ^+ candidate, and MipPion for the π^- candidate. This suggests that they are in fact too correlated to provide complementary PID cuts in this case. The resulting ECal PID cuts are:

- **Antimuon ECal PID:** If the μ^+ candidate track has one ECal segment associated with it, a cut is applied to the ECal EM energy divided by length (E/L) for that segment, accepting events for which:

  $$E/L \leq 8.8 \text{ MeV/cm}. \tag{4.12}$$

  This cut is waived if zero or multiple ECal segments are associated with the μ^+ candidate.

- **Pion ECal PID:** If the π^- candidate track has one ECal segment associated with it, a cut is applied to the ECal MipPion variable for that segment, accepting events for which:

  $$\text{MipPion} \geq 1.0. \tag{4.13}$$

  This cut is waived if zero or multiple ECal segments are associated with the μ^+ candidate.

4.3.3 Improved selection performance

The combination of the modified selection with the optimised ECal PID is referred to as the 'improved selection'. This begins with a pre-selection similar to that of the existing selection but lacking the leading track cut, comprising the following cuts:

- **Event quality**
- **Total multiplicity**
- **Track quality and fiducial**
- **Upstream background veto**
- **Broken track**

These are followed by the particle identification cuts:

- **Antimuon candidate selection loop**
- **One pion cut**
- **Antimuon ECal PID**
- **Pion ECal PID**

The performance of this selection was tested using the same MC sample and compared to that of the existing and modified selections. The overall performance of the three selections is summarised in Table 4.2, quantified by the significance ($S/\sqrt{S+B}$, where S is the number of signal events selected and B the number of backgrounds), the $\bar{\nu}_\mu$ CC1π^- purity, and the contamination by non-$\bar{\nu}_\mu$-CC backgrounds. A slight improvement over the existing selection and a substantial improvement over the modified selection can be seen in all cases, indicating that the performance of the ECal PID is sufficient to replace the leading track cut. Although the overall improvement in performance is small, the momentum dependence introduced by the leading track cut has been removed. This is demonstrated in Figure 4.7, which shows that events for which the π^- candidate has higher momentum than the μ^+ candidate are now accepted into the selection, and that the ECal PID cuts remove a large proportion of the resulting backgrounds while preserving most of the signal in that kinematic region. However, it should be noted that the contamination by non-$\bar{\nu}_\mu$-CC backgrounds remains large at 33.8% overall, and accepted events are dominated by backgrounds in the momentum region that would otherwise be rejected by the leading track cut.

Selection	**Significance**	$\bar{\nu}_\mu$ **CC1**π^- **purity**	**Non-**$\bar{\nu}_\mu$**-CC backgrounds**
Existing	39.5	47.8%	33.2%
Modified	34.7	30.6%	57.1%
Improved	39.6	47.5%	33.8%

Table 4.2: Summary of performance metrics for the existing, modified and improved selections.

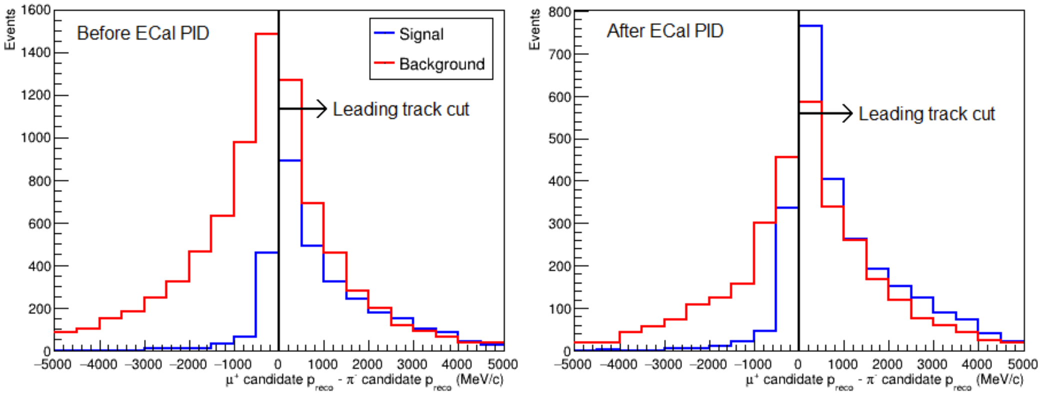


Figure 4.7: Histograms of the difference between the reconstructed momenta of the μ^+ and π^- in the modified (left) and improved (right) selections, i.e. before and after the application of ECal PID. Signal events (true $\bar{\nu}_\mu$ CC1π^-) are shown in blue and background events (all others) in red. The leading track cut from the existing selection is shown in black.

Table 4.3 compares the true particle content of the μ^+ and π^- candidate tracks in the existing, modified and improved selections. The purity of both tracks is substantially increased by the application of ECal PID (moving from the modified to the improved selection). The improved selection also shows similar or better track ID (greater purity and less contamination) compared to the existing selection in most cases, with the exception of π^+ contamination of the μ^+ candidate, which increases from 19.7% to 26.4%, though this is balanced out by the reduction in proton contamination. The effect of the PID changes on track purity can be seen in more detail in Figure 4.8, which shows the true particle content of the μ^+ candidate track as a function of its reconstructed momentum (similarly to Figure 4.2 for the existing selection). The peak in the proton contamination at around 1500 MeV/c is still present but reduced in size; at higher momenta the proton and pion contamination are greatly reduced. At lower momenta however, especially below 1000 MeV/c, the pion contamination is increased compared to the existing selection due to the removal of the leading track cut, resulting in lower μ^+ purity in this momentum region.

Selection	**μ^+ candidate true particle**			**π^- candidate true particle**	
	μ^+	π^+	p	π^-	μ^-
Existing	64.1%	19.7%	13.7%	67.7%	30.1%
Modified	41.6%	43.2%	13.1%	44.1%	54.0%
Improved	63.7%	26.4%	7.9%	70.1%	26.7%

Table 4.3: Summary of the true particle content of the μ^+ and π^- candidate tracks in the existing, modified and improved selections.

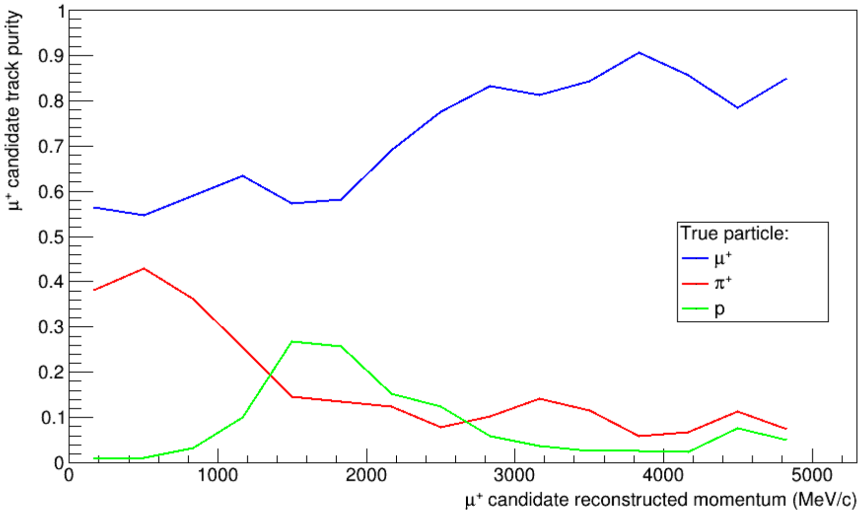


Figure 4.8: MC predicted purities of the μ^+ candidate track in the improved $\bar{\nu}_\mu$ CC1π^- selection, plotted as a function of reconstructed momentum (cf. Figure 4.2). Only the dominant three particle types (antimuons, positive pions and protons) are shown.

Figure 4.9 shows the purity and efficiency of the improved selection as a function of the μ^+ candidate reconstructed momentum and angle. The efficiency (defined with respect to the pre-selection) is largely stable with both variables. The purity exhibits some dependence on the momentum, being lower around 1500 MeV/c as a result of the proton contamination peak; and strong dependence on the angle, decreasing as the angle increases. This may be due to the effect of angle on track length: higher-angle tracks will traverse shorter distances in the TPC before exiting it, resulting in fewer data points for the dE/dx fit and thus poorer PID performance.

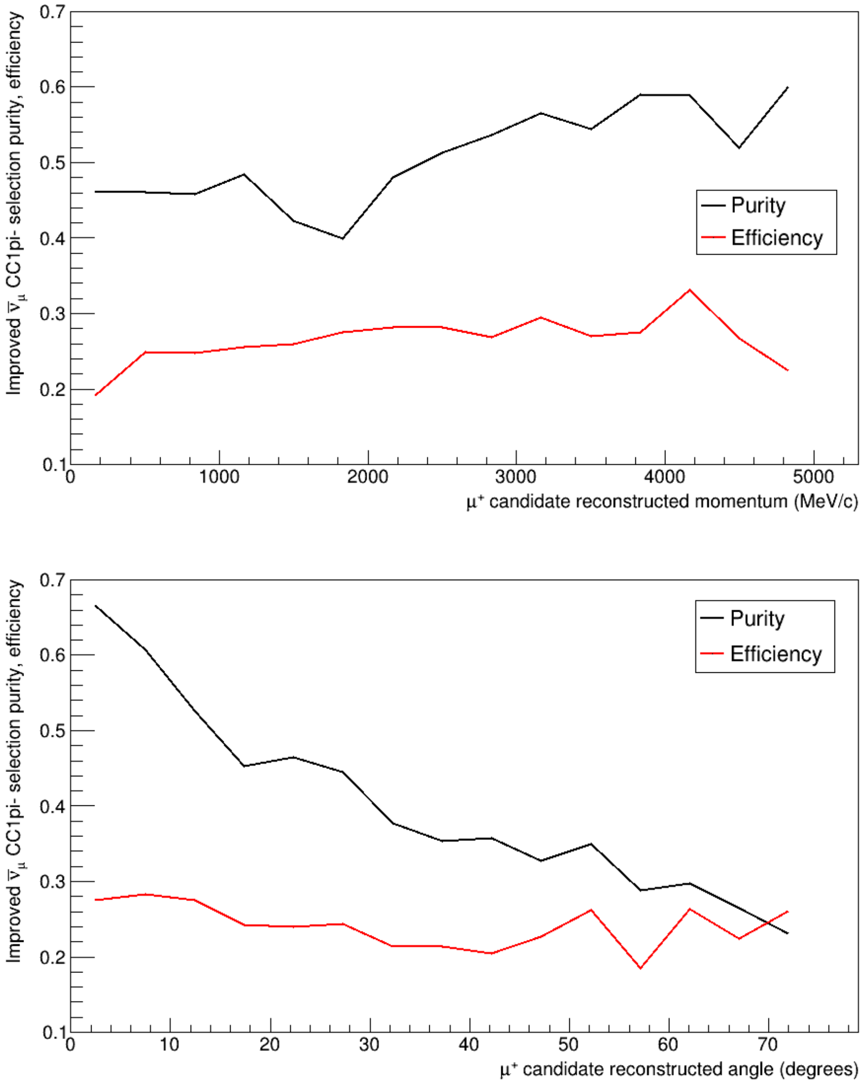


Figure 4.9: MC predicted purity and efficiency of the improved $\bar{\nu}_{\mu}$ CC1π^{-} selection, plotted as a function of the reconstructed momentum (top) and angle (bottom) of the μ^{+} candidate track with respect to the detector Z-axis. The efficiency is defined here as the number of signal events passing the full selection divided by the number of signal events passing the pre-selection.

Overall, these results indicate that the developed ECal PID offers slightly better $\bar{\nu}_{\mu}$

CC1π^- selection performance compared to the leading track cut of the existing selection, while avoiding the momentum dependence of the latter. The overall purity of the sample remains low at 47.5%, with around a third of μ^+ and π^- candidates being misidentified. To increase the selection purity without cutting on the event kinematics, further improvements to the PID are needed. This is indicative of a wider problem for event selections in ND280: particle identification currently relies on rectangular cuts on a small subset of the available PID information. A wide variety of PID variables are available from the various ND280 subdetectors, but rectangular cuts typically perform poorly for large numbers of variables, so the addition of further such cuts is unlikely to yield substantial performance improvements. To make efficient use of all the available PID information associated with a track, taking into account the values and correlations of the different variables, a multivariate approach is needed. The following chapter discusses the principles and methods of multivariate analysis.

Although multivariate PID variables such as the ECal MipPion have been developed, they each only include information from a single subdetector. A 'global' particle ID, combining variables from each subdetector a track crosses, has yet to be developed for ND280. The development of such a global PID tool is described in Chapter 6 and its application to the $\bar{\nu}_\mu$ CC1π^- selection in Chapter 7.

Chapter 5

Multivariate analysis techniques

Data analysis for modern high-energy physics experiments presents significant challenges. Data is the main output of these experiments and represents the result of large investment of funding and person-hours, so it is essential that we make as good use of it as possible. Advancements in readout electronics have enabled the recording of data at extremely high rates, yielding enormous data samples from which event types must be selected and physics quantities extracted. Modern computing technology enables very fast processing of large amounts of data, but requires sophisticated analysis methods to do so effectively [94].

Analysis tasks can be grouped into the categories of *classification* and *regression*. In classification, the task is to assign objects or events to one of a number of discrete classes. Regression is the process of extracting one or more variable parameters; this may involve fitting a known function, or deriving one empirically from the data. This chapter will mainly discuss multivariate analysis (MVA) in the context of classification problems, since they are very important in the HEP context and more relevant to the topic of this thesis. Nonetheless, in both cases functional approximation is the underlying task and many of the same multivariate analysis principles apply.

As more common physics processes become well-understood, we seek to analyse rarer categories of event. These can typically be mimicked by a wide variety of other processes, meaning a small number of signal events must be separated from a large number of backgrounds. The conventional approach to event selection has been to apply cuts on individual variables, but this is rarely optimal, particularly for complicated detectors in which many different variables can be recorded for a single event. Instead, a multivariate approach can be far more effective. This chapter describes the challenges in handling such

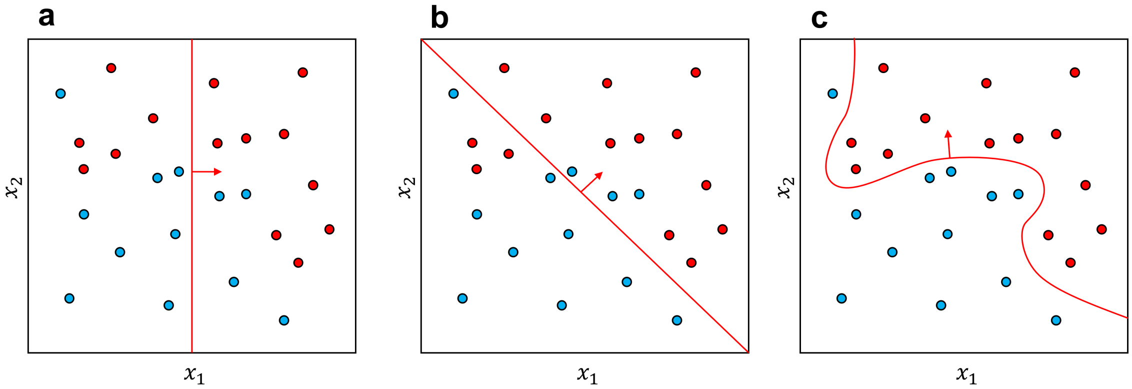


Figure 5.1: Illustration of a classification problem with multivariate data. Points belonging to two classes A and B are shown in red and blue respectively, plotted in two feature variables x_1 and x_2. Three example choices of a decision boundary to select class A are shown. A simple cut on x_1 as shown in **a** does not separate the classes well. A cut on a linear combination of x_1 and x_2 as shown in **b** gives better separation, while a nonlinear boundary as shown in **c** is required to fully resolve the two classes.

multivariate data, the principles involved in multivariate analysis, and a number of the methods that have been developed for this purpose.

5.1 Treatment of multivariate data

Each object or event in a data set is characterised by a number of quantities referred to as feature variables, which in most cases are correlated. In a multivariate treatment, the possible values of the feature variables are considered to form a d-dimensional feature space, in which each object or event is be represented by a vector $\mathbf{x} = (x_1, x_2, ..., x_d)$. Feature variables in a HEP data set may include such quantities as energy deposited in detector elements, track curvatures, reconstructed kinematic variables, etc.

Objects or events of a particular class should occupy specific contiguous regions in the feature space. In order to extract information about a particular class, we must be able to distinguish its members from those of other classes. The objective of a classification analysis is therefore to construct a function $y = f(\mathbf{x})$ that can form a useful decision boundary; that is, one that separates objects or events of one class from those of others. The ideal function y maximises some selection quality criterion. It is rarely feasible to calculate the ideal y

analytically, so in practice we typically attempt to approximate it by $\tilde{y} = f(\mathbf{x}, \mathbf{w})$, where $\mathbf{w}$ are one or more adjustable parameters. Examples of decision boundaries of varying classifying power are illustrated in Figure 5.1.

Before applying more advanced techniques, data may be preprocessed: the selected feature variables may be manipulated by applying a transformation to make them more useful for analysis. The transformation may be a simple scaling of the quantities, or something more sophisticated such as a decorrelation or the construction of more refined physics-motivated variables from a combination of the existing ones. These transformations alone may be sufficient to solve some problems, or may provide a starting point for a more advanced multivariate analysis.

5.2 Machine learning

In the conventional approach to finding the approximating function $\tilde{y} = f(\mathbf{x}, \mathbf{w})$, one chooses a mathematical model and derives its parameters analytically or numerically using an optimisation criterion. To be effective, this requires an appropriate choice of model, whether from a priori knowledge of the function or simply a good guess. Machine learning offers a powerful alternative, since the form of the approximating function can be inferred automatically from the data.

'Machine learning' refers to the use of algorithms that automatically learn from provided data in order to make predictions about future data; in other words, to find the approximating function $f(\mathbf{x}, \mathbf{w})$. While multiple approaches to machine learning exist (e.g. unsupervised learning, reinforcement learning), this text will discuss only the category of supervised learning. In supervised learning, a training data set is provided that contains both the feature vectors that form the input of the desired function, and the target outputs associated with them. In the training phase, the algorithm learns the relationship between inputs and desired outputs from the training data set, thus generating an approximating function f and its optimal parameters $\mathbf{w}$. In the testing phase, the learned function is applied to a testing data set to evaluate its performance. The testing data set should be statistically independent of the training data set, so as to avoid adaptation to statistical fluctuations in the training data. This is known as 'overtraining': the algorithm 'learning' features that do not exist in the true distributions in the feature space will typically lead to reduced performance.

Obtaining an optimal approximating function entails minimising the information loss incurred. This is quantified by a loss function $L\{y, f(\mathbf{x}, \mathbf{w})\}$, the average of which over the training data set is known as the risk $R(\mathbf{w})$. By minimising the risk function, a learning algorithm takes into account mistakes made in predictions and finds the best set of parameters $\mathbf{w}$. Depending on the specific method and problem, it may also be desirable for the optimisation to take into account some constraint, which can be added to the risk function. In this case the algorithm attempts to minimise the resulting cost function:

$$C(\mathbf{w}) = R(\mathbf{w}) + \lambda Q(\mathbf{w}) \tag{5.1}$$

where $Q(\mathbf{w})$ is the constraint to be imposed, and λ an adjustable parameter that determines the strength of the constraint. Constraints are typically used for regularisation, that is, controlling model complexity which if unchecked will result in overtraining. A machine learning algorithm attempts to find the global minimum of the risk (or cost) function in the parameter space, though in practice it is usually only possible to find a local minimum. Nevertheless, machine learning methods are virtually always superior to conventional ones when dealing with multivariate data.

5.3 Bayesian statistics

Many multivariate analysis techniques have their underpinnings in the Bayesian framework of statistical analysis [95]. The Bayesian approach is one of inductive inference: using prior knowledge and new data to update probabilities. The fundamental principle of Bayesian statistics is the Bayes theorem:

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)} \tag{5.2}$$

where $p(B)$ is referred to as the prior probability of B, $p(B|A)$ as the posterior probability, and $p(A|B)$ as the posterior likelihood. First let us take the simple case of a binary classification problem where data must be categorised as one of two classes, e.g. classifying events as either signal s or background b. It is natural to define a decision boundary such that a feature vector $\mathbf{x}$ is classified as signal if $p(s|\mathbf{x}) > p(b|\mathbf{x})$ (the Bayes rule), which minimises the probability of misclassification. Thus the Bayes discriminant is defined:

$$r(\mathbf{x}) = \frac{p(s|\mathbf{x})}{p(b|\mathbf{x})} = \frac{p(\mathbf{x}|s)p(s)}{p(\mathbf{x}|b)p(b)} \tag{5.3}$$

where $p(\mathbf{x}|s)$ and $p(\mathbf{x}|b)$ are the probabilities of obtaining a feature vector $\mathbf{x}$ from a signal and a background event respectively. The classification problem then becomes a task of calculating $r(\mathbf{x})$ or any one-to-one function thereof. With r defined thus, the posterior probability for the signal class can then be expressed

$$p(s|\mathbf{x}) = \frac{p(\mathbf{x}|s)p(s)}{p(\mathbf{x}|s)p(s) + p(\mathbf{x}|b)p(b)} = \frac{r}{1+r} \tag{5.4}$$

with the task then being to estimate $p(\mathbf{x}|s)$ and $p(\mathbf{x}|b)$. Alternatively, with a flexible enough $f(\mathbf{x}, \mathbf{w})$, $p(s|\mathbf{x})$ may be directly approximated. If $p(s)$ and $p(b)$ are not known, but $p(s)/p(b)$ is, the discriminant function

$$D(\mathbf{x}) = \frac{p(\mathbf{x}|s)}{p(\mathbf{x}|s) + p(\mathbf{x}|b)} \tag{5.5}$$

may be calculated. $D(\mathbf{x})$ can then be used to obtain $p(s|\mathbf{x})$ as follows:

$$p(s|\mathbf{x}) = \frac{D(x)}{D(x) + (1 - D(x))/k} \tag{5.6}$$

where $k = p(s)/p(b)$.

For classification into an arbitrary number N of classes C, the Bayes posterior probability for class C_k becomes

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{\sum_{i=1}^{N} p(\mathbf{x}|C_i)p(C_i)} \tag{5.7}$$

and the Bayes rule is to assign the object to the class with the highest posterior probability.

5.4 Multivariate analysis methods

Over the past decades, numerous methods have been developed that can make more efficient use of multivariate data than conventional ones. They vary in complexity and power, and each has its benefits and drawbacks. This section outlines a number of methods that are particularly useful and popular in high-energy physics.

5.4.1 Grid searches

Returning to the typical problem of separating signal events from background events, the conventional approach is to apply a simple cut on each feature variable, selecting events for which $x_1 > z_1, x_2 > z_2, ..., x_d > z_d$ where z are the cut values. This corresponds to a set of hyperplanar decision boundaries in the feature space, oriented parallel to the axes. This method is known as ‘rectangular cuts’, or the ‘cut-based method’.

While the cut values are often chosen through trial and error based on knowledge of the underlying physics of the variables, this will not necessarily yield the optimal set of cuts, especially for larger numbers of variables. Instead, the optimal set of rectangular cut values can be found by a grid search: a systematic search over the feature space in which many candidate sets of cuts are sampled to form a grid. The performance of the cuts defined by each point in the grid is tested to find the best set.

The choice of grid is an important aspect of this approach. Searching over a regular grid, while thorough, is inefficient: many of the points will lie in regions containing low numbers of signal or background points, wasting processing time. More problematic still is the ‘curse of dimensionality’: the number of grid points required M^d grows rapidly with the bin count in each variable M and the number of feature variables d. Fortunately, more efficient options exist. One example is a random grid search, in which the grid points are generated from a random distribution. Other methods such as genetic algorithms and simulated annealing can also be used [96].

Even when fully optimised, rectangular cuts are not a true multivariate method. Each decision boundary is a hyperplane parallel to the feature space axes, so they cannot take into account correlations between the feature variables. A set of rectangular cuts can only be competitive when feature variables exist with excellent separating power, making an MVA unnecessary. Consequently the following multivariate methods will always perform at least as well as, and typically better than, a grid search. However, grid searches are relatively simple and fast to compute, so they can still be useful to compare the separating power of individual variables, or provide a performance benchmark for more sophisticated methods.

5.4.2 Linear methods

One of the simplest improvements that can be made upon rectangular cuts is to allow decision boundaries rotated with respect to the axes, as illustrated by Figure 5.2. To achieve this, one may use a linear combination of the feature variables:

$$\tilde{y}(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + ... = \sum_i w_i x_i = \mathbf{w}\mathbf{x} \tag{5.8}$$

that is, a model where the parameters **w** are a set of weights on the feature variables. Methods for finding the optimal set of weights are known as linear methods.

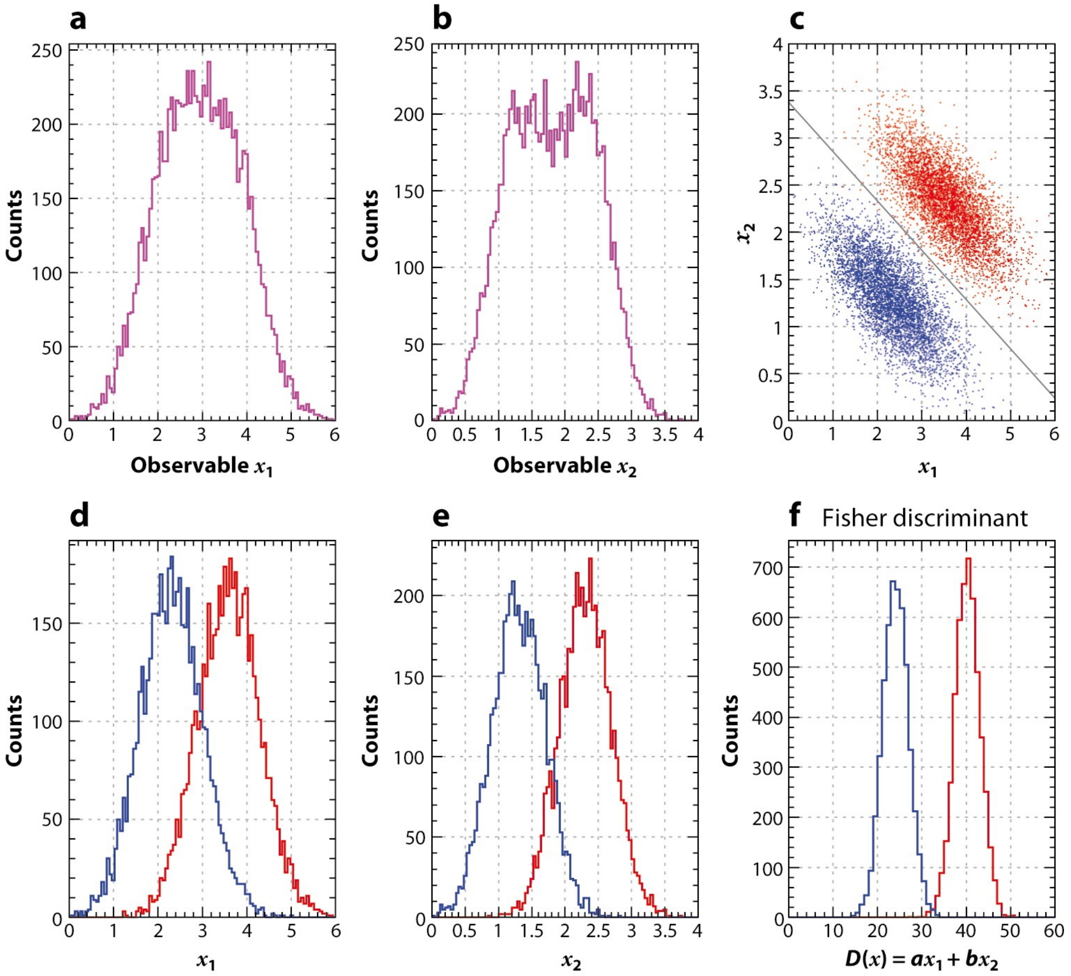


Figure 5.2: Example of a classification problem where linear methods are highly applicable [94]. Two classes (shown in red and blue) have bivariate Gaussian densities in feature variables x_1 and x_2 shown in (c). Rectangular cuts on the one-dimensional densities (d and e) would give poor discrimination between the two classes, whereas a linear discriminant function (f) almost fully separates them. The optimal decision boundary derived from a cut on the linear discriminant is shown in panel c.

The quintessential linear method is Fisher's discriminant [97]. In the Fisher method, one seeks to to find a set of weights **w** that 'pushes' the signal and background classes away from each other (maximising the distance between their means), while 'pulling' the events of a particular class close together (minimising their variances). Thus the Fisher discriminant to be maximised is defined:

$$F(\mathbf{w}) = \frac{(\mu_s - \mu_b)^2}{\sigma_s^2 + \sigma_b^2} \tag{5.9}$$

where μ_s and μ_b are the means of the signal and background classes along the **w** direction, and σ_s and σ_b the corresponding variances. This yields a set of weights such that

$$\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_s - \boldsymbol{\mu}_b) \tag{5.10}$$

where Σ is the common covariance matrix, and $\boldsymbol{\mu}_s$ and $\boldsymbol{\mu}_b$ the means of the signal and background classes in the feature variables.

When a variable has the same mean for signal and background in the sample, the Fisher method cannot discriminate at all, regardless of the shapes of the distributions; this may however be mitigated by applying an appropriate transformation to the input variables. Fisher discriminants perform well when the variables are Gaussian distributed with linear correlations, but are not competitive with more sophisticated methods otherwise [96].

5.4.3 Naïve Bayes classifier

A common approach to finding the Bayes discriminant (Equation 5.5) is to assume that, within each class, the distributions (densities) of the feature variables are independent, and therefore that the multivariate densities can be written as products of one-dimensional densities without loss of information. The Bayes discriminant then becomes

$$D(\mathbf{x}) = \frac{\Pi_i s_i(x_i)}{\Pi_i s_i(x_i) + \Pi_i b_i(x_i)} \tag{5.11}$$

where $s_i(x_i)$ and $b_i(x_i)$ are the densities of the ith variable in the signal and background classes respectively. The task then simplifies to finding the univariate densities, which are generally easy to estimate. This is known as the 'naïve' Bayes classifier [98], since in practice the independence assumption usually does not hold, but classifiers constructed in this way are often competitive with more sophisticated methods nonetheless. It has been found that the performance of a naïve Bayes classifier does not directly depend on how independent the feature variables are, but rather on how much information is lost as a result of the assumption.

5.4.4 Kernel-based methods

Multivariate densities can be estimated by counting the number of events in the data sample in some small regions of the feature space. The most obvious way to do so is to take a d-dimensional histogram, binning the data in M^d bins similarly to the regular grid search; but like the grid search, this suffers from the curse of dimensionality. The granularity of the histogram (i.e. the number of bins M) must be chosen such that the structure of the density is adequately resolved: too few bins and the structure will be washed out, too many and the estimation will be spiky due to statistical noise. With a suitable number of bins, a huge amount of data will be needed to fill them sufficiently.

Instead, it is more efficient to take each individual data point as the centre from which to sample the density in a surrounding small region of the feature space: its 'neighbourhood'. The neighbourhood is defined by a kernel function H, describing the contribution of each other data point to the density estimate for the neighbourhood based on its proximity. The size of the neighbourhood is controlled by a smoothing parameter h known as the 'bandwidth'.

A basic example of a kernel function is a hypercube of side h placed with its centre at the point $\mathbf{x}$: all points that lie within the hypercube contribute equally to the density estimate at $\mathbf{x}$, and those that lie outside do not contribute. This is still similar to the d-dimensional histogram described above and shares many of its shortcomings; it is better to weight the data points based on their proximity to the central point, as this yields smoother and more robust density estimates. This is achieved by using a smooth functional form for the kernel, such as a multivariate Gaussian, in which case the bandwidth h is the Gaussian width.

As with the granularity in the histogram approach, good resolution of the density depends on a good choice of bandwidth. In standard kernel methods the bandwidth is the same for all points, which can result in over-smoothing in high-density regions and spiky estimation in low-density ones. This can be addressed by use of the adaptive kernel method: scaling the bandwidth based on the local density. Instead of the global width h, a local width $h_i = \lambda_i h$ is defined, where λ_i is a scaling factor determined by the number of sample points in the locality. Alternatively, the volume of each kernel can be varied such that it contains some fixed number of points K, and the density at a point $\mathbf{x}$ is then

$$\tilde{p}(\mathbf{x}) = \frac{K}{NV} \tag{5.12}$$

where V is the volume of the kernel and N is the total number of points in the data sample. This is known as the K-nearest neighbour method, and is generally the most robust type of kernel methods.

Kernel-based methods can perform well when the separation between signal and background has irregular features that cannot be easily approximated by parametric learning methods, but tend to perform poorly for larger numbers of input variables [96].

5.4.5 Neural networks

Artificial neural networks (ANNs or simply NNs) are a powerful and popular class of methods inspired by the human brain. An ANN consists of a simulated collection of neurons (nodes) with connections between them, with each neuron receiving signals (input) and producing response (output). Inputs may be received either from external stimuli (data) or the response of other neurons; outputs may be transmitted to other neurons along connections, or form the overall output of the network. The relationship between the input received by a node and the output it transmits is governed by some activation (or transformation) function. An ANN therefore acts as a mapping from a space of input variables $\mathbf{x}$ to a space of output variables $\mathbf{y}$, which is nonlinear provided that at least one neuron has a nonlinear response to its input.

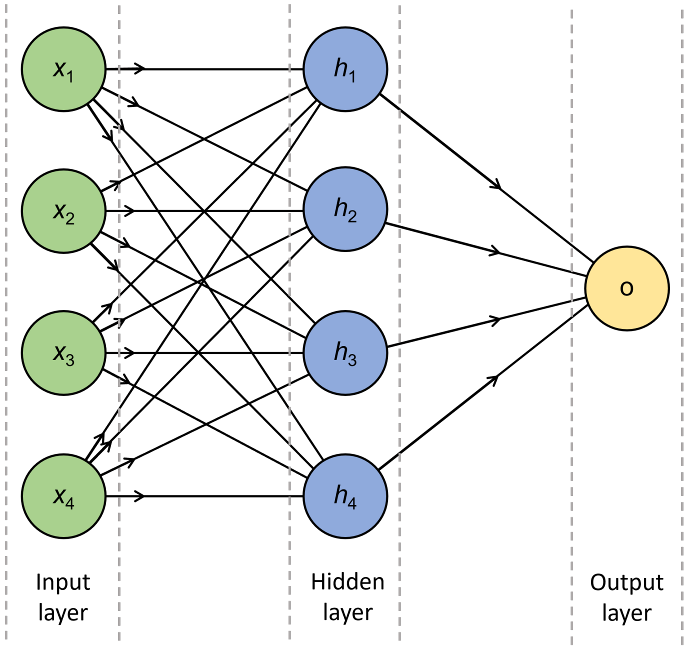


Figure 5.3: Diagram of an example MLP neural network with a single layer of hidden nodes.

Multilayer perceptrons (MLPs), can be considered the basic type of neural network. In an MLP the nodes are arranged in layers, the first of which receives the external inputs (i.e. the feature variables) and outputs them to the next layer. The final layer outputs the final response of the network as a whole. Between these are one or more 'hidden layers': layers that receive information from the preceding layer of nodes, and output to the next. The activation function g may be modified by some bias or threshold parameters θ, and the interconnections between nodes may be characterised by weights w; together these form the parameters of the MLP which are to be optimised in the training phase. The output of a node k is given by

$$O_k = g\left(\theta_k + \sum_j w_{jk} I_j\right) \tag{5.13}$$

where I_j are the inputs to that node and w_{jk} are the interconnection weights on the inputs. Hence we can express the output of a MLP as a nonlinear function. For example, for a MLP as shown in Figure 5.3 with input nodes i, one hidden layer of nodes h, and a single output node o, the response of the hidden nodes is given by

$$O_h = g\left(\theta_h + \sum_i w_{ih} x_i\right) \tag{5.14}$$

and hence the output of the network as a whole is

$$f(\mathbf{x}, \mathbf{w}) = O_o = g\left(\theta_o + \sum_h w_{ho} O_h\right) \tag{5.15}$$

which can model the posterior probability to arbitrary accuracy.

For a given configuration (number of nodes and hidden layers), the MLP must be trained in order to optimise the network parameters and thus ‘learn’ the decision function. A number of algorithms exist for NN training; in each case an error function is minimised iteratively in order to find the optimal set of parameters, typically with hundreds or thousands of iterations. After some number of iterations the error will stop decreasing, or may begin to increase as the NN overfits to the training data: at this point the training should be stopped.

The choice of configuration, or structure, is an important part of finding the ideal MLP for a particular problem. The hidden nodes are critical to the modelling of the function, so the number required for optimal performance depends on the density of the data. If not enough hidden nodes are provided, the flexibility of the network will be too low and result in underfitting; too many, and the flexibility will be too high, leading to overfitting. This can be addressed via structure stabilisation: optimising the size of the network by starting either with a large or a small network and pruning or adding nodes respectively as needed. Alternatively, one may employ regularisation: penalising network complexity by adding a term to the risk function. The network structure must be defined before training (and therefore performance testing) can be done, so each candidate configuration must be trained individually, making the overall optimisation of the network potentially a very slow process. Nonetheless, when completed it will typically result in a powerful classifier.

Various more sophisticated forms of neural network have been developed. A ‘deep’ neural network (DNN) is a NN with several hidden layers and a large number of neurons in each layer. With a sufficiently large training sample to avoid overfitting, such neural networks can learn complex and highly non-linear relations [96]. In convolutional neural networks (CNNs), the input data is treated as an image: instead of connecting every neuron of a particular layer with every neuron of the previous layer, parameters are learned only for a set of small kernels (typically 3×3 or 5×5 squares) at a time, sliding over the input. This reduces the number of learnable parameters and can substantially outperform conventional neural network methods. Generally, such advanced NN methods can outperform MLPs, but will require more work to optimise their configuration for the data.

Rather than attempting to find a single "best" network, the concept of neural networks can be expanded using Bayesian principles to work with the space of possible NN parameters: this is the concept of Bayesian neural networks (BNNs) [99]. A probability density can be assigned to each point $\mathbf{w}$ in the NN parameter space (including the meta-parameters describing the network structure e.g. number of hidden nodes), and used to perform a weighted average over all points (all possible networks). This probability density is given by the Bayes theorem:

$$p(\mathbf{w}|T) = \frac{p(T|\mathbf{w})p(\mathbf{w})}{p(T)} \tag{5.16}$$

where T refers to the training data $T = \{y, \mathbf{x}\}$. The average over the posterior distribution for a given input vector $\mathbf{x}$ then estimates $y(\mathbf{x})$ from the output of all possible networks:

$$\tilde{y}(\mathbf{x}) = \int f(\mathbf{x}, \mathbf{w})p(\mathbf{w}|T)d\mathbf{w} \tag{5.17}$$

where f is the NN output for the input vector $\mathbf{x}$ with parameters $\mathbf{w}$. In practice this is difficult to implement, since the parameter space typically has large dimensionality; the only feasible way to obtain (or rather estimate) the integral is to sample the density $p(\mathbf{w}|T)$ and approximate using the average:

$$\tilde{y}(\mathbf{x}) \approx \frac{1}{K}\sum_{k=1}^{K} f(\mathbf{x}, \mathbf{w}_k) \tag{5.18}$$

where K is the number of points sampled.

BNNs help address the issue of network configuration and overtraining. Since they average over many network configurations, they generally offer a more robust classifier. By assigning lower probability densities to larger networks, the need to limit the number of hidden nodes (and therefore flexibility) is removed, since unnecessarily large networks are effectively pruned away automatically. The main drawback is computing time: the more points to be sampled, and therefore the better the estimate of the integral in Equation 5.17, the more individual NNs will have to be trained.

5.4.6 Decision trees

The decision tree (DT) is a simple but useful algorithm. A decision tree consists of a sequence of cuts with more flexibility than the basic rectangular cuts method described above: the decision flow splits at each cut, forming a node with two branches connecting to further nodes. The decision flow terminates in 'leaf' nodes, which output the classification decision of the tree. This is illustrated in Figure 5.4. The event selections described in Chapter 4 are examples of (non-optimal) decision trees.

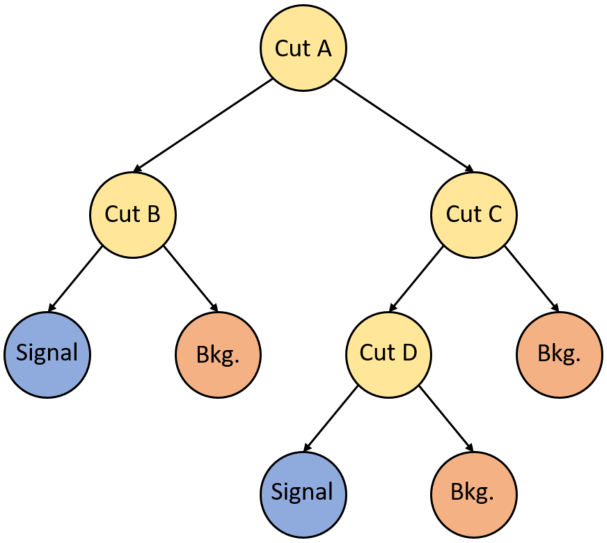


Figure 5.4: Diagram of a simple decision tree of maximum depth 3. Cut nodes are shown in yellow. Leaf nodes are shown in blue and orange for signal and background respectively, representing the output decision of the tree.

To optimise a decision tree, a criterion is defined to quantify the performance of a cut (generally some measure of the reduction of impurity). At each node, the best performing cut on each feature variable is found from the training data; of these, the best cut overall is used to split the data at that node. This is repeated, finding the optimal cut at each node, until some terminal criterion is reached: this may be when the improvement in purity becomes negligible, when insufficient events remain, or when the tree depth (the maximum number of nodes a datapoint may pass through) reaches a predefined maximum. At this point, the splitting stops and the node becomes a leaf with an output response based on the data that have reached it.

This is equivalent to partitioning the feature space into a d-dimensional histogram with bins of varying size (represented by the leaves), and a response value assigned to each bin. In other words, they form a piecewise constant approximation to the function being modelled. As the training data sample becomes arbitrarily large (also the number of bins, albeit slower than the size of the training sample) and the bin sizes arbitrarily small, the predictions of a decision tree approach those of the target function.

Decision trees have many advantages, and are very popular methods as a result. They are simple and transparent to understand and implement; they have high tolerance to missing variables in the training and testing data; they are insensitive to irrelevant variables; and they are invariant to one-to-one transformations of variables, making such preprocessing techniques unnecessary. However, these advantages are offset by several limitations: they are unstable with respect to the training sample, potentially producing a very different tree from slightly different data; predictions are constant within each bin and discontinuous at the boundaries, resulting in suboptimal performance; and as a consequence of recursive splitting, fewer and fewer training data can be used at each branch, increasing the effect of statistical fluctuations and therefore the risk of overtraining for trees with large depth.

Fortunately, the shortcomings of the individual decision tree can be overcome with the use of ensemble learning techniques, which are discussed in the following section.

5.5 Ensemble learning

Many individually weak classifiers can be combined to produce a single much more powerful and robust classifier. This is known as ‘ensemble learning’. Each individual classifier in the ensemble is optimised differently: they may receive different subsets or weightings of the training sample, or have access to different subsets of the feature variables. The output of the ensemble is obtained by a (weighted) average of the outputs of the individual classifiers:

$$\tilde{y}(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m y_m(\mathbf{x}, \mathbf{w}_m) \tag{5.19}$$

where y_m and $\mathbf{w}_m$ are the output and parameters of the mth classifier, and α_m a set of weighting coefficients for the classifiers. The coefficients α_m depend on the ensemble learning algorithm(s) used.

While in principle any classifier can be used as the building block for ensemble methods, decision trees are the most common choice due to their simplicity and flexibility. An ensemble of decision trees is known as a ‘forest’.

5.5.1 Boosting

A natural concept for iteratively training ensemble classifiers is the idea of learning from mistakes: the more poorly previous attempts performed in a particular region, the more subsequent attempts should focus on that region. This is applied in a technique known as ‘boosting’: when optimising each individual classifier, the events in the training sample are assigned weights based on the performance of the previous classifier(s). The worse the performance, the higher the weight; thus the training focuses more and more with each iteration on the events that previous iterations failed to classify.

A boosted ensemble of decision trees is known as a ‘boosted decision tree’ (BDT), a far more robust and powerful classifier than an individual DT. Since boosting works well for weak individual classifiers, the maximum depth of each tree can be kept small with minimal loss of overall performance. This ensures that a good amount of training data will be used at each split, thus reducing the effect of statistical fluctuations on tree structure (improving stability). Limiting tree depth to two or three has been found to almost completely eliminate overtraining [96].

A number of different boosting algorithms exist. In the most popular, known as AdaBoost (adaptive boost) [100], a boost weight α is defined for each tree:

$$\alpha = \ln \frac{1 - r_{err}}{r_{err}} \tag{5.20}$$

where r_{err} is the misclassification rate of the tree. After training each tree, the event weights are updated: the pre-existing weight for each event misclassified by that tree is multiplied by β, and the weights of the entire training sample are renormalised to keep their sum constant. This yields a new set of event weights which are used in training the next tree. When all trees have been trained, the boost weights β determine the weight on the output of their corresponding trees, such that the overall classification is given by:

$$\tilde{y}(\mathbf{x}) = \frac{1}{N} \sum_{i}^{N} \alpha_i y_i(\mathbf{x}) \tag{5.21}$$

where y_i and α_i are the output and boost weight of tree i, and N is the total number of trees in the BDT. The performance of AdaBoost can be further enhanced by enforcing 'slow learning': reducing the learning rate of the BDT and thus allowing a larger number of boost steps, by modifying the boost weight as $\alpha \to \alpha^l$ where $0 < l < 1$.

The underlying loss function in AdaBoost is exponential loss, $L(y, f) = e^{-F(\mathbf{x})y}$, which lacks robustness in the presence of outliers or mislabelled data points. This can be mitigated by modifying the loss function to, for example,

$$L(y, f) = \ln\left(1 + e^{-2f(\mathbf{x})y}\right) \tag{5.22}$$

which is the loss function used in the TMVA *GradientBoost* algorithm [96]. However, obtaining the corresponding boosting algorithm analytically is non-trivial, so instead a steepest-descent approach is used for the loss function minimisation. Like AdaBoost, GradientBoost works best with weak classifiers (i.e. trees of depth 2–4) and thus avoids overtraining, and its learning rate can be reduced to potentially improve accuracy.

Boosted decision trees are sometimes referred to as the best "out of the box" classifiers [96]: like neural networks, they are capable of learning complex relationships, but require comparatively little tuning to obtain good classification results. The NN architecture has many adjustable parameters, and needs to be adapted to specific problems in order to obtain full efficacy and robustness. BDTs on the other hand have fewer adjustable parameters, and generally perform well provided that a sensible configuration is chosen [101]. Neural networks are in principle more powerful than BDTs, and with sufficient tuning work (and recent advancements such as deep learning) they may eventually yield the superior classifier, as demonstrated in [102]. However the example of [102] also shows that the improvement in performance may be small, and can be limited by external factors such as how the data has been processed. Thus, for practical purposes, these gains may not justify the greater investment of development time. It is for these reasons that the multivariate particle ID development described in Chapter 6 used a BDT rather than a neural network.

5.5.2 Bagging

In 'bagging' (bootstrap aggregating), each classifier in an ensemble is trained on a randomly selected subset (bootstrap sample) of the training data. The ensemble classifier output is then the simple average of the outputs of the individual classifiers. While this does not emphasise improving the classification performance directly, it has the effect of smearing over statistical representations of the training data. This stabilises the response with respect to statistical fluctuations.

Bagging can easily be used in combination with other ensemble learning techniques. For example, the GradientBoost algorithm can benefit from the addition of a bagging-like resampling, in which each tree is grown only from a random subsample of the training data; this is known as ‘stochastic gradient boosting’ [96].

5.5.3 Random forests

Rather than selecting the datapoints at random when training an individual classifier, one may instead select a random subset of the feature variables: for decision trees this technique is known as ‘random forests’. In a random forest, the feature variables are randomly sampled at each DT node when choosing the ‘best’ cut. Like bagging, this improves robustness with respect to statistical fluctuations in the training sample [103], and can be added to boosting techniques such as GradientBoost.

5.6 Performance metrics

To evaluate the performance of a classifier and thereby optimise its parameters, it is necessary to choose a performance metric. A performance metric is a figure of merit for the classification output, which is to be maximised to obtain the optimal set of parameters. A great many performance metrics have been defined for classification tasks; these can measure very different things, and consequently evaluating performance with different choices of metric can yield very different results [104]. The suitability of a particular performance metric when developing a classifier depends on the requirements of the intended application, and there is not necessarily an objectively ‘best’ choice of metric for any given problem. Rather, performance metrics are often constructed intuitively to balance multiple desirable features.

Datasets in HEP typically contain large numbers of events of various different types, so a classification algorithm is needed in order to select events of a particular signal process and reject its backgrounds. When developing such an algorithm, it is usual to test its performance using Monte Carlo simulated data. For a given sample size, this provides predictions of the total number of events that will be selected N_{sel}. From the truth information, this can be broken down into signal and background such that $N_{\mathrm{sel}} = S_{\mathrm{sel}} + B_{\mathrm{sel}}$, where S_{sel} and B_{sel} are the numbers of signal and background events selected respectively. The performance metric for this algorithm will typically be a function of these predictions, constructed in such a way as to minimise the expected uncertainty of the intended physics measurement.

One of the most important aspects of an event sample is its purity, that is, the fraction of the selected events that originate from the signal process:

$$P = \frac{S_{\text{sel}}}{N_{\text{sel}}} \tag{5.23}$$

which quantifies the degree to which the sample is contaminated by backgrounds. The more background events selected (and therefore treated as signal in analysis of the sample), the greater the uncertainty of the analysis result, so a high purity is usually desirable. However, increasing the purity requires applying more stringent cuts on the classifier, which is often subject to diminishing returns: at the high end of the purity scale, more and more signal events will be rejected for smaller and smaller purity gains. This may be acceptable if signal events are abundant in the dataset, in which case the purity may be suitable as the (primary) performance metric. But otherwise, we are throwing away large numbers of potentially useful signal events, resulting in a small sample and consequently a large statistical uncertainty.

For this reason, we usually also consider the efficiency, that is, the fraction of the total available signal events that are actually selected:

$$E = \frac{S_{\text{sel}}}{S_{\text{tot}}} \tag{5.24}$$

where S_{tot} is the total number of signal events in the MC dataset. The efficiency on its own is a poor performance metric (no cut at all yields 100% efficiency) but can be combined with other quantities. To balance the different sources of uncertainty, a compromise is needed between accepting more signal (increasing efficiency) and rejecting more backgrounds (increasing purity). This is often done by simply multiplying the efficiency and purity to form a combined quantity EP, which is often used as a performance metric in HEP analyses and usually yields good results. However, this metric has its shortcomings in some situations, since it gives equal weight to the constituent quantities regardless of the size of the resulting sample. When the number of recorded events is small, and particularly when the numbers of signal and background events are similar, maximising EP may select very few events. This will cause a large statistical uncertainty on any measurement made using the sample.

A more statistically-motivated performance metric can be derived by attempting to maximise the significance of the signal, $S/\Delta N$, where ΔN is the uncertainty on the number of selected events. We can estimate ΔN by treating the signal and background as Poisson processes. Assuming that the systematic uncertainty on the number of background events is small compared to the statistical uncertainty, by Poisson statistics the uncertainties on S and B are then $\Delta S = \sqrt{S}$ and $\Delta B = \sqrt{B}$ respectively, and this leads to

$$\Delta N = \sqrt{(\Delta S)^2 + (\Delta B)^2} = \sqrt{(\sqrt{S})^2 + (\sqrt{B})^2} = \sqrt{S + B}. \tag{5.25}$$

The significance then becomes $S/\sqrt{S+B}$, so we can use the performance metric

$$Z = \frac{S_{\mathrm{sel}}}{\sqrt{S_{\mathrm{sel}} + B_{\mathrm{sel}}}} \tag{5.26}$$

or, when $B \gg S$, a simplified version $S/\sqrt{B}$. Z estimates the statistical significance with which we expect to establish the existence of the signal process, given the model used to predict S_{sel}. Maximising the significance is desirable when attempting to establish discovery of a process, so these are commonly-used performance metrics in HEP [105].

The decision of which performance metric to use in a particular context is ultimately subjective, but the properties of the data set and the needs of the experiment should be taken into account. Maximising EP can yield good results when N is large and it is not necessary to maximise discovery significance (e.g. when the goal is to measure a rate parameter rather than establish discovery), whereas Z is generally a better choice for discovery searches and/or when N is small. The ND280 data contains only small numbers of candidate $\bar{\nu}_\mu$ CC1π^- events, resulting in small sample sizes and therefore large statistical uncertainties, so Z was chosen as the overall performance metric for the selection development described in Chapters 4 and 7.

5.7 MVA implementation

Since most of the above methods are complex but broadly applicable, the typical user need not implement them from scratch; instead it is common practice to make use of one of many available software packages. These packages provide tools for the training, testing and application of MVA methods, allowing the user to define the input and configuration while automating the required algorithms. Some are dedicated to a particular method, such as the variants of the neural network. Others are more general and offer many different methods, often with the option to test multiple methods simultaneously and compare their performance.

For the analysis described in this thesis, the TMVA package [96] was an obvious choice due to its flexibility and integration with the ROOT framework on which the Highland analysis software is built. TMVA was used for the training and application of the BDT described in Chapter 6, using the GradientBoost BDT algorithm in multiclassification mode. Since the version of TMVA (ROOT 5.34.34) currently compatible with Highland offers only limited performance evaluation for multiclassification analyses, more rigorous testing was conducted by applying the BDT to a sample of testing events in Highland.

5.7.1 The TMVA package

The TMVA package [96] is a general multivariate analysis package provided as part of the ROOT framework [69], written in C++. It offers object-oriented implementations of a large variety of methods, utilities such as parameter fitting and transformations, and a system of user interfaces for evaluation of input variables and MVA output. Training, testing, performance evaluation, and application can all be automated via the TMVA tools.

The methods implemented in TMVA include (but are not limited to): rectangular cut optimisation, linear and nonlinear discriminants, kernel methods, ANNs, and BDTs. All TMVA methods fall under the category of supervised learning, and thus require both input variables and desired outputs for each event. All methods can be used for binary classification, and in some cases classification into more than two output classes (‘multiclassification’) and/or regression. The implementations are abstract and object-oriented, allowing the user to define input variables and event classes and adjust method configuration options as needed. TMVA analyses are separated into two phases: *training* and *application.* In the training phase, the MVA methods chosen by the user are trained, tested and evaluated. Following this, in the application phase, the trained methods are applied to data to perform the required classification (or regression) task.

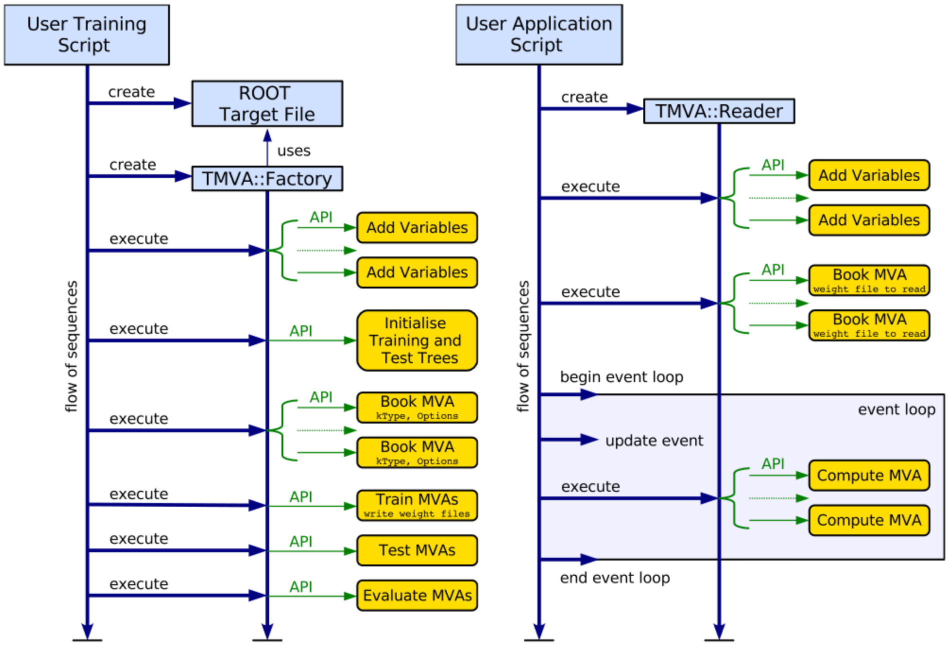


Figure 5.5: Flow (top to bottom) of the typical TMVA training (left) and application (right) sequences [96].

Interaction between the user and the multivariate methods for the training phase is managed by a *Factory* class. This provides member functions to specify the data sets for training and testing, to register the input variables to be considered (alongside any desired transformations/preprocessing), and to ‘book’ the desired MVA methods. The Factory object performs preanalysis of the data sets, computing useful information regarding the input variables such as their correlation coefficients and a preliminary ranking according to their 1D separation power. The requested preprocessing is applied, and the booked MVA methods are trained, tested and evaluated. ‘Weight’ files are created containing the optimised parameters for each method, and various graphical representations of the evaluation information are produced for the user’s perusal. Using this information (and/or any further testing the user may wish to perform by applying the methods themselves) the method configuration options can be tuned; that is, the training can be re-run with different options in order to improve performance.

For the application phase, a *Reader* class is provided. This reads and interprets the weight files of the trained methods. Taking the input variables from an event, it can then compute the MVA response values. This enables the user to apply the chosen MVA method to the data set to be analysed. Alternatively, standalone response classes can be generated for integration into any C++ application without dependencies on ROOT or the weight files.

5.7.2 BDT configuration

The BDT implementation offered by TMVA was used for the PID development described in Chapter 6. Since there are multiple particle types to identify, it made sense to use a multiclassification BDT, for which GradientBoost was the only algorithm available. For BDT training a separation criterion must be defined; that is, the performance metric to be maximised at each tree node. For multiclassification GradientBoost, the only option available was the Gini index $p \cdot (1-p)$ where p is the purity obtained by the node cut. Other parameters for BDT training include:

- nCuts: the number of grid points used to find the optimal cut in node splitting. A high value provides high granularity.

- MaxDepth: the maximum allowed depth of each decision tree. This determines the extent to which each individual decision tree can adapt to the training data. Thus greater values of MaxDepth can be expected to yield greater performance, but setting this too high may lead to overtraining. For this reason, the TMVA documentation [96] recommends values between 2 and 4.

- MinNodeSize: the minimum percentage of training events required in a leaf node. This is implemented to avoid overtraining.

- NTrees: the number of trees in the forest. More trees enable more learning, but increase training and evaluation time.

- Shrinkage: reduction in the algorithm learning rate. Smaller values can improve the accuracy of the prediction in some settings, but require more trees [96].

- BaggedSampleFraction: the size of the bagged event sample relative to the full training sample (when using bagging). Bagging does not aim to improve the performance of the BDT directly, but rather its stability with respect to statistical fluctuations. TMVA recommends values between 0.5 and 0.8.

- UseNVars: determines the size of the subset of input variables used at each node split (when using random forests). This may be a fixed number or the mean of a Poisson distribution for random number generation. Similarly to bagging, random forests aims to improve stability with respect to statistical fluctuations.

Different values of MaxDepth, NTrees, Shrinkage, BaggedSampleFraction and UseNVars were tested in order to maximise performance. Other parameters were not tuned: nCuts was kept fixed at 2000 in order to provide adequate granularity (provided that the value chosen is high enough to adequately separate the training data in each input variable, this parameter is unlikely to benefit from tuning) and MinNodeSize was kept fixed at 5% to avoid overfitting. The BDT parameter tuning process is described in Section 6.5.1.

5.7.3 Highland integration

Developing the BDT required interfacing the tools provided by TMVA with the Highland analysis framework used by ND280. The input events for the TMVA training phase had to be extracted from the output of a Highland analysis package processing Monte Carlo simulated data. Following training, testing and evaluation within TMVA, the trained BDTs had to be applied within Highland in order to test performance with various event samples for a number of specific particle identification tasks.

The training dataset was generated from Monte Carlo simulations of particles traversing the ND280 detector, which were processed using the ND280 reconstruction algorithms and then a Highland analysis package. The Highland package applied a sequence of pre-selection cuts to select appropriate reconstructed tracks, and saved various event and track variables to file. These included the reconstructed variables identified as candidates for the MVA input variables, and the true particle identities of the tracks which defined the desired PID outputs. With the pre-selection applied and the required variables saved, the output was then used as input to a TMVA training macro.

The trained BDTs were applied within the Highland analysis package by instantiating a Reader object and reading in the TMVA-produced weight files. This enabled the Highland package to compute the BDT response as part of the event selection and analysis, and use it to apply cuts to the testing data. The performance of these cuts for particle ID tasks was used to evaluate the various BDT configurations tested. The details of this work are described in the following chapter.

Chapter 6

Particle ID development with boosted decision trees

As demonstrated in Chapter 4, the ND280 $\bar{\nu}_\mu$ CC1π^- event selection suffers from low purity, due in large part to the poor performance of its track PID in distinguishing muons, pions and protons. Although the addition of further rectangular cuts has been shown to help mitigate this, the improvement in performance is limited. This is part of a wider issue for ND280 event selections: as T2K physics analyses begin to consider rarer processes (such as $\bar{\nu}_\mu$ CC1π^-) for which statistics are low, selection efficiency becomes more of an issue, so PID performance must be as high as possible. Unfortunately, the conventional approach does not yield sufficiently high-performing PID, so a move towards multivariate methods is needed.

A large number of complementary PID variables are derived from the ND280 subdetector data, and may have correlations with other quantities such as the particle kinematics and each other, but the conventional approach to track identification has been to apply cuts on a small number of individual PID variables. This cut-based method cannot efficiently take into account all of these variables and their correlations, whereas a multivariate approach does. Additionally, the conventional PID cuts are not consistent: different variables and cut values have been developed for different individual selections. This motivates the development of a single 'global' PID for tracks in ND280, combining information from all subdetectors such that it can be used consistently to identify tracks in any event selection. This chapter presents the development of such a PID using multivariate methods.

6.1 Charged particle ID in ND280 event selections

Reconstruction of charged particle tracks is one of the core functions of ND280, and identifying them is an essential part of any event selection. Charged particles traversing the detector deposit ionisation energy in its sensitive volumes, which is recorded and used to reconstruct a track. The spatial and temporal distribution of this energy depends on the particle species and the subdetector, and this forms the basis of PID methods. The commonly-occurring charged particle species that need to be distinguished are muons/antimuons ($\mu^{\pm}$), charged pions ($\pi^{\pm}$), electrons/positrons ($e^{\pm}$), and protons (p).

Tracks selected for analysis generally originate from a vertex in one of the FGDs. The particle may be absorbed or decay before leaving the FGD (this is known as an ‘FGD-iso’ track), in which case PID can only be performed via FGD information. Otherwise the track will go on to traverse other subdetectors. Only the TPCs can measure a particle’s momentum, so a reconstructed TPC segment is often a requirement for non-FGD-iso tracks in event selections. As described in Section 4.2.1, the rate of energy deposit recorded in the TPC is compared to that expected for the μ, π, p and e hypotheses to construct the TPC PID variables. Similarly, for FGD-iso tracks, FGD PID variables are constructed by comparing the energy deposit and track length to that expected for the μ, π and p hypotheses as described in Section 4.2.1.

6.1.1 ECal PID

Particle identification in the ECal is based on the shape of the charge cluster, which differs according to whether or not a shower is produced and which interaction (strong or electromagnetic) dominates in the shower. Three topologies are defined: MIP-like tracks (referred to simply as ‘tracks’), electromagnetic (EM) showers, and hadronic showers [92].

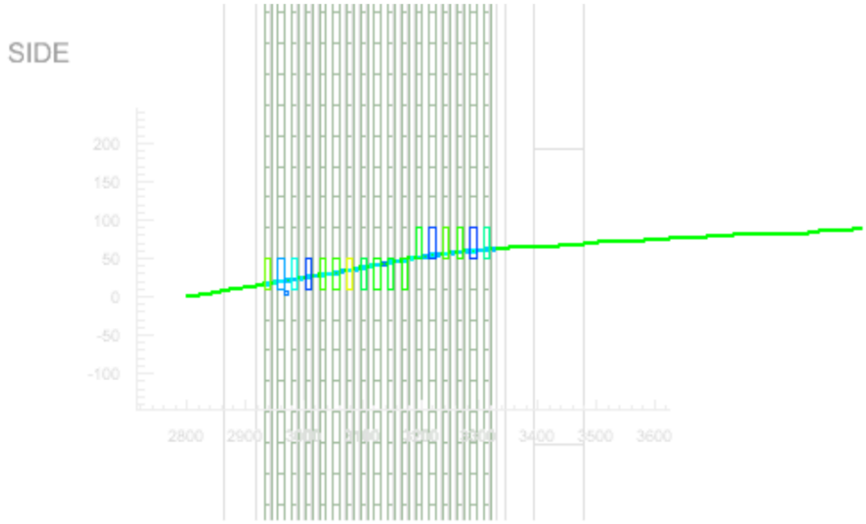


Figure 6.1: Event display 2D side view of a MIP-like track produced in the DS-Ecal by a muon [92]. The green line shows the true path of the muon. The coloured boxes represent the hits recorded in each bar, with the colour indicating the measured charge.

Tracks are produced by particles that behave as MIPs: they do not shower, and deposit energy along their path according to a Landau distribution. This results in a very narrow track with a relatively small spread in charge deposits. Low-energy MIPs may stop in the ECal, but tracks will otherwise tend to span the full depth of the module. An example of a MIP-like track reconstructed in the ECal is shown in Figure 6.1.

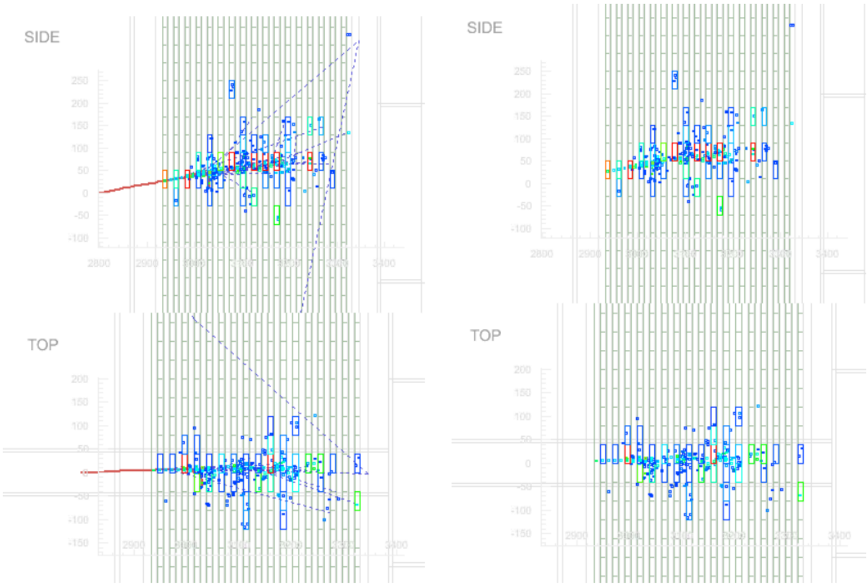


Figure 6.2: Event display 2D views of an EM shower produced in the DS-Ecal by a medium energy electron, with track information (left) and without (right) [92]. The red line shows the true path of the electron. The coloured points represent the hits generated by the simulation, and the coloured boxes the hits recorded in each bar, with the colour indicating the hit charge.

Electromagnetic showers are produced by electrons/positrons and photons. Although the granularity of the ECal does not provide good resolution of the characteristic cone shape of a typical EM shower, they can be identified by a charge cluster with a large width compared to length and high variation in hit charges. EM showers tend to be centred on the inner part of the ECal since the particle will usually shower immediately upon contact with the ECal lead layers. An example of an EM shower cluster reconstructed in the ECal is shown in Figure 6.2.

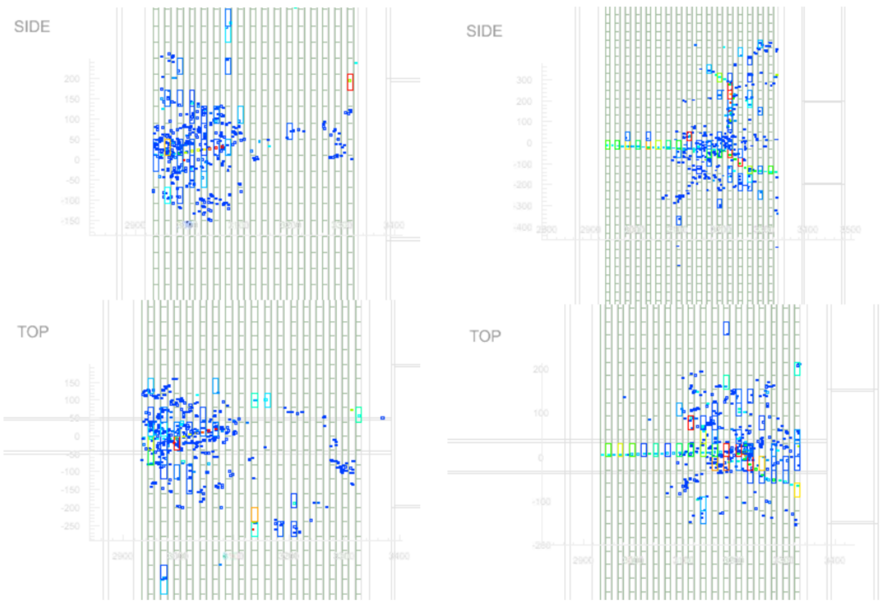


Figure 6.3: Event display 2D views of hadronic showers produced in the DS-Ecal [92]. The shower on the left converts soon after entering the detector, while the shower on the right passes through much of the ECal before converting. The coloured points represent the hits generated by the simulation, and the coloured boxes the hits recorded in each bar, with the colour indicating the hit charge.

Compared to tracks, hadronic showers are very similar to EM showers. They too are typified by a charge cluster with large width compared to length and high variation in hit charges, but do exhibit some differences by which they may be distinguishable. Hadronic showers tend to have a more spherical shape than EM showers, although this difference is subtle at the ECal resolution. Additionally, hadronic showering particles (particularly charged pions) may pass through multiple lead layers before they shower, resulting in a short MIP-like section and a cluster centred further from the inner edge. Examples of hadronic shower clusters reconstructed in the ECal are shown in Figure 6.3.

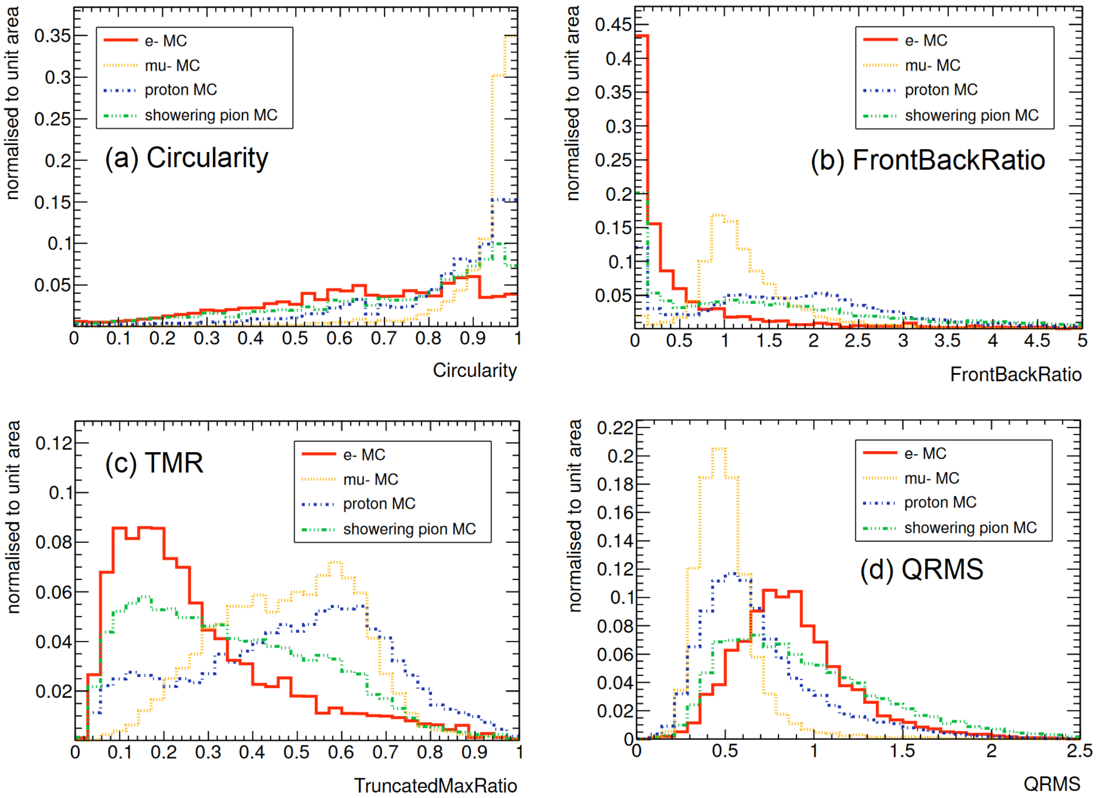


Figure 6.4: Distributions of the low-level ECal variables for particles entering the DsECal in the T2K neutrino beam, simulated by particle gun MC reweighted to the kinematic distribution predicted by the full neutrino event MC [93].

The existing high-level PID variables developed for the ECal (MipEm, EmHip, MipPion) are designed to distinguish between MIP-like, EM-shower-like, hadronic-shower-like and showering pion-like charge distributions. They are derived from the low-level variables, which quantify different aspects of the charge distribution and are designed to be dimensionless and insensitive to the overall charge of the cluster [93]. The low-level variables are as follows:

- **Circularity:** a measure of the transverse spread of the charge distribution relative to its length. It is calculated from a combination of results from 2D principle component analysis (PCA) of hits in both 2D views, weighted by their measured charge. The circularity in each view i is a function of the 2nd principal component, and the combined circularity variable is the product of the circularities in the x and y views.

This distinguishes between track-like (long and thin) clusters which peak near 1, and shower-like (short and wide) clusters that produce lower values.

- **FrontBackRatio:** compares charge deposited at either end of a track. The cluster is divided along its principal axis into four quarters of equal length. The FrontBackRatio is defined as the ratio of the total charge in the back and front quarters. This distinguishes between MIP-like tracks which deposit charge uniformly with distance and thus peak around 1, stopping tracks which deposit more charge in a Bragg peak at the end of the track and thus produce values greater than one, and EM showers which deposit more energy towards the front end and thus produce values less than one.

- **TruncatedMaxRatio:** compares per-layer charge deposits, exploiting the differences in the longitudinal charge profiles of showering and MIP-like particles. The hits are ordered by charge and the top 20% and bottom 20% are removed to truncate the distribution and thus mitigate sensitivity to noise and very high hit charges. Then the charge deposited in each layer of the ECal is calculated, and the TruncatedMaxRatio is defined as the ratio of the lowest and highest per-layer charge deposits. This distinguishes between MIP-like tracks which have more uniform per-layer charge deposit and thus produce values closer to 1, and showering and stopping particles which have greater variation in hit charge and thus produce values closer to zero.

- **QRMS:** derived from the variance of the hit charge distribution to provide distinction between MIPs which deposit charge uniformly, and other track types which do not. QRMS is defined

$$q_{\text{RMS}} = \frac{1}{\bar{q}}\sqrt{\sum_{i}^{N}\frac{(q_i - \bar{q})^2}{N}} \tag{6.1}$$

  where q_i is the charge of hit i, $\bar{q}$ the mean hit charge, and N the number of hits within the cluster. Showering particles tend to have larger QRMS than non-showering ones.

Figure 6.4 shows distributions of the low-level ECal variables. The high-level ECal variables are constructed as log-likelihood ratios (LLRs) derived from the low-level variables, that is,

$$\ln|\lambda(x)| = \ln|P(\mathbf{x}|H_0)| - \ln|P(\mathbf{x}|H_1) \tag{6.2}$$

where $\lambda(x)$ is the likelihood ratio, $\mathbf{x}$ are the low-level variables, H_0 and H_1 the hypotheses considered, and P the probability density functions (PDFs). The PDFs are factorised by assuming the input variables are independent:

$$P(\mathbf{x}|H_0) = \prod_i^N P(x_i|H_0). \tag{6.3}$$

These PDFs are estimated by histograms generated from particle gun MC for four hypotheses: MIPs modelled with muons, EM showers modelled with electrons, showering pions modelled with stopping charged pions, and highly ionising particles (HIPs) modelled with stopping protons. The high-level ECal variables are LLRs comparing three pairs of these hypotheses:

- **MipEm:** MIP hypothesis vs EM shower hypothesis.
- **MipPion:** MIP hypothesis vs showering pion hypothesis.
- **EmHip:** EM shower hypothesis vs HIP hypothesis.

Distributions of these variables for a particle gun sample can be found in Section 6.4.4 below. The LLRs are the main form of ECal PID currently used in ND280 event selections (if any), but in some cases a comparison of the energy measured in the ECal and the momentum measured in the TPC is used. The calculation of the total EM energy of a cluster (E_{EM}) assumes an EM shower. When this assumption is correct, the relationship between E_{EM} and p_{reco} should be approximately one-to-one. In the case that the assumption is wrong (i.e. for MIP-like tracks or hadronic showers) this relationship will not hold, so comparing E_{EM} and p_{reco} can provide additional information as to whether the ECal segment is EM shower-like or otherwise. This is done by cutting on their ratio, E_{EM}/p_{reco}, also referred to simply as E/p (shown in Figure 6.19c below).

6.1.2 Current usage

PID implementations differ between event selections. As outlined in Section 4.2.1, the $\bar{\nu}_\mu$ charged-current selection takes the highest-momentum positive track as the antimuon candidate and applies PID in the form of an optimised cut on the LLRs ($L_\mu > 0.1$,

$L_{\mathrm{MIP}} > 0.9$ if $p < 500$ MeV/c) from the TPC immediately downstream of the vertex. Secondary tracks are then identified as π, p or e by comparing the corresponding TPC likelihoods as described in Section 4.2.1. The ν_μ CC selection PID is similar, but uses a different optimised cut ($L_\mu > 0.05$, $L_{\mathrm{MIP}} > 0.8$ if $p < 500$ MeV/c) to identify the muon candidate, since different rates of signal and background events are expected between the ν_μ and $\bar{\nu}_\mu$ cases. As discussed in Section 4.2.2, this fully TPC-dependent approach fails to distinguish $\mu^\pm$ from $\pi^\pm$, and also μ^+ from protons at certain momenta.

The $\nu_e/\bar{\nu}_e$ selections, on the other hand, use a much more complicated cut flow for PID. These selections have been developed with much more stringent background rejection, since ν_e events are much rarer than ν_μ and have very large backgrounds by comparison. Rather than rely entirely on information from a single TPC, PID is applied to $e^\pm$ candidates using information from TPC2, TPC3 and the ECal. Additionally, the cut flow splits depending on the presence or absence of ECal information: tracks with an ECal segment have ECal cuts applied, and those without an ECal segment are subject to tighter cuts on the TPC pulls. It also splits based on the momentum, in order to take into account the momentum-dependent performance of certain variables. The $\nu_e/\bar{\nu}_e$ selection PID cut flow for primary electron/positron candidates is as follows:

- **Initial TPC2 electron pull cut:** accept if $-2.0 < \delta_e < 2.5$.
- **TPC2 pull cuts (tracks without ECal segments):** accept if $-1.0 < \delta_e < 2.0$, reject if $-2.5 < \delta_\mu < 2.5$ or $-2.5 < \delta_\pi < 2.5$.
- **TPC2 pull cut (tracks with ECal segments and < 36 TPC hits)**: reject if $-2.5 < \delta_\pi < 2.5$.
- **ECal PID cuts (tracks with ECal segments):** split according to momentum. If $p > 1000$ MeV/c, then accept if the ECal EM energy $E > 1100$ MeV. If $p \leq 1000$ MeV/c, then accept if ECal MipEm > 0.
- **TPC3 pull cut:** differs between the ν_e and $\bar{\nu}_e$ selections. For ν_e, reject if $-2.5 < \delta_\mu < 2.5$. For $\bar{\nu}_e$, if $600 < p < 1650$ MeV/c and the track has no ECal segment, then accept if $-3.0 < \delta_e < 3.0$.
- **ECal proton rejection ($\bar{\nu}_e$ selection only):** only applied if $p > 600$ MeV/c and the track has an ECal segment. Reject if ECal EmHip > 0. If $p < 1650$ MeV/c, then reject if $E/p < 0.65$. If $p \geq 1650$ MeV/c, then reject if $E/p < 0.15$.

This PID process makes use of information from multiple subdetectors, and adapts to the momentum of the track and the presence or absence of an ECal segment. However, it is still a simple decision tree and so would likely be outperformed by a MVA approach that could fully take into account the correlations between kinematic and subdetector PID variables. The existing ν_e and $\bar{\nu}_e$ selections have been found to suffer from relatively low efficiency ($\sim 36\%$ for both). Although the main background is photons, which do not produce TPC tracks, these selections also suffer from proton and pion backgrounds and require stringent muon rejection [106]. Improvements to $e^{\pm}$ identification could increase the signal efficiency and/or reduce p, $\pi^{\pm}$ and $\mu^{\pm}$ contamination, improving the performance of these selections.

6.2 Towards a global PID with boosted decision trees

Improving PID performance in ND280 requires the development of new tools to make more efficient use of the available information. The goal of the work presented in this chapter was to develop a PID tool fulfilling three main criteria:

- **Global**: Information from each subdetector crossed by the track should be considered in the PID decision-making.
- **Multivariate**: The PID should use multivariate analysis methods to make efficient use of the many variables available.
- **Versatile**: The PID tool should be generally applicable to tracks in ND280, rather than designed for one event selection in particular, and use methods that can be easily extended to include new input variables.

As described in Chapter 5, multivariate classification methods enable the user to combine many input variables into a single powerful classifier. To fulfil the ‘global’ criterion, a number of candidate input variables were identified, drawing from each subdetector that a forward-going FGD1-originated track may cross: the FGDs, TPCs, ECals, and SMRDs. The variables considered include existing PID variables from the TPCs and ECals, as well as other quantities known to have some separating power such as the ECal E/L. Since the behaviour of many of the considered variables depends on the kinematics of the track, kinematic variables were also included as inputs. The candidate input variables are discussed in Section 6.4.

Of the many multivariate methods available, boosted decision trees were chosen in order to guarantee a good baseline of classification performance. BDTs generally work well ‘out of the box’, offering high classification power with little tuning required. Although a deep neural network might yield greater performance, the tuning required to optimise it would likely be substantially more time-consuming than the relatively minimal tuning required by a BDT. Considering the project’s time constraints, a BDT was judged to be a safer choice than a NN. Conventional ND280 PID processes are themselves decision trees, so a BDT can be seen as a natural evolution (via ensemble learning) of the existing methods. The use of a BDT also contributes to the ‘versatility’ of the tool: new input variables can easily be added to train a new version of the BDT. Given their characteristic stability, a BDT can be expected to immediately exploit the performance improvements offered by new variables.

To fulfil the ‘versatility’ requirement that the tool be generally applicable, the desired output must be considered. ND280 PID needs to distinguish between four main species of charged particles: $\mu^{\pm}$, $\pi^{\pm}$, $e^{\pm}$, p. This can be achieved by using a multiclassification BDT: one that classifies objects into multiple categories (as opposed to binary classification, which considers only signal and background). The multiclassification BDT algorithm offered by TMVA achieves this by outputting a value between 1 and 0 for each category, representing the estimated probability that the object is a true member of that category. The outputs must therefore sum to 1, and the classification decision of the BDT is represented by the category with the largest output value. The categories chosen for the PID tool were μ-like, π-like, p-like and e-like, since the candidate input variables should have minimal sensitivity to the charge of the track.

6.3 Training and testing samples

Training and testing the BDT required a sample of reconstructed tracks. Tracks reconstructed from Monte Carlo simulations are the natural choice for this purpose, since their true particle identities are known. Although these tracks could have been taken from the ND280 neutrino interaction event MC samples, the decision to include kinematic variables among the inputs would introduce model bias, since the kinematics of neutrino interaction products depend strongly on the interaction model. To avoid this, ‘particle gun’ MC was used instead to generate dedicated training and testing samples for the BDT, with flat prior distributions of momentum and angle. A particle gun (PG) generator enables the generation of events containing individual simulated particles with user-defined identity, starting position, direction and energy. Thus the distributions of energy and direction, as well as the relative numbers of each particle type, can be chosen so as to avoid any bias from neutrino interaction models or specific event selection use cases.

Subsamples of each positive particle type (μ^+, π^+, e^+, p) were produced, each comprising 5×10^5 particle gun events. Each subsample was further divided in half to yield statistically independent training and testing samples each comprising 2.5×10^5 events per particle type. The particle gun was configured to generate simulated events with the following distributions:

- Starting position: uniformly distributed throughout FGD1.
- Momentum: uniformly distributed between limits of 150 MeV/c and 2000 MeV/c.
- Direction: uniformly distributed in a cone of opening angle 65°.

The limits of these distributions were chosen to contain a wide range of possible FGD1-originating tracks, while avoiding devoting processing time to kinematic regions where low statistics are expected. Since the simulated distributions are uniform, the lower the expected statistics in a kinematic region, the more inefficient it is to simulate tracks in that region.

The simulated data generated from the particle gun events was processed using the ND280 software, with the same method used for full neutrino event MC. The propagation of each particle and the detector response were modelled with Geant4. The reconstruction algorithms were applied, and the resulting reconstructed tracks analysed in Highland to apply pre-selection cuts and extract the variables to be inputted into the BDT. The procedure used to select tracks suitable for BDT training and testing is as follows. First, for consistency with the usual ND280 analysis pre-selection, the following cuts (as defined in Section 4.2) are applied:

- **Event quality**
- **Total multiplicity**
- **Track quality and fiducial**
- **Upstream background veto**
- **Broken track**

Events for which the true particle type associated with the reconstructed track is not the same as the original generated by the particle gun are cut from the training sample. This is done to avoid training on events where a secondary particle of different type to the original is produced in the FGD and reconstructed as the main track. Additionally, cuts are applied to the reconstructed momentum and angle of the tracks to remove those close to the bounds of the particle gun distributions. Due to detector resolution and other factors, there is a degree of 'smearing' between the true and reconstructed values of momentum (Figure 6.5) and angle. For this reason I define a region of validity for the particle gun sample in each of these values, in order to avoid training or application of the BDT on tracks with kinematics in the smeared regions. The validity region is defined by a momentum range of $200 < p_{\mathrm{reco}} < 1500$ MeV/c and angle $\theta_{\mathrm{reco}} < 60°$. Tracks outside the validity region are cut from the training and testing samples. The distributions of true and reconstructed momentum for each subsample following the kinematic cuts are shown in Figure 6.6.

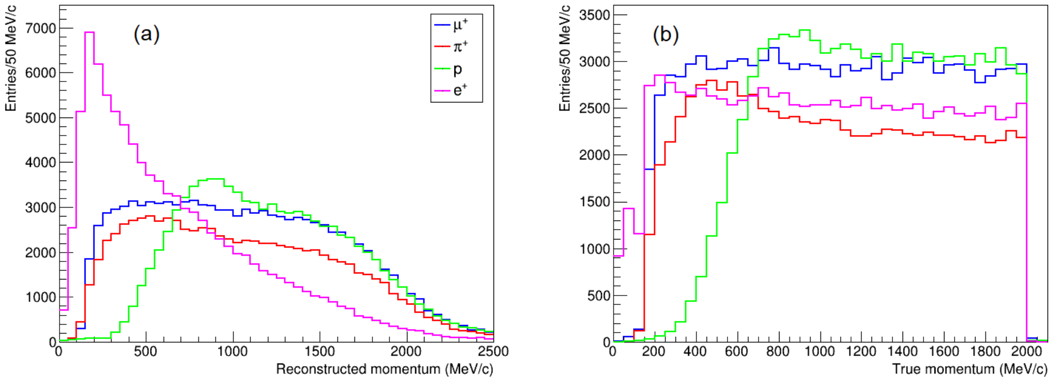


Figure 6.5: Distributions of reconstructed (a) and true (b) momentum in the original particle gun training sample for each of the true particle types following the pre-selection cuts. The kinematic cuts are not applied. The smearing effect can be seen at both ends of the reconstructed momentum spectra.

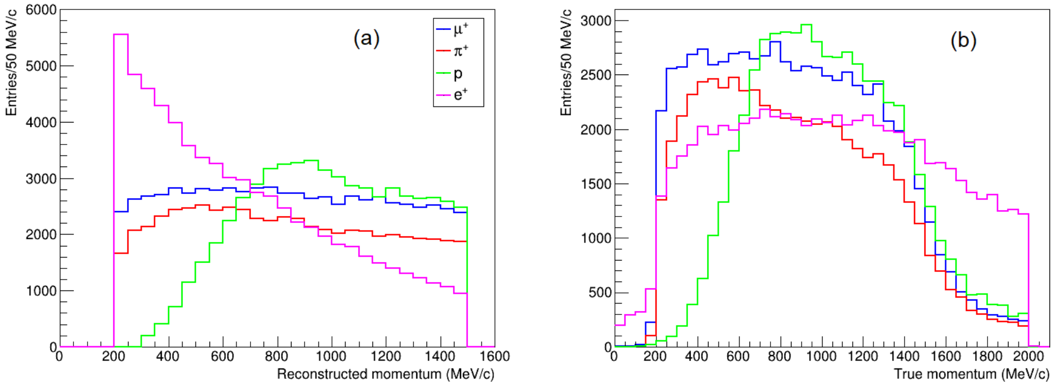


Figure 6.6: Distributions of reconstructed (a) and true (b) momentum in the original particle gun training sample for each of the true particle types following the pre-selection and kinematic cuts. Note the drop-off in proton statistics at low momenta.

Particle gun events passing these cuts are considered valid for input into the BDT. Half of the overall sample is designated as the training sample to be used for developing the BDT in TMVA: this is split in half again between the training and TMVA's built-in testing functions. The other half of the overall sample is used as the particle gun testing sample to evaluate the BDT's performance outside of TMVA.

6.3.1 Event weighting and momentum sensitivity

For the BDT training, an overall weighting is applied to the four particle type subsamples to equalise their total sizes. Additionally, it was found that the identical particle gun momentum distributions do not yield similarly identical distributions in reconstructed momentum, as can be seen in Figure 6.6. In particular, there is a drop-off in proton statistics at low momenta (around 500 MeV/c and below), most likely due to detector acceptance (lower-momentum protons are more likely to be absorbed before reaching the TPC); and the momenta of positrons are skewed heavily towards lower values, possibly due to brehmsstrahlung energy loss which is not accounted for in the momentum reconstruction. If the BDT has access to these momentum distributions its outputs will be directly momentum-dependent, which is undesirable: the momentum should be used only to better interpret the values of other variables. To avoid this, the events are weighted so as to make each subsample's momentum distribution appear uniform and identical in the training. These weights are calculated from histograms of the reconstructed momenta for each subsample:

$$W_{preco}(i) = \frac{1}{N(i)} \tag{6.4}$$

where $W_{preco}(i)$ is the weight for events in reconstructed momentum bin i, and $N(i)$ is the total number of events in that bin.

A further problem arises as a result of this weighting, however. As Figure 6.6 shows, the proton subsample exhibits a sharp drop-off in statistics at low momenta (around 500 MeV/c and below); this is due to the detector threshold. In this region, the lower the momentum of a proton, the less likely it is to escape the FGD and produce a track. Under the weighting scheme in Equation 6.4 this causes very large weights to be given to proton track events in the low-momentum region, which will result in large statistical fluctuations from these events. To avoid this, additional low-momentum proton tracks were generated and added to the training sample in order to 'pad out' the low momentum region, resulting in the distributions shown in Figure 6.7. Even with these additional events, proton statistics were still very small at the lowest momenta, so a cut was applied to exclude protons below 300 MeV/c from the training sample. Due to the detector threshold we would not expect to identify tracks below 300 MeV/c as protons anyway, so the lack of proton training data in this region should not be an issue. This is the only significant difference between the subsample momentum distributions in the weighted training sample. The angular distributions are not weighted, since they are judged to be sufficiently similar between subsamples (see Figure 6.8). The total numbers of training and testing events for each subsample following the above cuts and 'padding' are listed in Table 6.1.

Subsample	Training events	Testing events
μ^+	68976	69031
π^+	56005	58424
p	63759	57848
e^+	67418	67350

Table 6.1: Numbers of events selected for the training and testing samples for each particle type.

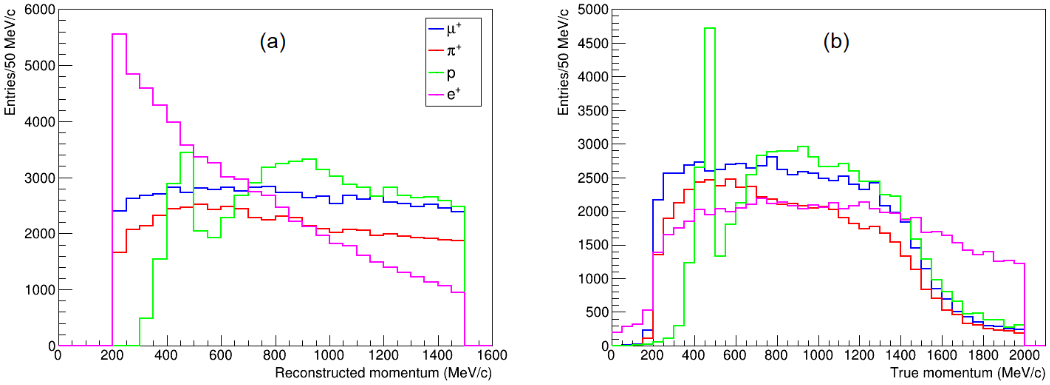


Figure 6.7: Distributions of reconstructed (a) and true (b) momentum in the 'padded' particle gun training sample for each of the true particle types following the pre-selection and kinematic cuts.

6.4 Candidate BDT input variables

A number of candidate input variables were considered based on their known or expected discriminating power for particle ID, drawing from each of the relevant subdetectors: the FGDs, TPCs, ECal and SMRD. These variables are extracted from the particle gun MC Highland analysis files following the application of the above cuts. The rationale for including each candidate input variable is given in this section, along with their distributions in each of the training subsamples as 'seen' by the BDT (i.e. with the above cuts and weightings applied). In some cases, multiple representations of the same information are available, such as the TPC energy loss which may be represented by the dE/dx truncated mean, pulls, or likelihoods. In these cases, each representation was considered for the BDT input, and the relative performance yielded by the different representations is explored in

Section 6.5.2.

In the case of variables from FGD2, TPC3 and the ECal, a reconstructed segment in these subdetectors is not required by the selection. When such a segment does not exist no value exists for the related variables, but the BDT must be provided with a value nonetheless. This is handled by assigning a default value, chosen in each case to be outside the normal range for the variable. Additionally, some variables may be assigned very large positive or negative values by the reconstruction in a small number of cases. This is undesirable since the BDT uses a set number of bins to determine split points; such large outlier values will effectively reduce the granularity of the variable distributions as seen by the BDT. For this reason, a limited range is chosen for each variable, with outliers placed in overflow/underflow bins. The default and overflow/underflow values are not shown in the histograms presented in this section.

6.4.1 Kinematic variables

Many PID variables are affected by the kinematics of the particle: correlations may exist between a PID variable and one or more kinematic variables, and these correlations may differ between different particle types. One can therefore expect the BDT to perform better if it can take the relevant kinematic variables into account in its decision-making. For this reason, the reconstructed momentum measured in the TPCs (p_{reco}) and the starting angle of the track with respect to the detector Z-axis (θ_{reco}) were considered as candidate input variables. As described in Section 6.3.1, the training sample was constructed with flat priors in momentum and angle; the four subsamples were produced and weighted to have approximately identical distributions in p_{reco} and θ_{reco}. The BDT therefore should not cut on the prior distributions of these variables, only on the correlations between them and the other inputs.

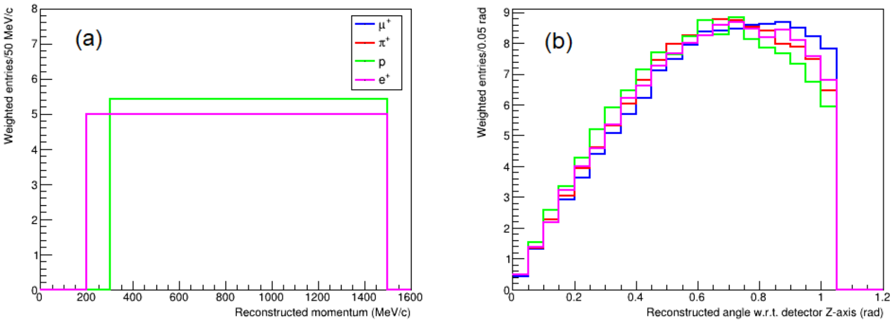


Figure 6.8: Distributions of the reconstructed momentum (a) and angle (b) with respect to the detector Z-axis in the Monte Carlo training sample for each of the true particle types. The distributions are shown weighted by the reconstructed momentum. The number of bins used here for the momentum is a multiple of that used for the weighting histogram, so the distributions appear perfectly flat, and the μ^+ and π^+ momentum distributions are identical to that of $e^\pm$ (and are therefore obscured). The proton momentum distribution appears different due to the higher threshold.

The momentum of a particle is crucial to interpreting its rate of energy loss, so variables derived from the dE/dx (the truncated means in the TPC, and the E/L in the FGDs and ECals) should be more useful to the BDT in conjunction with p_{reco}. Additionally, the separating power of a variable can depend on the momentum: for example, the dE/dx curves for different particles overlap in certain momentum regions, so dE/dx-based PID will perform poorly for distinguishing those particles in those regions. With access to p_{reco}, the BDT could prioritise other input variables (e.g. ECal information) in those regions. The reconstructed momentum is therefore expected to be a very useful input variable for the BDT.

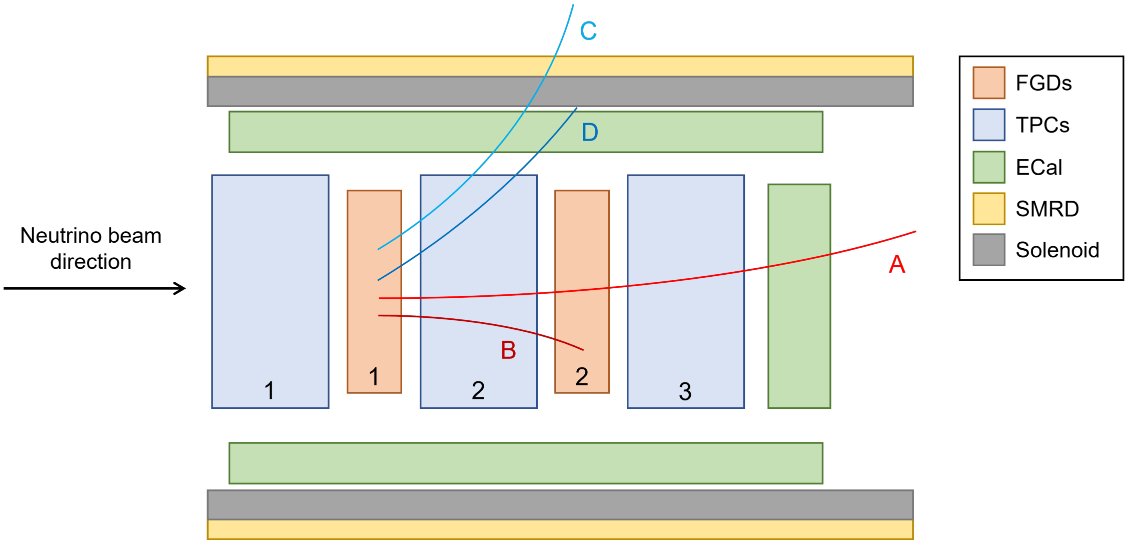


Figure 6.9: Diagram of ND280 YZ side view with examples of charged particle trajectories. The subdetector sensitive volumes are shown: FGDs (orange), TPCs (blue), ECal (green), and SMRD (yellow). The solenoid material between the ECal and SMRD is shown in grey. Regions of the detector upstream of TPC1 are not shown. Four example particles are shown, illustrating different directions of travel and stopping behaviour. As required by the pre-selection, all originate in FGD1 and enter TPC2. Particle A is forward-going, having large momentum and small angle with respect to the detector Z-axis, and so passes through FGD2, TPC3 and the DS-ECal. Particle B is also forward-going but stops in FGD2. Particle C has large angle, and thus moves upward through the barrel-ECal, traversing the magnet and SMRD. Particle D is also high-angle but stops in the solenoid.

As illustrated in Figure 6.9, the direction in which the particle travels will also have an effect on certain variables; this can be simply represented by the reconstructed angle θ_{reco} with respect to the ND280 Z-axis, as measured at the beginning of the track. Only forward-going tracks (small θ_{reco}) can reach FGD2, and only high-angle tracks (large θ_{reco}) can reach the SMRD; thus θ_{reco} can indicate whether a segment in each of these subdetectors should be expected based on the particle's expected range, which differs according to particle type and the material traversed (for example, $\mu^{\pm}$ are the most likely to traverse the dense material of the ND280 magnet and reach the SMRD). The weighted distributions of p_{reco} and θ_{reco} in the training sample are shown in Figure 6.8.

6.4.2 TPC particle ID

The TPC energy loss PID variables described in Section 4.2.1 were included among the candidate inputs. These offer good separation between MIPs, protons and electrons; but as discussed above, they cannot distinguish between muons and pions and have momentum-dependent performance. The pulls and likelihoods contain a built-in comparison to the expected energy loss for each particle type, but the BDT may be able to infer these relationships from those in the MC provided that it has access to the momentum, so each representation is included. These variables are extracted from the TPC2 segment and, if one exists, the TPC3 segment as well. From TPC2, the dE/dx truncated mean and the derived pulls and likelihoods are all included as candidate inputs. From TPC3, the truncated mean and the pulls are included (since the likelihoods are not calculated for TPC3 from an FGD1 vertex). As part of the selection, a track quality cut (requiring 19 or more reconstructed nodes) is applied to the TPC2 segment, but no such cut is applied for the TPC3 segment, so the information quality of the TPC3 variables may be poor in some cases. To test the effect of this, two variants of the TPC3 variables were considered: with and without the track quality check applied. For the versions with the quality check applied, if the segment has 18 nodes or fewer, the reconstructed value is replaced with the default value. These 'good quality' versions of the TPC3 variables are denoted by 'GQ'. The distributions of the TPC2 and TPC3 PID variables in the weighted training sample are shown in Figures 6.10–6.14.

Additionally, the number of TPC segments associated with the track, here referred to as 'nTPCs', is equivalent to the presence/absence of a TPC3 segment (since a TPC2 segment is guaranteed) and shows some separation between particle types, so was also included. The distributions of nTPCs are shown in Figure 6.15.

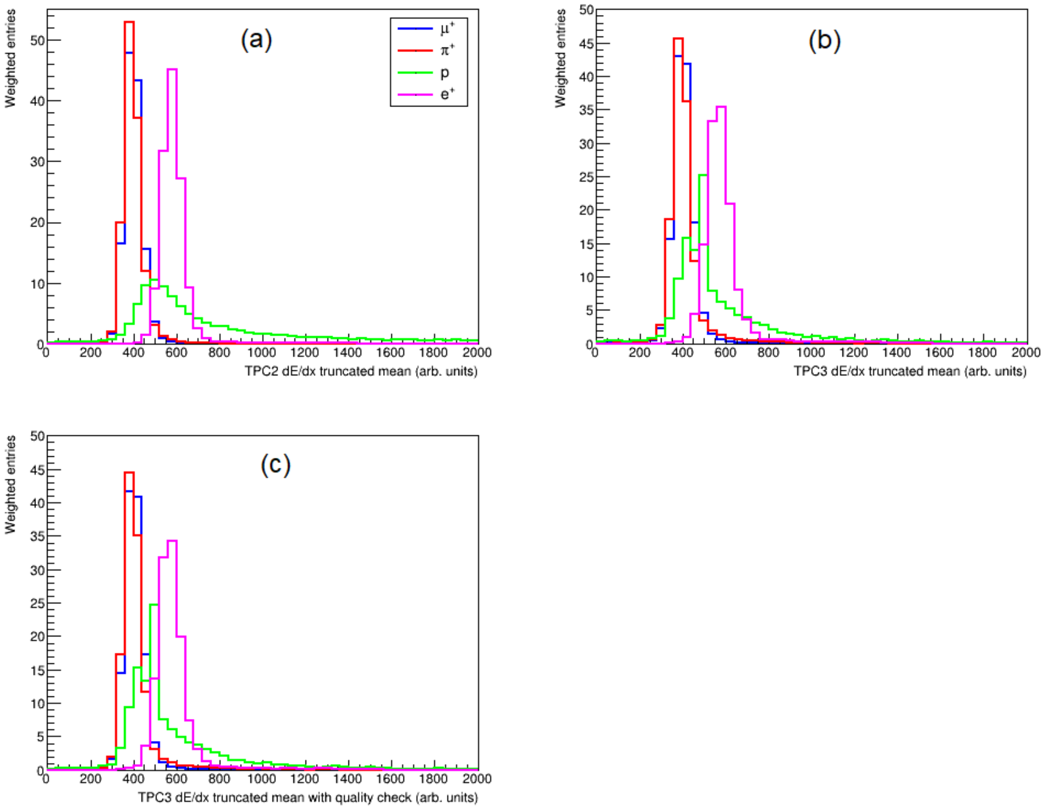


Figure 6.10: Distributions of the dE/dx truncated mean recorded by TPC2 (a) and TPC3 (b, c) in the Monte Carlo training sample for each of the true particle types. The distributions are shown weighted by reconstructed momentum and subsample size. In the case of TPC3, the version with the data quality check is shown (c) in addition to the version without (b).

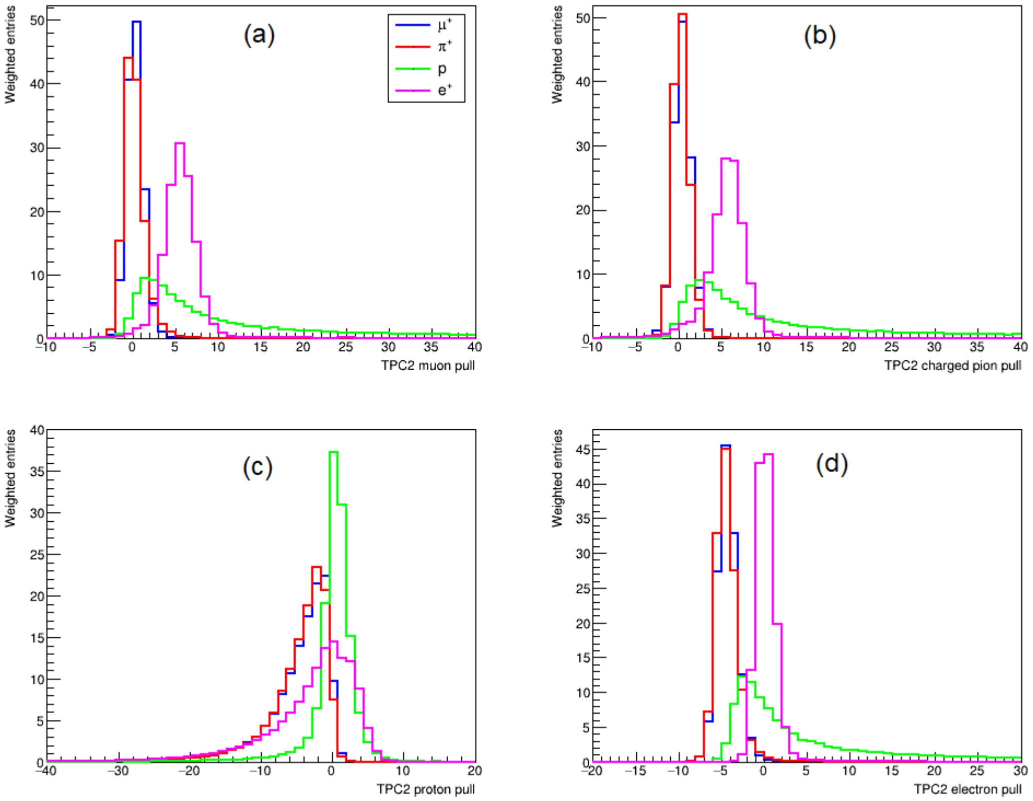


Figure 6.11: Distributions of the TPC2 PID pulls in the Monte Carlo training sample for each of the true particle types. The muon-like (a), charged pion-like (b), proton-like (c) and electron-like (d) pulls are shown. The distributions are shown weighted by reconstructed momentum and subsample size.

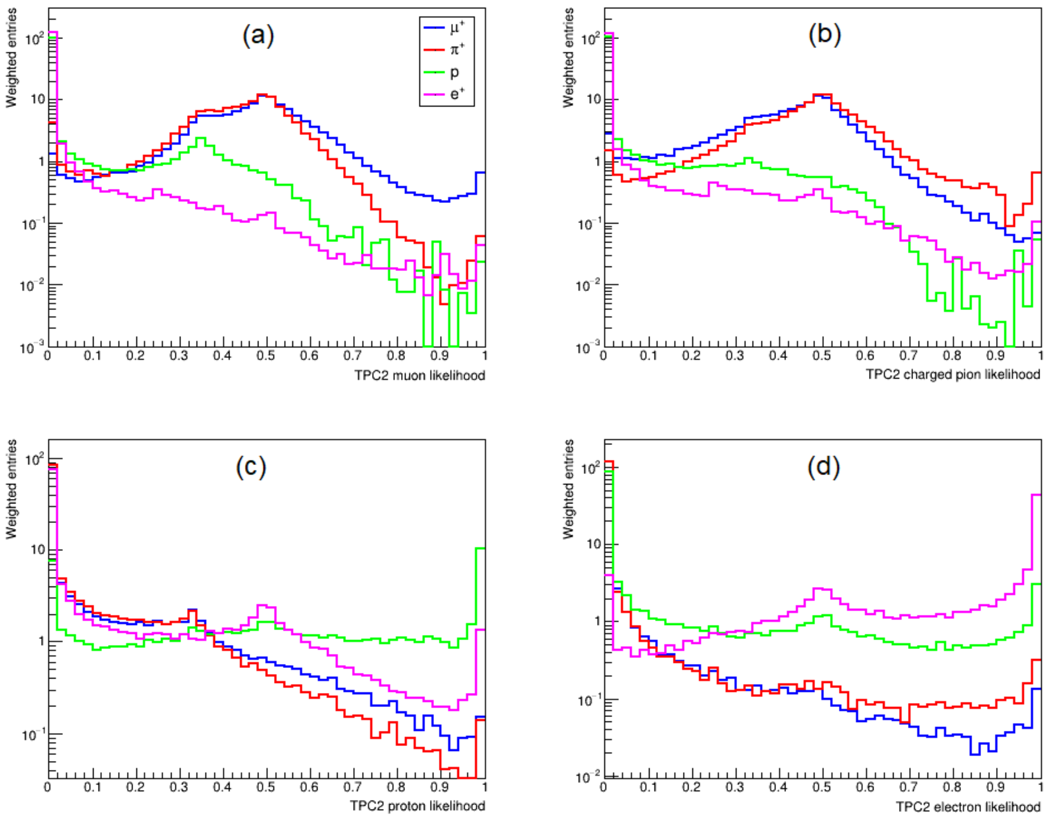


Figure 6.12: Distributions of the TPC2 likelihood variables in the Monte Carlo training sample for each of the true particle types. The muon-like (a), charged pion-like (b), proton-like (c) and electron-like (d) likelihoods are shown. The distributions are shown weighted by reconstructed momentum and subsample size.

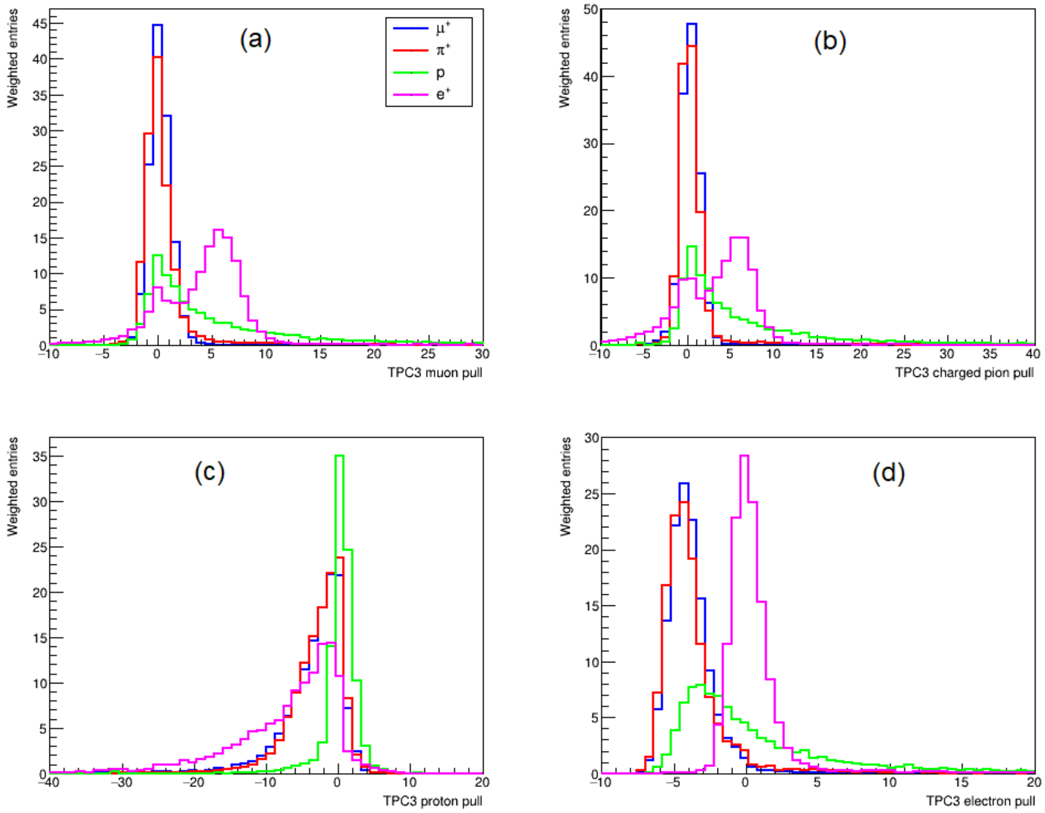


Figure 6.13: Distributions of the TPC3 PID pulls (without the track quality check) in the Monte Carlo training sample for each of the true particle types. The muon-like (a), charged pion-like (b), proton-like (c) and electron-like (d) pulls are shown. The distributions are shown weighted by reconstructed momentum and subsample size.

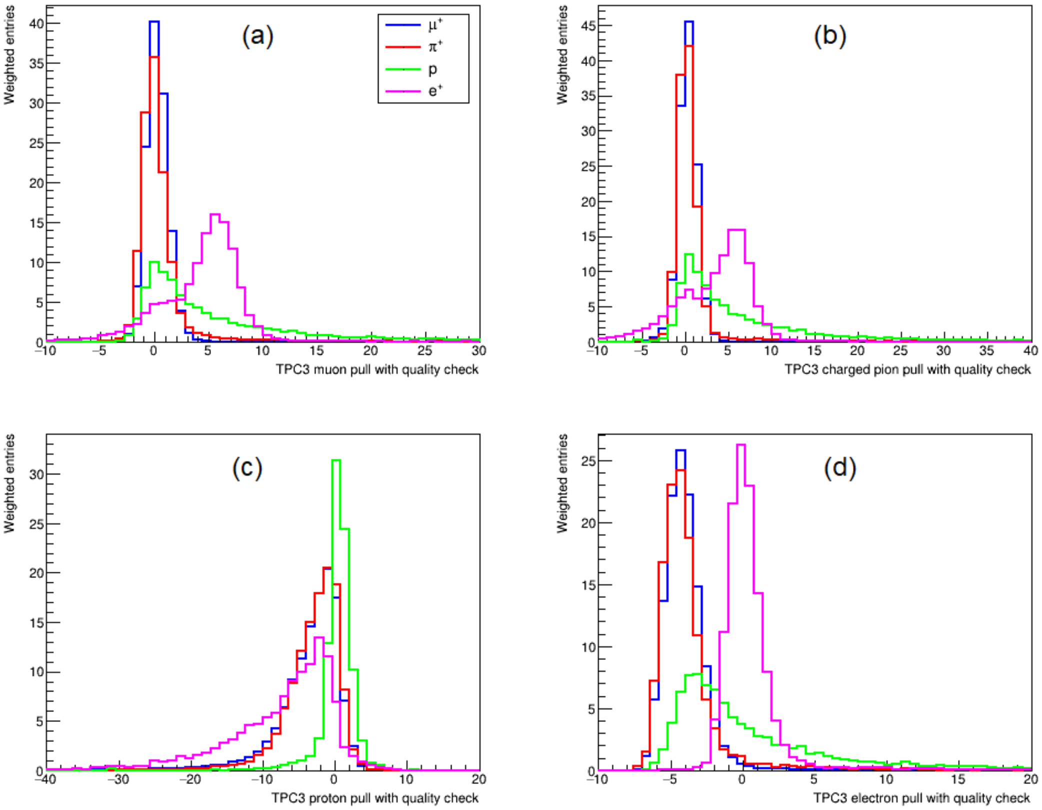


Figure 6.14: Distributions of the TPC3 PID pulls (with the track quality check) in the Monte Carlo training sample for each of the true particle types. The muon-like (a), charged pion-like (b), proton-like (c) and electron-like (d) pulls are shown. The distributions are shown weighted by reconstructed momentum and subsample size.

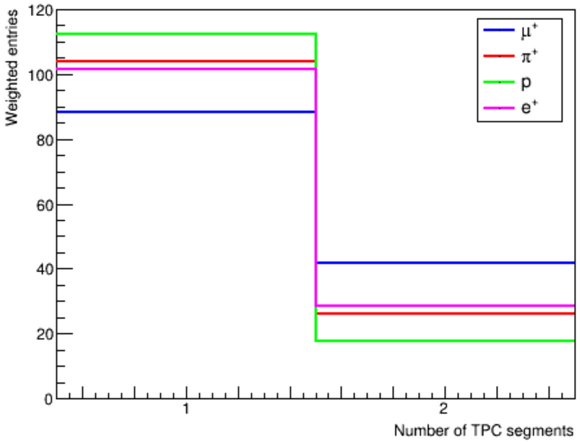


Figure 6.15: Distributions of the number of TPC segments in the Monte Carlo training sample for each of the true particle types. The distributions are shown weighted by reconstructed momentum and subsample size.

6.4.3 FGD particle ID

The FGD PID variables described in Section 4.2.1 rely on the expected range of particles, and are therefore only applicable for FGD-contained tracks. As such they are not suitable inputs for this tool. However, the FGDs energy loss information can still be incorporated by considering the total energy and length of an FGD segment.

The ND280 reconstruction software clusters FGD hits from the two 2D views into a 3D track. Each hit contains the charge deposit measured by scintillation light, which is corrected for WLS fibre attenuation and Birks saturation [107], and converted to an energy deposit in MeV. The total energy deposit E of an FGD segment is the sum of the energy deposited in each hit. The length L of the FGD segment is calculated as the length of the straight line between the final and initial 3D positions, which are obtained by fitting a straight line for each 2D projection [90]. Thus a simple PID variable E/L can be constructed, representing the mean energy loss along the track. Per the Bethe-Bloch equation, in a given medium this quantity depends only on the mass and momentum of the particle. Hence, if the BDT also has access to the reconstructed momentum of the track, E/L may be a useful input variable. In addition, a segment in FGD2 is not guaranteed, and its presence or absence (represented by the default value for the FGD2 E/L) may provide some useful information e.g. about the range of the particle.

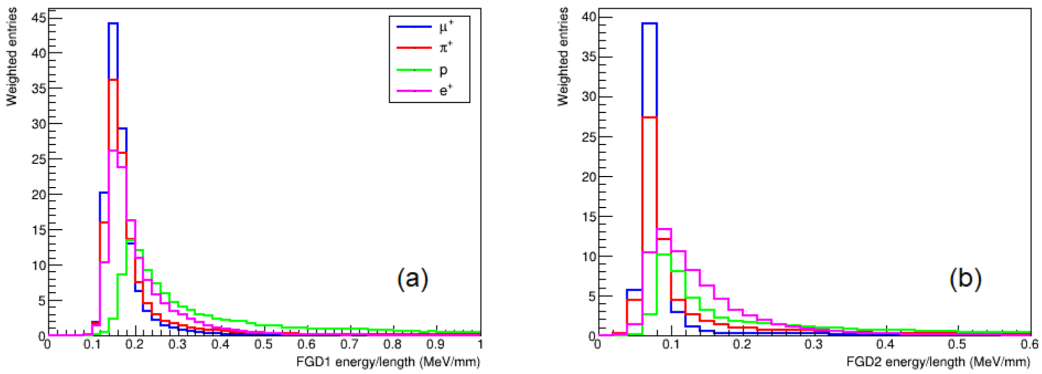


Figure 6.16: Distributions of the FGD1 (a) and FGD2 (b) energy/length variables in the Monte Carlo training sample for each of the true particle types. The distributions are shown weighted by reconstructed momentum and subsample size.

The values of E/L measured in FGD1 and FGD2 were considered as BDT input variable candidates. The distributions of these variables in the training sample are displayed in Figure 6.16. Little separation is seen between the particle types, but as previously argued, their main PID potential is in combination with the reconstructed momentum.

6.4.4 ECal particle ID

As discussed in Section 6.1.1, ECal PID variables are computed from the shape of the charge cluster, and are therefore independent of the TPC and FGD variables which measure energy loss per unit length. This makes ECal PID variables highly complementary to those of the TPCs and FGDs, and therefore valuable inputs to the BDT (provided that an ECal segment exists).

Both the high-level and the low-level ECal variables were considered as candidate input variables for the BDT. Although the high-level variables have been shown to provide good distinction in the contexts they are designed for, the low-level variables may contain additional information that can be extracted by the BDT but is lost in the construction of the LLRs. The distributions of the low-level variables in the training sample are shown in Figure 6.17, and the high-level variables in Figure 6.18.

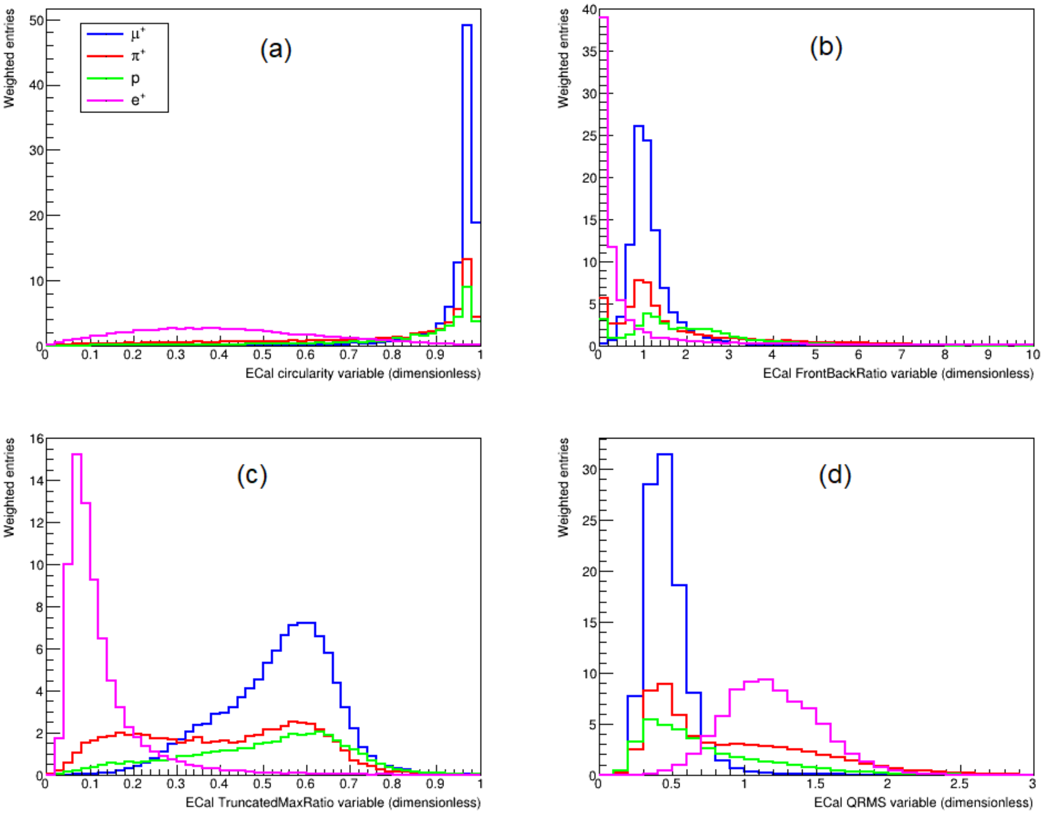


Figure 6.17: Distributions of the ECal low-level PID variables in the Monte Carlo training sample for each of the true particle types. The low-level variables comprise the Circularity (a), FrontBackRatio (b), TruncatedMaxRatio (c), and QRMS (d). The distributions are shown weighted by reconstructed momentum and subsample size.

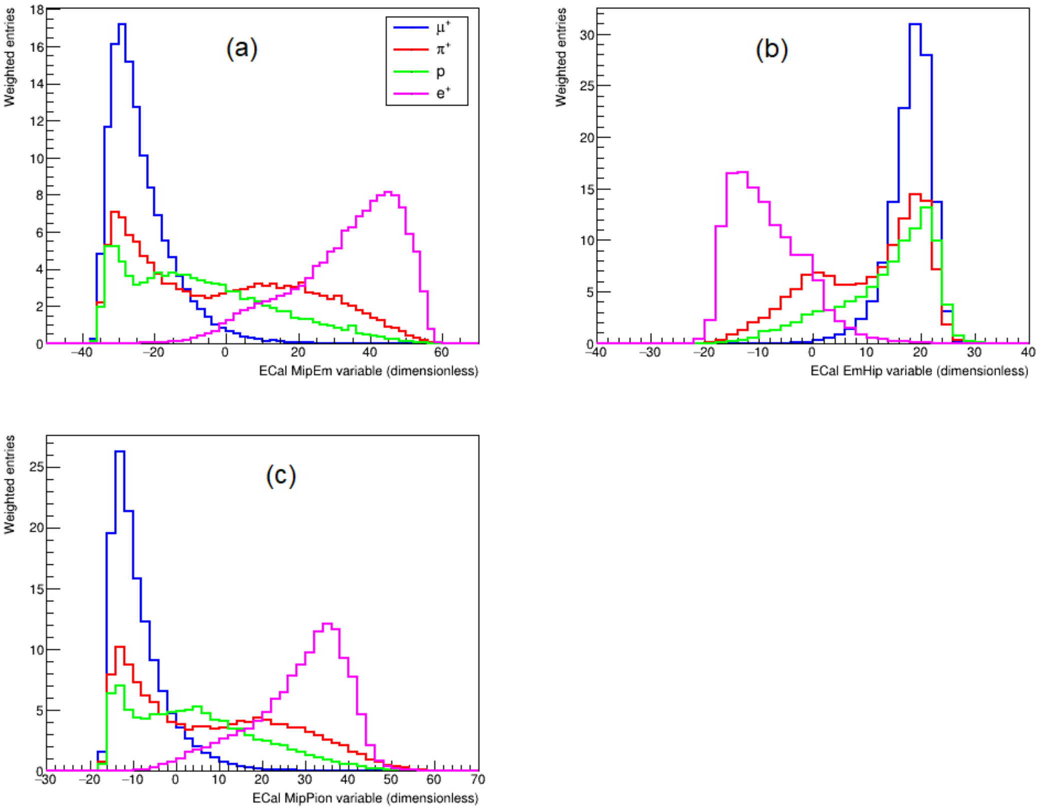


Figure 6.18: Distributions of the ECal high-level PID variables in the Monte Carlo training sample for each of the true particle types. The high-level variables comprise MipEm (a), EmHip (b), and EmHip (c) The distributions are shown weighted by reconstructed momentum and subsample size.

As described in Section 6.1.1 above, the relationship between E_{EM} and p_{reco} offers PID information, for which they are usually used in the form E/p. With p_{reco} already included, E_{EM} was also considered as a candidate input for the BDT. Additionally, the EM energy of the cluster divided by its length (E/L, referred to in this chapter as 'E_{EM}/L' to distinguish it from the FGD E/L) is a good discriminator between MIP-like and showering particles (as demonstrated in Section 4.3.2), and is therefore included as a candidate input variable. The distributions of E_{EM} and E_{EM}/L for tracks in the training sample are shown in Figure 6.19.

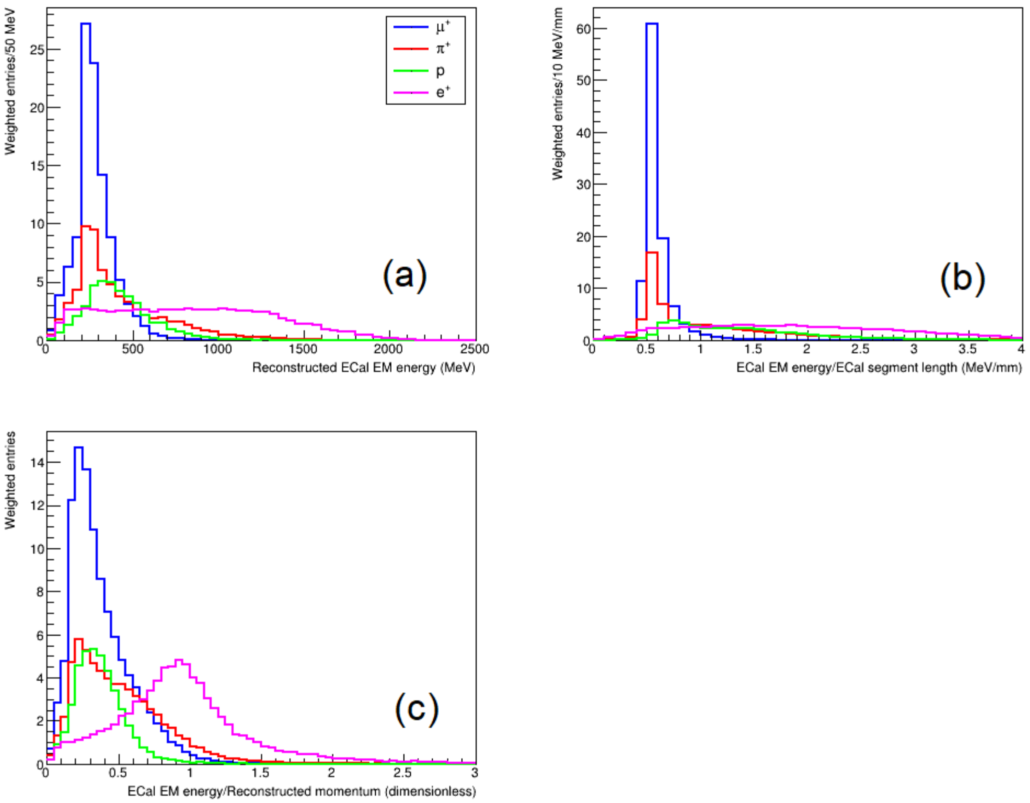


Figure 6.19: Distributions of the ECal E_{EM} (a) and E_{EM}/L (b) variables in the Monte Carlo training sample for each of the true particle types. The distributions of E_{EM}/p_{reco} (c) are also shown to illustrate the separating power of E_{EM} when considered in conjunction with the reconstructed momentum. The distributions are shown weighted by reconstructed momentum and subsample size.

6.4.5 SMRD particle ID

Although the SMRD is too coarse to provide high-level PID information, it can still help identify tracks via the simple presence or absence of a reconstructed SMRD segment. The SMRD is located within the UA1 magnet yoke, so to reach it, particles originating in the FGDs must penetrate a large amount of dense material in the yoke and the magnet itself. Due to their high penetrating power, muons are much more likely to reach the SMRD than the other particle species considered in this PID development. Hence any track with a reconstructed SMRD segment is likely to be a muon. For this reason, the number of SMRD segments associated with the track (‘nSMRDs’) was considered as a candidate input variable for the BDT. The distributions of nSMRDs in the training sample are shown in Figure 6.20.

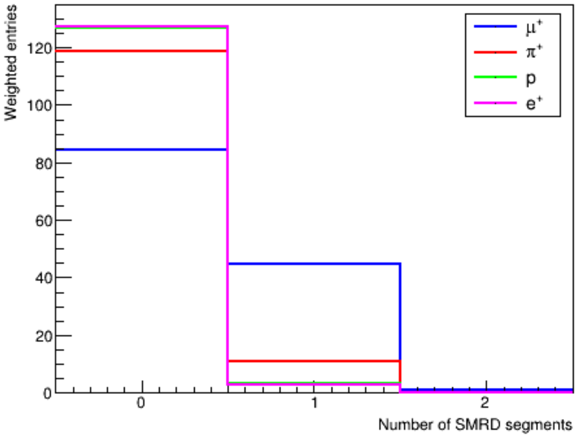


Figure 6.20: Distributions of the number of SMRD segments in the Monte Carlo training sample for each of the true particle types. The distributions are shown weighted by reconstructed momentum and subsample size.

6.5 Performance analysis and optimisation

Having identified the methods and inputs to be used for the development of the PID tool, the next step was to optimise the choice of BDT parameters and input variables. This was achieved by training different configurations of the BDT on the training sample, and applying each version to the tracks in the training sample to test the performance. The space of possible configurations (values for each of the BDT parameters and combinations of input variables) is very large, and training and testing each BDT required non-negligible computing time, so testing a wide variety of them was not feasible. Instead the optimisation was performed sequentially: a number of decision points (choice of parameter value, choice of variable set, etc) were identified and tested in turn. At each decision point, the optimal choice was identified and used in all following training/testing.

The performance of the BDT is here assessed by computing its efficiency when selecting each particle type; that is, the fraction of tracks of each true particle type identified by the BDT as each particle type. For each track in the testing sample, the BDT outputs are compared and the largest taken as the PID decision for the track (this will be referred to as the BDT 'preference'). The efficiencies for correctly identifying each particle type in this way (μ^+ as μ^+, π^+ as π^+ etc) are computed, yielding four primary performance metrics by which each BDT configuration can be judged. The nine 'background' efficiencies for incorrectly identifying each particle type as each other particle type (μ^+ as π^+, e^+ as p etc) are also computed and considered, in order to better interpret the four 'signal' efficiencies. A perfect PID would have signal efficiencies of 1, and background efficiencies of 0.

6.5.1 BDT training parameters

The BDT parameters MaxDepth, NTrees, Shrinkage, BaggedSampleFraction and UseNVars were considered for tuning. These parameters are defined in Section 5.7.2. As a starting point, each of these parameters was set equal or close to its default value. MaxDepth was set to 3 and NTrees to 1000. Shrinkage was set to 1.0 (i.e. no reduction in learning rate) and the bagging and random forests functionalities were turned off.

When tuning the BDT parameters, a 'default' set of input variables was chosen. This comprised all candidate input variables, except those that are different representations of the same information (e.g. the TPC dE/dx and pulls), in which cases the lowest-level representations were chosen as defaults. The default input variables were as follows:

- p_{reco}
- θ_{reco}
- FGD1 E/L
- FGD2 E/L
- TPC2 dE/dx
- TPC3 dE/dx
- nTPCs
- ECal Circularity
- ECal FrontBackRatio
- ECal TruncatedMaxRatio
- ECal QRMS
- ECal E_{EM}
- ECal E_{EM}/L
- nSMRDs

For MaxDepth, values from 2 to 5 were tested, in the expectation that the effects of overtraining (if any) would be seen in diminishing returns above 4. The results of this testing are shown in Figure 6.21: the μ, π and p efficiencies appear relatively stable, while the e efficiency increases with MaxDepth (though even here the overall change is small). The increase in e efficiency between 4 and 5 is smaller than between 3 and 4, which may indicate the expected diminishing returns due to overtraining. On this basis, a value of MaxDepth = 4 was chosen for maximum e efficiency while keeping within the recommended bounds.

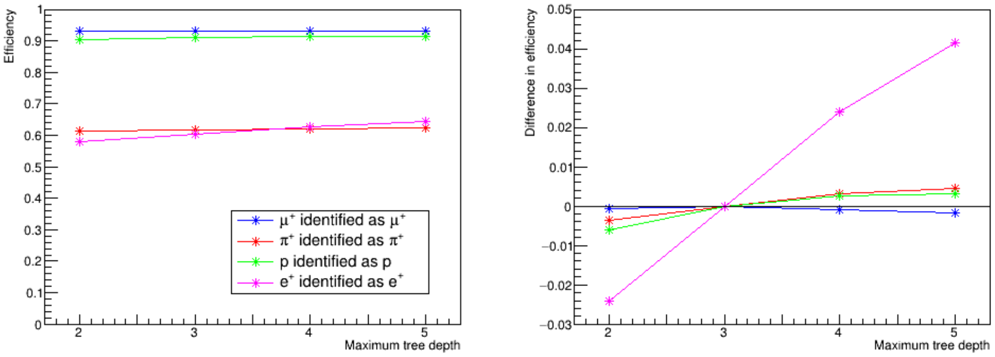


Figure 6.21: Particle identification efficiencies of BDT configurations with different values of the MaxDepth parameter: the maximum allowed depth of each decision tree. The overall efficiencies are shown on the left-hand plot. The right-hand plot shows the differences in efficiency relative to MaxDepth = 3.

The number of trees in the forest NTrees determines the overall learning capacity of the BDT (as well as the computing time required for training). Thus we can expect an increase in performance with NTrees until the BDT is extracting most of the useful information from the input, beyond which additional trees become increasingly redundant leading to diminishing returns. A range of NTrees values between 500 and 2500 were tested; it was found that above 2500 the training time became prohibitively long for the purposes of further tuning. The results of this testing are shown in Figure 6.22. Again, the greatest variation between configurations is seen in the e efficiency, which increases with NTrees at low values. Diminishing returns set in between 1000 and 1500, with little gain in e efficiency and a growing loss in μ efficiency above 2000. On this basis, a value of NTrees = 1500 was chosen.

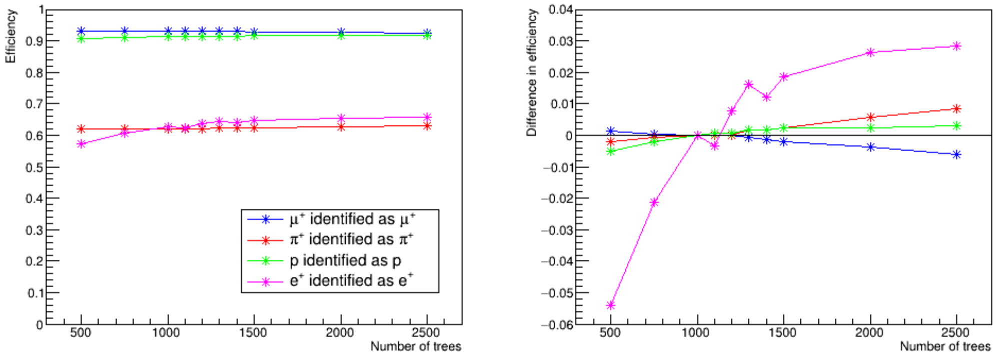


Figure 6.22: Particle identification efficiencies of BDT configurations with different values of the NTrees parameter: the number of decision trees in the forest. The overall efficiencies are shown on the left-hand plot. The right-hand plot shows the differences in efficiency relative to NTrees = 1000.

Values of Shrinkage between 0.1 and 1.0 were tested, the results of which are shown in Figure 6.23. Reducing the learning rate resulted in a sharp drop-off in the e efficiency and negligible improvements in the others, so the original value of Shrinkage = 1.0 was maintained.

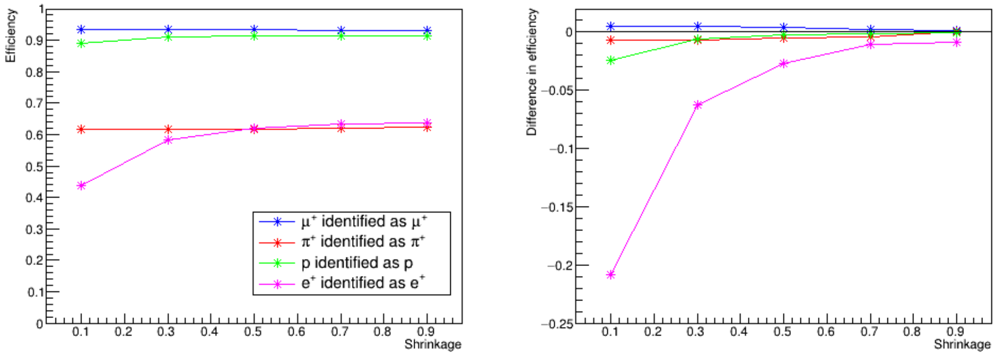


Figure 6.23: Particle identification efficiencies of BDT configurations with different values of the Shrinkage parameter: the learning rate of the BDT. The overall efficiencies are shown on the left-hand plot. The right-hand plot shows the differences in efficiency relative to Shrinkage = 1.0 (i.e. no shrinkage of learning rate).

Values of BaggedSampleFraction between 0.1 and 0.9 were tested; TMVA recommends values between 0.5 and 0.8. The results are shown in Figure 6.24. The performance for each particle type appears largely unaffected by bagging, with the exception of the lowest value BaggedSampleFraction = 0.1 which causes a drop in all four efficiencies. On this basis, bagging was not used.

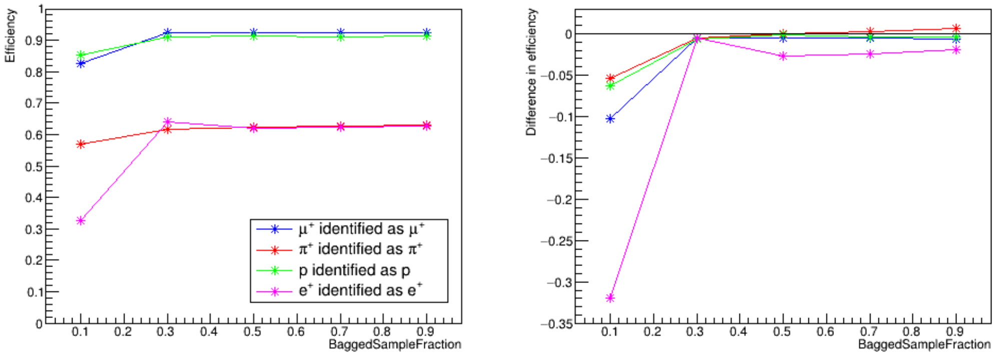


Figure 6.24: Particle identification efficiencies of BDT configurations with different values of the BaggedSampleFraction parameter: the size of the subsample used in bagging relative to that of the full training sample. The overall efficiencies are shown on the left-hand plot. The right-hand plot shows the differences in efficiency relative to the configuration without bagging enabled.

Values of NVars (set as the mean of a Poisson distribution) between 4 and 12 were tested, from the total of 14 default input variables. The results are shown in Figure 6.25. The efficiency for each particle type appears largely unaffected by NVars. Since it appears to have negligible effect on performance, and to avoid adding unnecessary complexity, the random forests method was not used. Together with the results for bagging, this may indicate that the effect of statistical fluctuations in the training sample is minimal.

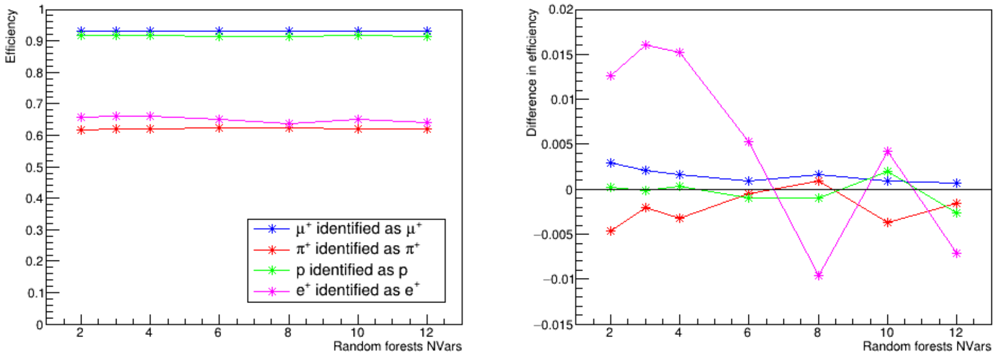


Figure 6.25: Particle identification efficiencies of BDT configurations with different values of the UseNVars parameter: the mean of a Poisson distribution used to determine the number of variables considered at each node splitting in the 'random forests' method. The overall efficiencies are shown on the left-hand plot. The right-hand plot shows the differences in efficiency relative to the configuration without random forests enabled.

6.5.2 Comparing representations

To represent certain PID information, multiple options exist among the candidate input variables, for example the TPC dE/dx truncated mean itself and the pulls and likelihoods which are calculated by comparing it to predictions for the four particle hypotheses. Where these multiple options exist, including more than one representation of the same information will most likely be inefficient, so a single representation should be chosen. However it is not obvious which representations will be optimal: the more mature 'higher-level' variables may present the PID information in a form more accessible to the BDT for decision-making by incorporating predictions (or similar), or alternatively, information lost in the process of constructing the higher-level variables may still be accessible to the BDT in the lower-level variables. Additionally, different representations often comprise different numbers of variables, so if the same information can be 'packaged' more efficiently in a smaller number of variables this will be favourable. These options should therefore be tested in order to find the optimal choice in each case. This was determined similarly to the BDT parameter choice studies: BDT configurations were trained with each choice of representation, and evaluated using the particle identification efficiencies defined in the previous section. In each case the representation judged to offer the best overall performance was selected.

In TPC2, the choice is between the dE/dx truncated mean, the pulls, and the likelihoods. The truncated mean is a single variable rather than a correlated set of four, but the pulls and likelihoods also contain a comparison to the expected energy loss; with only the dE/dx, the BDT will have to learn the energy loss distributions for the four particle types from the training data. The efficiencies for these variable choices are shown in Figure 6.26. The pulls and likelihoods both outperform the truncated mean; of these two, the likelihoods offer slightly better e efficiency, but this is outweighed by the greater efficiencies in π and p efficiency offered by the pulls. On this basis, the pulls were chosen as the representation for the TPC2 energy loss information.

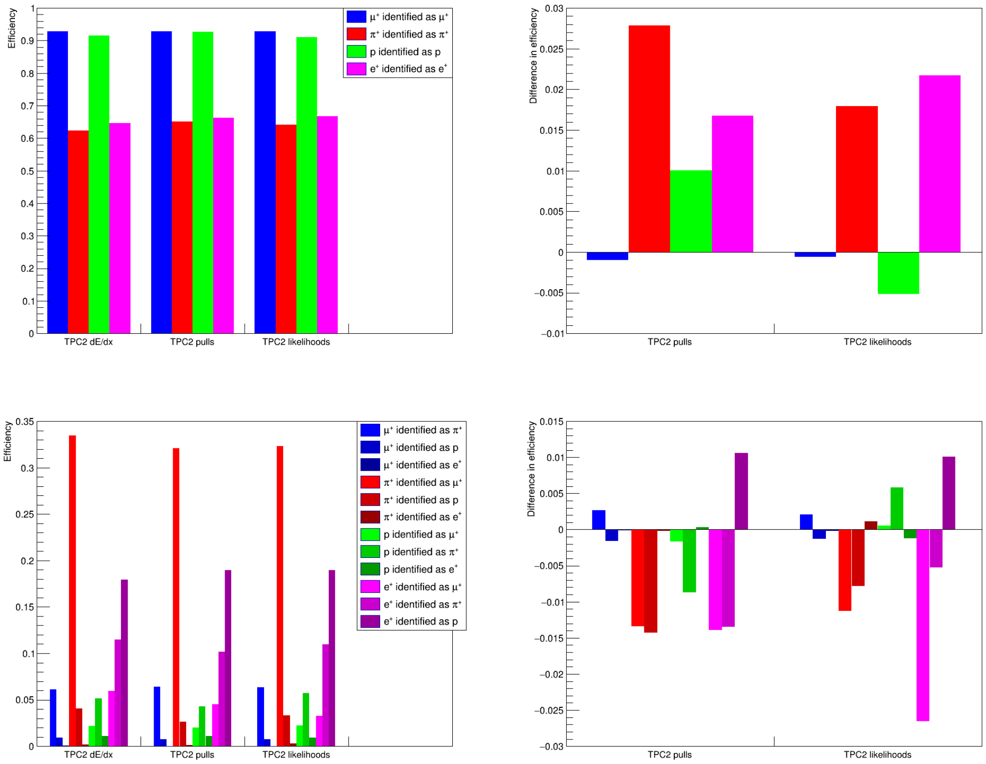


Figure 6.26: Particle identification efficiencies of BDT configurations with different representations of the TPC2 energy loss information: the dE/dx truncated mean itself, the pulls, and the likelihoods. The overall efficiencies are shown on the left-hand plots. The right-hand plots show the differences in efficiency relative to the configuration with the truncated mean. The top and bottom plots show the 'signal' and 'background' efficiencies respectively.

In TPC3, the choice is between the dE/dx truncated mean and the pulls; additionally, versions of each of these with a data quality check were considered, since including information from poor quality TPC3 tracks (i.e. those with a small number of nodes) may impact the performance of the BDT. The data quality check requires at least 19 nodes in the track; otherwise a default value is assigned. The efficiencies for these variable choices are shown in Figure 6.27. The data quality check makes a substantial difference to the e efficiency, showing an improvement of almost 30% between the two configurations with the truncated mean. This brings the e efficiency, which was previously relatively poor, up to a level similar to the μ and p efficiencies. Comparing the ‘good quality’ versions of the truncated mean and the pulls, there is little difference in performance: the truncated mean was chosen for its slightly better μ and p efficiencies, and for the sake of simplicity (a single input variable instead of four).

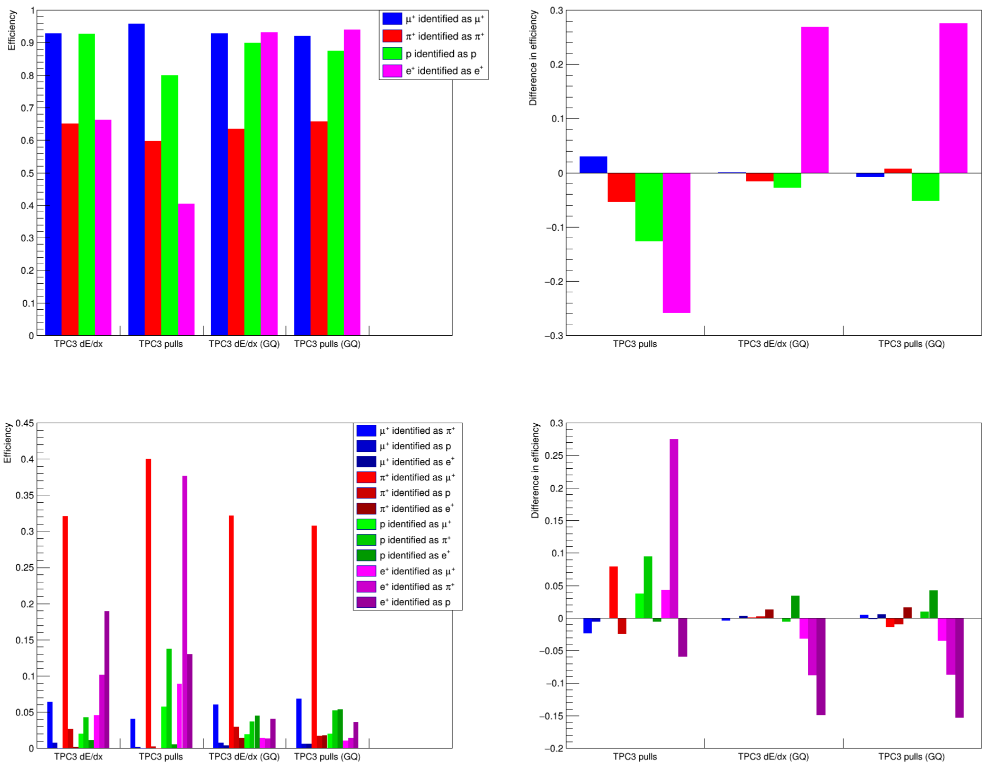


Figure 6.27: Particle identification efficiencies of BDT configurations with different representations of the TPC3 energy loss information: the dE/dx truncated mean, the pulls, and versions of these with the data quality check imposed (denoted ‘GQ’). The overall efficiencies are shown on the left-hand plots. The right-hand plots show the differences in efficiency relative to the configuration with the truncated mean and no data quality check. The top and bottom plots show the ‘signal’ and ‘background’ efficiencies respectively.

In the ECal, the choice is between the low-level variables (Circularity, FrontBackRatio, TruncatedMaxRatio, QRMS) and the high-level LLRs derived from them (MipEm, MipPion, EmHip). The high-level variables are constructed to compare four specific ECal track types

(MIP-like, HIP-like, EM shower, pion shower), but only include three of the 12 possible comparisons, so some information may be lost. The efficiencies for these variable choices are shown in Figure 6.28. The main difference between the two representations is the e efficiency, which is almost 40% lower when the high-level variables are used. This makes the low-level variables the clear choice.

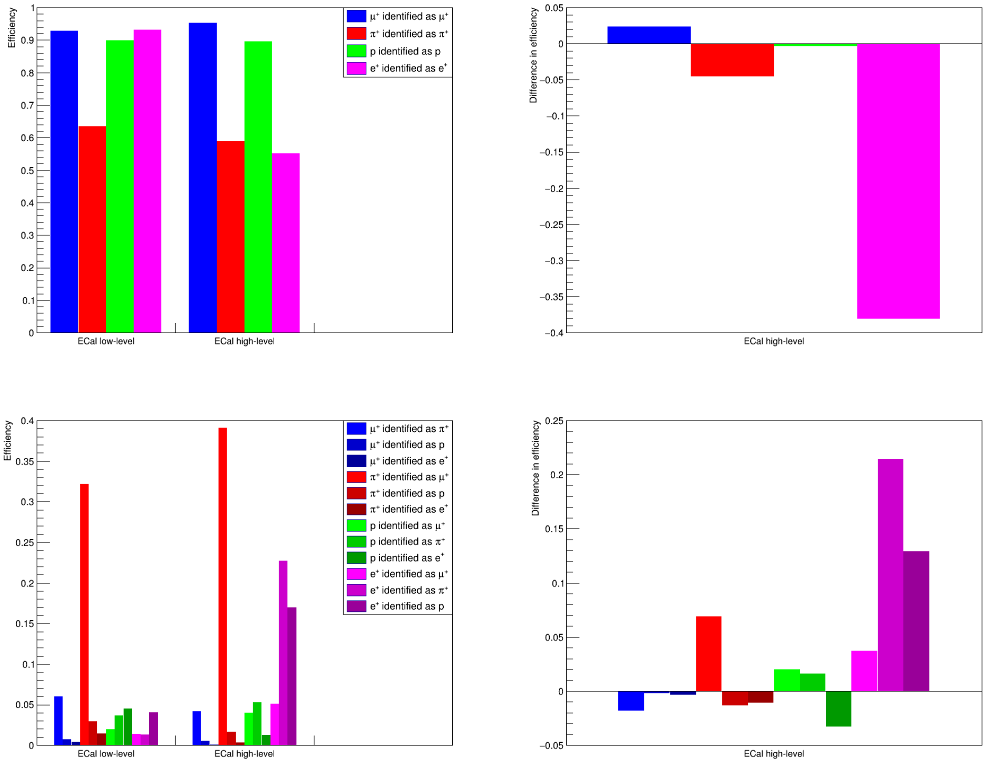


Figure 6.28: Particle identification efficiencies of BDT configurations with different representations of the ECal charge distribution information: the low-level variables and the high-level LLRs. The overall efficiencies are shown on the left-hand plots. The right-hand plots show the differences in efficiency relative to the configuration with the low-level variables. The top and bottom plots show the 'signal' and 'background' efficiencies respectively.

6.5.3 Variable removal

Following the decisions described in the previous section, the candidate input variables were:

- p_{reco}
- θ_{reco}
- FGD1 E/L
- FGD2 E/L
- TPC2 muon pull
- TPC2 charged pion pull
- TPC2 proton pull
- TPC2 electron pull
- TPC3 dE/dx (GQ)
- nTPCs
- ECal E_{EM}
- ECal E_{EM}/L
- ECal Circularity
- ECal FrontBackRatio
- ECal TruncatedMaxRatio
- ECal QRMS
- nSMRDs

To test the importance of each of these variables to the performance of the BDT, a series of configurations were tested, each with a single input variable removed. These are referred to as ‘N-1 studies’. By comparing the performance of the BDT with and without each variable, variables that contribute little to (or indeed reduce) the BDT performance can be identified and removed from the input. Only the kinematic variables p_{reco} and θ_{reco} were not considered for removal, due to their expected importance for interpreting other variables.

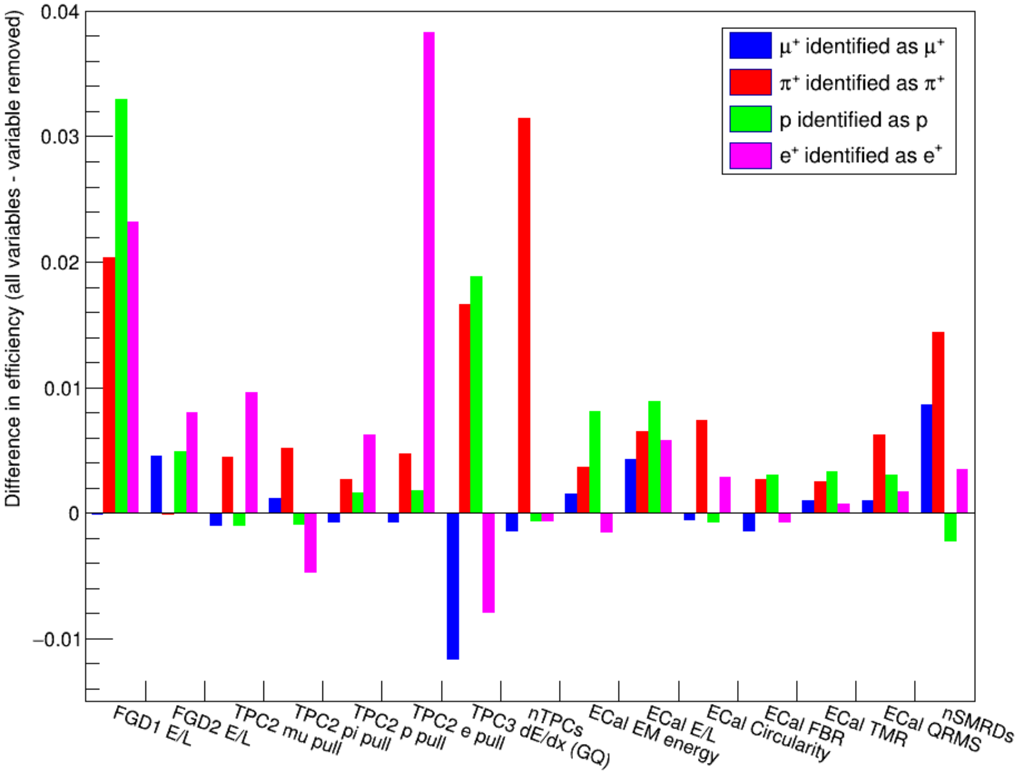


Figure 6.29: Particle identification efficiencies of BDT configurations with each candidate input variable removed. The X-axis labels denote the variable removed in each configuration. The Y-axis shows the difference in efficiencies relative to the configuration with all variables present.

The results of the N-1 studies are shown in Figure 6.29. The more positive the difference in efficiency for each particle type, the more useful the variable can be said to be; conversely, the more negative the difference, the more of a detriment the variable is to the BDT's ability to correctly identify that particle type. Almost all variables show a positive contribution to two or more of the four efficiencies, with minimal loss in the others. One exception is the TPC charged pion pull, for which the improvement in π efficiency is roughly balanced by the loss in e efficiency. Another is the TPC3 dE/dx truncated mean, which yields relatively large gains of about 2% in the π and p efficiencies, but also losses of about 1% in μ and e. Although the differences are small overall (all less than 4%, and many below 1%) it is interesting to note the various contributions of the different variables. Some make significant contributions to a single efficiency: for example, nTPCs makes a 3% difference to the π efficiency while having negligible effect on the others, and similarly the TPC2 electron pull for the e efficiency. Others offer smaller contributions but in multiple efficiencies, such as the ECal E/L which adds modestly to each of the four particle types. The overall highest-performing individual variable is the FGD1 E/L, which adds at least 2% to each of the π, p and e efficiencies, with negligible change in μ; while the lowest-performing variable is the TPC2 pion pull, for which the increase in π efficiency is offset by the loss in e efficiency. On the basis of these efficiency changes, the TPC2 pion pull was removed from the input variables. All other variables were kept, including the TPC3 dE/dx: the increase it offers in π and p efficiency was deemed to outweigh the losses in μ and e.

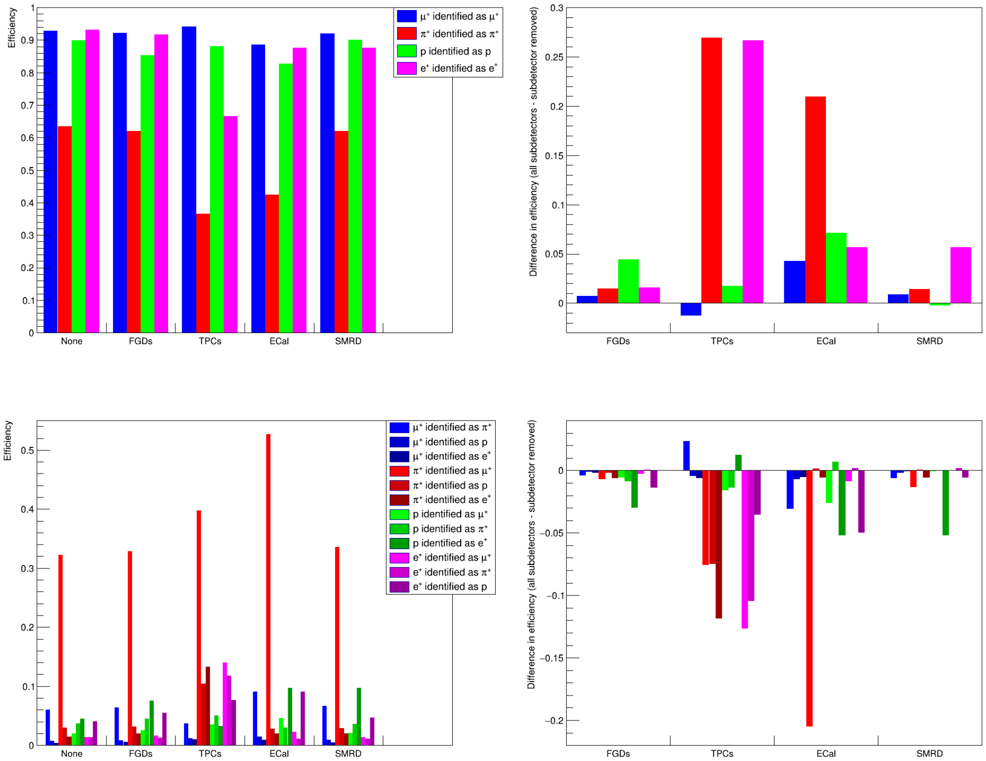


Figure 6.30: Particle identification efficiencies of BDT configurations with variables from each subdetector removed. The kinematic variables p_{reco} and $theta_{reco}$ were kept in all configurations. The X-axis labels denote the subdetector removed in each configuration. The overall efficiencies are shown on the left-hand plots. The right-hand plots show the differences in efficiency relative to the configuration with all variables included. The top and bottom plots show the ‘signal’ and ‘background’ efficiencies respectively.

Additionally, to demonstrate the importance of each subdetector to the BDT performance, a similar study was performed by removing all variables from each subdetector in turn. The results of this are shown in Figure 6.30. From this it can be seen that the greatest overall contribution comes from the TPC, impacting more on the π, p and e efficiencies when removed than any other subdetector. Conversely, the greatest loss in the μ efficiency comes from removing the ECal. The smallest overall contributions come from the FGDs and SMRD, which is unsurprising as they each contribute only one or two variables and have inferior resolution compared to the TPCs and ECal. Nevertheless, each subdetector makes a significant positive contribution to the BDT performance, which demonstrates the value of a global PID.

6.6 Final configuration performance

After the above tuning process, the final BDT configuration uses the following parameters:

- MaxDepth = 4
- NTrees = 1500
- Shrinkage = 1.0
- nCuts = 2000
- MinNodeSize = 5%

and the following input variables:

- p_{reco}
- θ_{reco}
- FGD1 E/L
- FGD2 E/L
- TPC2 muon pull
- TPC2 proton pull
- TPC2 electron pull
- TPC3 dE/dx (GQ)
- nTPCs
- ECal E_{EM}
- ECal E_{EM}/L
- ECal Circularity
- ECal FrontBackRatio
- ECal TruncatedMaxRatio
- ECal QRMS
- nSMRDs

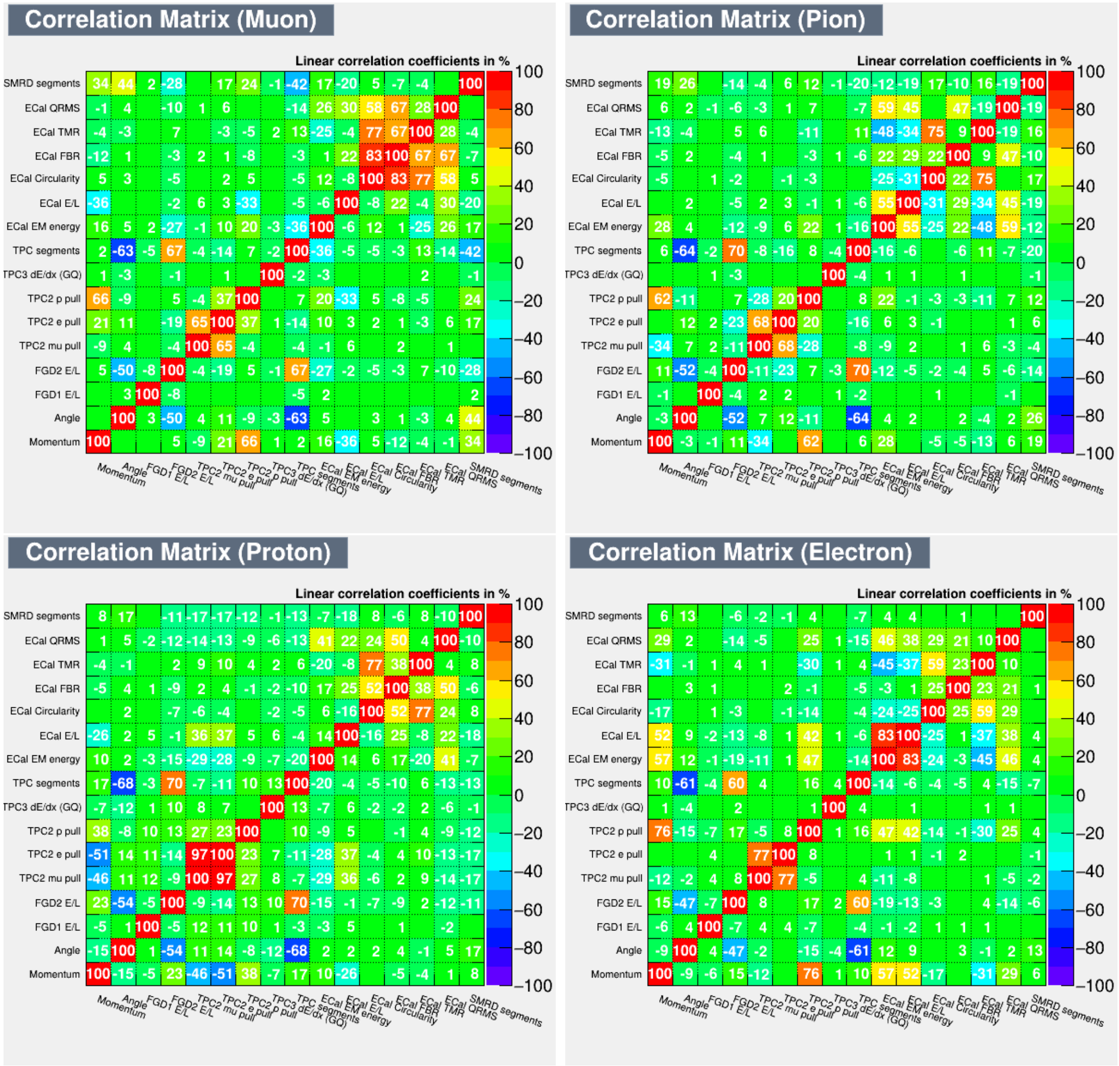


Figure 6.31: Correlation matrices of the input variables used for the final BDT configuration, in each of the training subsamples (clockwise from top left): muon-like (μ^+), charged pion-like (π^+), electron-like (e^+), proton-like (p). These were generated using only tracks with ECal segments, since otherwise the default values cause the correlation coefficients between ECal variables to appear inflated (that is, for tracks without an ECal segment, all ECal variables will have their default values).

The correlations between the input variables are displayed in Figure 6.31. A lot of variation can be seen in the correlation coefficients between different pairs of variables, as well as in the different subsamples, and these complex correlations support the use of a

Rank	**Variable**	**Importance**
1	TPC2 electron pull	0.09737
2	TPC3 dE/dx	0.09116
3	TPC2 muon pull	0.09101
4	p_{reco}	0.09024
5	θ_{reco}	0.08089
6	TPC2 proton pull	0.07659
7	FGD1 E/L	0.07350
8	ECal EM energy	0.05713
9	ECal E/L	0.05265
10	ECal TruncatedMaxRatio	0.05251
11	ECal QRMS	0.05136
12	ECal Circularity	0.04880
13	FGD2 E/L	0.04731
14	ECal FrontBackRatio	0.04508
15	nTPCs	0.02563
16	nSMRDs	0.01877

Table 6.2: Importance ranking of the final BDT configuration input variables as computed by TMVA.

BDT as opposed to lower-level MVA methods. Although some pairs of variables have high correlation coefficients ($>\sim$ 80%) in certain subsamples, none have greater than 65% in all four. This, along with the N-1 study, confirms that each variable contributes some amount of information that is independent of that contributed by other variables. An estimate of the 'importance' of each input variable is computed by TMVA during the training phase: this is derived by counting how often each variable is selected for node splitting, and weighting each split by the separation gain-squared it achieves and the number of events in the node [96]. The importance values for the input variables are displayed in Table 6.2. Although some variation can be seen, most of the importance values are similar and all lie within an order of magnitude of each other, indicating that the input variables are all contributing to a relatively similar extent to the overall functionality of the BDT.

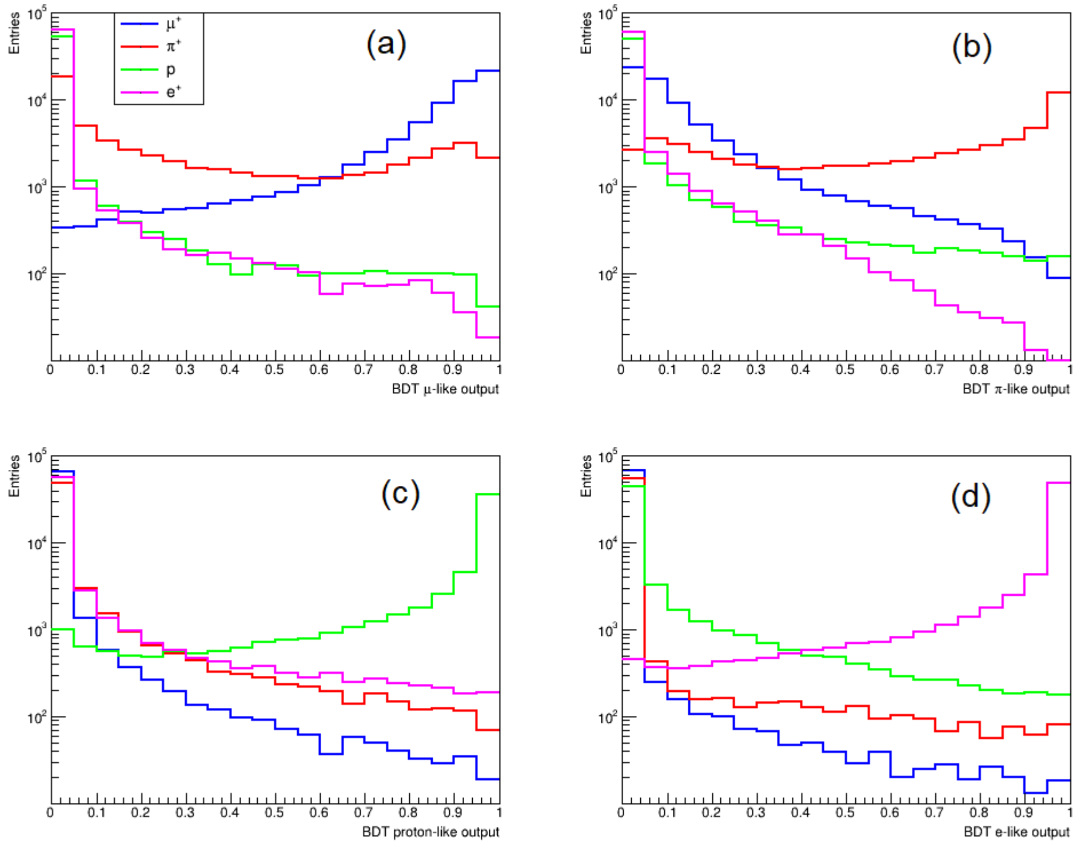


Figure 6.32: Distributions of the muon-like (a), charged pion-like (b), proton-like (c) and electron-like (d) outputs of the final BDT configuration for tracks in the testing sample. The y-axis is shown with a logarithmic scale for clarity.

The distributions of the BDT outputs for tracks in the testing sample are displayed in Figure 6.32. As expected, each particle type is most likely to have values near 1 for its corresponding BDT output (and 0 for the others), with the probability decreasing as the output moves further from 1 (0). The exceptions are the muon- and pion-like outputs for π^+, which each have a broader distribution of values with a second peak at the opposite end of the spectrum; this is due to the large proportion of pions that exhibit MIP-like behaviour and are thus interpreted by the BDT as appearing partly or fully muon-like. The BDT outputs can be used for PID in multiple ways: in the preference method, all four are compared, selecting the hypothesis corresponding to the largest as the PID decision. Alternatively, cuts can be applied to the individual outputs. The advantage of the preference is that it can be applied to multiple tracks consistently and unambiguously: each track will be assigned one and only one PID identity. This makes it suitable for PID of secondary tracks. On the other hand, when a particular particle type is sought for selection purposes (i.e. primary PID), a cut on a single output can be more useful e.g. to obtain greater purity. In this case, the cut can be optimised to maximise a figure of merit in the usual way.

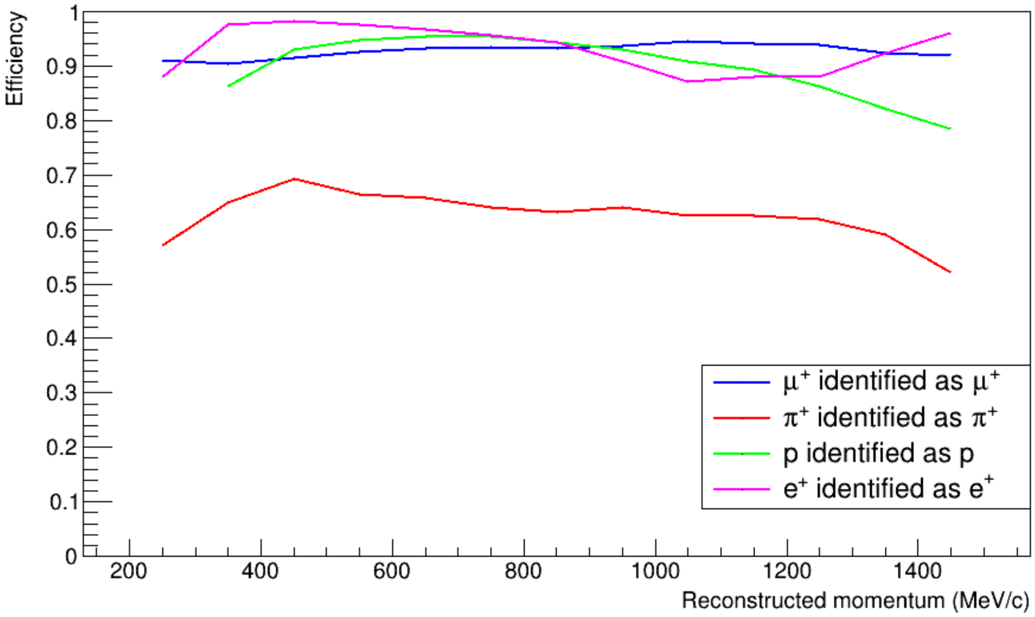


Figure 6.33: Particle identification 'signal' efficiencies of the final BDT configuration preference cuts as a function of the reconstructed momentum.

The performance of the final configuration is here assessed using both the preference and single-output cut methods. The performance with the preference is assessed by computing the particle identification efficiencies as a function of the reconstructed momentum p_{reco}. These efficiencies are calculated by binning the PG testing sample events by the reconstructed momentum, before and after application of the BDT PID cut in question, and calculating the PID efficiencies for each of the momentum bins. This helps reveal the extent to which the BDT decision-making is influenced by p_{reco}, and enables a meaningful comparison to the existing PID methods while still using the particle gun testing sample. The 'signal' and 'background' efficiencies (i.e. misidentification rates) for the BDT preference cuts are shown in Figures 6.33 and 6.34 respectively. Overall, the signal efficiencies appear high: all but the π^+ efficiency remain above 80% across the momentum spectrum, and the π^+ efficiency remains between 50% and 70% (since a large proportion of charged pions behave near-identically to muons, this can still be considered good performance). The main confusions appear to be between muons and pions, which is to be expected since both can exhibit very similar MIP-like behaviour; and between protons and positrons, particularly at higher momenta. This too is to be expected since their TPC dE/dx curves cross at around 1 GeV/c, and both produce showers in the ECal which can be difficult to distinguish.

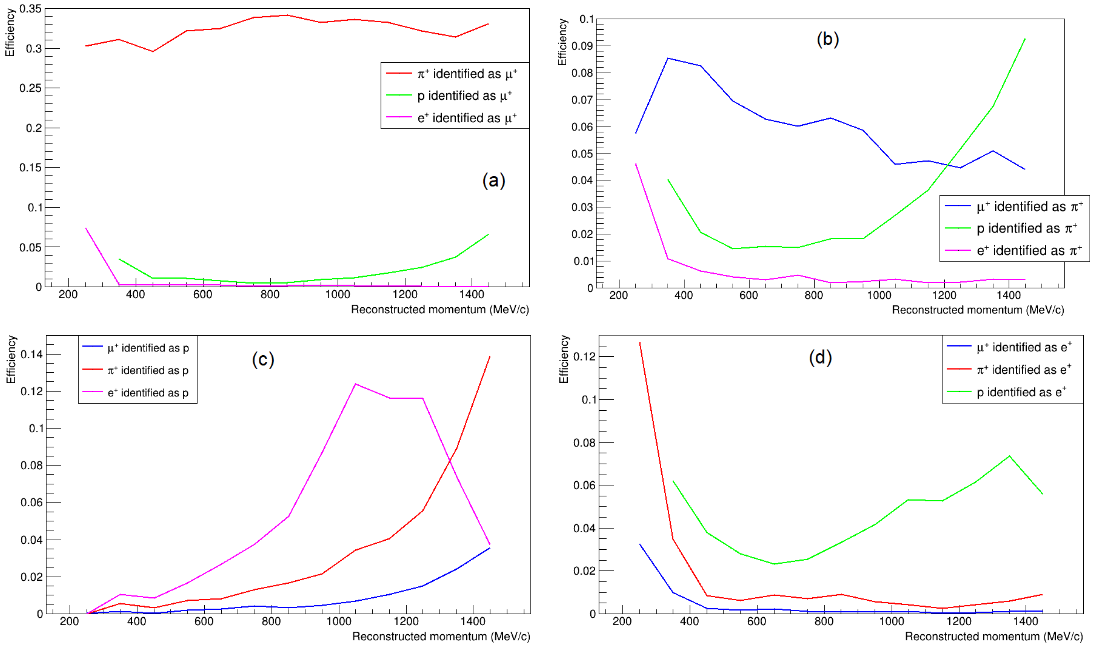


Figure 6.34: Particle identification 'background' efficiencies of the final BDT configuration preference cuts as a function of the reconstructed momentum. The efficiencies are grouped for each plot by the intended selection: μ-like (a), π-like (b), p-like (c), and e-like (d).

Although the signal efficiencies do not show much momentum-dependence, most of the background efficiencies (Figure 6.34) do: in several cases, such as π^+ being misidentified as e^+, the misidentification rate is much higher at the very lowest momentum bin. Conversely, the misidentification rate for π^+ as protons increases with momentum. Others peak in particular momentum ranges, such as e^+ being misidentified as protons between 1 and 1.3 GeV/c. Regions of higher misidentification generally coincide with overlapping in the dE/dx curves. However, as will be seen later in this section, these momentum dependencies are generally less pronounced than those of the existing PID methods.

Unlike the preference, the performance of a cut on an individual variable depends on the choice of cut values, which is generally a tradeoff between efficiency (correctly accepting the desired 'signal' particle type) and purity (rejecting the unwanted 'background' particle types). The 'tighter' the cut value, the higher the purity and the lower the efficiency, and vice versa. It is therefore useful to examine the full range of efficiency and purity values that can be obtained by cuts on the classifier in question. This has been achieved here by testing a series of cut values on each BDT output with the testing sample, attempting to select the corresponding particle type and reject the others, and plotting a receiver operating characteristic (ROC) curve of the resulting efficiency and purity values. The efficiency is here defined as the proportion of signal tracks in the sample that are selected by the cut, and the purity as the proportion of tracks selected by the cut that belong to the signal particle type. The closer the ROC curve comes to the top and right edges of the plot (that is, the closer the efficiency and purity both come to 1), the better the classifier can be said to perform. Hence the top-right corner (efficiency = purity = 1) represents perfect performance, and the closer the ROC curve approaches it, the better. For each BDT output, 50 cuts were tested in a range between 0 (total acceptance) and 0.98 with equal increments of 0.02. The resulting ROC curves are shown in Figure 6.35. The best performance is obtained when selecting positrons: the curve for the e-like output comes closest to the top-right corner. This is followed by protons (p-like output), and then by antimuons (μ-like output) and pions (π-like output). This can be explained by considering the similarities and differences in the behaviour of the four particle types: positrons exhibit the most characteristic behaviour in the ND280 subdetectors, having a TPC dE/dx curve that is generally well-separated from the others (Figure 4.1) and reliably producing EM showers in the ECal, so it is unsurprising that they should be the most readily identifiable to the BDT. Protons generally produce hadronic showers in the ECal, but these may also be produced by pions, and their dE/dx curve overlaps with those of pions and antimuons, so there is somewhat more ambiguity to their behaviour and hence the PID performance is poorer. As previously discussed, antimuons and pions are difficult to distinguish; the TPC dE/dx cannot be relied upon and pions may produce MIP-like tracks in the ECal as well, so it is to be expected that the performance is poorest for these particle types. It is interesting to note that the curves for the μ-like and π-like outputs cross: at higher efficiencies, better purity is obtained for μ^+ than π^+, whereas the reverse is true for lower efficiencies. This can be understood by considering the variety of pion behaviours in the ND280 subdetectors: some π^+ will behave entirely as MIPs as they traverse the detector, and thus be effectively indistinguishable from μ^+, so a selection of μ^+ will always contain some amount of π^+ impurity. Conversely, by rejecting all MIP-like behaviour, a selection of π^+ can eliminate μ^+ and thus reach higher purity, albeit with large efficiency cost.

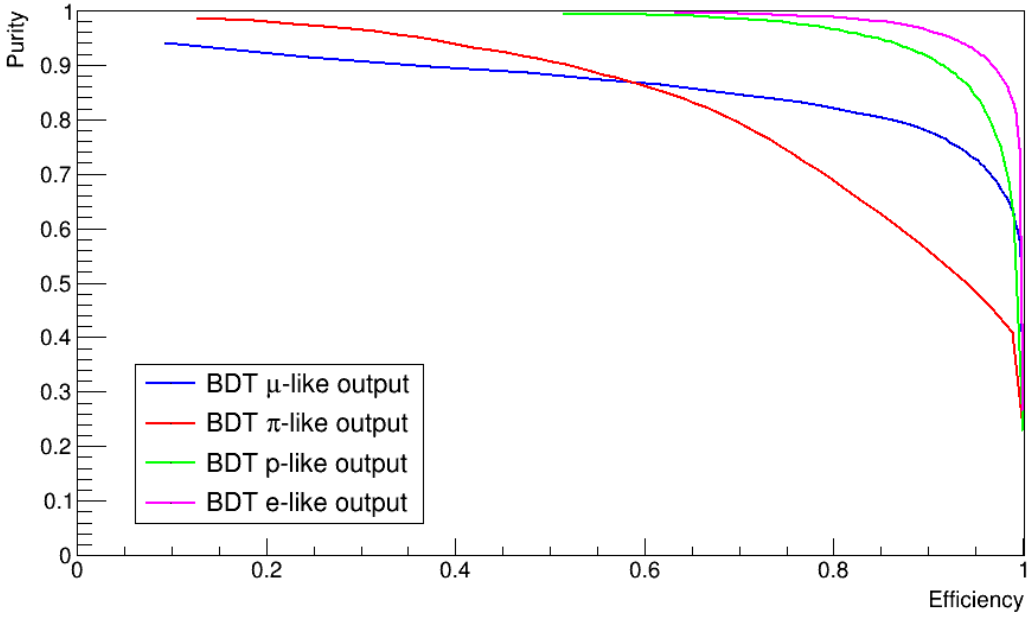


Figure 6.35: ROC curves for selecting each particle type from the testing sample via cuts on the corresponding BDT output, showing the efficiency and purity that can be obtained with a range of 50 cut points in each case.

6.6.1 Comparisons to existing PID

To contextualise the performance of the BDT, existing ND280 PID methods were applied to the particle gun testing sample following the same pre-selection cuts used for the BDT, and their performance for selecting each particle type was measured in the same ways as that of the BDT. The signal and background efficiencies of the existing PID methods were computed as a function of p_{reco}, and compared to those of the BDT PID as shown in Figures 6.33 and 6.34. Although the event selection scenarios for which the existing PID methods are designed have different kinematic distributions to the particle gun sample, the selected events are often binned by the reconstructed momentum of the primary track, so the efficiencies as a function of p_{reco} can be usefully compared to assess the momentum-sensitivity of the performance. Similarly, although the proportions of particle types differ between the particle gun samples and the selection use cases, by considering the relevant background efficiencies as well as the signal, a useful comparison can be made. For each particle selection, the performance of the existing PID was compared to that of the BDT preference and/or an example cut on the relevant BDT output, depending on which made the most appropriate comparison (this will be explained in greater detail below in each case). Furthermore, the ROC curves shown in Figure 6.35 were compared to equivalent ROC curves for the TPC likelihoods, as well as the efficiency and purity values obtained with existing PID cut flows.

For μ^+ selection, the BDT was compared to the primary PID from the $\bar{\nu}_\mu$ CC-inclusive selection in both its existing form ($L_\mu > 0.1$ and $L_{\mathrm{MIP}} > 0.9$ if $p_{reco} < 500$ MeV/c) and the improved version described in Chapter 4 (the existing cuts plus the ECal cut $E/L < 8.8$ MeV/cm). Pion and (to a lesser extent) proton rejection were found to be major issues for this PID in the CC1pi selection, so a focus is placed on these backgrounds; positron contamination is less of an issue so is not considered in this case. Figure 6.36 shows the signal and background efficiencies for μ^+ selection as a function of p_{reco}, comparing the existing and improved $\bar{\nu}_\mu$ selection primary PID to the BDT preference, as well as a tighter cut on the BDT μ-like output (> 0.8) to illustrate how higher purity can be obtained. It can be seen that the BDT preference cut greatly improves pion and proton rejection with little efficiency loss, and the tighter cut selects a much purer sample with moderate efficiency loss. Additionally, the BDT PID efficiencies show much less sensitivity to the momentum than those of either conventional PID. Further comparisons are made in Figure 6.37, which shows the ROC curves of efficiency and purity for cuts on the BDT μ-like output and the TPC μ likelihood, as well as the efficiency and purity yielded by the conventional primary PID cuts. The ROC curve for the BDT output lies well above that of the TPC likelihood and both points representing the conventional PID cuts, showing that a cut on the BDT output offers much higher purity for the same efficiency (and vice versa). These results show that the BDT PID greatly outperforms these existing PID methods at selecting μ^+ tracks.

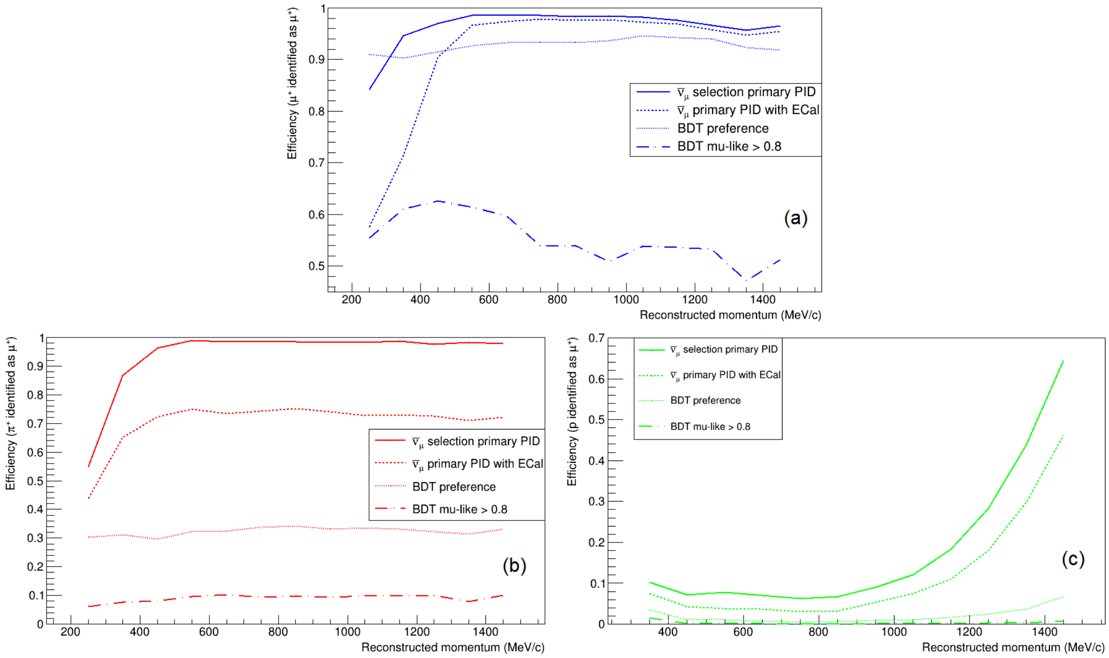


Figure 6.36: Particle identification efficiencies for μ^+ selection as a function of the reconstructed momentum, comparing conventional $\bar{\nu}_\mu$ selection primary PID methods to cuts on the BDT output. The efficiencies for correctly identifying μ^+ (a) are shown, along with those for misidentifying π^+ (b) and protons (c) as μ^+. Although the existing PID offers high signal efficiency (except in the lowest-momentum bin), it fails entirely to reject π^+ at all but the lowest momenta. Above 500 MeV/c, the improved conventional PID rejects $\sim 30\%$ of π^+ while maintaining signal efficiency very close to that of the original, but at lower momenta, the signal efficiency suffers. Both conventional PIDs offer similarly poor proton rejection above 1200 MeV/c. The BDT preference yields slightly reduced μ^+ efficiency, but successfully rejects $\sim 70\%$ of π^+ across the momentum spectrum and improves proton rejection to a similar extent. The tight BDT cut, on the other hand, rejects $\sim 90\%$ of π^+ at the cost of $\sim$ 40–50% of μ^+ efficiency, and rejects at least 98% of protons at all momenta.

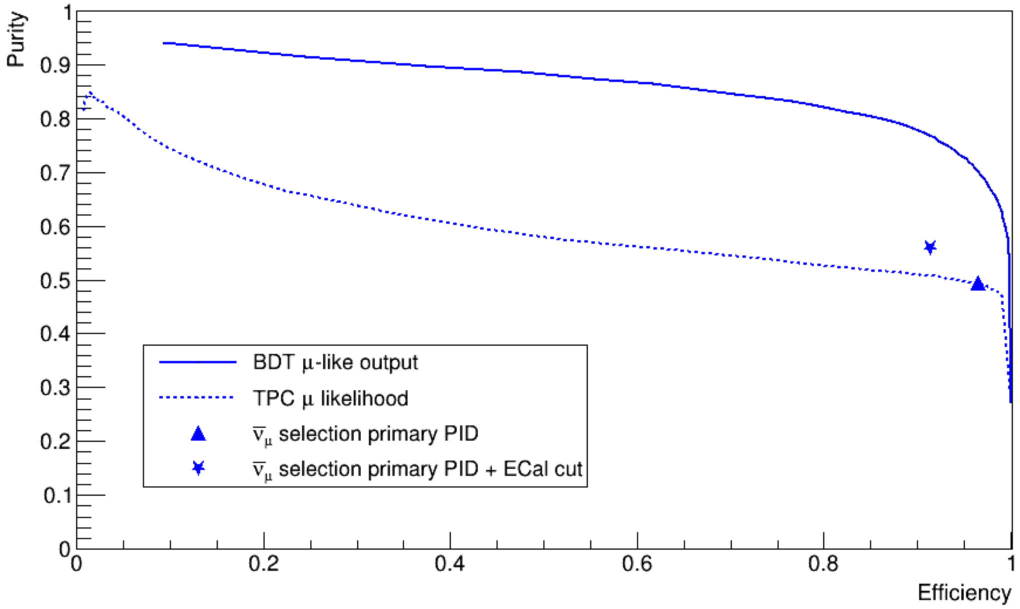


Figure 6.37: ROC curves for selecting μ^+ from the PG testing sample, comparing the efficiency and purity obtained with a range of 50 cut points on the BDT μ-like output (solid line) and TPC μ likelihood (dashed line). The specific values of efficiency and purity obtained with the existing and improved $\bar{\nu}_\mu$ selection primary PID cuts are also shown as individual points. It can be seen that the curve for the BDT PID lies well above that of the TPC likelihood and the points representing the conventional selection PID. This shows that, for a given efficiency, the μ^+ purity that can be obtained with the BDT PID is significantly higher than with the TPC likelihood in all cases. The point for the $\bar{\nu}_\mu$ selection primary PID lies on the curve of the TPC likelihood, since it consists entirely of cuts on the TPC likelihood variables. The addition of the ECal cut increases the performance somewhat with respect to the TPC curve, but this improvement is much smaller than that offered by the BDT PID.

For π^+ selection, the BDT preference was compared to the existing secondary track PID from the $\bar{\nu}_\mu$ CC multiple pion selections and the preference of the TPC likelihoods (i.e. taking the largest of the likelihoods as the PID decision). The secondary pion PID is itself a TPC likelihood preference cut, albeit one that does not consider the μ-like hypothesis; that is, a track is identified as a π^+ if $L_\pi > L_p$ and $L_\pi > L_e$. Figure 6.38 shows the signal and background efficiencies for π^+ selection as a function of p_{reco}, comparing the existing secondary pion PID to the BDT preference, as well as the TPC preference with all hypotheses considered. It can be seen that the BDT preference far outperforms the conventional PID at rejecting μ^+ tracks, removing over 90% across the momentum range, at only moderate π^+ efficiency cost. Proton rejection is also improved, and the efficiencies again appear less momentum-dependent for the BDT than for the conventional PID. The ROC curves for cuts on the BDT π-like output and the TPC π likelihood are compared in Figure 6.39. The ROC curve for the BDT output again lies well above that of the TPC likelihood, demonstrating the superior performance of the former. These results show that the BDT outperforms the conventional TPC PID for selecting π^+, and together with the results for μ^+ selection, demonstrates greatly improved muon-pion discrimination. Based on this, the BDT PID can be expected to greatly reduce the wrong-sign background in the $\bar{\nu}_\mu$ CC1pi selection (as is investigated directly in Chapter 7).

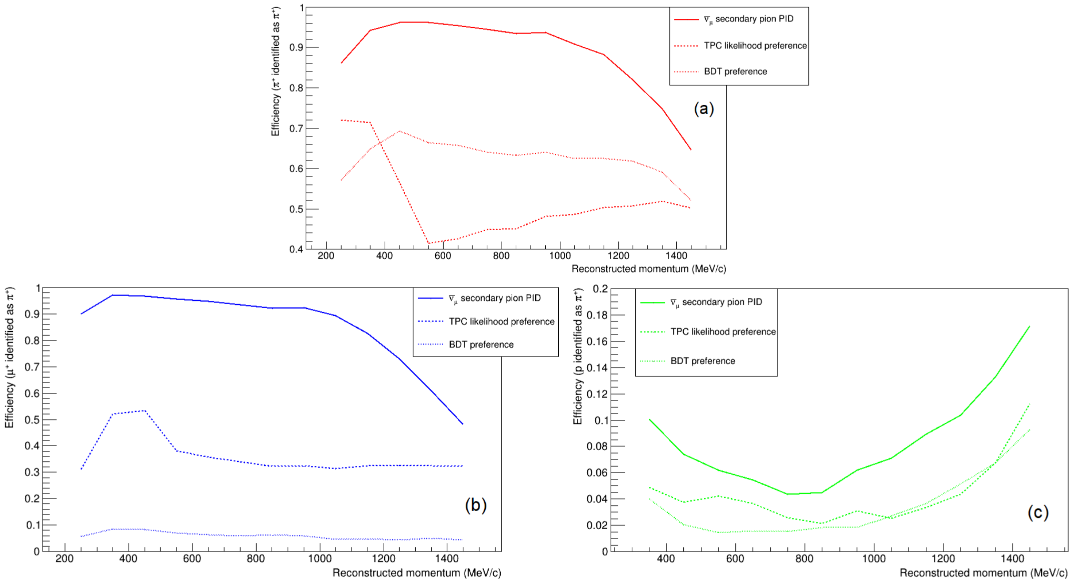


Figure 6.38: Particle identification efficiencies for π^+ selection as a function of the reconstructed momentum, comparing the results for the conventional $\bar{\nu}_\mu$ selection secondary PID and the TPC likelihood preference to the BDT PID preference. The efficiencies for correctly identifying π^+ (a) are shown, along with those for misidentifying μ^+ (b) and protons (c) as π^+. Although the BDT preference only offers about two-thirds the signal efficiency of the existing secondary pion PID, it far outperforms both conventional options in background rejection. Whereas the secondary PID does not attempt to reject muons, the BDT preference does so with over 90% efficiency at all momenta. The BDT preference also suffers less signal efficiency drop-off at higher momenta, and rejects protons with significantly greater efficiency than the secondary PID (and slightly greater than the TPC preference).

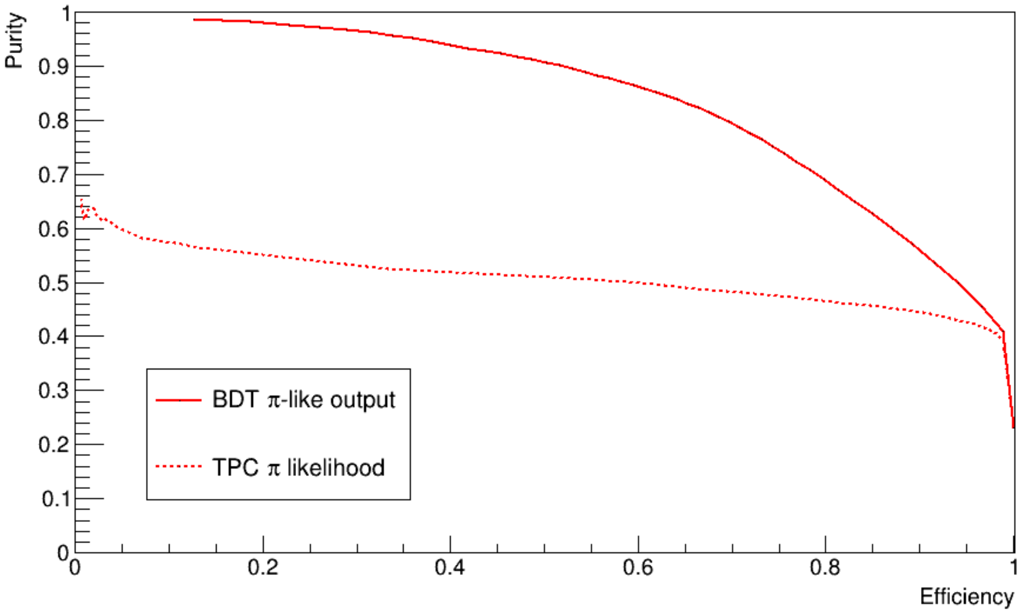


Figure 6.39: ROC curves for selecting π^+ from the PG testing sample, comparing the efficiency and purity obtained with a range of 50 cut points on the BDT π-like output (solid line) and TPC π likelihood (dashed line). The TPC likelihood shows very little purity improvement as the efficiency is reduced, and the highest purity obtainable barely passes 60%, showing that a large proportion of backgrounds cannot be removed at all by a cut on this variable. By contrast, the BDT PID shows much higher purity at all but the highest efficiency levels, increasing to near 1 as the efficiency is reduced.

Figure 6.40 shows the signal and background efficiencies for proton selection as a function of p_{reco}, comparing the $\bar{\nu}_\mu$ selection secondary PID (a cut on the TPC proton likelihood $L_p > 0.5$) and the TPC and BDT preferences similarly to the π^+ selection testing. It can be seen that the BDT preference offers significantly better signal efficiency, as well as similar or better rejection for each background, and again the BDT efficiencies shows less momentum dependence. The ROC curves for cuts on the BDT proton-like output and the TPC proton likelihood are compared in Figure 6.41, and again the curve for the BDT PID is seen to lie well above that of the TPC likelihood. These plots demonstrate that the BDT PID offers greatly improved identification of proton tracks, and could therefore be of value to event selections for which the number of protons is important, such as $\bar{\nu}_\mu$ CC one-pion-one-proton.

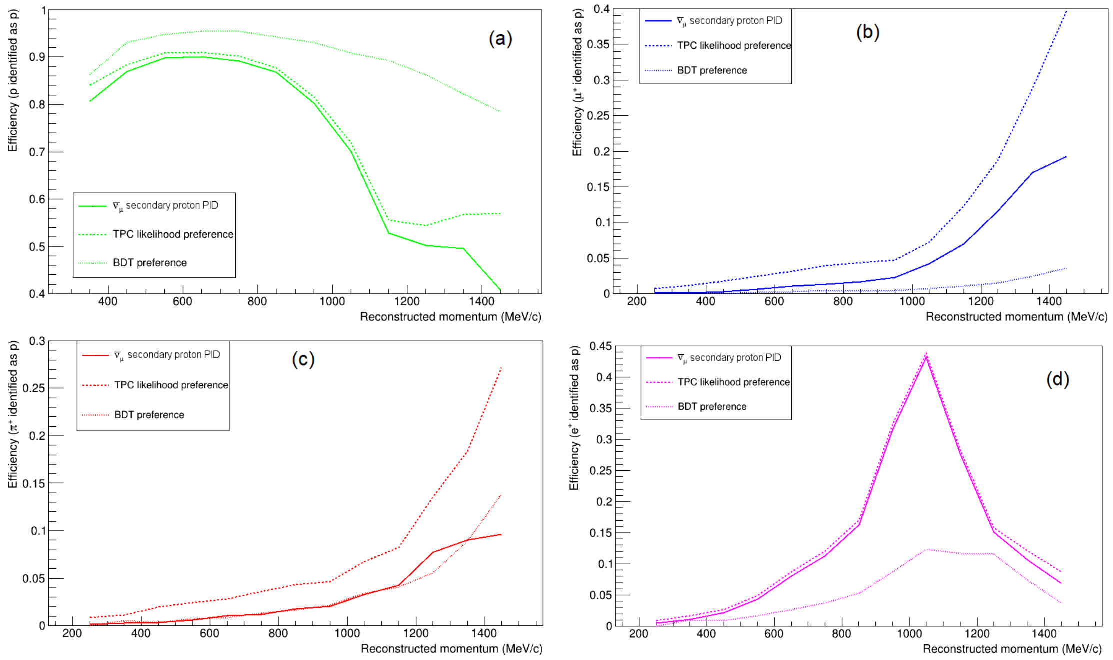


Figure 6.40: Particle identification efficiencies for proton selection as a function of the reconstructed momentum, comparing the results for the conventional $\bar{\nu}_\mu$ selection secondary PID and the TPC likelihood preference to the BDT PID preference. The efficiencies for correctly identifying protons (a) are shown, along with those for misidentifying μ^+ (b), π^+ (c) and e^+ (d) as protons. The BDT preference offers greater signal efficiency across the full momentum range, especially at at momenta above 1 GeV/c where there is a significant drop-off in the efficiency of the conventional options, most likely due to the proton dE/dx curve crossing those of μ and π. The BDT preference also rejects μ^+ and e^+ much more efficiently than either conventional option, and has similar π^+ rejection efficiency to the existing secondary PID.

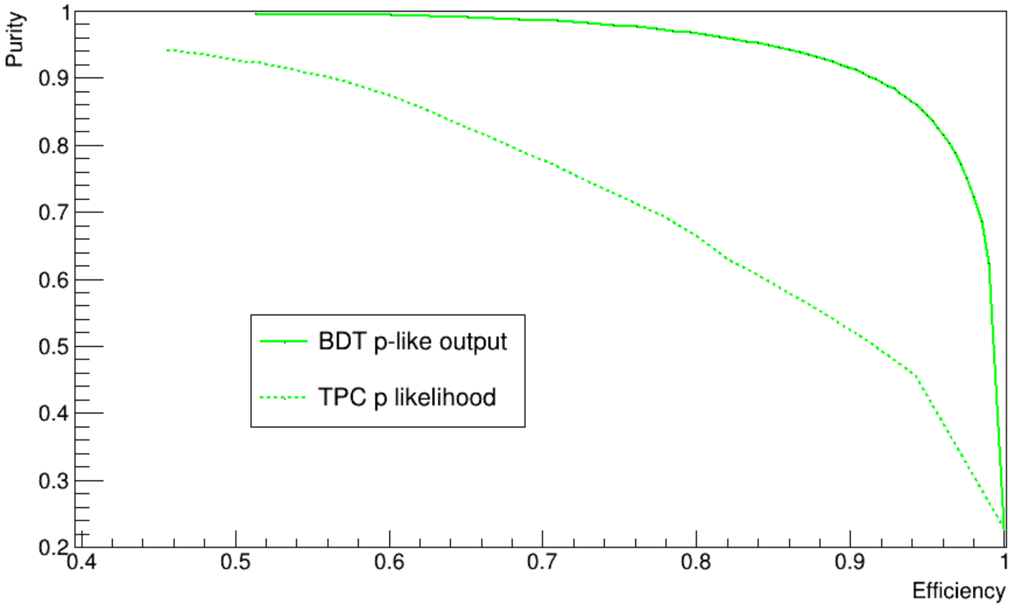


Figure 6.41: ROC curves for selecting protons from the PG testing sample, comparing the efficiency and purity obtained with a range of 50 cut points on the BDT p-like output (solid line) and TPC p likelihood (dashed line). The curve for the BDT output lies well above that of the TPC likelihood, showing that the BDT PID offers much greater purity for the same efficiency (and vice versa).

For e^+ selection, the BDT was compared to the primary PID from the $\bar{\nu}_e$ CC-inclusive selection (see 6.1.2 for the relevant cuts). Figure 6.42 shows the signal and background efficiencies for e^+ selection as a function of p_{reco}, comparing the performance of the existing $\bar{\nu}_e$ primary PID to that of the BDT PID. This selection requires very efficient rejection of μ^+ and protons, so the BDT preference is not a suitable comparison; instead, it is compared to tight cuts on the BDT output (> 0.8 and > 0.9). It can be seen that the BDT PID offers both higher signal efficiency and better background rejection than the existing PID, particularly above 600 MeV/c, where the existing PID incorporates tighter cuts in an attempt to reject protons, resulting in a severe drop in signal efficiency (and a large peak of proton contamination can still be seen). By contrast, the BDT PID offers several times better proton rejection and suffers no such drawback, retaining high e^+ efficiency across the momentum range. This impression is reinforced by Figure 6.43, which compares the ROC curves for cuts on the BDT e-like output and the TPC e likelihood as well as the efficiency and purity obtained with the existing $\bar{\nu}_e$ selection primary PID. The BDT

PID again compares very favourably to the existing primary PID cuts: at the same purity of $\sim 92\%$, the efficiency of the existing cuts is $\sim 50\%$, whereas that of the BDT PID is much higher at $\sim 98\%$. For the same efficiency, the purity with the BDT PID is close to 1. These results indicate that the BDT PID could significantly improve both the efficiency and purity of the $\bar{\nu}_e$ CC-inclusive selection.

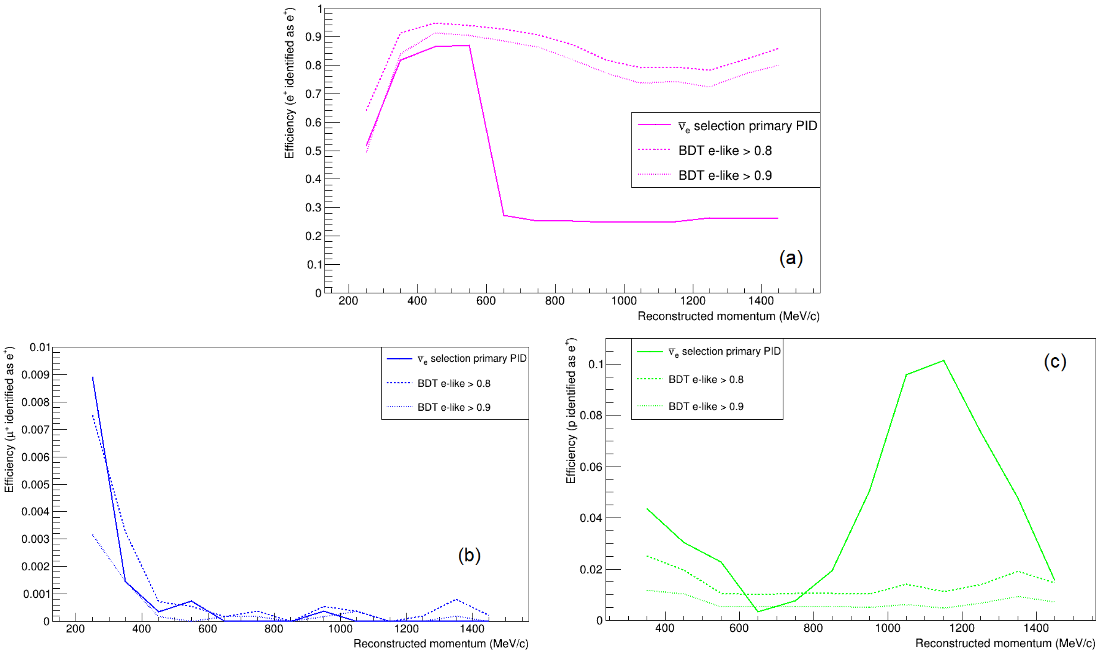


Figure 6.42: Particle identification efficiencies for e^+ selection as a function of the reconstructed momentum, comparing the results for conventional $\bar{\nu}_e$ selection primary PID to cuts on the BDT output. The efficiencies for correctly identifying e^+ (a) are shown, along with those for misidentifying μ^+ (b) and protons (c) as e^+. Below 600 MeV/c, the BDT cuts at 0.9 and 0.8 offer similar and slightly better signal efficiency respectively compared to the $\bar{\nu}_e$ primary PID. Above 600 MeV/c, the efficiency of the $\bar{\nu}_e$ primary PID drops sharply, whereas that of the BDT remains high. This drop is due to the differences in cuts in the $\bar{\nu}_e$ primary PID above and below 600 MeV/c: above 600 MeV/c, tighter cuts are applied in an attempt to reduce the proton background, but this also results in a loss of signal efficiency. For μ^+ rejection, the the BDT cut at 0.9 offers better performance than the existing PID; for proton rejection, both BDT cuts outperform the existing PID at almost all momenta, particularly around 1.1 GeV/c.

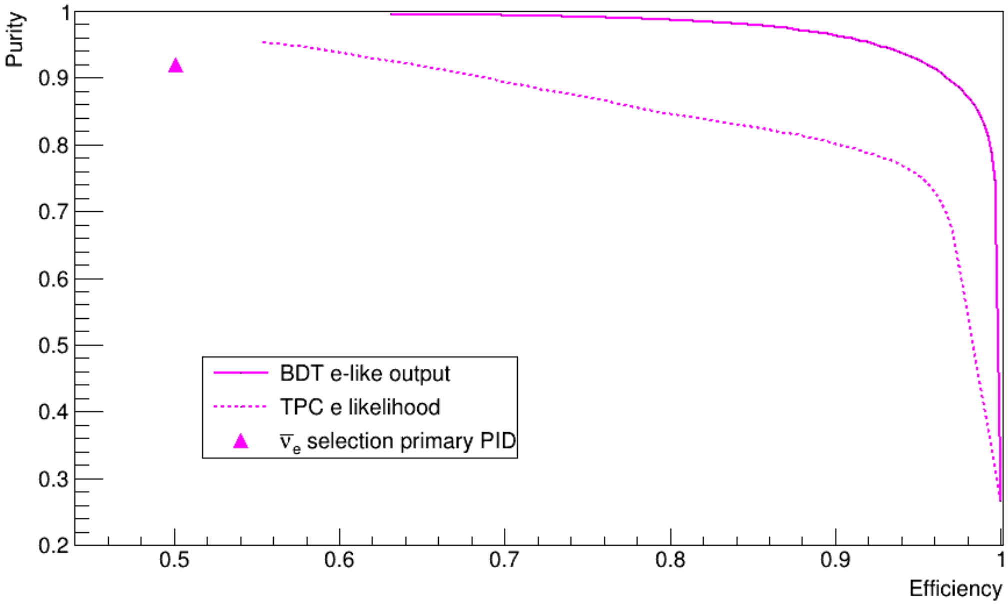


Figure 6.43: ROC curves for selecting e^+ from the PG testing sample, comparing the efficiency and purity obtained with a range of 50 cut points on the BDT e-like output (solid line) and TPC e likelihood (dashed line). The specific values of efficiency and purity obtained with the existing $\bar{\nu}_e$ selection primary PID cuts are also shown as an individual point. The curve for the BDT output lies well above that of the TPC likelihood, showing that the BDT PID offers greater purity for the same efficiency (and vice versa). The curve for the BDT output also lies particularly far to the right of the point representing the $\bar{\nu}_e$ selection primary PID, showing that the BDT PID can achieve similar (or indeed higher) purity while retaining much higher signal efficiency.

Overall, the BDT appears to substantially outperform the existing PID methods tested. In each particle selection case, the BDT preference or a single-output cut offers significantly improved signal efficiency and/or background rejection, while also exhibiting less momentum dependence than the existing methods. Furthermore, the ROC curves show that cuts on the BDT outputs can achieve much greater efficiency and purity than is possible with the TPC likelihoods or the existing $\bar{\nu}_\mu$ and $\bar{\nu}_e$ selection primary PID cuts.

6.6.2 Overtraining

Overtraining is an important factor to consider in applications of machine learning. Since the BDT was trained on a limited sample of simulated tracks, it may have adapted to statistical fluctuations in the training dataset. Although steps were taken to avoid this (keeping the TreeDepth and MinNodeSize parameters within the ranges recommended by TMVA) it is still important to evaluate the level of overtraining in the final BDT configuration. This can be done by comparing the performance of the BDT when applied to the statistically-independent training and testing samples: the more overtraining has occurred, the greater the difference in performance will be.

Initially, this was only tested by comparing the BDT preference signal efficiencies for the training and testing samples as in Figure 6.44. For each subsample, the efficiencies appear largely similar, though that of the training sample is generally somewhat higher for μ^+ and π^+, indicating a small amount of overtraining. There is also a larger discrepancy for protons at low momentum ($p_{reco} < 600$ MeV/c), increasing as the momentum decreases, up to $\sim 10\%$ for the lowest momentum bin. This is likely due to the low statistics of the proton training sample at low momenta, even with the 'padding' described in Section 6.3.1. However, the low detector efficiency for low-momentum protons also means that the tool would rarely be applied to proton tracks with such low momenta, so the overtraining in this region should have little effect in practice. On the basis of this plot, the overtraining of the BDT was initially judged to be minimal, so no further steps were taken to reduce it before its use in the performance testing above and in Chapter 7. However, with subsequent investigation, the degree of overtraining became more evident. Figure 6.45 compares the ROC curves for the BDT outputs (as in Figure 6.35) between the training and testing samples. The effect of overtraining can be clearly seen when selecting μ^+ with the BDT μ-like output: the performance is consistently higher for the training sample than the testing sample. This impression is confirmed by the results of a two-sample Kolmogorov-Smirnov (K-S) test [108][109], which estimates the probability that two samples come from the same distribution. The ROOT [69] implementation of the two-sample K-S test was performed on the distributions of each BDT output for the training and testing samples. The results are shown in Table 6.3. Since the Kolmogorov probability is close to zero in all cases, this test indicates that the training and testing sample BDT outputs are not consistent (i.e. very unlikely to be drawn from the same distribution) and hence that overtraining has occurred despite the precautions taken in setting the training parameters.

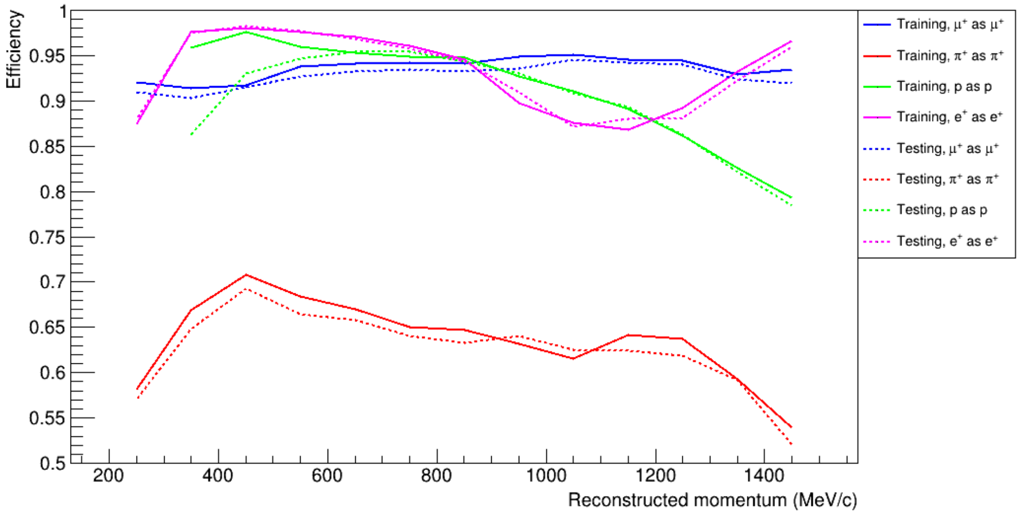


Figure 6.44: Particle identification 'signal' efficiencies of the final BDT configuration preference cuts as a function of the reconstructed momentum, comparing the results for the training and testing samples, denoted by solid and dashed lines respectively. Some evidence of overtraining can be seen where the performance is higher for the training sample than the testing sample, particularly at low proton momenta (likely due to the low statistics and hence larger weights for individual events in this region).

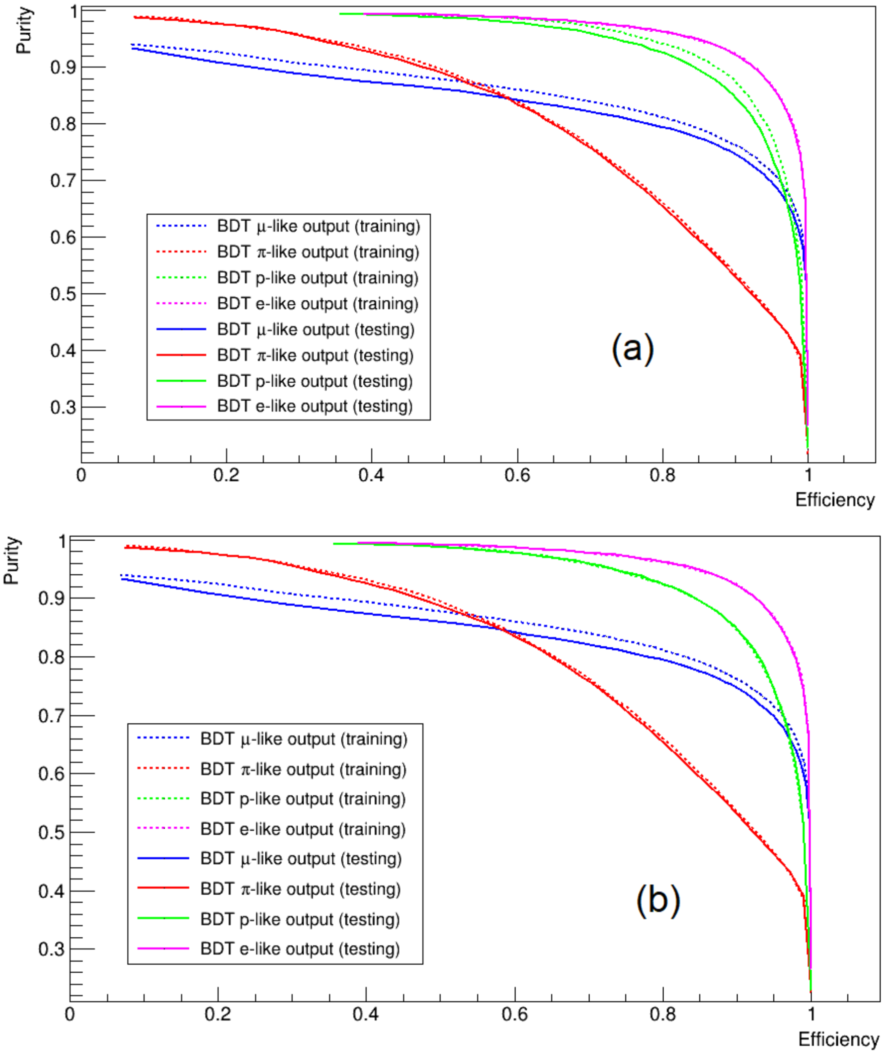


Figure 6.45: ROC curves for selecting each particle type via a cut on the corresponding BDT output as in Figure 6.35, comparing the results for the training and testing samples. Versions are shown with (a) and without (b) the extra low-momentum protons in the training sample; the testing sample does not contain such protons, so in the latter case the samples are more consistent. The effect of overtraining can be seen in the higher μ^+ selection performance with the training sample compared to the testing sample in both plots. Higher training sample performance is also seen for protons in (a), but not in (b), so this may be an artifact of the low-momentum protons rather than overtraining.

BDT output	**Kolmogorov probability for training and testing samples**	
	Full training sample	**Extra low-mom. protons removed**
μ-like	1.84452×10^{-11}	2.8786×10^{-12}
π-like	4.06449×10^{-09}	7.46376×10^{-10}
p-like	1.18527×10^{-31}	1.3937×10^{-33}
e-like	2.54713×10^{-09}	1.05733×10^{-09}

Table 6.3: Two-sample Kolmogorov-Smirnov test results obtained by applying the ROOT [69] `TMath::KolmogorovTest` function to the BDT output distributions of the training and testing samples. Versions of the training sample with and without the extra low-momentum protons have been tested, since the latter case should be more consistent with the testing sample (which does not contain extra low-momentum protons). The Kolmogorov probability estimates the probability that both samples are drawn from the same distribution.

Since the discrepancy between the training and testing samples in Figures 6.44 and 6.45 appears relatively small, the impact of this overtraining on the BDT is likely also small. Nevertheless, the overtraining should be removed in future development if possible. The most obvious way to do this would be to choose more conservative values of certain BDT training parameters known to help eliminate overtraining: larger values of MinNodeSize and/or smaller values of TreeDepth should be tested. Indeed it is possible that eliminating the overtraining may improve the classification performance of the BDT somewhat.

6.7 Conclusions

Despite some overtraining, the BDT PID outperforms the existing conventional PID methods in nearly all cases tested above. Thus it appears that the development goal of 'versatility' has been achieved: a single PID tool has been developed which nonetheless outperforms the existing PID methods in a variety of use cases, often very significantly and despite the latter having been developed and optimised specifically for those use cases. These results illustrate the power of multivariate methods to make efficient use of the large numbers of PID variables recorded by ND280, and the limitations of conventional methods by comparison, and strongly motivate the use of a global BDT PID tool in ND280 event selections. This will be further demonstrated in the following chapter, in which the BDT PID tool is applied directly within a $\bar{\nu}_\mu$ CC1pi selection.

Chapter 7

Muon-antineutrino CC1pi selection with BDT PID

In this chapter, the finalised BDT PID tool is applied within a $\bar{\nu}_\mu$ CC1π^- event selection in place of conventional PID methods, and the performance is compared with the selections described in Chapter 4. Versions of the $\bar{\nu}_\mu$ CC1π^- selection that use the BDT PID will be referred to collectively as 'BDT selections'. Three BDT selection variants were tested to evaluate the effect of different ways the BDT PID may be applied, and thus determine which will yield the best selection performance.

As outlined in the previous chapter, the BDT preference may be used to classify tracks, taking the hypothesis corresponding to the highest of the outputs as the PID decision. Depending on the class of track, better performance may be obtained by excluding certain hypotheses: for example, we do not expect to detect antiprotons, so the proton-like output can be ignored for negative tracks. Although we do expect large numbers of μ^- due to the wrong-sign background, true π^- that do not shower in the ECal will closely resemble them, so rejecting muon-like negative tracks may not be desirable — a large proportion of π^- from signal events will be rejected. To examine this, selections were prepared with and without the muon-like hypothesis included when evaluating the preference for negative tracks, and their performance is compared below.

Additionally, depending on the relative signal and background rates for the selection in question, cuts on individual BDT outputs may substantially outperform the preference provided a suitable cut value is chosen. To test this for the $\bar{\nu}_\mu$ CC1π^- selection, a pair of cuts on the μ-like and π-like BDT outputs for the muon and pion candidate tracks respectively have been optimised and applied in a further BDT selection variant.

7.1 Track identification by BDT preference

The BDT selections presented in this chapter use the same starting (‘pre-selection’) cuts as in the ‘improved’ selection as described in Chapter 4:

- **Event quality**
- **Total multiplicity**
- **Track quality and fiducial**
- **Upstream background veto**
- **Broken track**

For antimuon candidate PID, the loop over positive tracks is used, with the TPC likelihood cuts replaced by a cut on the output values of the BDT. For the initial comparisons, the preference was used (that is, identifying a positive track as a μ^+ if the BDT muon-like output is the highest of the four). A version with optimised cuts to select μ^+ and π^- tracks is presented later in this chapter.

For PID of secondary tracks, the BDT preference replaces the decision flow described in Section 4.2.1. For positive tracks, the pion, proton and positron hypotheses are considered (since additional μ^+-like tracks are rejected as part of the primary PID loop), with positrons being considered evidence of π^0 as in the existing selection. For negative tracks, two versions were tested: a ‘full’ version which attempts to identify and reject μ^- tracks, and one which does not (since a large proportion of π^- will appear MIP-like and therefore muon-like). In the former case, the muon, pion and electron hypotheses are considered; in the latter, only the pion and electron hypotheses are. Having identified each secondary track, single-pion events are then selected using the same criteria as in the **one pion cut**. No further PID cuts are applied.

To properly assess the performance of the BDT PID, its kinematic region of validity should be considered. The BDT can be expected to more reliably identify tracks with reconstructed momentum and angle within the range provided in the training phase (200 MeV/c $< p_{reco} <$ 1500 MeV/c, $\theta_{reco} < 60°$). To account for this when comparing the performance of the BDT selections to that of the conventional selections, a **BDT validity cut** was added to each of them, selecting events in which both the μ^+ and π^- candidate tracks satisfy the validity criteria.

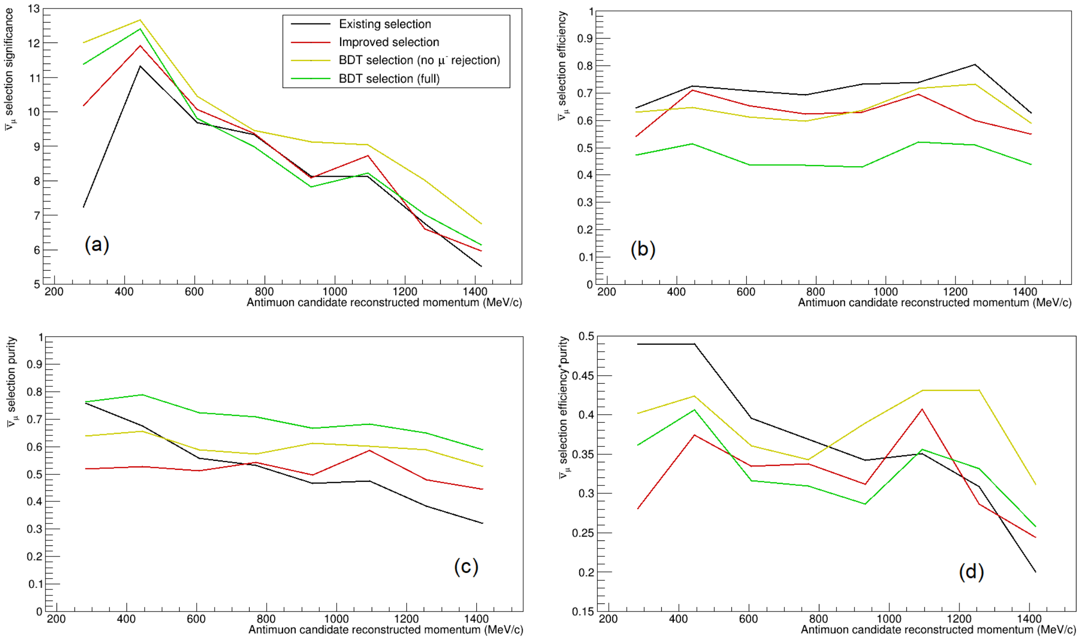


Figure 7.1: Significance (a), efficiency (b), purity (c), and efficiency*purity (d) for $\bar{\nu}_\mu$ CC1π^- event selections as a function of antimuon candidate momentum. The existing (black) and improved (red) selections with conventional PID are compared to selections using the BDT PID tool, with (green) and without (yellow) considering the μ^- hypothesis for negative tracks. The kinematics of the μ^+ and π^- candidate tracks have been limited to the region of validity of the BDT.

The selection significance ($S/\sqrt{S+B}$) is again taken as the primary performance metric, though the efficiency, purity and their product are also taken into consideration. Each of these is plotted as a function of the μ^+ candidate p_{reco} in Figure 7.1. The BDT selection without μ^- rejection offers the best significance, outperforming the other options across the momentum spectrum. Although the version with μ^- rejection has the greatest purity of the selections tested, it also has the lowest efficiency, resulting in a significance mostly similar to that of the conventional selections. It is interesting to note that the existing selection has the highest efficiency*purity at momenta below 900 MeV/c while the same is not seen in the significance; this is because the efficiency*purity is not sensitive to the overall number of events selected, whereas the significance is. These results show that BDT PID can be used to build event selections that outperform conventional ones, even without optimisation, provided that the PID is applied to tracks with kinematics within those of the training data set.

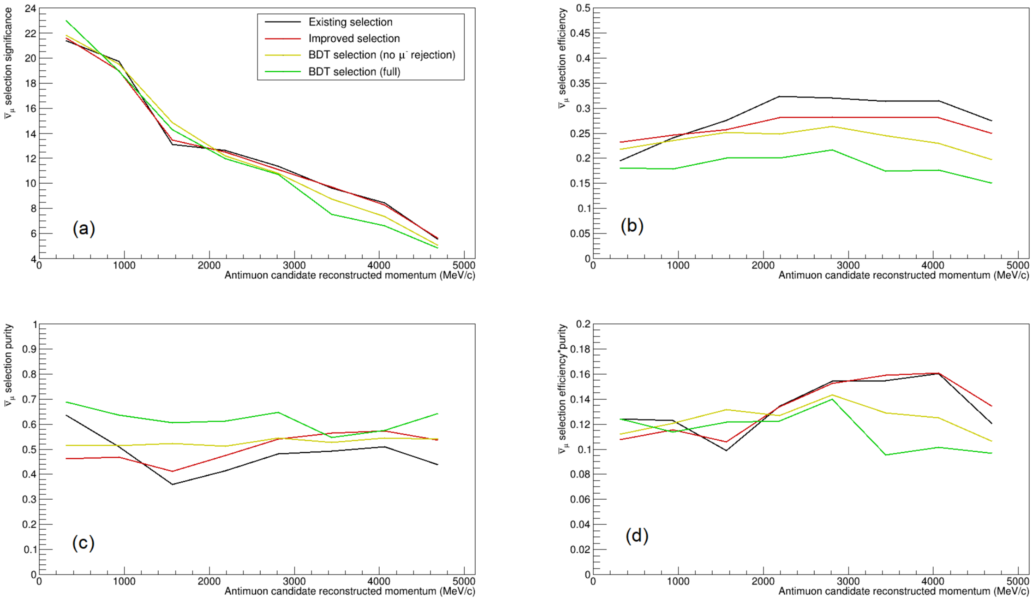


Figure 7.2: Significance (a), efficiency (b), purity (c), and efficiency*purity (d) for $\bar{\nu}_\mu$ CC1π^- event selections as a function of antimuon candidate momentum, as in Figure 7.1 but without restrictions on the track kinematics for BDT validity (i.e. removing the BDT validity cut).

The importance of kinematics for BDT validity is demonstrated by Figure 7.2, which shows the same metrics as in Figure 7.1 but without the BDT validity cut applied. The significance of the BDT selections is considerably lower here, often similar or worse than the conventional selections. Although an improvement in purity can still be seen, it is offset by the loss of efficiency, resulting in poor overall performance. This indicates that the BDT PID does not perform well when applied to tracks that do not meet its validity criteria. For practical applications within event selections, a larger training data sample with broader kinematic distributions may be needed, in order to train the BDT on a set of tracks with momenta and angles reflecting the full range of tracks to which we will wish to apply PID.

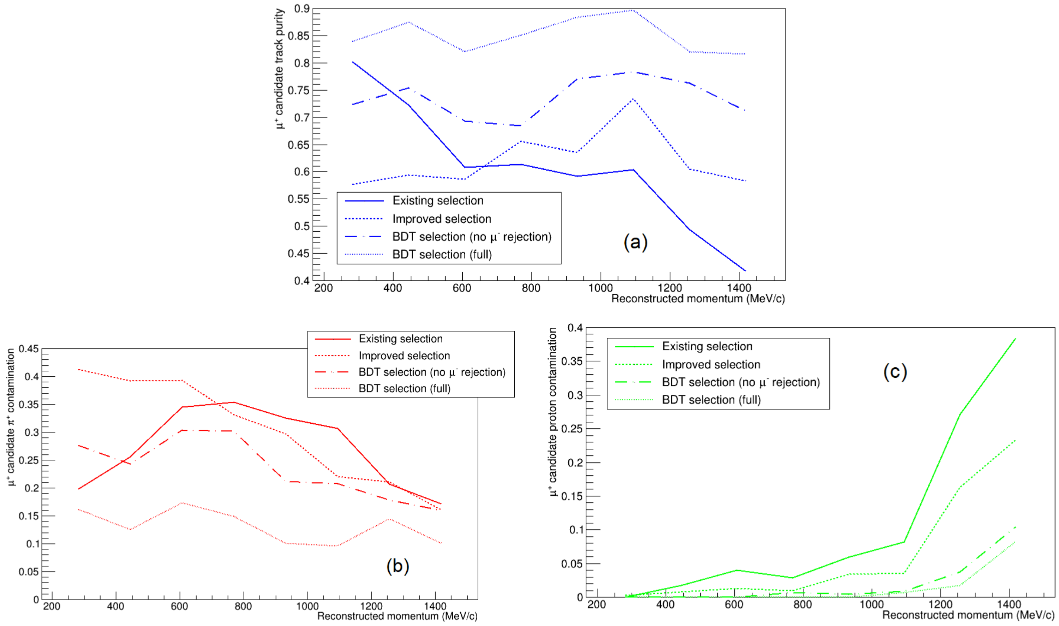


Figure 7.3: Purity of μ^+ (a) and contamination of π^+ (b) and protons (c) as a function of reconstructed momentum for antimuon candidates in $\bar{\nu}_\mu$ CC1π^- event selections. The kinematic restrictions for BDT validity have been applied to both the μ^+ and π^- candidate tracks for the events shown.

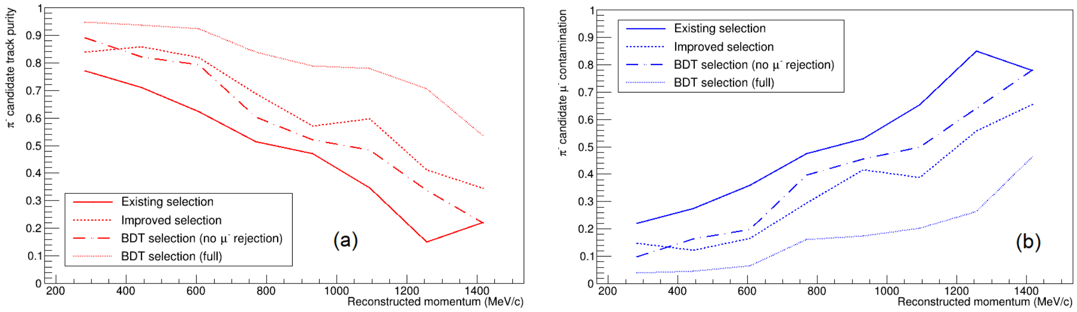


Figure 7.4: Purity of π^- (a) and contamination of μ^- (b) as a function of reconstructed momentum for negative pion candidates in $\bar{\nu}_\mu$ CC1π^- event selections. The kinematic restrictions for BDT validity have been applied to both the μ^+ and π^- candidate tracks for the events shown.

The purities and main background contaminations of the antimuon and pion candidate tracks for each selection version are shown in Figures 7.3 and 7.4 respectively. Broadly speaking, the ‘full’ BDT selection has the highest track purities, followed by the BDT selection without muon rejection and the improved conventional selection, and the existing selection has the lowest. The improved selection offers greater π^- track purity than the BDT selection without muon rejection because it includes an ECal cut to reject μ^-, but as noted previously, this is offset by the greater efficiency of the BDT selection. The background contaminations can all be seen to be lower with the BDT selection than the existing selection, with the exception of the π^+ contamination at low momentum ($\sim$ 300 MeV/c), but this is most likely due to the absence of the leading track cut in the BDT selections rather than the performance of the PID itself. These comparisons show improvements of as much as a factor of 2 (μ^+ purity) and 5 (π^- purity), demonstrating that the BDT PID can greatly improve the accuracy with which the main tracks in this selection are identified.

7.2 Optimised cuts

To improve selection performance further, optimal cuts on the BDT outputs for the antimuon and pion candidate tracks can be found and applied. To demonstrate this, a simple grid search was performed across the space of cuts on the μ^+ candidate μ-like output and the π^- candidate pi-like output, here referred to as B_μ and B_π respectively. The same pre-selection from the previous section (event quality to broken track cut) was used, with an additional requirement that selected events have only one positive and one negative TPC track (**two-track cut**) since allowing proton tracks would complicate the process. The BDT output cuts were tested with 20 different values each in increments of 0.02, ranging between 0 (no cut) and 0.98, for a total of 400 cut pairs. For each point in the cut space the significance was computed, and the point with the highest significance overall was taken as the optimised cut pair. This is visualised in Figure 7.5, and yields a cut pair of $B_\mu > 0.26$, $B_\pi > 0.12$. It can be seen that looser cut values are favoured, particularly for B_π, and cut values approaching 1 result in a sharp drop-off in significance.

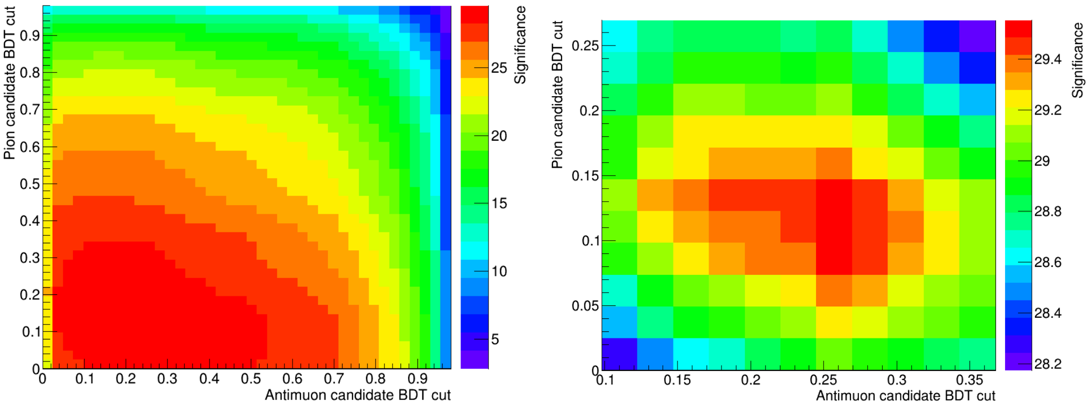


Figure 7.5: Optimisation plot for cuts on the BDT outputs of the CC1π^- selection antimuon and pion candidates, showing the full cut space (left) and the peak region (right). The significance ($S/\sqrt{(S+B)}$) is shown on the Z-axis as a function of cuts on the antimuon candidate μ-like output and the pion candidate π-like output, rejecting events below the cut value. The optimal cut point is found for a cut pair of $B_\mu > 0.26$, $B_\pi > 0.12$, with a significance of 29.6.

The selection comprising the pre-selection cuts and these optimised cuts on the BDT outputs will be referred to as the 'optimised BDT selection'. The performance of this selection is compared to that of the existing selection and the BDT selection using the preference (without muon rejection) in Table 7.1 and Figure 7.6, with the two-track cut and BDT validity criteria added to all selections. The optimised BDT selection improves performance further compared to the preference selection, yielding equal or better significance, with the difference being greater at lower momentum values. The optimised selection has similar efficiency to the existing selection, while that of the preference selection is lower; and similar purity to the preference selection, while that of the existing selection is lower (except in the lowest momentum bin). The efficiency*purity is here also higher for the optimised selection than the existing selection across the momentum spectrum.

Selection	**Significance**	$\bar{\nu}_\mu$ **CC1**π^- **purity**	**Non-**$\bar{\nu}_\mu$**-CC backgrounds**
Existing	22.7	51.9%	38.8%
BDT (no μ^- rejection)	27.4	66.5%	23.6%
BDT (optimised)	29.6	65.5%	22.2%

Table 7.1: Summary of performance metrics for the existing selection and the preference-based and optimised BDT selections. Only events with one positive and one negative TPC track have been selected, and each required to have kinematics in the region of validity of the BDT.

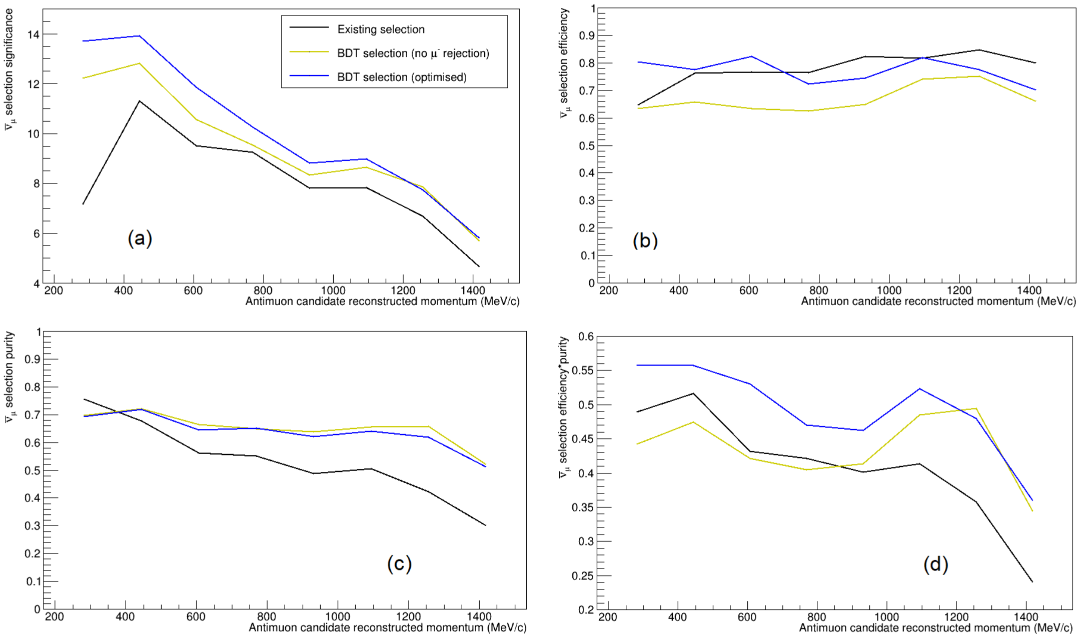


Figure 7.6: Significance (a), efficiency (b), purity (c), and efficiency*purity (d) for $\bar{\nu}_\mu$ CC1π^- event selections as a function of antimuon candidate momentum. The existing (black) selection with conventional PID is compared to BDT selections using the preference (yellow) and the optimised cut pair (blue). Only events with one positive and one negative TPC track have been selected, and each required to have kinematics in the region of validity of the BDT.

Table 7.2 compares the overall track purities similarly to Table 4.3, again with the two-track cut and BDT validity criteria applied. Compared to the existing selection, the BDT selections show an increase from 59.8% to $\sim$ 78% in the μ^+ candidate purity and a particularly striking reduction in the proton contamination, which drops from 12.6% to less than 2%. The purity of π^- candidates is similarly improved. The two BDT selections have similar μ^+ candidate purity and background contaminations, while for π^- candidates the purity is about 5% better in the optimised selection. These track purity improvements combine to yield the increases in overall selection purity seen in Table 7.1, and demonstrate that the BDT PID is making a substantial improvement to the identification of the main tracks in the $\bar{\nu}_\mu$ CC1π^- selection.

Selection	μ^+ **candidate true particle**			π^- **candidate true particle**	
	μ^+	π^+	p	π^-	μ^-
Existing	59.8%	26.7%	12.6%	66.6%	32.2%
BDT (no μ^- rejection)	77.8%	20.3%	1.5%	75.8%	20.3%
BDT (optimised)	77.6%	20.2%	1.8%	79.3%	18.6%

Table 7.2: Summary of the true particle content of the μ^+ and π^- candidate tracks in the existing selection and the preference-based and optimised BDT selections. Only events with one positive and one negative TPC track have been selected, and each required to have kinematics in the region of validity of the BDT.

The potential performance of the BDT PID for this selection is further assessed in Figures 7.7 and 7.8, which show the track identification efficiency and purity of various PID methods for the antimuon and pion candidates respectively. The ROC curves for the respective BDT outputs and and TPC likelihoods are shown, as well points representing the performance of the cuts used in the existing and improved selections. Similarly to comparisons made with the particle gun sample in Chapter 6, it can be seen that the curves for the BDT outputs lie well above those of the TPC likelihoods, showing that the BDT offers track identification performance that is greatly superior to that of TPC PID alone, as expected. In Figure 7.7, the points representing the existing and improved selection μ^+ candidate cuts both lie above the ROC curve for the TPC likelihood, showing the improvements in performance yielded by the leading track cut and the ECal E/L cut respectively; but these points also lie well below the curve for the BDT output, showing that a cut on the BDT PID can greatly outperform both. In Figure 7.8, however, the positions of the points representing the existing and improved selection π^- candidate cuts are more surprising. The point for the existing selection π^- PID lies *below* the TPC likelihood ROC curve, indicating that the leading track cut causes a loss in π^- identification performance rather than a gain. Conversely, the point for the improved selection π^- PID lies *above* the curve for the BDT output, implying that better performance is achieved with simple TPC and ECal cuts than with the BDT! Thus it appears that, while offering much improved PID for the μ^+ candidate, the BDT is actually underperforming for the π^- candidate. It is not immediately clear why this should be the case, but it may indicate that the assumption that the BDT will perform well for negative tracks (despite only having been trained on positive tracks) does not hold, perhaps due to the inclusion of the proton hypothesis. This motivates further testing of the BDT PID performance for negative tracks (for example, with a particle gun sample of μ^-, π^- and e^-). If the performance is found to be generally poorer for such tracks, this could be mitigated by the inclusion of negative tracks in the training sample (as well as the track charge information that would then be necessary) or the separate training of a second BDT specifically for negative tracks.

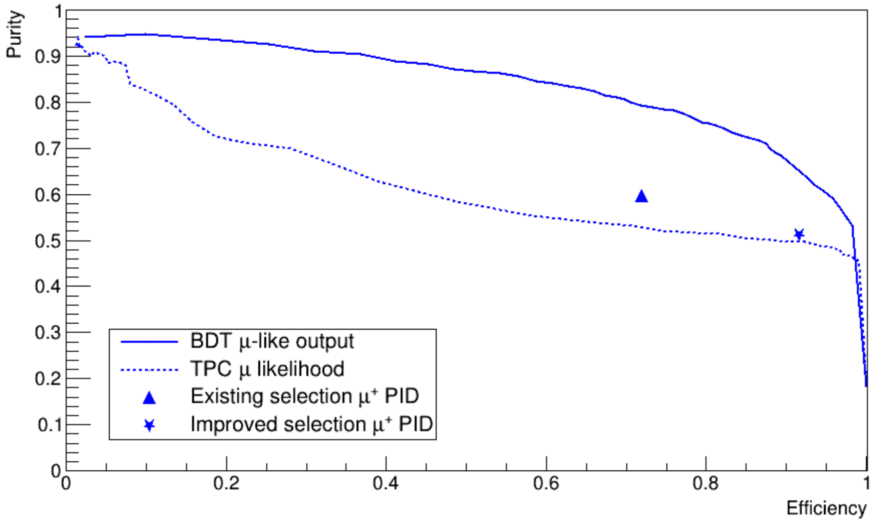


Figure 7.7: ROC curves for PID of the μ^+ candidate in $\bar{\nu}_\mu$ CC1pi selections, comparing the track identification efficiency (proportion of true μ^+ accepted by the preselection that are then accepted by the μ^+ PID) and purity (proportion of tracks accepted by the μ^+ PID that are true μ^+) obtained with a range of 50 cut points. The curves for the BDT μ-like output and TPC μ likelihood are shown, as well as points representing the performance of the μ^+ candidate cuts of the existing selection PID (including the leading track cut) and the improved selection PID. The BDT validity and two-track cuts are applied in all cases.

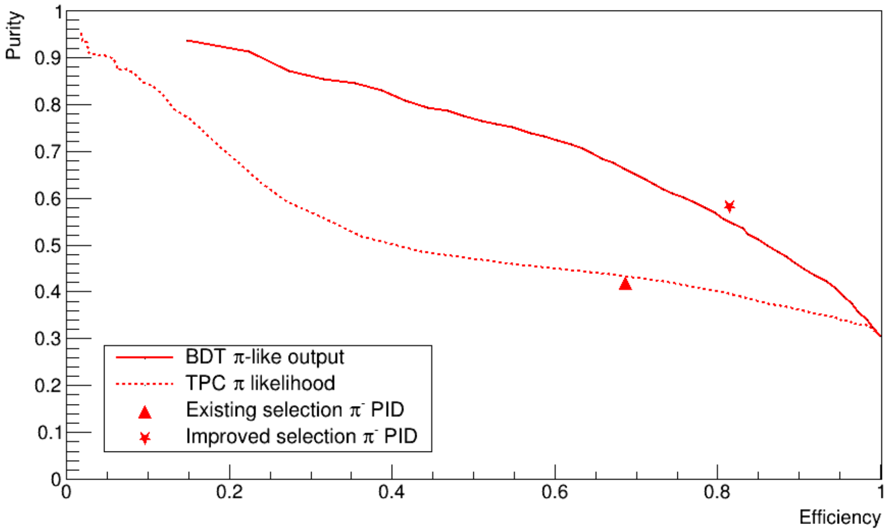


Figure 7.8: ROC curves for PID of the π^- candidate in $\bar{\nu}_\mu$ CC1pi selections, comparing the track identification efficiency (proportion of true π^- accepted by the preselection that are then accepted by the π^- PID) and purity (proportion of tracks accepted by the π^- PID that are true π^-) obtained with a range of 50 cut points. The curves for the BDT π^--like output and TPC π^- likelihood are shown, as well as points representing the performance of the π^- candidate cuts of the existing selection PID (including the leading track cut) and the improved selection PID. The BDT validity and two-track cuts are applied in all cases.

7.3 Comparisons to real data

To verify the above selection performance predictions, the event selections were applied to a sample of real T2K data (see Section 4.1) and the selected events in the data and MC samples were compared to assess their consistency. Comparisons were made between the pre-selection (comprising again the series of cuts quoted in Section 7.1, followed by the BDT validity and two-track cuts), the selection using the existing PID, and the optimised BDT selection.

Selection	**Data events**	**MC events (unscaled)**	**MC events (POT scaled)**
Pre-selection	1136	11041	1113
Existing	220 (19.3%)	1922 (17.4%)	194 (17.4%)
BDT (optimised)	207 (18.2%)	2036 (18.4%)	205 (18.4%)

Table 7.3: Overall numbers of events accepted by each selection, comparing real data and MC. The statistics for MC are given both in full and scaled by POT to the data. The percentages of events selected relative to the pre-selection are also given. A greater data-MC discrepancy is seen for the existing selection: this may be due to greater model-dependency in that selection as a result of the leading track cut.

The overall numbers of events accepted by each selection are summarised in Table 7.3, and appear similar between data and MC when the latter is scaled to the same POT. However these values alone are not very illustrative; the degree of consistency between data and MC can be assessed in more detail from histograms of the event kinematics, which are shown in Figures 7.9 and 7.10. Figure 7.9 shows the distributions of the reconstructed momentum of the μ^+ candidate, which appear largely consistent between data and MC for each selection. The main discrepancy is seen in the lowest momentum bin for each selection, in which the content for real data is somewhat higher than MC. The reason for this excess is not obvious, but its presence for the pre-selection shows that it is not a result of model dependencies of either PID; rather it may be due to underestimation of low-momentum μ^+ by NEUT.

Figure 7.10 shows similar broad agreement between data and MC for the momentum of the π^- candidate, but some discrepancies can be seen. The lowest-momentum bin again shows a discrepancy in each case, but here the bin content appears underestimated for the pre-selection and existing selection but overestimated for the BDT selection. It is not clear why this difference should arise, though given the size of the error bars it may be a statistical effect.

These plots also provide further illustration of the performance of the selections in question, particularly in Figure 7.9. The wrong-sign background contamination is very evident in the pre-selection[1], and while the existing PID does succeed in removing a large proportion of it, the accompanying loss of $\bar{\nu}_\mu$ CC1π^- signal for lower-momentum μ^+ can be clearly seen. By contrast, the BDT selection shows a still greater reduction in the

[1] The pre-selection used here results in a particularly large wrong-sign contamination since the two-track cut requires a negative TPC track, which will often be a μ^- from a ν_μ event.

wrong-sign background, and much more acceptance of signal events at low μ^+ momenta. These results indicate that the selection behaviour with MC simulated data is for the most part consistent with real data, so we can reasonably expect the BDT PID to improve selection performance as predicted by the MC, though the discrepancies (particularly for low-momentum tracks) may need to be investigated.

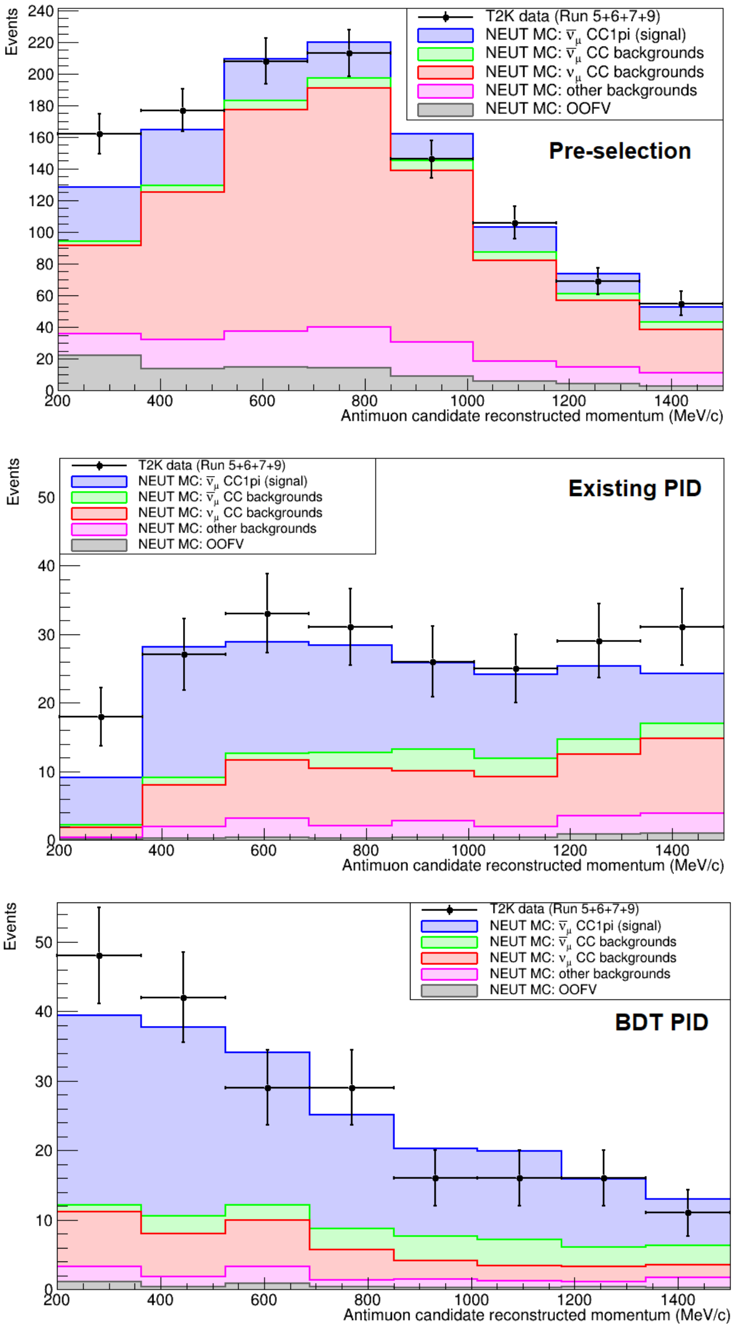


Figure 7.9: Reconstructed momentum of the μ^+ candidate for $\bar{\nu}_\mu$ CC1π^- event selections, comparing between real data (black) and MC, which is shown as a histogram stacked by topology categories: the $\bar{\nu}_\mu$ CC1π^- signal (blue), other $\bar{\nu}_\mu$ CC topologies (green), the wrong-sign ν_μ CC background (red), any other backgrounds (magenta), and out-of-fiducial-volume events (grey). The MC histogram bin content has been scaled by POT to the data. The uncertainty in the number of entries N for each momentum bin is taken to be $\sqrt{N}$.

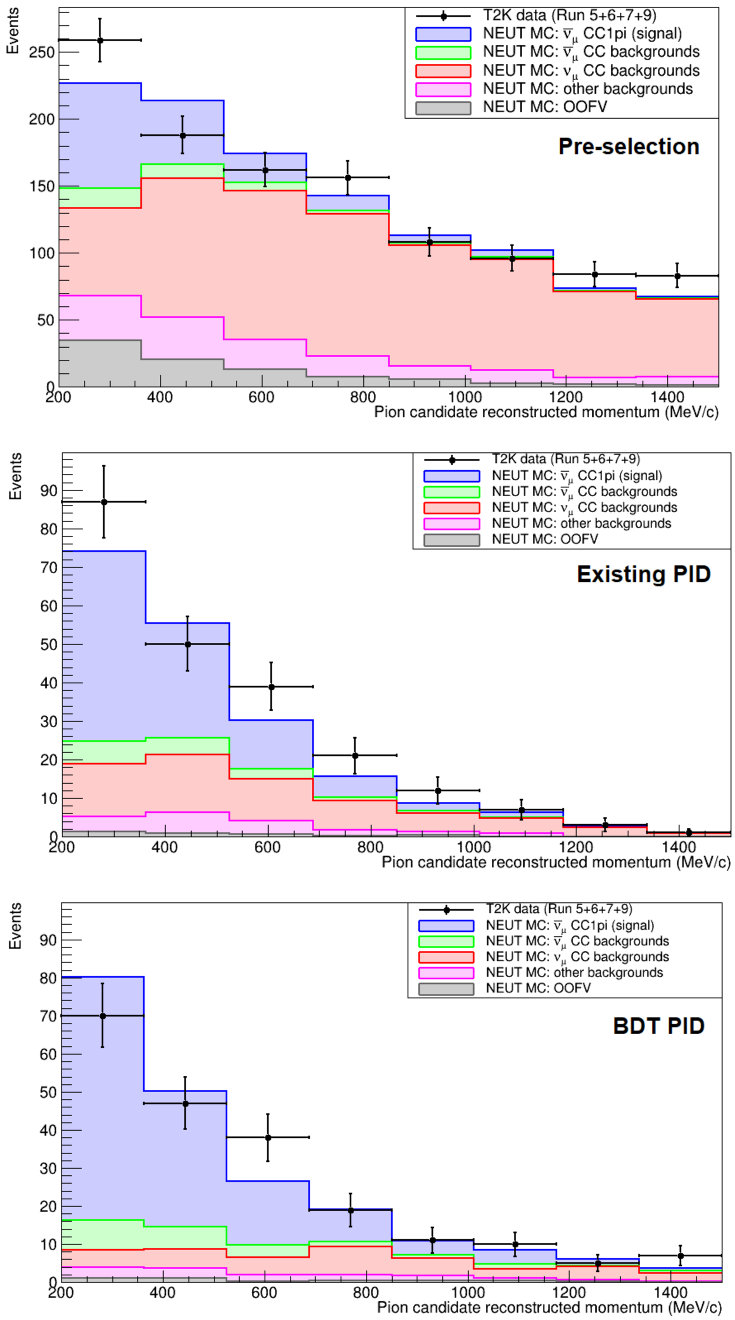


Figure 7.10: Reconstructed momentum of the π^- candidate for $\bar{\nu}_\mu$ CC1π^- event selections, comparing between real data (black) and MC, which is shown as a histogram stacked by topology categories: the $\bar{\nu}_\mu$ CC1π^- signal (blue), other $\bar{\nu}_\mu$ CC topologies (green), the wrong-sign ν_μ CC background (red), any other backgrounds (magenta), and out-of-fiducial-volume events (grey). The MC histogram bin content has been scaled by POT to the data. The uncertainty in the number of entries N for each momentum bin is taken to be $\sqrt{N}$.

Comparisons between real data and MC can also help verify the predicted behaviour of the BDT outputs more directly, by comparing their distributions. Figures 7.11–7.14 show histograms of each of the four BDT outputs for μ^+ candidate tracks in the pre-selection. A version with a logarithmic scale has been provided in each case so that the bins with low content can be more easily compared. The MC predicted distributions appear consistent with real data for the most part, with the main discrepancy being seen in the lowest and highest bins of the proton-like output. It is difficult to draw conclusions from these plots in isolation, since the relative quantities of each particle type are model-dependent and so may differ between real data and MC regardless of the BDT behaviour, but together with the above momentum histograms, a general consistency can be seen which suggests that the MC predictions are close to reality.

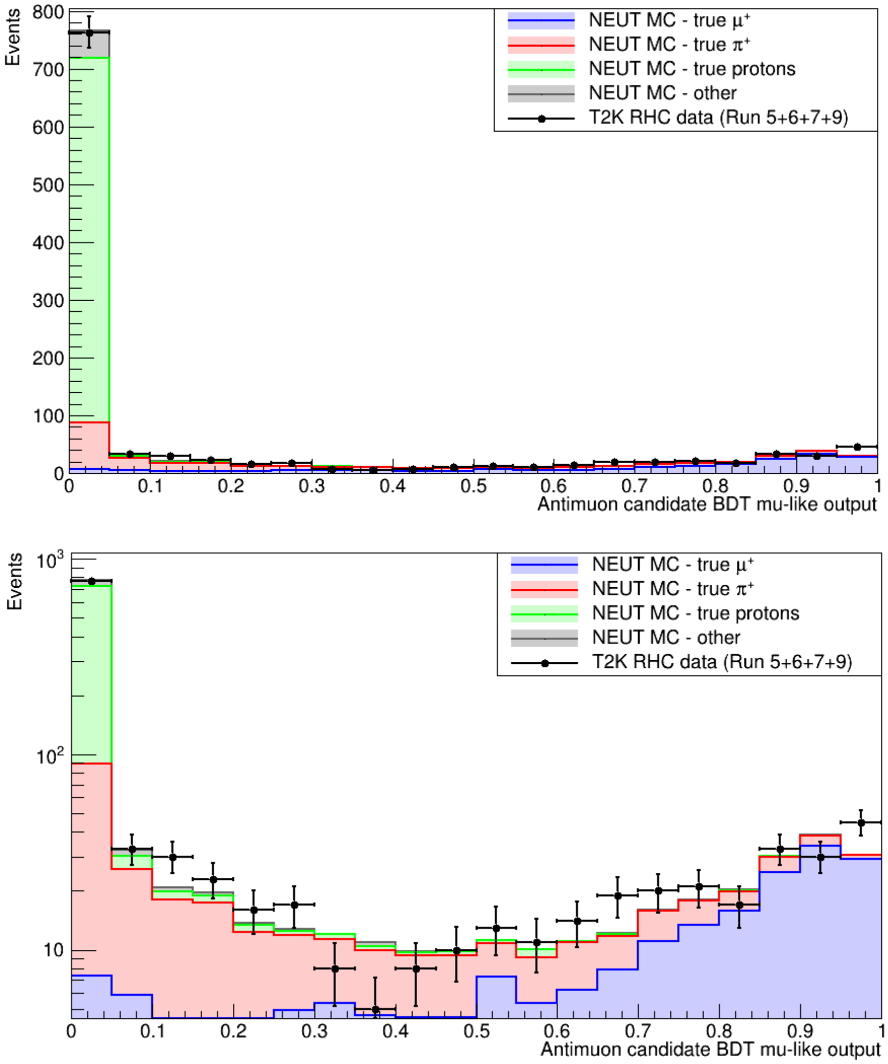


Figure 7.11: Histograms of the BDT μ-like output for μ^+ candidates in the $\bar{\nu}_\mu$ CC1π^- pre-selection sample, comparing between real data (black) and MC, which is shown stacked by the true particle type: antimuons (blue), π^+ (red), protons (green), and any other particles (grey). Versions with linear (top) and logarithmic (bottom) scales are shown. The MC histogram bin content has been scaled by POT to the data. The uncertainty in the number of entries N for each momentum bin is taken to be $\sqrt{N}$.

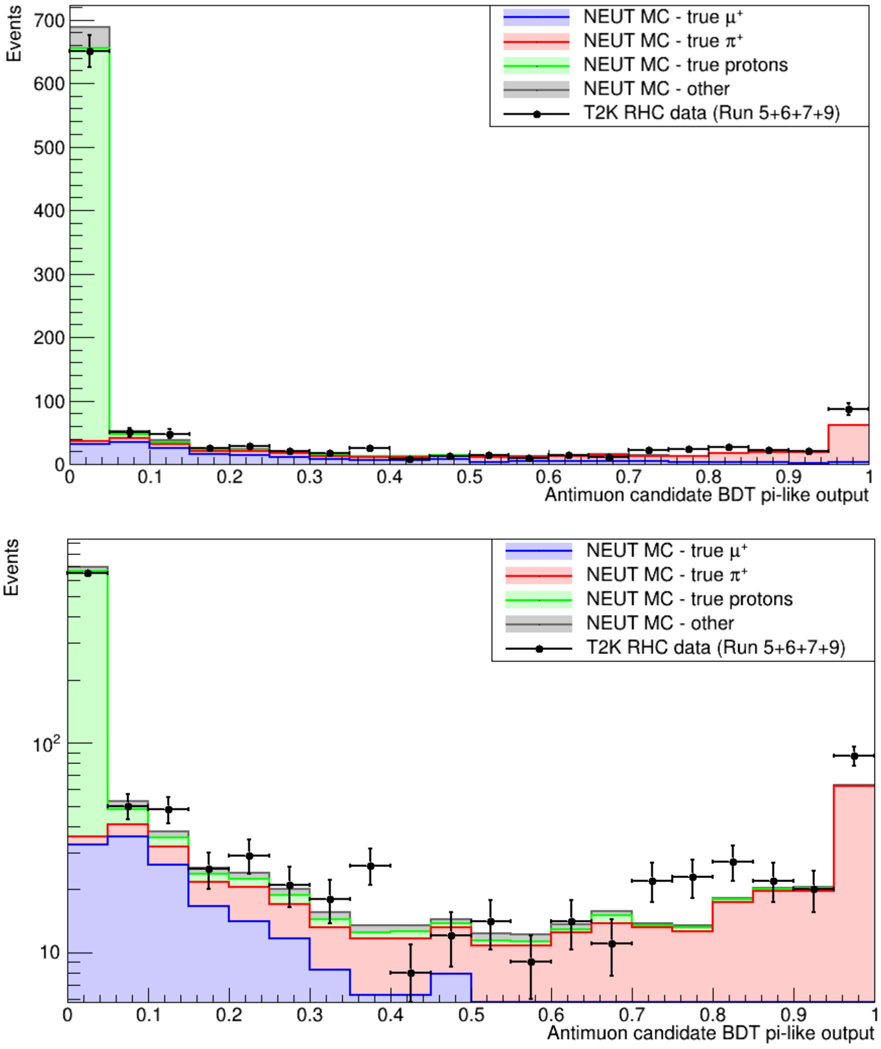


Figure 7.12: Histograms of the BDT π-like output for μ^+ candidates in the $\bar{\nu}_\mu$ CC1π^- pre-selection sample, comparing between real data (black) and MC, which is shown stacked by the true particle type: antimuons (blue), π^+ (red), protons (green), and any other particles (grey). Versions with linear (top) and logarithmic (bottom) scales are shown. The MC histogram bin content has been scaled by POT to the data. The uncertainty in the number of entries N for each momentum bin is taken to be $\sqrt{N}$.

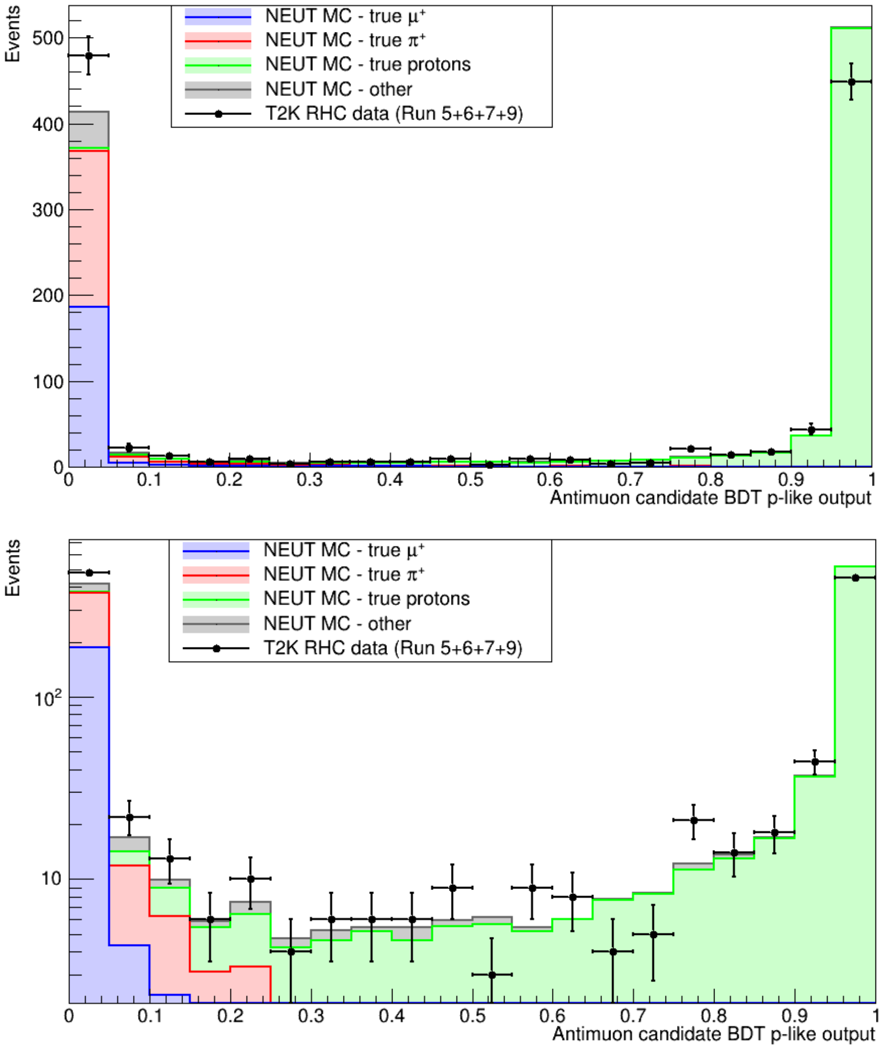


Figure 7.13: Histograms of the BDT p-like output for μ^+ candidates in the $\bar{\nu}_\mu$ CC1π^- pre-selection sample, comparing between real data (black) and MC, which is shown stacked by the true particle type: antimuons (blue), π^+ (red), protons (green), and any other particles (grey). Versions with linear (top) and logarithmic (bottom) scales are shown. The MC histogram bin content has been scaled by POT to the data. The uncertainty in the number of entries N for each momentum bin is taken to be $\sqrt{N}$.

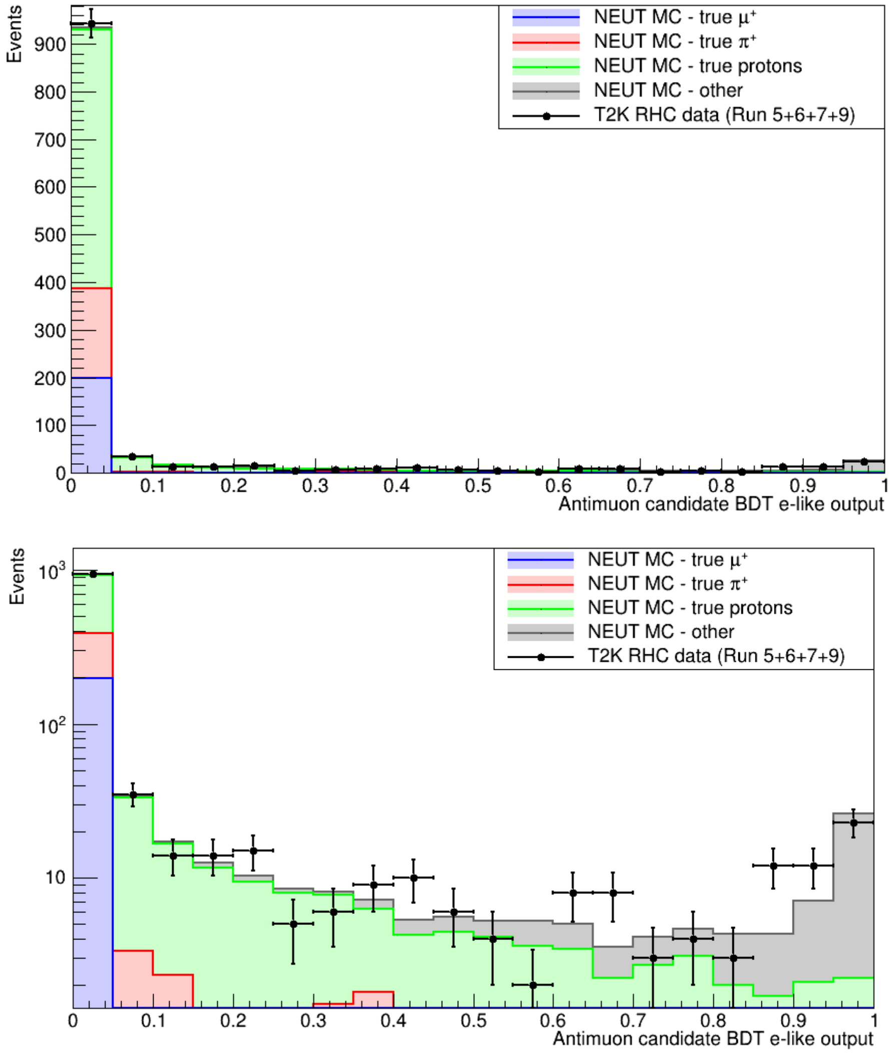


Figure 7.14: Histograms of the BDT e-like output for μ^+ candidates in the $\bar{\nu}_\mu$ CC1π^- pre-selection sample, comparing between real data (black) and MC, which is shown stacked by the true particle type: antimuons (blue), π^+ (red), protons (green), and any other particles (grey). Versions with linear (top) and logarithmic (bottom) scales are shown. The MC histogram bin content has been scaled by POT to the data. The uncertainty in the number of entries N for each momentum bin is taken to be $\sqrt{N}$.

7.4 Conclusions and outlook

In summary, the above MC studies show that the BDT PID tool can be applied in the ND280 $\bar{\nu}_\mu$ CC1π^- event selection to improve identification of μ^+, π^- and proton tracks and thus increase the selection performance significantly, and these predictions appear largely consistent with real data. Good performance can be obtained by simply using the BDT preference to identify tracks, but finding optimal cuts on the BDT outputs for the main tracks yields higher performance still. The purity of the μ^+ and π^- candidate tracks is improved greatly, resulting in higher overall CC1π^- purity. Provided that the tracks have kinematics within the region of validity of the BDT, the loss of efficiency is more than outweighed by the purity gains, and consequently the selection significance is improved across the momentum spectrum compared to conventional selections. As the T2K experiment continues operations and the size of its data sample increases — an increase of around a factor of 5 is currently expected by the end of the experiment's lifetime [110] — the purity will have greater and greater impact on selection significance compared to the efficiency, so these improvements will become more and more valuable over time.

Given the substantial increases in track identification efficiency for each particle type seen in the testing presented in Chapter 6, and the corresponding improvement in selection performance demonstrated above for $\bar{\nu}_\mu$ CC1π^-, it seems likely that BDT PID will yield similar improvements if applied in other event selections such as $\bar{\nu}_e$. This in turn will enable more precise measurements of the corresponding interaction rates, improving our understanding of neutrino-nucleus cross-sections and of neutrino oscillations as part of the wider T2K experiment.

The work presented in this thesis demonstrates the power and value of global multivariate PID for ND280 event selections, and sets out a development process that yields good results and can be replicated and expanded upon to develop a fully usable tool for T2K. The tuning results for the BDT parameters and the choice of input variables should provide a strong starting point (provided that the issue of overtraining is first addressed, as discussed in Section 6.6.2), and an obvious next step will be to expand the kinematic region of validity — that is, to re-train the BDT with a new particle gun sample with a wider range in momentum and angle. Additionally, it would be ideal for the tool to support tracks originating in FGD2, high-angle tracks which do not pass through the TPC, or backward-going tracks, in order to be applicable in as many ND280 event selections as possible. The performance for negative tracks should be assessed, and if found to be substantially poorer than for positive tracks, steps should be taken to address this (such as including negative tracks in the training sample). An understanding of the systematic uncertainties on the BDT outputs is also needed, and will need to take into account those of the various input variables and how they interact within the complex structure of the BDT. Improvements in performance might be obtained by increasing the statistics of the training samples, and/or by using a neural network instead of boosted decision trees, so these options should be investigated. With further development, the multivariate global PID methods explored in this thesis should yield a broadly-applicable tool that will enable significant performance improvements in multiple ND280 event selections.

The ND280 upgrade, in which the PØD is replaced by a new suite of subdetectors, presents an important opportunity in this context. The new subdetectors, which include a 'super-FGD' scintillator detector with significantly improved granularity and angular acceptance compared to the existing FGDs [111], will yield new PID information and thus new input variables for the BDT which will substantially improve its performance; and conversely, any new PID information will be more effective as part of a global multivariate PID than on its own. T2K will soon start taking data with the upgraded ND280 detector, the analysis of which would benefit greatly from global BDT PID, so we recommend that such methods be pursued and adopted by T2K as soon as possible.

Bibliography

[1] Wikipedia contributors. Standard model — Wikipedia, the free encyclopedia, 2019. URL `https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg`. [Online; accessed 03-February-2022].

[2] G. Cowan and T. Gershon. *Tetraquarks and Pentaquarks.* 2399-2891. IOP Publishing, 2018. ISBN 978-0-7503-1593-7. arXiv:1808.04153.

[3] T. Endoh et al. CP Violation in Neutrino Oscillations and Leptogenesis. *Phys. Rev. Lett.*, 89:231601, 2002.

[4] H. Becquerel. On the rays emitted by phosphorescence. *Compt. Rend. Hebd. Seances Acad. Sci.*, 122(8):420–421, 1896.

[5] E. Rutherford. XV. The magnetic and electric deviation of the easily absorbed rays from radium. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 5(26):177–187, 1903. URL `https://doi.org/10.1080/14786440309462912`.

[6] J. Chadwick. Intensitätsverteilung im magnetischen Spectrum der β-Strahlen von radium B + C. *Verhandl. Dtsc. Phys. Ges.*, 16:383, 1914.

[7] W. Pauli. Dear radioactive ladies and gentlemen. *Phys. Today*, 31N9:27, 1978.

[8] J. Chadwick. The existence of a neutron. *Proc. R. Soc. Lond A*, 136:672–708, 1932.

[9] E. Fermi. Versuch einer Theorie der β-Strahlen. *Zeitschrift für Physik*, 88(3):161–177, 1934.

[10] C. L. Cowan et al. Detection of the Free Neutrino: a Confirmation. *Science*, 124 (3212):103–104, 1956. URL https://www.science.org/doi/abs/10.1126/science.124.3212.103.

[11] G. Danby et al. Observation of High-Energy Neutrino Reactions and the Existence of Two Kinds of Neutrinos. *Phys. Rev. Lett.*, 9:36–44, Jul 1962. URL https://link.aps.org/doi/10.1103/PhysRevLett.9.36.

[12] M. L. Perl et al. Evidence for Anomalous Lepton Production in e+ - e- Annihilation. *Phys. Rev. Lett.*, 35:1489–1492, 1975.

[13] K. Kodama et al. Observation of tau neutrino interactions. *Physics Letters B*, 504(3):218–224, Apr 2001. ISSN 0370-2693. URL http://dx.doi.org/10.1016/S0370-2693(01)00307-0.

[14] J. Bahcall et al. Solar Neutrino Flux. *Astrophysical Journal*, 137:344–346, January 1963.

[15] B. T. Cleveland et al. Measurement of the Solar Electron Neutrino Flux with the Homestake Chlorine Detector. *Astrophysical Journal*, 496:505, 1998.

[16] B. Pontecorvo. Neutrino Experiments and the Problem of Conservation of Leptonic Charge. *Sov. Phys. JETP*, 26:984, 1968.

[17] Q. R. Ahmad et al. Direct Evidence for Neutrino Flavor Transformation from Neutral-Current Interactions in the Sudbury Neutrino Observatory. *Phys. Rev. Lett.*, 89: 011301, 2002.

[18] S. Fukuda et al. Solar B-8 and hep neutrino measurements from 1258 days of Super-Kamiokande data. *Phys. Rev. Lett.*, 86:5651–5655, 2001.

[19] K. S. Hirata et al. Experimental study of the atmospheric neutrino flux. *Physics Letters B*, 205(2):416–420, 1988.

[20] Y. Fukuda et al. Evidence for Oscillation of Atmospheric Neutrinos. *Phys. Rev. Lett.*, 81:1562–1567, 1998.

[21] KamLAND collaboration. First results from KamLAND: Evidence for reactor antineutrino disappearance. *Phys. Rev. Lett.*, 90:021802, 2003.

[22] The T2K Collaboration. Constraint on the matter-antimatter symmetry-violating phase in neutrino oscillations. *Nature*, 580:339–344, 2020.

[23] G. Mention et al. Reactor antineutrino anomaly. *Physical Review D*, 83(7), Apr 2011. ISSN 1550-2368. URL `http://dx.doi.org/10.1103/PhysRevD.83.073006`.

[24] MicroBooNE collaboration. Search for an Excess of Electron Neutrino Interactions in MicroBooNE Using Multiple Final State Topologies, arXiv:2110.14054, 2021. URL `https://doi.org/10.48550/arXiv.2110.14054`.

[25] S. F. King. Neutrino Mass. *Contemporary Physics*, 48(4):195–211, 2007.

[26] M. H. Pinsonneault J. Bahcall. What Do We (Not) Know Theoretically about Solar Neutrino Fluxes? *Phys. Rev. Lett.*, 92:121301, 2004.

[27] J. N. Bahcall and C. Peña-Garay. Solar models and solar neutrino oscillations. *New Journal of Physics*, 6:63, 2004. URL `https://doi.org/10.1088/1367-2630/6/1/063`.

[28] Y. Fukuda et al. Atmospheric muon neutrino, electron neutrino ratio in the multi-GeV energy range. *Physics Letters B*, 335(2):237–245, 1994.

[29] M. Cribier et al. Results of the whole GALLEX experiment. *Nuclear Physics B - Proceedings Supplements*, 70(1):284–291, 1999.

[30] M. Altmann et al. Complete results for five years of GNO solar neutrino observations. *Physics Letters B*, 616(3):174–190, 2005.

[31] J. N. Abdurashitov et al. Solar neutrino flux measurements by the Soviet-American gallium experiment (SAGE) for half the 22-year solar cycle. *Journal of Experimental and Theoretical Physics*, 95(2):181–193, 2002.

[32] B. Aharmim et al. Electron energy spectra, fluxes, and day-night asymmetries of B-8 solar neutrinos from measurements with NaCl dissolved in the heavy-water detector at the Sudbury Neutrino Observatory. *Phys. Rev. C*, 72:055502, 2005.

[33] A. Yu. Smirnov. The MSW effect and Solar Neutrinos, arXiv:hep-ph/0305106, 2003. URL `https://arxiv.org/abs/hep-ph/0305106`.

[34] C. V. Achar et al. Detection of muons produced by cosmic ray neutrinos deep underground. *Physics Letters*, 18(2):196–199, 1965.

[35] F. Reines et al. Evidence for High-Energy Cosmic-Ray Neutrino Interactions. *Phys. Rev. Lett.*, 15:429–433, 1965.

[36] ANTARES collaboration. Measuring the atmospheric neutrino oscillation parameters and constraining the 3+1 neutrino model with ten years of ANTARES data. *JHEP*, 06:113, 2019.

[37] IceCube collaboration. Measurement of Atmospheric Neutrino Oscillations at 6-56 GeV with IceCube DeepCore. *Phys. Rev. Lett.*, 120:071801, 2018.

[38] M. S. Athar et al. Status and Perspectives of Neutrino Physics. *Progress in Particle and Nuclear Physics*, 124:103947, may 2022. URL `https://doi.org/10.1016%2Fj.ppnp.2022.103947`.

[39] T. Araki et al. Measurement of neutrino oscillation with KamLAND, evidence of spectral distortion. *Phys. Rev. Lett.*, 94:081801, 2005.

[40] K. Abe et al. Solar neutrino measurements in Super-Kamiokande-IV. *Phys. Rev. D*, 94:052010, Sep 2016. URL `https://link.aps.org/doi/10.1103/PhysRevD.94.052010`.

[41] Daya Bay collaboration. Observation of electron-antineutrino disappearance at Daya Bay. *Phys. Rev. Lett.*, 108:171803, 2012.

[42] RENO collaboration. Observation of Reactor Electron Antineutrino Disappearance in the RENO Experiment. *Phys. Rev. Lett.*, 108:191802, 2012.

[43] JUNO collaboration. JUNO Physics and Detector. *Progress in Particle and Nuclear Physics*, 123:103927, Mar 2022. ISSN 0146-6410. URL `http://dx.doi.org/10.1016/j.ppnp.2021.103927`.

[44] P. Adamson et al. Measurement of the neutrino mixing angle θ_{23} in NOνA. *Phys. Rev. Lett.*, 118:181802, 2017.

[45] K. Abe et al. Evidence of Electron Neutrino Appearance in a Muon Neutrino Beam. *Phys. Rev. D*, 88(3):032002, 2013.

[46] R. Acciarri et al. Long-Baseline Neutrino Facility (LBNF) and Deep Underground Neutrino Experiment (DUNE) Conceptual Design Report Volume 1: The LBNF and DUNE Projects, arXiv:1601.05471, 2016. URL `https://doi.org/10.48550/arXiv.1601.05471`.

[47] Hyper-Kamiokande Working Group. A Long Baseline Neutrino Oscillation Experiment Using J-PARC Neutrino Beam and Hyper-Kamiokande, arXiv:1412.4673, 2015. URL `https://doi.org/10.48550/arXiv.1412.4673`.

[48] NOνA collaboration. First Measurement of Neutrino Oscillation Parameters using Neutrinos and Antineutrinos by NOνA. *Phys. Rev. Lett.*, 123:151803, 2019.

[49] G. P. Zeller. Low Energy Neutrino Cross Sections: Comparison of Various Monte Carlo Predictions to Experimental Data, arXiv:hep-ex/0312061, 2003. URL `https://doi.org/10.48550/arXiv.hep-ex/0312061`.

[50] NuSTEC collaboration. NuSTEC White Paper: Status and challenges of neutrino-nucleus scattering. *Prog. Part. Nucl. Phys.*, 100:1, 2018.

[51] Y. Hayato and L. Pickering. The NEUT neutrino interaction simulation program library. *The European Physical Journal Special Topics*, 340:4469–4481, 2021.

[52] C. Andreopoulos et al. The GENIE Neutrino Monte Carlo Generator: Physics and User Manual, arXiv:1510.05494, 2015. URL `https://doi.org/10.48550/arXiv.1510.05494`.

[53] T. Golan et al. Final State Interactions Effects in Neutrino-Nucleus Interactions. *Phys. Rev. C*, 86:015505, 2012.

[54] O. Buss et al. Transport-theoretical Description of Nuclear Reactions. *Phys. Rept.*, 512:1, 2012.

[55] R. Guenette. The ArgoNeuT experiment, arXiv:1110.0443, 2011. URL `https://doi.org/10.48550/arXiv.1110.0443`.

[56] H. Ray. The MiniBooNE Experiment: An Overview, arXiv:hep-ex/0701040, 2007. URL `https://doi.org/10.48550/arXiv.hep-ex/0701040`.

[57] L. Aliaga et al. Design, calibration, and performance of the MINERvA detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 743:130–159, Apr 2014. ISSN 0168-9002. URL `http://dx.doi.org/10.1016/j.nima.2013.12.053`.

[58] T. Le et al. Measurement of $\bar{\nu}_\mu$ charged-current single π^- production on hydrocarbon in the few-GeV region using MINERvA. *Physical Review D*, 100(5), Sep 2019. ISSN 2470-0029. URL `http://dx.doi.org/10.1103/PhysRevD.100.052008`.

[59] R. A. Smith and E. J. Moniz. Neutrino interactions on nuclear targets. *Nuclear Physics B*, 43:605, 1972.

[60] E. Hernández, J. Nieves, and M. J. V. Vacas. Single π production in neutrino nucleus scattering. *Physical Review D*, 87(11), Jun 2013. ISSN 1550-2368. URL `http://dx.doi.org/10.1103/PhysRevD.87.113009`.

[61] K. Abe et al. (T2K collaboration). The T2K experiment. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 659(1):106–135, 2011.

[62] K. Abe et al. T2K neutrino flux prediction. *Phys. Rev. D*, 87:012001, Jan 2013. URL `https://link.aps.org/doi/10.1103/PhysRevD.87.012001`.

[63] M. Miura et al. (Super-Kamiokande collaboration). Search for Nucleon Decay in Super-Kamiokande. *Nuclear and Particle Physics Proceedings*, 273–275:516–521, 2016.

[64] M. Ikeda et al. (Super-Kamiokande collaboration). Search for Supernova Neutrino Bursts at Super-Kamiokande. *The Astrophysical Journal*, 669(1):519–524, nov 2007.

[65] Y. Itow et al. The JHF-Kamioka neutrino project, arXiv:hep-ex/0106019, 2001. URL `https://arxiv.org/abs/hep-ex/0106019`.

[66] P. A. Amaudruz et al. The T2K Fine-Grained Detectors. *Nucl. Instrum. Meth. A*, 696:1–31, 2012, arXiv:1204.3666.

[67] T2K ND280 TPC collaboration. Time Projection Chambers for the T2K Near Detectors, arXiv:1012.0865, 2010. URL `https://doi.org/10.48550/arXiv.1012.0865`.

[68] S. Assylbekov et al. The T2K ND280 off-axis pi–zero detector. *Nucl. Instrum. Meth. A*, 686:48–63, 2012. ISSN 0168-9002. URL `https://www.sciencedirect.com/science/article/pii/S0168900212005153`.

[69] R. Brun and F. Rademakers. ROOT - An Object Oriented Data Analysis Framework. *Nucl. Inst. Meth. in Phys. Res. A*, 389:81, 1997.

[70] S. Agostinelli et al. Geant4—a simulation toolkit. *Nucl. Instru. Meth. A*, 506(3): 250–303, 2003. ISSN 0168-9002. URL `https://www.sciencedirect.com/science/article/pii/S0168900203013688`.

[71] S. Ritt, P. Amaudruz, and K. Olchanski. MIDAS (Maximum Integration Data Acquisition System), 2001.

[72] D. Allan et al. The Electromagnetic Calorimeter for the T2K Near Detector ND280. *JINST*, 8:P10019, 2013.

[73] K. Abe et al. (T2K collaboration). Scintillator ageing of the T2K near detectors from 2010 to 2021. *"JINST"*, 8:P10019, 2022. URL `https://arxiv.org/abs/2207.12982`.

[74] Kuraray Co., Ltd. Plastic Scintillating Fibers. URL `https://www.kuraray.com/uploads/5a717515df6f5/PR0150_psf01.pdf`. [Online; accessed 24-January-2023].

[75] K. Yamamoto et al. Development of Multi-Pixel Photon Counter (MPPC). *IEEE Nuclear Science Symposium Conference Record*, 2006.

[76] M. Yokoyama et al. Application of Hamamatsu MPPCs to T2K Neutrino Detectors. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 610(1):128—130, Oct 2009. ISSN 0168-9002. URL `http://dx.doi.org/10.1016/j.nima.2009.05.077`.

[77] A. Vacheret et al. The front end readout system for the T2K-ND280 detectors. *IEEE Nuclear Science Symposium Conference Record*, pages 1984–1991, 2007.

[78] A. Vacheret et al. Calibration of the ND280 Scintillator Detectors (Prod 4). *T2K Technical Note 037*, 2011.

[79] P. Paudyal. Probing low energy particle production in neutrino interactions through the vertex activity in the Fine Grained Detector at the T2K experiment. *PhD thesis, University of Liverpool*, 2019.

[80] M. Antonova et al. Aging studies of the ND280 scintillator detectors. *T2K Technical Note 308*, 2018.

[81] A. Hillairet et al. ND280 Reconstruction. *T2K Technical Note 072*, 2011.

[82] D. Brailsford et al. Study of the tracker ECal systematic uncertainties. *T2K Technical Note 279*, 2017.

[83] The T2K ND280 UK/ECal Group. T2K 2010a Analysis ND280 DsECAL Status Report. *T2K Technical Note 018*, 2010.

[84] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 1960, https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/82/1/35/5518977/35_1.pdf. ISSN 0021-9223. URL `https://doi.org/10.1115/1.3662552`.

[85] K. Fusshoeller et al. FHC muon neutrino charged current multiple pion samples in the ND280 tracker for the 2020 oscillation analysis inputs: Run 2+3+4+8 with P6T NEUT_D MC. *T2K Technical Note 407*, 2020.

[86] K. Fusshoeller et al. Muon antineutrino and neutrino Charged-Current multiple pion selections in antineutrino mode. *T2K Technical Note 273*, 2020.

[87] C. Giganti and M. Zito. Particle Identification with the T2K TPC. *T2K Technical Note 001*, 2009.

[88] S. Bordoni et al. The TPC Particle Identification algorithm with Production 6B. *T2K Technical Note 221*, 2015.

[89] R. L. Workman et al. (Particle Data Group). Review of Particle Physics. *PTEP*, 2022(8):083C01, 2022. Section 27: Passage of Particles through Matter.

[90] C. Licciardi and M. Barbi. Particle identification with the Fine Grained Detectors. *T2K Technical Note 103*, 2011.

[91] S. Jenkins. Reinvestigation of the ECal pi0 veto. *T2K Technical Note 392*, 2019.

[92] A. Carver. Particle Identification in the ND280 Electromagnetic Calorimeter. *T2K Technical Note 002*, 2009.

[93] G. Barker et al. Implementation of the Second Generation PID for the ND280 Tracker ECals. *T2K Technical Note 111*, 2012.

[94] P. C. Bhat. Multivariate Analysis Methods in Particle Physics. *Annual Review of Nuclear and Particle Science*, 61(1):281–309, 2011. URL `https://doi.org/10.1146/annurev.nucl.012809.104427`.

[95] A. O'Hagan and J. Forster. *Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Inference*. Arnold, 2004. ISBN 978-0-470-68569-3.

[96] K. Albertsson et al. TMVA - Toolkit for Multivariate Data Analysis, Users Guide, 2020. URL `https://root.cern.ch/download/doc/tmva/TMVAUsersGuide.pdf`.

[97] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.*, 7(2):179–188, 1936. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x`.

[98] I. Rish. An Empirical Study of the Naïve Bayes Classifier. *IJCAI 2001 Work. Empir. Methods Artif. Intell.*, 3, 2001.

[99] D. J. C. MacKay. Bayesian neural networks and density networks. *Nucl. Inst. Meth. in Phys. Res. A*, 354(1):73–80, 1995. ISSN 0168-9002. URL `https://www.sciencedirect.com/science/article/pii/0168900294009317`.

[100] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory*, pages 23–37. Springer Berlin Heidelberg, 1995. ISBN 978-3-540-49195-8.

[101] B. P. Roe et al. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 543 (2-3):577–584, May 2005. ISSN 0168-9002. URL `http://dx.doi.org/10.1016/j.nima.2004.12.018`.

[102] D. Stanev et al. Deep Neural Network as an alternative to Boosted Decision Trees for PID, arXiv:2104.14045, 2021. URL `https://doi.org/10.48550/arXiv.2104.14045`.

[103] L. Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

[104] C. Ferri et al. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38, 2009. ISSN 0167-8655. URL https://www.sciencedirect.com/science/article/pii/S0167865508002687.

[105] P. A. Zyla et al. (Particle Data Group). Review of Particle Physics. *PTEP*, 2020(8): 083C01, 2020. Section 40: Statistics.

[106] G. Christodolou et al. Selection of ν_e and $\bar{\nu}_e$ in the ND280 tracker using anti-neutrino beam data from Run5 and Run6. *T2K Technical Note 282*, 2016.

[107] J. B. Birks. Scintillations from Organic Crystals: Specific Fluorescence and Relative Response to Different Radiations. *Proc. Phys. Soc. A*, 64(10):874, 1951. URL https://dx.doi.org/10.1088/0370-1298/64/10/303.

[108] F. J. Massey Jr. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951, https://www.tandfonline.com/doi/pdf/10.1080/01621459.1951.10500769. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1951.10500769.

[109] J. W. Pratt and J. D. Gibbons. *Concepts of Nonparametric Theory, Chapter 7: Kolmogorov-Smirnov Two-Sample Tests*, pages 318–344. Springer New York, New York, NY, 1981. ISBN 978-1-4612-5931-2. URL https://doi.org/10.1007/978-1-4612-5931-2_7.

[110] K. Abe et al. Proposal for an Extended Run of T2K to 20×10^{21} POT, arXiv:1609.04111, 2016. URL https://arxiv.org/abs/1609.04111.

[111] K. Abe et al. T2K ND280 Upgrade – Technical Design Report, arXiv:1901.03750, 2019. URL https://arxiv.org/abs/1901.03750.