Advancing Neutrino Detection and Triggering in DUNE

Francisco Martínez López

Submitted in partial fulfillment of the requirements of the Degree of Doctor of Philosophy

School of Physical and Chemical Sciences
Queen Mary University of London

School of Physics and Astronomy
University of Southampton

December 2024

Statement of originality

I, Francisco Martínez López, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the University has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature: [can be digital signature]
Date:

Details of collaboration and publications:
[insert details here if applicable]

Abstract

Work in progress . . .

¡Oh memoria, enemiga mortal de mi descanso!

El ingenioso hidalgo don Quijote de la Mancha

Miguel de Cervantes Saavedra

Acknowledgements

Work in progress . . .

Contents

List of Figures

List of Tables

List of Abbreviations

ADC Analog to Digital Converter.

ALEPH Apparatus for LEP PHysics.

ALICE A Large Ion Collider Experiment.

BDT Boosted Decision Tree.

CC Charged Current.

DM Dark Matter.

DUNE Deep Underground Neutrino Experiment.

ECal Electromagnetic Calorimeter.

FD Far Detector.

FHC Forward Horn Current.

HPgTPC High Pressure gaseous Time Projection Chamber.

LBL Long BaseLine.

List of Abbreviations

MuID	Muon IDentification system.
NC	Neutral Current.
ND	Near Detector.
ND-GAr	Near Detector Gaseous Argon.
ND-LAr	Near Detector Liquid Argon.
PDG	Particle Data Group.
RHC	Reverse Horn Current.

Chapter 1

Introduction

Chapter 2

Neutrino physics

Ever since they were postulated in 1930 by Wolfgang Pauli to explain the continuous β decay spectrum [17] and later found by Reines and Cowan at the Savannah River reactor in 1953 [18], neutrinos have had a special place among all other elementary particles. They provide a unique way to probe a wide range of quite different physics, from nuclear physics to cosmology, from astrophysics to colliders. Moreover, there is compelling evidence to believe that the study of neutrinos may be key to unveil different aspects of physics beyond the SM, difficult to test elsewhere.

In this Chapter I will review the basics of neutrino physics, from its role within the SM to the main open questions related to the neutrino sector, paying special attention to the phenomenology of neutrino oscillations.

2.1 Neutrinos in the SM

By definition, in the SM there are no right-handed neutrino fields. A direct implication of this fact is that neutrinos are strictly massless within the SM. This follows from the experimental observation that all neutrinos produced via weak interactions are pure left-handed helicity states (and similarly antineutrinos are pure right-handed states). The hypothetical existence of right-handed neutrinos could be indirectly inferred from the observation of non-zero neutrino masses, nevertheless the existence neutrino masses is not a sufficient condition for the existence of such fields.

In the SM neutrinos appear in three flavours, namely ν_e, ν_μ and ν_τ. These are associated with the corresponding charged leptons e, μ and τ, in such a way that the charged current part of the Lagrangian coupling them is diagonal. As in the electroweak theory neutrinos are coupled to the Z boson in a universal way, by measuring the so-called invisible decay width of the Z we have an estimate of the number of light (i.e. lighter than the Z boson) neutrino flavours. This number was measured by LEP in a combined analysis of $e^+e^- \to \mu^+\mu^-$ and $e^+e^- \to$ hadrons to be $N_\nu = 2.9840 \pm 0.0082$ [19].

2.2 Neutrino oscillations

The evidence for neutrino oscillation [20], and therefore the existence of non-zero neutrino masses, constitutes one of the groundbreaking discoveries of modern Physics and has acted as driving force for Beyond the Standard Model (BSM) Physics. The minimal extension of the Standard Model (SM) we can do to address these phenomena is introducing distinct masses for at least two of the neutrinos. This way, we are left with three neutrino mass eigenstates ν_1, ν_2 and ν_3, with masses m_1, m_2 and m_3 respectively, which in general will not coincide with the flavour eigenstates ν_e, ν_μ and ν_τ.

The way to relate these two sets of neutrino eigenstates is via a 3×3 unitary matrix, called the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix [21, 22], as:

$$|\nu_\alpha\rangle = \sum_{i=1}^{3} U^*_{\alpha i} |\nu_i\rangle \,, \tag{2.1}$$

where the Greek index α denotes the flavour $\{e, \mu, \tau\}$ and the Latin index i the associated masses $\{1, 2, 3\}$. This leptonic mixing matrix may be parametrized in terms of 6 parameters, 3 of which are mixing angles θ_{12}, θ_{13} and θ_{23}, one CP-violating phase δ_{CP}

and 2 Majorana phases α and β:

$$U = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_{23} & s_{23} \\ 0 & -s_{23} & c_{23} \end{pmatrix} \begin{pmatrix} c_{13} & 0 & s_{13}\mathrm{e}^{-i\delta_{CP}} \\ 0 & 1 & 0 \\ -s_{13}\mathrm{e}^{-i\delta_{CP}} & 0 & c_{13} \end{pmatrix} \begin{pmatrix} c_{12} & s_{12} & 0 \\ -s_{12} & c_{12} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \mathrm{e}^{i\alpha} & 0 \\ 0 & 0 & \mathrm{e}^{i\beta} \end{pmatrix}, \tag{2.2}$$

where $c_{ij} \equiv \cos\theta_{ij}$ and $s_{ij} \equiv \sin\theta_{ij}$. This matrix is analogous to the Cabibbo-Kobayashi-Maskawa (CKM) matrix in the quark sector. If neutrinos are Dirac fermions, we can drop the Majorana phases in the PMNS matrix. But, in any case, these phases play no role on the neutrino oscillations.

2.2.1 Oscillations in vacuum

Consider the case where a neutrino of flavour α is produced at $t = 0$, and then it propagates through vacuum. Such a state will evolve in time according to the relation:

$$|\nu_\alpha(t)\rangle = \sum_{i=1}^{3} U^*_{\alpha i}\mathrm{e}^{-iE_i t}\,|\nu_i(t=0)\rangle\,, \tag{2.3}$$

as the mass eigenstates are also eigenstates of the free Hamiltonian. Now, if we express the mass eigenstates as a superposition of flavour eigenstates, the last expression can be rewritten as:

$$|\nu_\alpha(t)\rangle = \sum_{i=1}^{3} U_{\beta i}\mathrm{e}^{-iE_i t}U^*_{\alpha i}\,|\nu_\beta\rangle\,. \tag{2.4}$$

This way, the probability for the neutrino to transition from flavour α to flavour β will be given by:

$$P(\nu_\alpha \to \nu_\beta) = |\langle\nu_\beta|\nu_\alpha(t)\rangle|^2 = \left|\sum_{i=1}^{3} U_{\beta i}\mathrm{e}^{-iE_i t}U^*_{\alpha i}\right|^2\,. \tag{2.5}$$

A usual approximation to take at this point is to consider ultra-relativistic neutrinos, i.e. $E \approx |\vec{p}|$, so we can write the dispersion relations as:

$$E_i = \sqrt{p^2 + m_i^2} \approx E + \frac{m_i^2}{2E}, \tag{2.6}$$

so we can write the oscillation probability as:

$$\begin{aligned} P(\nu_\alpha \to \nu_\beta) &= \sum_{i,j} U^*_{\alpha i} U_{\beta i} U_{\alpha j} U^*_{\beta j} \mathrm{e}^{-i\frac{\Delta m^2_{ij}}{2E}t} \\ &= \delta_{\alpha\beta} - 4\sum_{i<j} \mathfrak{Re}\left[U^*_{\alpha i} U_{\beta i} U_{\alpha j} U^*_{\beta j}\right] \sin^2\left(\frac{\Delta m^2_{ij}}{4E}t\right) \\ &\quad + 2\sum_{i<j} \mathfrak{Im}\left[U^*_{\alpha i} U_{\beta i} U_{\alpha j} U^*_{\beta j}\right] \sin\left(\frac{\Delta m^2_{ij}}{2E}t\right), \end{aligned} \tag{2.7}$$

where Δm^2_{ij} is the difference of the squared masses of the jth and ith neutrino mass eigenvalues. At this point, it is usual to write the phase responsible for the oscillations as (under the approximate assumption $t \approx L$):

$$\Delta_{ij} \equiv \frac{\Delta m^2_{ij}}{4E} L \simeq 1.27 \frac{\Delta m^2_{ij}}{(\mathrm{eV}^2)} \frac{L}{(\mathrm{km})} \frac{(\mathrm{GeV})}{E}. \tag{2.8}$$

Notice that, in the case of antineutrinos the only difference would be the sign of the last term in the oscillation probability. This way, one can write the CP asymmetry as:

$$\begin{aligned} A^{\alpha\beta}_{CP} &= P(\nu_\alpha \to \nu_\beta) - P(\bar{\nu}_\alpha \to \bar{\nu}_\beta) \\ &= 4\sum_{i<j} \mathfrak{Im}\left[U^*_{\alpha i} U_{\beta i} U_{\alpha j} U^*_{\beta j}\right] \sin 2\Delta_{ij}. \end{aligned} \tag{2.9}$$

2.2.2 Oscillations in matter

When neutrinos propagate through matter, their oscillation can be affected in mainly two ways. First, neutrinos can inelastically scatter with nuclei, thus destroying the coherent propagation of their quantum state. Nevertheless, in most cases this effect is negligible (even in very dense mediums like the core of the Sun). Second, neutrinos can also experience coherent or forward scatterings, that can affect their oscillation but not lose the coherent propagation of the state.

The first proposed model to account for neutrino oscillations in matter was proposed by Mikhaev, Smirnov and Wolfenstein (MSW) [23]. It relies on the fact that, as the only charged lepton present in ordinary matter is the electron, electron neutrinos can undergo both charged and neutral-current interactions with matter whereas for muon and tau neutrinos just neutral currents are possible.

2.2.3 Current status of neutrino oscillations

A wide range of neutrino experiments provide experimental input to the neutrino oscillation framework, both using natural or synthetic neutrino sources. The results from one of the neutrino global fit analyses, shown in Tab. 2.1 [1], summarise well our current understanding of the different oscillation parameters.

Solar neutrino experiments detect neutrinos produced in thermonuclear reactions inside the Sun, mainly from the so-called *pp* chain and the CNO cycle. These neutrinos have a typical energy in the range from 0.1 to 20 MeV. These experiments (Homestake [24], GALLEX [25], SAGE [26], Borexino [27], Super-Kamiokande [28] and SNO [29]) provide the best sensitivities to θ_{12} and Δm^2_{21}.

Atmospheric neutrino experiments detect the neutrino flux produced when cosmic rays scatter with particles in Earth's atmosphere. These collisions generate particle showers that eventually produce electron and muon neutrinos (and antineutrinos). Their energies range from few MeV to about 10^9 GeV. Experiments, like Super-Kamiokande [30] and IceCube [31] use atmospheric neutrinos to measure oscillations and are specially sensitive to θ_{23} and Δm^2_{32}.

Reactor neutrino experiments look for the $\bar{\nu}_e$ spectrum produced by nuclear reactors, with energies in the MeV scale. Depending on the distance to the source, long-baseline experiments like KamLAND [32] are sensitive to the solar mass splitting Δm^2_{21} whereas much shorter baseline experiment such as RENO [33] or DayaBay [34] measure θ_{13} and Δm^2_{31}.

Accelerator experiments measure neutrino fluxes generated in particle accelerators. Usually mesons are produced in the accelerator to be focused into a beam, then some decay to muon neutrinos and the rest are absorbed by a target. Depending on the

[1] These are the results reported during M. Tórtola's talk at Neutrino 2024 (see this link). I need to keep an eye and see if they publish these or other updated results in the near future.

Table 2.1: Summary of neutrino oscillation parameters determined in the Neutrino Global Fit of 2020 [16].

Parameter	Best fit $\pm$ 1σ	3σ range
Δm^2_{21} [eV^2 $\times$ 10^{-5}]	$7.55^{+0.22}_{-0.20}$	$6.98 - 8.19$
$\lvert\Delta m^2_{31}\rvert$ [eV^2 $\times$ 10^{-3}] (NO)	$2.51^{+0.02}_{-0.03}$	$2.43 - 2.58$
$\lvert\Delta m^2_{31}\rvert$ [eV^2 $\times$ 10^{-3}] (IO)	$2.41^{+0.03}_{-0.02}$	$2.34 - 2.49$
$\sin^2\theta_{12}/10^{-1}$	3.04 ± 0.16	$2.57 - 3.55$
$\sin^2\theta_{23}/10^{-1}$ (NO)	$5.64^{+0.15}_{-0.21}$	$4.23 - 6.04$
$\sin^2\theta_{23}/10^{-1}$ (IO)	$5.64^{+0.15}_{-0.18}$	$4.27 - 6.03$
$\sin^2\theta_{13}/10^{-2}$ (NO)	$2.20^{+0.05}_{-0.06}$	$2.03 - 2.38$
$\sin^2\theta_{13}/10^{-2}$ (IO)	$2.20^{+0.07}_{-0.04}$	$2.04 - 2.38$
δ_{CP}/π (NO)	$1.12^{+0.16}_{-0.12}$	$0.76 - 2.00$
δ_{CP}/π (IO)	$1.50^{+0.13}_{-0.14}$	$1.11 - 1.87$

configuration one can obtain a beam made of mostly neutrinos or antineutrinos. The typical energies of these neutrinos are in the GeV range. Experiments such as NOvA [35], T2K [36], MINOS [37], OPERA [?] and K2K [38] (and in the future DUNE [39]) are primarily sensitive to θ_{13}, θ_{23} and Δm^2_{32}. Also, in the coming years DUNE [39] and Hyper-Kamiokande [40] will be sensitive to δ_{CP}.

2.3 Open questions in the neutrino sector

A crucial question that remains open these days, and is of vital importance for oscillation phenomena, is whether the mass eigenvalue ν_3 is the heaviest (what we call normal ordering) or the lightest (refered to as inverted ordering) of the mass eigenstates. In other words, this means that we do not know the sign of Δm^2_{32}, so we can either have $m_1 < m_2 < m_3$ (NO) or $m_3 < m_1 < m_2$ (IO).

Another big puzzle is related to the value of δ_{CP}. Nowadays it is poorly constrained, with all values between π and 2π being consistent with data. A prospective measurement different from $\delta_{CP} = 0, \pi$ will predict CP-violation in the leptonic sector, and thus contribute along with the one measured in the quark sector to the total amount of CP-violation. Although it is true that these two contributions by themselves are not enough to explain the matter anti-matter asymmetry in our universe, the amount of CP-violation in the leptonic sector can be key to explain such imbalance.

Both of these questions, because of their nature, could be understood thanks to future oscillation experiments.

Notwithstanding, there are other mysteries that can not be unveiled just by conducting oscillation experiments, as certain quantities do not influence these phenomena. Among these there is the question of the absolute values of the neutrino masses. Depending on the value of the lightest of the neutrino masses we can have different mass spectra, from hierarchical $m_1 \ll m_2 < m_3$ (NO) or $m_3 \ll m_1 < m_2$ (IO) to quasi-degenerate $m_1 \simeq m_2 \simeq m_3$.

Other open question concerns the nature itself of the neutrinos. If neutrinos are Dirac particles then their mass term can be generated through the usual Higgs mechanism by adding right-handed neutrino fields. However, if they are Majorana particles and therefore their own antiparticles, there is no need to add extra fields to have the mass term in the Lagrangian. Experiments like SuperNEMO [**?**], SNO+ [**?**] and NEXT [**?**], which search for neutrino-less double beta decay, will be able to determine whether neutrinos are Dirac or Majorana.

Chapter 3

The Deep Underground Neutrino Experiment

The Deep Underground Neutrino Experiment (DUNE) is a next generation long-baseline neutrino experiment [1]. It will aim to address several questions in neutrino physics, study neutrinos from astrophysical sources and search for beyond the standard model physics.

This chapter reviews the main goals of the DUNE experiment, the design of the far detector modules and their data acquisition (DAQ) system, and the role that the near detector plays in the physics program of DUNE.

3.1 Overview

The main physics goals of DUNE are:

- measure the neutrino mass hierarchy, the amount of CP violation in the leptonic sector and the θ_{23} octant,
- detect rare low energy neutrino events, like neutrinos from supernova bursts, and
- search for proton decay and other beyond the standard model phenomena.

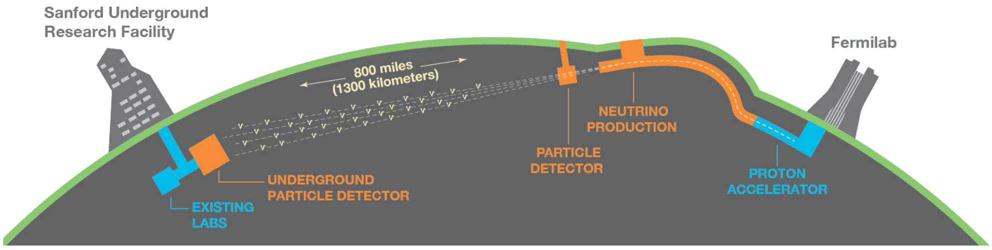


Figure 3.1: Schematic diagram of the DUNE experiment and the LBNF beamline [1].

The design of DUNE has bee tailored with these goals in mind. It will consist of two neutrino detectors. A near detector (ND) complex will be placed in Fermilab, 574 m downstream of the neutrino production point, whereas a larger far detector (FD) will be built in the Sandford Underground Research Facility (SURF), South Dakota, approximately 1300 km away. Fig. 3.1 shows a simplified view of the various components of DUNE (not to scale).

The beam neutrinos to be used in DUNE will be provided by the LBNF beamline, the multi-megawatt wide-band neutrino beam planned for Fermilab. First, an intense proton beam is extracted from the Fermilab Main Injector. Then, these protons with energies between 60 GeV and 120 GeV collide with a high-power production target and produce charged mesons. Two magnetic horns allow to focus the mesons and perform a sign selection (thus having the capability to switch between neutrino and antineutrino mode). Soon after that, the mesons decay and produce neutrinos (or antineutrinos) which are then aimed to SURF.

Before arriving to the FD, the neutrino beam meets the ND complex, which serves as the experiment's control. Its role is to measure the unoscillated neutrino energy spectra. From these we can predict the unoscillated spectra at the FD, which can be compared to the spectra measured at the FD in order to extract the oscillation parameters. Therefore, the design of the DUNE ND is mainly driven by the needs of the oscillation physics program.

The liquid Argon time projection chamber (LArTPC) technology has been chosen for

Table 3.1: Summary of the two-phased plan for DUNE. Adapted from Ref. [41].

Parameter	Phase I	Phase II	Benefit
FD mass	20 kt fiducial	40 kt fiducial	FD statistics
Beam power	up to 1.2 MW	2.4 MW	FD statistics
ND config.	ND-LAr, TMS, SAND	ND-LAr, ND-GAr, SAND	Systematic constraints

the FD modules of DUNE. Its four modules will record neutrino interactions from the accelerator-produced beam arriving at predictable times. As it also aims at recording rare events, the FD requires trigger schemes which can deal with both kinds of physics, and also maximum uptime.

DUNE is planned to be built using a staged approach consisting on two phases, which are summarised in Tab. 3.1. Phase I consists of a FD with 50% of the total fiducial mass, a reduced version of the ND complex and a 1.2 MW proton beam. It will be sufficient to achieve some early physics goals, like the determination of the neutrino mass ordering. For its Phase II, DUNE will feature the full four FD modules, a more capable ND and a 2.4 MW proton beam. The physics milestones for the two phases are given in Tab. 3.2, in a staging scenario which assumes that Phase II is completed after 6 years of operation.

A summary of the DUNE science program can be found in the DUNE FD Technical Design Report (TDR) Volume I [1]. For a detailed discussion on the two-phased approach the reader is referred to the DUNE Snowmass 2021 report [41].

3.2 Physics goals of DUNE

As noted in the literature (see for instance Ref. [16] for a review), the parameter space of the neutrino oscillation phenomena within the three-flavour picture is quite constrained by current experimental data. However, there are still crucial open questions, like the mass ordering, the value of δ_{CP} or the θ_{13} octant. One of the main goals of DUNE is to shed some light on the values of these parameters [42].

To address these questions DUNE can look to the subdominant oscillation channel

Table 3.2: Exposure and time required to achieve the different physics milestones of the two phases. The predictions assume a Phase II staging scenario where FD modules 3 and 4 are deployed in years 4 and 6 and both the beam and ND are upgraded after 6 years. Adapted from Ref. [41].

Stage	Physics milestone	Exposure (kt-MW-years)	Years (staged)
Phase I	5σ MO ($\delta_{CP} = -\pi/2$)	16	1-2
	5σ MO (100% of the δ_{CP} values)	66	3-5
	3σ CPV ($\delta_{CP} = -\pi/2$)	100	4-6
Phase II	5σ CPV ($\delta_{CP} = -\pi/2$)	334	7-8
	δ_{CP} resolution of 10 degrees ($\delta_{CP} = 0$)	400	8-9
	5σ CPV (50% of the δ_{CP} values)	646	11
	3σ CPV (75% of the δ_{CP} values)	936	14
	$\sin^2(2\theta_{13})$ resolution of 0.004	1079	16

$\nu_\mu \to \nu_e$ ($\bar{\nu}_\mu \to \bar{\nu}_e$) and study the energy dependence of the ν_e ($\bar{\nu}_e$) appearance probability. When we focus on the antineutrino channel $\bar{\nu}_\mu \to \bar{\nu}_e$ there is a change in the sign of δ_{CP}, thus introducing CP-violation. Moreover, due to the fact that there are no positrons in the composition of Earth, there is a sign difference for the matter effect contribution when looking to the antineutrino channel. This asymmetry is proportional to the baseline length L and is sensitive to the sign of Δ_{31}, and thus to the neutrino mass ordering.

Another of the main physics goals of DUNE is the search for baryon-number violating processes. Specifically, it will try to answer the question of whether protons are stable or not. There is no symmetry argument that forbids protons from decaying, but its apparent stability seems to suggest that baryon number is conserved [43]. However, proton decay is a usual feature of grand-unified theories, where electromagnetic, weak and strong interactions are unified above a certain energy scale [44].

As the energy deposition scale for this kind of searches is nearly the same as the one for long-baseline neutrino oscillations, DUNE will be able to look for them. It has several advantages over other experiments, such as excellent imaging and particle identification, which can be translated to lower backgrounds.

The last of the main objectives of DUNE is the detection of neutrinos originated in supernovae explosions, what is called a supernova neutrino burst (SNB). These neutrinos carry with them information about the core-collapse process, from the progenitor to the explosion and the remnant; but also may have information about new exotic physics. So far, the only neutrino events ever recorded from such a process were a few dozens of $\bar{\nu}_e$ events from the 1987A supernova located in the Magellanic Cloud, 50 kpc away from Earth [45, 46].

DUNE aims to collect also some SNB events. Although these are quite rare, as the expected supernovae explosion events are about one every few decades for our galaxy and Andromeda, the long lifetime of the experiment (around a few decades as well) makes it reasonable to expect some. Nowadays the main sensitivity to SNB of most experiments is to the $\bar{\nu}_e$ through inverse beta decay. One of the advantages of DUNE is its expected sensitivity to ν_e, since the dominant channel will be ν_e CC scattering.

Moreover, due to the stringent requirements that the main physics goals set for DUNE, it will allow also to perform searches for all kind of BSM physics. Among others, DUNE will be able to look for: active-sterile neutrino mixing, non-unitarity of the PMNS matrix, non-standard interactions, Lorentz and CPT violations, neutrino trident production, light-mass DM, boosted DM and heavy neutral leptons. The reader is referred to the DUNE FD TDR Volume II [42] for a full discussion of the physics scope of DUNE.

3.3 Far Detector

The so-called DUNE FD complex will sit 1.5 km underground at SURF, South Dakota. Two caverns will host the four FD modules, two of them per cavern, each embedded in cryostats of dimensions 18.9 m (w) $\times$ 17.8 m (h) $\times$ 65.8 m (l). A central, smaller cavern will host the cryogenic system.

Three out of the four modules will be liquid argon (LAr) time projection chamber detectors, often refer to as LArTPCs, with a LAr fiducial mass of at least 10 kt each. The first and second FD modules, FD-1 and FD-2, will use a Horizontal Drift (HD) technology, whereas the third module, FD-3, will have a Vertical Drift (VD) direction. The technology for the fourth module is still to be decided,

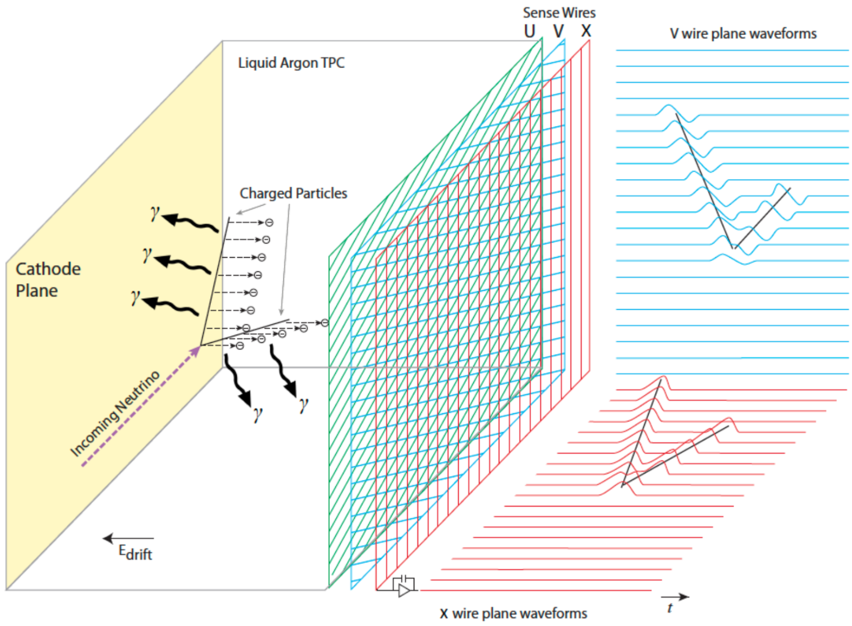


Figure 3.2: Schematic diagram showing the operating principle of a LArTPC with wire readout. Figure taken from Ref. [1].

For each event, with energies ranging from a few MeV to several GeV, these detectors collect both the scintillation light and the ionisation electrons created when the charged particles produced in neutrino-nucleus interactions ionise the argon nuclei. In both HD and VD designs the characteristic 128 nm scintillation light of argon is collected by a photon detection system (PDS). This light will indicate the time at which electrons start to drift, thus enabling reconstruction over the drift coordinate when compared to the time when the first ionisation electron arrives to the anode. Reconstruction of the topology in the transverse direction is achieved using the charge readout. Fig. 3.2 illustrates the detection principle described, for the case of a HD detector with a wire readout.

3.3.1 Horizontal Drift

Within the HD design the ionisation electrons produced as charged particles traverse the LAr drift horizontally towards the anode planes, made out of three layers of wire readout, due to the effect of an electric field. This design, previously known as single-phase (SP), was tested by the ProtoDUNE-SP detector at CERN. The prototype collected data from

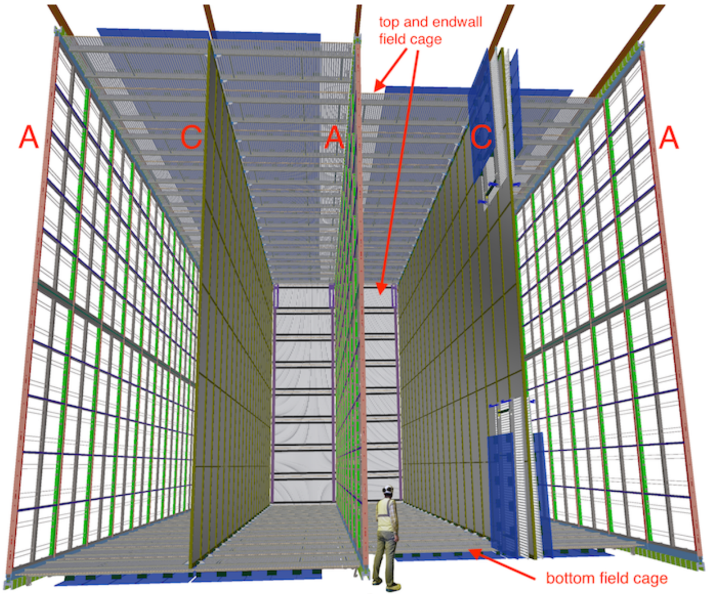


Figure 3.3: Proposed design for the FD-1 and FD-2 modules following the HD principle. Figure taken from Ref. [1].

a hadron beam and cosmic rays, providing high-quality data sets for calibration studies and proving the excellent performance of this design.

Each FD HD detector module is divided in four drift regions, with a maximum drift length of 3.5 m, by alternating anode and cathode walls. The surrounding field cage ensures the uniformity of the 500 V/cm horizontal electric field across the drift volumes. The three anode walls, which constitute the charge readout of the detector, are built by stacking anode plane assemblies (APAs), 2 high times 25 wide. The design of the HD modules is shown in Fig. 3.3.

Each APA is made of 2560 active wires arranged in three layers, plus an extra grid layer, wrapped around a metal frame. The two induction wire planes, U and V, sit at $\pm 35.7°$ to the vertical on each side of the APA. The collection and shielding plane wires, X and G, run parallel to the vertical direction. The ionisation electrons drift past the induction planes, generating bipolar signals on those wires, and are collected by the collection plane, producing a monopolar positive signal. The spacing between the wires is ~ 5 mm, and it defines the spatial resolution of the APA.

The front-end readout electronics, or cold electronics as they are immerse in the LAr,

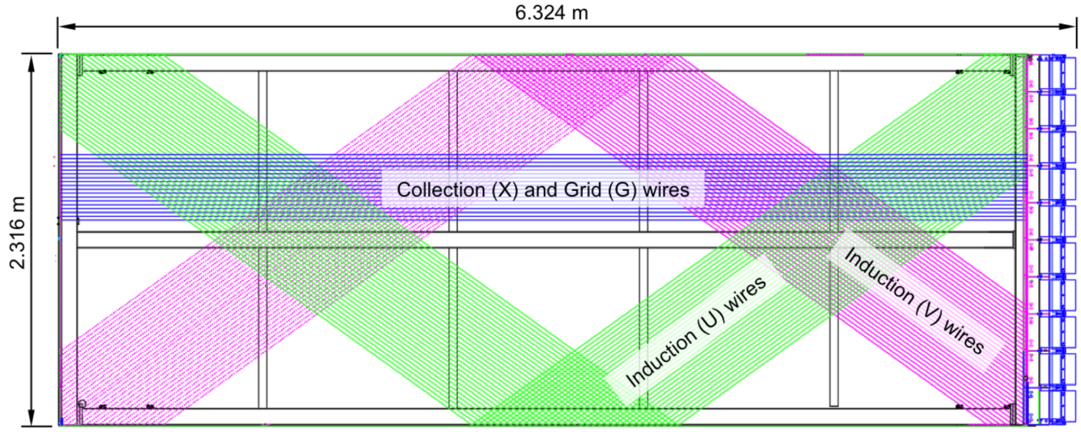


Figure 3.4: Schematic representation of an APA. The black lines represent the APA steel frame. The green and magenta lines correspond to the direction of the U and V induction wires respectively. The blue lines indicate the direction of the X collection wires and the wire shielding G. Figure taken from Ref. [1].

are attached to the top of the up APAs and the bottom of the down APAs. Mounted on the front-end mother boards we have a series of ASICs that digitize the signals from the collection and induction planes. Each wire signal goes to a charge-sensitive amplifier, then there is a pulse-shaping circuit and this is followed by the analogue-to-digital converter. This part of the process happens inside the LAr to minimise the number of cables penetrating the cryostat. The digitised signals come out finally via a series of high-speed serial links to the warm interface boards (WIBs), from where the data is sent to the back-end DAQ through optical fibers.

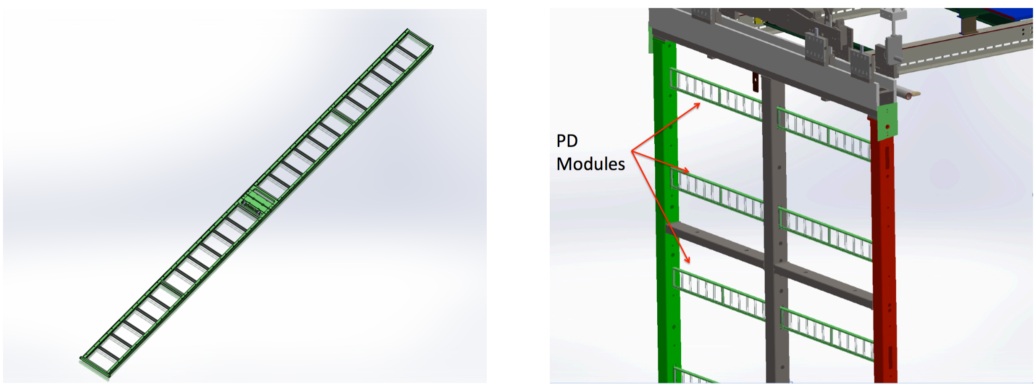


Figure 3.5: A PDS module containing 24 X-ARAPUCAs (left) and the location of the modules on the APAs (right). Figure taken from Ref. [1].

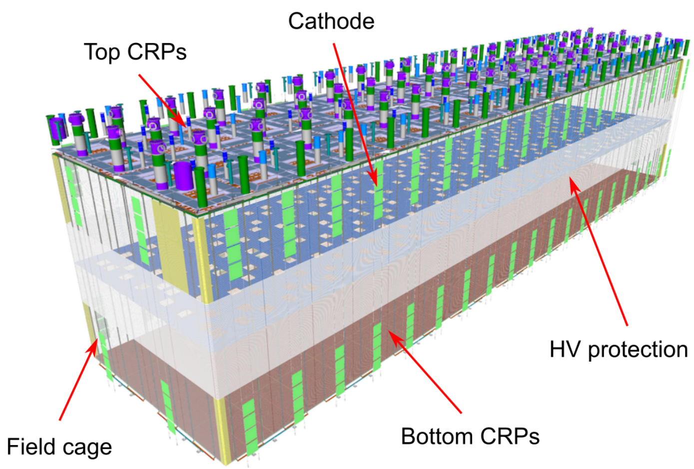


Figure 3.6: Proposed design for the FD-3 module following the VD principle. Figure adapted from Ref. [47].

The PDS uses modules of X-ARAPUCA devices, mounted on the APA frames between the wire planes. Each X-ARAPUCA consists of layers of dichroic filter and wavelength-shifter. They shift the VUV scintillation light into the visible spectrum, sending then the visible photons to silicon photomultiplier (SiPM) devices. The PDS modules are 209 cm $\times$ 12 cm $\times$ 2 cm bars, containing 24 X-ARAPUCAs. There are 10 of these PDS modules per APA. Fig. 3.5 shows a PDS module (left) and the placement of the modules on the APAs (right).

3.3.2 Vertical Drift

In the VD case the ionisation electrons will drift vertically until they meet a printed circuit board-based (PCB) readout plane. It is based on the original dual-phase (DP) design deployed at CERN, known as ProtoDUNE-DP, used a vertical drift design with an additional amplification of the ionization electrons using a gaseous argon (GAr) layer above the liquid phase. The VD module incorporates the positive features of the DP design without the complications of having the LAr-GAr interface.

The current design of the FD VD module counts with two drift chambers with a maximum drift distance of 6.5 cm. A cathode plane splits the detector volume along the

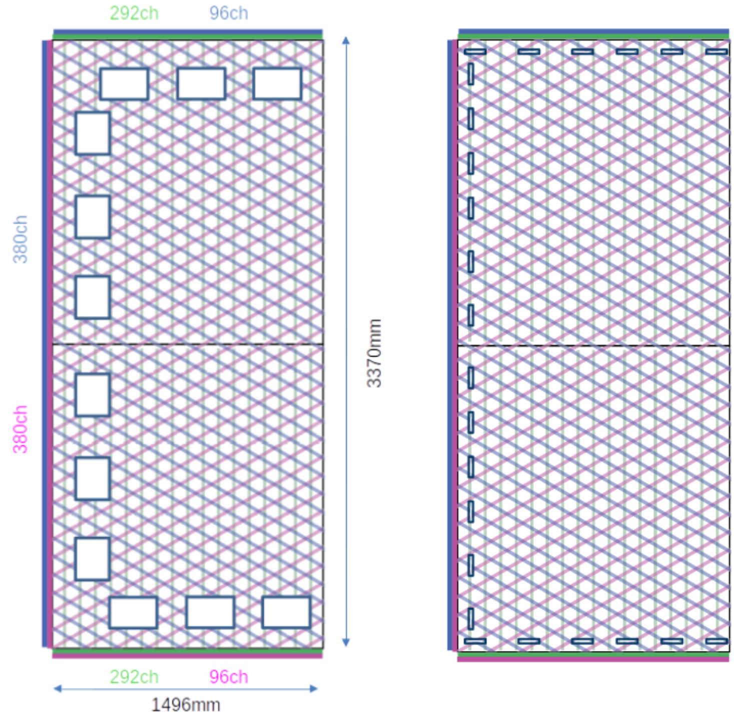


Figure 3.7: Schematic representation of the electrode strip configuration for a top (left) and bottom (right) CRU. Figure taken from Ref. [47].

drift direction while the two anode planes are connected to the bottom and top walls of the detector. The layout of the VD module is shown in Fig. 3.6. Compared with the HD design, the VD option offers a slightly larger instrumented volume and a more cost-effective solution for the charge readout.

As in the HD design, each drift volume features a 500 V/cm electric field and a field cage that ensures its uniformity. The anode planes are arrays of 3.4 m $\times$ 3 m charge-readout planes (CRPs). These are formed by a pair of charge-readout units (CRUs), which are built from two double-sided perforated PCBs, with their perforations aligned. The perforations allow the drift electrons to pass between the layers.

The PCB face opposite to the cathode has a copper guard plane which acts as shielding, while its reverse face is etched with electrode strips forming the first induction plane. The outer PCB has electrode strips on both faces, the ones facing the inner PCB form the second induction plane while the outermost ones form the collection plane. Fig. 3.7 shows the layout of the electrode strips for the top (left) and bottom (right) CRUs.

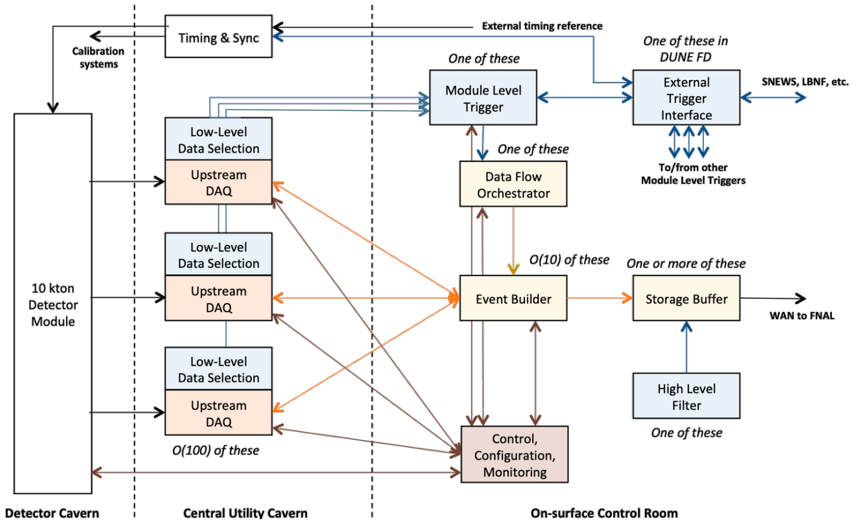


Figure 3.8: Detailed diagram of the DUNE FD DAQ system. Figure taken from Ref. [2].

The magenta and blue lines represent the first and second induction planes respectively, and the green lines correspond to the collection plane.

The PDS in the VD module will use the same X-ARAPUCA technology developed for the HD design. The plan is to place the PDS modules on the cryostat walls and on the cathode, in order to maximise the photon yield.

3.3.3 FD Data Acquisition System

The task of the data acquisition (DAQ) system is to receive, process and store data from the detector modules. In the case of DUNE the DAQ architecture is designed to work for all FD modules interchangeably, except some aspects of the upstream part which may depend on the specific module technology.

The enormous sample rate and the number of channels in TPC and PD readouts will produce a very large volume of data. These pose really strong requirements and challenges to the DUNE FD DAQ architecture. It will be required to read out data of the order of ten thousand or more channels at rates of a few MHz. In order to cope with the huge data volume, segmented readouts and compression algorithms are used to reduce the data rate to manageable levels.

The DAQ system of the DUNE FD is composed of five different subsystems. The first one is the upstream DAQ, which receives the raw data from the detector, buffers it and perform some low-level pre-processing. The minimally processed data is then fed into a hierarchical data selection system, which then performs a module level trigger decision. In case of a positive decision a trigger command is produced and executed by the data flow orchestrator, located in the back-end (BE) DAQ subsystem. Subsequently the DAQ BE retrieves the relevant data from the buffers located in the upstream DAQ, adds all the data into a cohesive record and saves it to permanent storage. Watching over all the other subsystems we also have the control, configuration and monitoring subsystem and the time and synchronization subsystem. Fig. 3.8 shows a schematic diagram of the DAQ system, showing the different subsystems and their relations.

A notorious challenge for the DUNE DAQ system comes from the its broad physics goals. We must be prepared to process events spanning a wide range of time windows (from 5 ms in the case of beam and cosmic neutrinos and nucleon decay to 100 s in the case of SNBs) and therefore this requires a continuous readout of the detector modules. Moreover, because of the off-beam measurements we need to ensure the capabilities of online data processing and self-triggering. Having this into account, together with the technical constraints, the DUNE FD DAQ faces a series of challenges: it needs to be fault tolerant and redundant to reduce downtime, accommodate new components while it keeps serving the operational modules, have large upstream buffers to handle SNB physics, be able to support a wide range of readout windows and last reduce the throughput of data to permanent storage to be at most 30 PB/year.

3.4 Near Detector

In order to estimate the oscillation parameters we measure the neutrino energy spectra at the FD. This reconstructed energy arises from a convolution of the neutrino flux, cross section, detector response and the oscillation probability. Using theoretical and empirical models to account for the other effects, one can extract the oscillation probability using

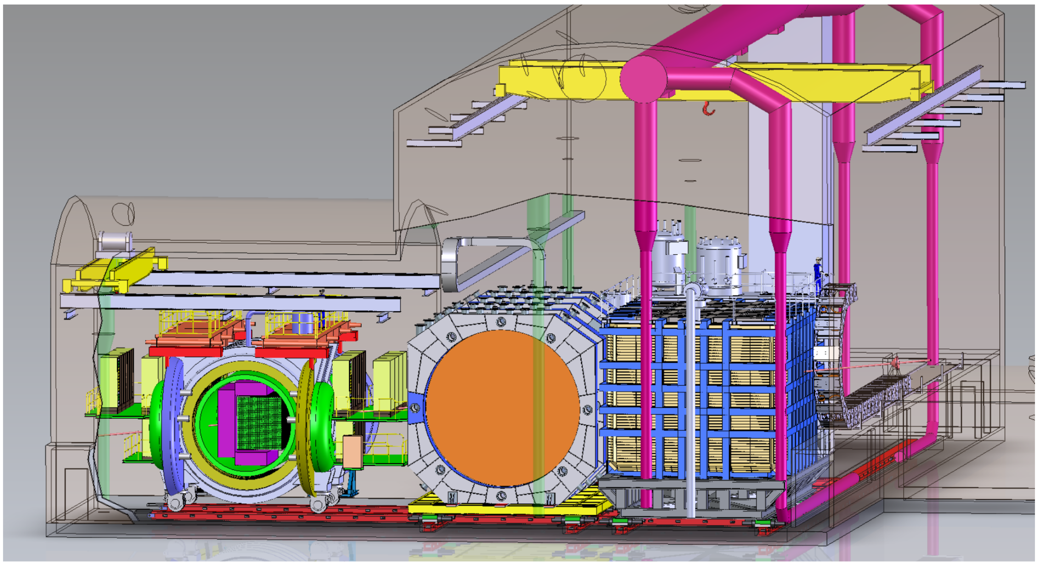


Figure 3.9: Representation of the ND hall in Phase II, showing the different subcomponents. From right to left, in the direction of the beam, we have ND-LAr, ND-GAr and SAND. Figure taken from Ref. [48].

the measurement. However, these models have associated a number of uncertainties that are then propagated to the oscillation parameters.

One of the main roles of the ND is to measure the neutrino interaction rates before the oscillation effects become relevant, i.e. close to the production point. By measuring the ν_μ and ν_e energy spectra, and that of their corresponding antineutrinos, at the ND we can constrain the model uncertainties. A complete cancellation of the uncertainties when taking the ratio between the FD and ND measurements is not possible, as that would require both detectors to have identical designs and the neutrino fluxes to be the same. Because of the distance, the flux probed by the FD will have a different energy and flavour composition than that at the ND, as neutrinos oscillate and the beam spreads. The differences in the flux also determine the design of the detectors, therefore the ND is limited in its capability to match the FD design.

Nevertheless, having a highly capable ND DUNE can minimise the systematic uncertainties affecting the observed neutrino energy. The ND data can be used to tune the model parameters by comparison with the prediction. Then, one uses the tuned model to predict the unoscillated FD spectra. Comparing the prediction with the

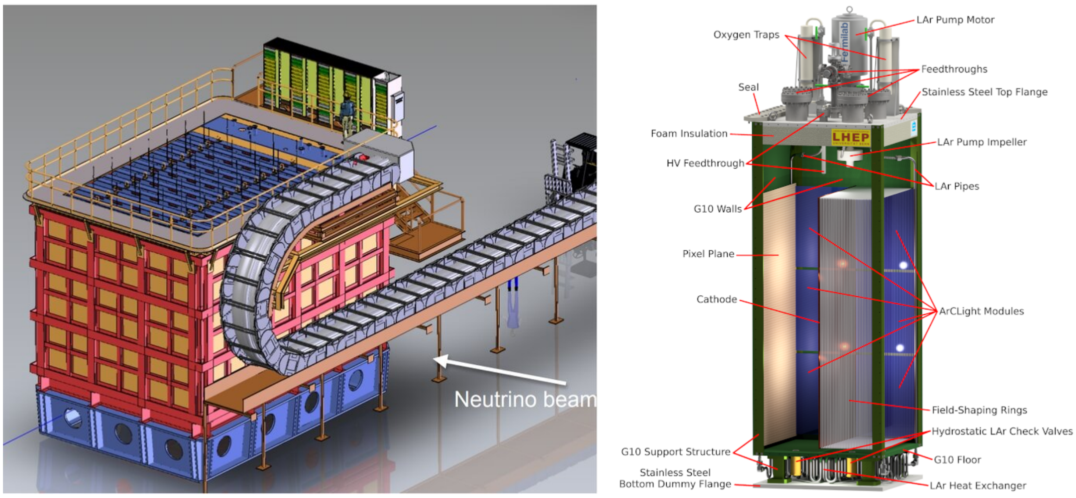


Figure 3.10: Schematic representation of the external components of ND-LAr, including the cryostat and the PRISM movable system (left) and detailed drawing of one ArgonCube module (right). Figure adapted from Ref. [1].

measured spectra it is possible to extract the oscillation parameters.

Additionally, the ND will have a physics program of its own. In particular, it will measure neutrino cross sections that will then be used to constrain the model used in the long-baseline oscillation analysis. It will also be used to search for BSM phenomena such as heavy neutral leptons, dark photons, millicharged particles, etc.

The DUNE ND can be divided in three main components, a LArTPC known as ND-LAr, a magnetised muon spectrometer, which will be the Temporary Muon Spectrometer (TMS) in Phase I and ND-GAr in Phase II, and the System for on-Axis Neutrino Detection (SAND). The layout of the Phase II DUNE ND can be seen in Fig. 3.9. The first two components of the ND will be able to move off-axis, in what is called the Precision Reaction-Independent Spectrum Measurement (PRISM) concept. More details on the purpose and design of the ND can be found in the DUNE ND Conceptual Design Report (CDR) [48].

3.4.1 ND-LAr

ND-LAr is a LArTPC, as the ND needs a LAr component in order to reduce cross section and detector systematic uncertainties in the oscillation analysis. However, its

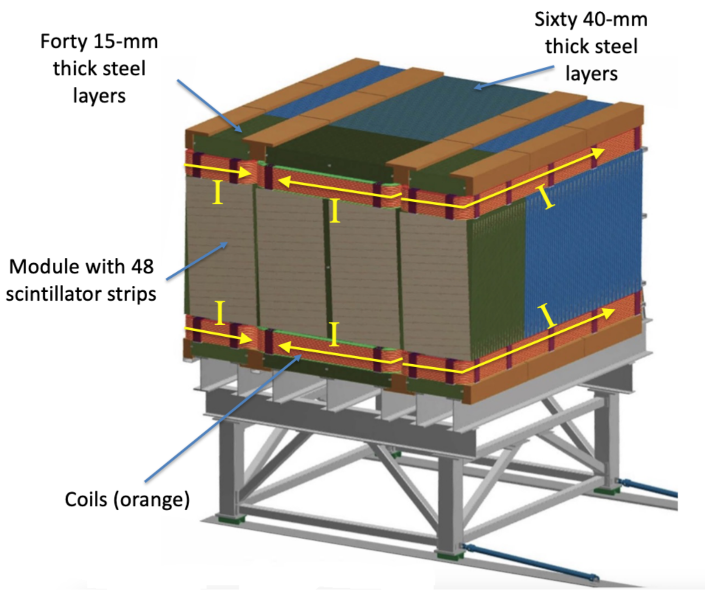


Figure 3.11: Schematic view of the TMS detector, highlighting its main parts. Figure adapted from Ref. [1].

design differs significantly from those proposed for the FD modules. Because of the high event rates at the ND, approximately 55 neutrino interaction events per 10 μs spill, ND-LAr will be built in a modular way. Each of the modules, based on the ArgonCube technology, is a fully instrumented, optically isolated TPC with a pixelated readout. The pixelisation allows for a fully 3D reconstruction and the optical isolation reduces the problems due to overlapping interactions. Fig. 3.10 shows a representation of the external parts of ND-LAr (left) and a detailed diagram of an ArgonCube module (right).

With a fiducial mass of 67 t and dimensions 7 m (w) $\times$ 3 m (h) $\times$ 5 m (l), ND-LAr will be able to provide high statistics and contain the hadronic systems from the beam neutrino interactions, but muons with a momentum higher than 0.7 GeV will exit the detector.

3.4.2 TMS/ND-GAr

In order to accurately estimate the neutrino energy, the momentum of the outgoing muons needs to be determined. That is the reason why a muon spectrometer is needed downstream of ND-LAr.

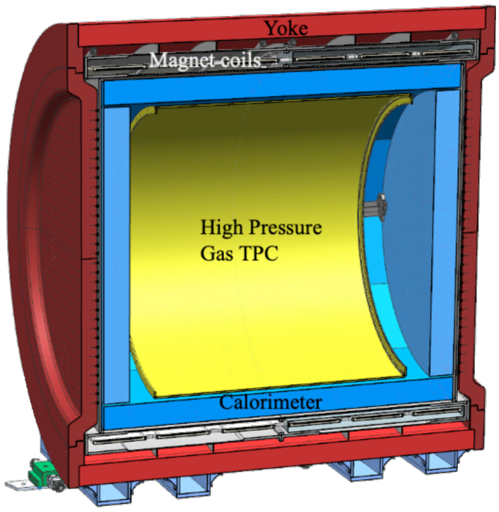


Figure 3.12: Cross section of the ND-GAr geometry, showing the HPgTPC, ECal and magnet. Figure adapted from Ref. [1].

In Phase I that role will be fulfilled by TMS. It is a magnetised sampling calorimeter, with alternating steel and plastic scintillator layers. Fig. 3.11 shows a schematic view of the TMS detector. The magnetic field allows a precise measurement of the sign of the muon, so one can distinguish between neutrino and antineutrino interactions.

After the Phase II upgrade, TMS will be replaced with ND-GAr. This detector is a magnetised, high-pressure GAr TPC (often denoted as HPgTPC) surrounded by an electromagnetic calorimeter (ECal) and a muon tagger. A cross section of its geometry can be seen in Fig. 3.12. ND-GAr will be able to measure the momenta of the outgoing muons while also detect neutrino interactions inside the GAr volume. This allows ND-GAr to constrain the systematic uncertainties even further, as it will be able to accurately measure neutrino interactions at low energies thanks to the lower tracking thresholds of GAr.

3.4.3 PRISM

In general, the observed peak neutrino energy of a neutrino beam decreases as the observation angle with respect to the beam direction increases. This feature has been used in other long-baseline neutrino experiments, like T2K (2.5° off-axis) and NOvA

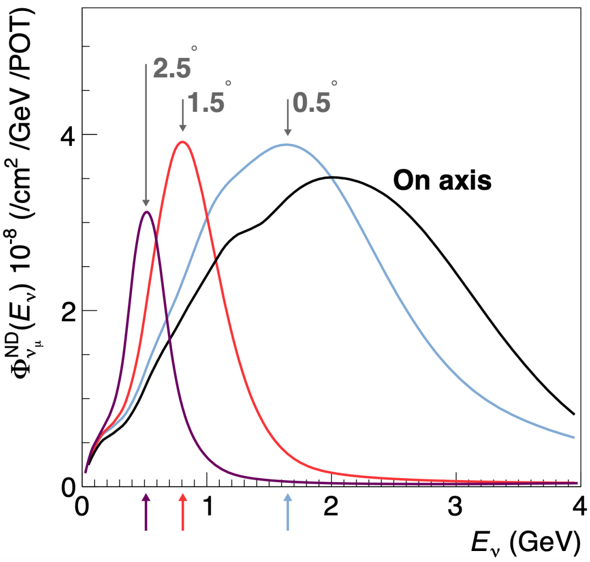


Figure 3.13: Predicted beam muon neutrino flux at the ND location for different off-axis positions. Figure taken from Ref. [48].

(0.8° off-axis), in order to achieve narrower energy distributions. The DUNE PRISM concept exploits this effect using a movable ND. Within PRISM both ND-LAr and the muon spectrometer (TMS in Phase I and ND-GAr in Phase II) can be moved up to 3.2° off-axis, equivalent to move the detectors 30.5 m laterally through the ND hall.

This allows to record additional data samples with different energy compositions. Fig. 3.13 compares the on-axis muon neutrino flux at the ND with the fluxes at different off-axis positions. As the off-axis position increases the neutrino flux becomes closer to a monoenergetic beam with a lower peak energy. These samples can be used to perform a data-driven determination of the relation between true and reconstructed neutrino energy, in order to reduce the dependence on the interaction model. The off-axis samples are linearly combined to produce a narrow Gaussian energy distribution centered on a target true energy. From the combination coefficients one can build a sample of reconstructed neutrino events that will determine the energy mapping.

The PRISM samples can also be used to form a flux at the ND location similar in shape to the oscillated flux measured by the FD. This method can be used to extract the oscillation parameters with minimal input from the neutrino interaction model.

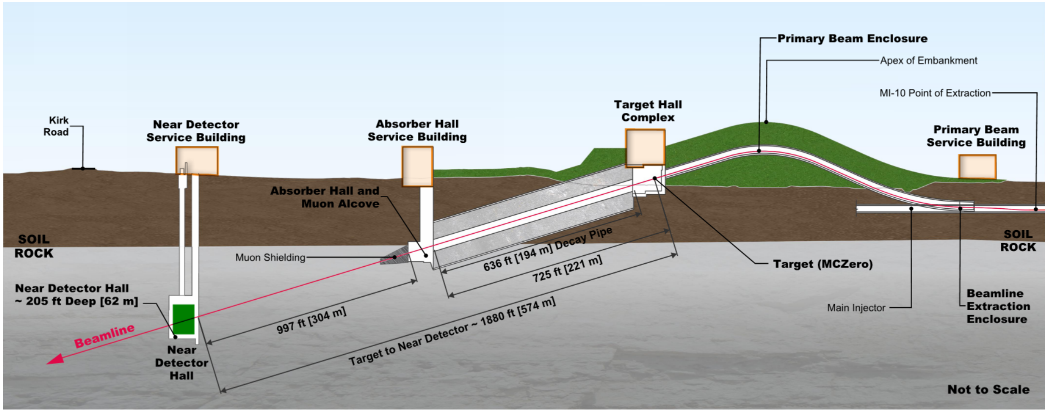


Figure 3.14: Schematic longitudinal section of the LBNF beamline at Fermilab (not to scale). Figure taken from Ref. [49].

3.4.4 SAND

The role of SAND is to monitor the beam stability by measuring the on-axis neutrino energy spectra. As the PRISM program requires that ND-LAr and its downstream muon spectrometer spend about half of the time in off-axis positions, it is not possible to monitor the stability with the movable detectors. Moreover, for the success of PRISM it is essential to have a stable beam configuration, or, at least, a quick assessment and modeling of the distortions.

The SAND detector is magnetised, and it counts with an inner low density tracker, a LAr target with optical readout and surrounding sampling calorimeter.

3.5 LBNF beamline

The Long-Baseline Neutrino Facility (LBNF) project is responsible for producing the neutrino beam for the DUNE detectors. A detailed discussion of the LBNF program can be found in the DUNE/LBNF CDR Volume III [49].

The LBNF beamline will provide a high-intensity neutrino beam within the adequate energy range in order to meet the long-baseline oscillation physics goals of DUNE. A schematic diagram of the longitudinal section of the LBNF beamline is shown in Fig. 3.14. First, a beam of 60 – 120 GeV protons is extracted from the Fermilab Main

Injector. This beam is aimed towards the target area, where it collides with a cylindrical graphite target to produce pions and kaons.

The diffuse, secondary beam of particles is focused by a pair of magnetic horns. These select the positively charged particles when operated in Forward Horn Current (FHC) mode, or the negatively charged ones when the current is reversed, also known as Reverse Horn Current (RHC) mode. The focused secondary beam then enters a 194 m decay pipe where the pions and kaons will predominantly produce $\mu^+\nu_\mu$ pairs when in FHC mode (or $\mu^-\bar{\nu}_\mu$ in RHC mode).

At the end of the decay pipe a hadron absorber removes the undecayed hadrons and muons from the beam, which reduces the ν_e ($\bar{\nu}_e$) and $\bar{\nu}_\mu$ (ν_μ) contamination coming from the μ^+ (μ^-) decays. The resulting neutrino flux at the FD is shown in Fig. 3.15, both for FHC (left) and RHC (right) modes. These predictions show the intrinsic $\overset{(-)}{\nu}_e$ contamination and wrong sign component from wrong sign and neutral meson decays, as well as muons decaying before reaching the absorber.

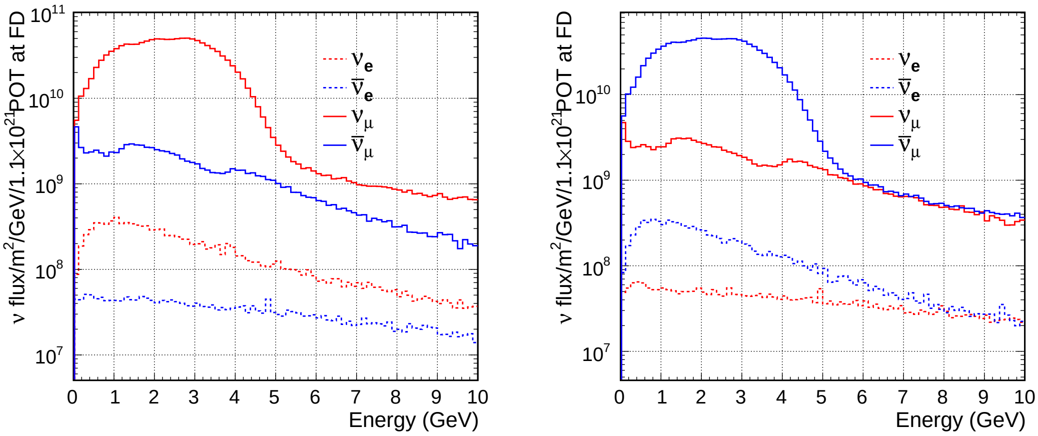


Figure 3.15: Predicted neutrino fluxes at the FD in FHC mode (left panel) and RHC mode (right panel). Figures taken from Ref. [42].

Chapter 4

ND-GAr

ND-GAr is a magnetised, high-pressure gaseous argon TPC (HPgTPC), surrounded by an electromagnetic calorimeter (ECal) and a muon detector (commonly refer to as μID). A detailed discussion on the requirements, design, performance and physics of ND-GAr can be found in the DUNE ND CDR [48] and the ND-GAr whitepaper (cite).

In DUNE Phase II ND-GAr will fulfill the role of TMS, measuring the momentum and sign of the charged particles exiting ND-LAr. Additionally, it will be able to measure neutrino interactions inside the HPgTPC, achieving lower energy thresholds than those of the ND and FD LArTPCs. By doing so ND-GAr will allow to constrain the relevant systematic uncertainties for the LBL analysis even further.

The goal of the present chapter is to review the requirements that the physics program of DUNE impose on ND-GAr, present the current status of its design and describe the GArSoft package, its simulation and reconstruction software.

4.1 Requirements

The primary requirement for ND-GAr is to the measure the momentum and charge of muons from ν_μ and $\bar{\nu}_\mu$ CC interactions in ND-LAr, in order to measure their energy spectrum. To achieve the sensitivity to the neutrino oscillation parameters described in the DUNE FD TDR Volume II [42] ND-GAr should be able to constrain the muon energy within a 1% uncertainty or better. The main constraint will come from the calibration of the magnetic field, performed using neutral kaon decays in the HPgTPC.

Another requirement for ND-GAr is the precise measurement of neutrino interactions on argon for the energies relevant to the neutrino oscillation program. The goal is to constrain the cross section systematic uncertainties in the regions of phase space that are not accessible to ND-LAr. This requires the kinematic acceptance for muons in ND-GAr to exceed that of ND-LAr, being comparable to the one observed in the FD.

ND-GAr should also be able to the relationship between true and reconstructed energy from neutrino interactions on argon with low thresholds, being sensitive to particles that are not observed or may be misidentified in ND-LAr. In particular, ND-GAr needs to have low tracking thresholds in order to measure the spectrum of pions and protons produced in final-state interactions (FSI). It also must be able to accurately measure the pion multiplicity in 1, 2 and 3 pions final states, to inform the pion mass correction in the LArTPCs.

4.2 Reference design

The final design of ND-GAr is still under preparation. However, a preliminary baseline design was in place at the time of the ND CDR. This section summarises the main features of that design, as it is also the one used for the default geometry in our simulation. A DUNE Phase II whitepaper, discussing the different options under consideration for the ND-GAr design, is in progress.

4.2.1 HPgTPC

The reference design for the ND-GAr HPgTPC follow closely that of the ALICE TPC. It is a cylinder with a central high-voltage cathode, generating the electric field for the two drift volumes, with a maximum drift distance of 2.5 m each. The anodes will be instrumented with charge readout chambers. The original design repurposed the multi-wire proportional readout chambers of ALICE, however the current R&D efforts

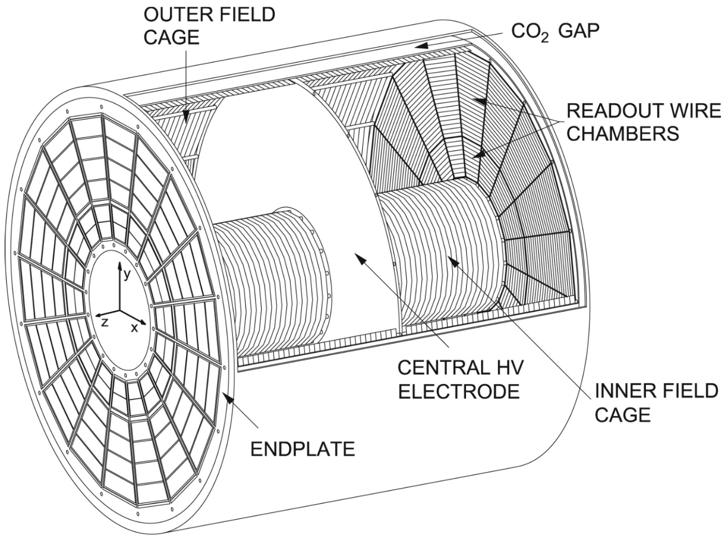


Figure 4.1: Diagram of the ALICE TPC, showing the two drift chambers, inner and outer field cages and readout chambers. Figure taken from Ref. [1].

focus on a gas electron multiplier option instead. Fig. 4.1 shows a schematic diagram of the ALICE TPC design. The basic ND-GAr geometry will resemble this, except for the inner field cage.

It will use a 90-10 molar fraction argon-CH_4 mixture at 10 bar. With this baseline gas mixture light collection is not possible, as the quenching gas absorbs most of the VUV photons. Additional R&D efforts are underway, to understand if different mixtures allow for the light signal to be used to provide a t_0 while maintainig stable charge gain.

4.2.2 ECal

The main role of the ND-GAr ECal is the calorimetric measurement of the electron energies and the reconstruction of photons, in particular those from neutral pion decays. Also, the ECal is able to provide a t_0 timestamp for neutrino interactions, by associating its activity to the tracks in the HPgTPC. The ECal will also be able to perform neutron reconstruction using time of flight and reject external backgrounds, thanks to its sub-nanosecond time resolution.

The ECal design features three independent subdetectors, two end caps at each side and a barrel surrounding the HPgTPC. Each of the detectors is divided in modules,

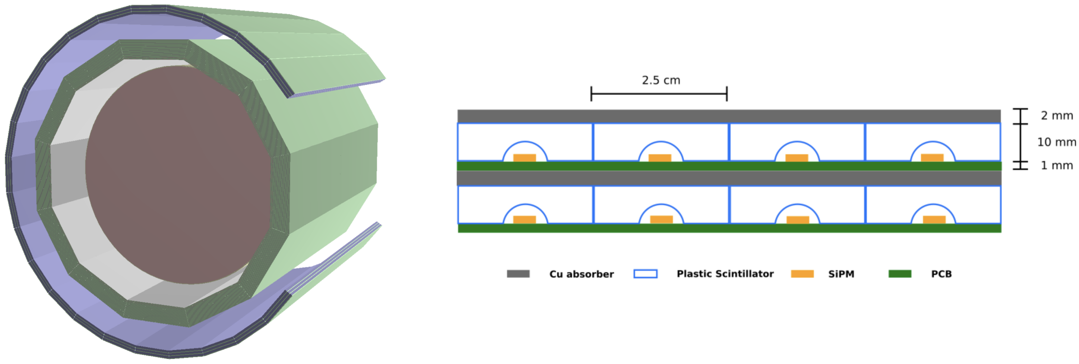


Figure 4.2: View of the 12-sided ECal barrel and outer muon tagger geometries (left) and layout of the ECal tile layers for the 2 mm Cu, 10 mm scintillator option (right). Figure adapted from Ref. [1].

which combine alternating layers of plastic scintillator and absorber material readout by SiPMs. The inner scintillator layers consist of 2.5×2.5 cm^2 high-granularity tiles, whereas the outer ones are made out of 4 cm wide cross-strips spanning the whole module length. The current barrel geometry consists of 8 tile layers and 34 strip layers, while the end caps feature 6 and 36 respectively. The thickness of the scintillator layers is 7 mm and 5 mm for the Pb absorber layers. The 12-sided geometry of the ECal barrel (left) and the layout of the tile layers (left)[1] can be seen in Fig. 4.2.

4.2.3 Magnet

The ND-GAr magnet design, know as the Solenoid with Partial Yoke (SPY), consists of two coupled solenoids with an iron return yoke. The idea behind the design is to have a solenoid as thin as possible, as well as a return yoke mass distribution that minimises the material budget between ND-LAr and ND-GAr. The magnet needs to provide a 0.5 T field in the direction perpendicular to the beam, parallel to the drift electric field. It needs to host the pressure vessel and the surrounding ECal, which points to a inner diameter of ~ 6.4 m.

The solenoid is a single layer coil, based on niobium titanium superconducting

[1] The figure shows the layout of the tile layers for a previous design with 2 mm Cu absorber and 10 mm plastic scintillator, as mentioned in the text the current choice is 5 mm Pb absorber and 7 mm scintillator.

Rutherford cable. The total length of the coil is 7.5 m. The bobbin will be split in four segments grouped in pairs with two identical cryostats, connected in series. The iron yoke features an aperture in the upstream side to allow the muons coming from ND-LAr. Still, its material will be enough to reduce the magnetic field reaching SAND, and also stop the charged pions produced inside the HPgTPC.

4.2.4 Muon system

The design of the ND-GAr muon system is still in a preliminary stage. Its role is to distinguish between muons and pions punching through the ECal. This is especially important for wrong-sign determination, to separate these from neutral current events.

In its current form, the muon system consists of three layers of longitudinal sampling structures. It alternates 10 cm Fe absorber slabs with 2 cm plastic scintillator strips. The transverse granularity required is still under study.

4.3 GArSoft

GArSoft is a software package developed for the simulation and reconstruction of events in ND-GAr. It is inspired by the LArSoft toolkit used for the simulation of LArTPC experiments, like the DUNE FD modules. It is based on `art`, the framework for event processing in particle physics experiments [?]. Other of its main dependencies are `ROOT`, `NuTools`, `GENIE` and `Geant4`. It allows the user to run all the steps of a generation-simulation-reconstruction workflow using FHiCL configuration files.

4.3.1 Event generation

The standard generator FHiCLs in GArSoft run the event generation and particle propagation simulation (i.e. Geant4) in the same job by default. However, it is possible to split them up if needed. The current version of GArSoft provides five different event generators, each of them producing `simb::MCTruth` products defined in `NuTools`. The available modules are:

- `SingleGen`: particle gun generator. It produces the specified particles with a given distribution of momenta, initial positions and angles.
- `TextGen`: text file generator. The input file must follow the `hepevt` format[2], the module simply copies this to `simb::MCTruth` data products.
- `GENIEGen`: GENIE neutrino event generator. The module runs the neutrino-nucleus interaction generator using the options specified in the driver FHiCL file (flux file, flavour composition, number of interactions per event, t_0 distribution, ...). Current default version is `v3_04_00`.
- `RadioGen`: radiological generator. It produces a set list of particles to model radiological decays. Not tested.
- `CRYGen`: cosmic ray generator. The module runs the CRY event generator with a configuration specified in the FHiCL file (latitude and altitude of detector, energy threshold, ...). Not tested.

The module `GArG4` searches for all the generated `simb::MCTruth` data products, using them as inputs to the Geant4 simulation with the specified detector geometry. A constant 0.5 T magnetic field along the drift coordinate is assumed. The main outputs of this step are `simb::MCParticle` objects for the generated Geant4 particles, `gar::EnergyDeposit` data products for the energy deposits in the HPgTPC and `gar::CaloDeposit` data products for the energy deposits in the ECal and muon system.

4.3.2 Detector simulation

The standard detector simulation step in GArSoft is all run with a single FHiCL, but the different modules can be run independently as well. First the `IonizationReadout`

[2] In brief, each event contains at least two lines. The first line contains two entries, the event number and the number of particles in the event. Each following line contains 15 entries to describe each particle. The entries are: status code, pdg code for the particle, entry of the first mother for this particle, entry of the second mother for this particle, entry of the first daughter for this particle, entry of the second daughter for this particle, x component of the particle momentum, y component of the particle momentum, z component of the particle momentum, energy of the particle, mass of the particle, x component of the particle initial position, y component of the particle initial position, z component of the particle initial position and time of the particle production.

module simulates the charge readout of the HPgTPC, and later the `SiPMReadout` module runs twice, once for the ECal and then for the muon system, with different configurations.

The `IonizationAndScintillation` module collects all the `gar::EnergyDeposit` data products, to compute the equivalent number of ionization electrons for each energy deposit. The `ElectronDriftAlg` module simulates the electron diffusion numerically both in the longitudinal and transverse directions and applies an electron lifetime correction factor. The induced charge on the nearest and neighbouring readout pads is modeled using the provided pad response functions. The digitisation of the data is then simulated with the `TPCReadoutSimAlg` module. By default, the ADC sampling rate used is 50.505 MHz. The resulting raw waveforms for each channel are stored with zero-suppression, in order to save memory and CPU time. The algorithms keep blocks of ADC values above a certain threshold, plus some adjustable additional early and late tick counts. The results of these three steps are `gar::raw::RawDigit` data products.

For the ECal and the muon system the `SiPMReadout` module calls either the `ECALReadoutSimStandardAlg` or `MuIDReadoutSimStandardAlg` modules. These take all the `gar::CaloDeposit` data products in the corresponding detector and do the digitisation depending on whether the hit was in a tile or strip layer. They include single photon statistics, electronic noise, SiPM saturation and time smearing. The resulting objects are `gar::raw::CaloRawDigit` data products.

4.3.3 Reconstruction

The reconstruction in GArSoft is also run as a single job by default. It first runs the hit finding, clustering, track fitting and vertex identification in the HPgTPC, followed by the hit finding and clustering in the ECal and muon system. After those it produces the associations between the associations between the tracks and the ECal clusters.

Focusing first on the HPgTPC reconstruction, the `CompressedHitFinder` module takes the zero-suppressed ADCs from the `gar::raw::RawDigit` data products. The reconstructed hits largely correspond to the above threshold blocks, however the hit finder identifies waveforms with more than one maximum, diving them in multiple hits if they dip below a certain threshold. The data products produced are of the form `gar::rec::Hit`. These are the inputs to the clustering of hits in the `TPCHitCluster` module. Hits close in space and time are merged, and the resulting centroids are found. This module outputs `gar::rec::TPCClusters` objects and associations to the input hits.

The following step prior to the track fitting is pattern recognition. The module called `tpcvechitfinder2` uses the `gar::rec::TPCClusters` data products to find track segments, typically called vector hits. They are identified by performing linear 2D fits to the positions of the clusters in a 10 cm radius, one fit for each coordinate pair. A 3D fit defines the line segment of the vector hit, using as independent variable the one whose sum of (absolute value) slopes in the 2D fits is the smallest. The clusters are merged to a given vector hit if they are less than 2 cm away from the line segment. The outputs are `gar::rec::VecHit` data products, as well as associations to the clusters. The `tpcpatrec2` module takes the `gar::rec::VecHit` objects to form the track candidates. The vector hits are merged together if their direction matches, their centers are within 60 cm and their direction vectors point roughly to their respective centers. Once the clusters of vector hits are formed they are used to make a first estimation of the track parameters, simply taking three clusters along the track. The module produces `gar::rec::Track` data products and associations between these tracks and the clusters and vector hits.

The track is fitted by means of a Kalman filter in the `tpctrackfit2` module, using the position along the drift direction as the independent variable. Two different fits are performed per track, a forward and a backwards fit, each starting from one of the track ends. The Kalman filter state vector $(y, z, R, \phi, \tan\lambda)$ is estimated at each point along the track using a Bayesian update. The track parameters reported in the forward and backwards fits are the ones computed at the opposite end where the fit started. The main outputs of the track fit are the `gar::rec::Track` objects. Additionally, the module stores the fitted 3D positions along the track in the `gar::rec::TrackTrajectory` data products and the total charge and step sizes for each point also get stored in the form of `gar::rec::TrackIonization` objects.

After the tracking step, the `vertexfinder1` module looks at the reconstructed `gar::rec::Track` products, creating vertex candidates with the track ends that are within 12 cm of each other. The vertices are then fitted using linear extrapolations from the different track ends associated. The results are `gar::rec::Vertex` data products, and associations to the tracks and corresponding track ends.

For the ECal and muon tagger, the `SiPMHitFinder` module runs twice with different configurations, adapted to the particular capabilities of both. The module simply takes the `gar::raw::CaloRawDigit` products, applies a calibration factor to convert the ADC counts to MeV and for the strip layer hits it calculates the position along the strip using the times recorded of both SiPMs. This module produces `gar::rec::CaloHit` data products. Next, these objects are used as inputs to the `CaloClustering` module. It merges the hits based on a simple nearest neighbours (NN) algorithm. For the resulting clusters it also computes the total energy and position of the centroid. The results are stored as `gar::rec::Cluster` data products, with associations to the hits.

The last step in the reconstruction is associating the reconstructed tracks in the HPgTPC to the clusters formed in the ECal and muon system. The `TPCECALAssociation` module checks first the position of the track end points, considering only the points that are at least 215 cm away from the cathode or have a radial distance to the center greater than 230 cm. The candidates are propagated up to the radial position, in the case of clusters in the barrel, or the drift coordinate position, for the end cap cluster, of the different clusters in the collection using the track parameters computed at the end point. The end point is associated to the cluster if certain proximity criteria are met. This module creates associations between the tracks, the end points and the clusters. The criteria for the associations are slightly different for the ECal and the muon tagger.

Chapter 5

FWTPG offline software

Chapter 6

Matched Filter approach to induction wire Trigger Primitives

6.1 Motivation

The filter implemented in the firmware of the upstream DUNE FD DAQ is a $32nd$-order low-pass finite impulse-response (FIR) filter. The output of such filter for a discrete system can be written as:

$$y[i] = \sum_{j=0}^{N} h[i]x[i-j], \tag{6.1}$$

where N is the order of the filter, y is the output sequence, x is the input sequence and h is the set of coefficients of the filter. The current implementation within `dtp-firmware` [50] uses a set of 16 non-zero integer coefficients.

Filtering is a vital step in the hit finder chain. It helps to suppress the noise and enhance the signal peaks with respect to the noiseless baseline. A good filtering strategy allows us to use lower thresholds when forming the trigger primitives (TPs) and thus increasing the sensitivity of our detector to low energy physics events. In such events, the hits produced by the ionisation electrons tend to have lower amplitudes than those of interest to the baseline physics programme of the DUNE experiment.

This is particularly important for the induction planes. In general, signal peaks in

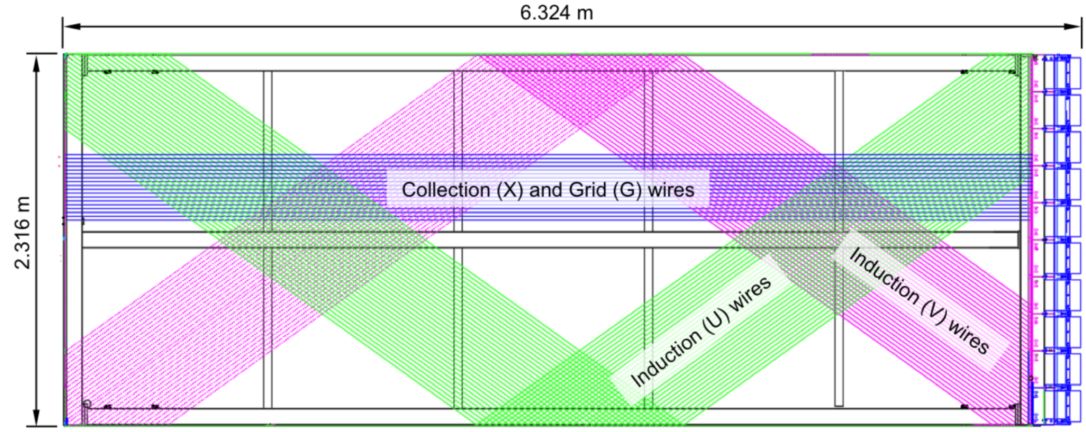
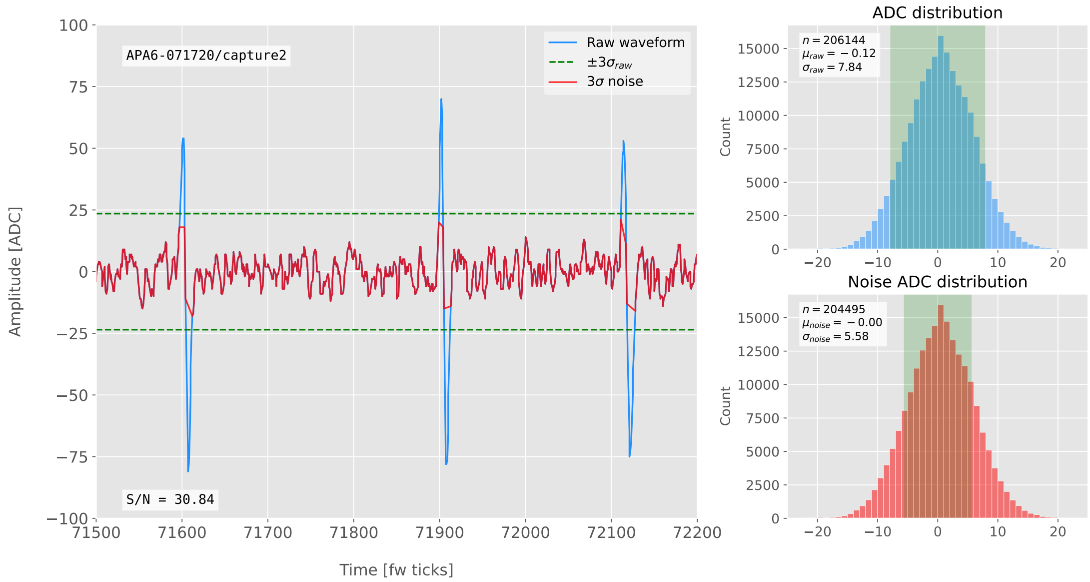


Figure 6.1: *Schematic representation of an APA. The black lines represent the APA steel frame. The green and magenta lines correspond to the direction of the U and V induction wires respectively. The blue lines indicate the direction of the X collection wires and the wire shielding G.* REP?

the induction wires have smaller amplitude than the ones in the induction plane (2.). This, together with the fact that the pulse shapes are bipolar, reduces our capacity to detect the hits on these channels. The inefficiency of detecting TPs in the induction planes (denoted as U and V planes) leads trigger algorithms to focus mainly on the TPs from the collection plane (so-called X plane). As a result, the possibility of making trigger decisions based on the coincidence of TPs across the three wire planes remains nowadays unexploited in DUNE. Fig. 6.1 shows a schematic view of an anode plane assembly (APA), with the different wire plane orientations highlighted.

A possible improvement of the current hit finder chain could require optimising the existing or choosing a new filter implementation. A filter strategy which improves the induction signals may ~~be able to~~ enhance the detection efficiency of TPs from the induction planes and ideally make it comparable to that of the collection plane.

The goal is to implement a better finite-impulse response filter design and to evaluate its performance relative to the current filter. To do so, we need to take into account the limitations of the firmware: the FIR filter shall have maximum 32 coefficients (so-called taps) whose values are 12-bit unsigned integers. Although it is technically possible to include non-integer coefficients, it would be a technical challenge as we have 40 FIR



THESE LOOK LIKE THE SAME DISTRIBUTION? I WOULD HAVE EXPECTED THE NOISE ONE TO BE WITHIN THE 3σ BAND

Figure 6.2: *Left panel: Zoomed unfiltered waveform corresponding to channel* 7840 *from the ProtoDUNE-SP raw data capture* `felix-2020-07-17-21:31:44` *(blue line). The green dashed lines mark the region* $\pm 3\sigma_{raw}$*. The resulting noise waveform is also shown (red line). Top right panel: ADC distribution for channel* 7840*, where the green shaded region represents* $\pm\sigma_{raw}$*. Bottom right panel: noise ADC distribution for channel* 7840*, where the green shaded region represents* $\pm\sigma_{noise}$*.*

instances per APA, as there are 4 FIR per optical link and 10 optical links per APA. With these restrictions, the task is to provide a set of 32 coefficients which yield an optimal filter performance for the induction wires.

6.2 Signal-to-noise ratio definition

I introduce the signal to noise ratio (S/N) as a measure of the FIR filter performance and demonstrate how to extract its value for a set of ProtoDUNE-SP data. The S/N metrics allow us to compare different filter implementations and serve as a basis for more detailed studies presented later in this document. Specifically, I use the ADC capture `felix-2020-07-17-21:31:44` (data capture taken for firmware validation purposes). I defined S/N as the height of the signal peaks relative to the size of the noise peaks. To quantify this quantity channel by channel one first need to estimate the standard deviation of the ADC data for each channel, σ_{ADC}. Then, I define the corresponding noise waveform to be the ADC values in the range $\pm 3\sigma_{ADC}$. From this new noise data one can estimate ~~again~~ the mean and standard deviation, μ_{noise} and σ_{noise}, so I can write the S/N for any given channel as:

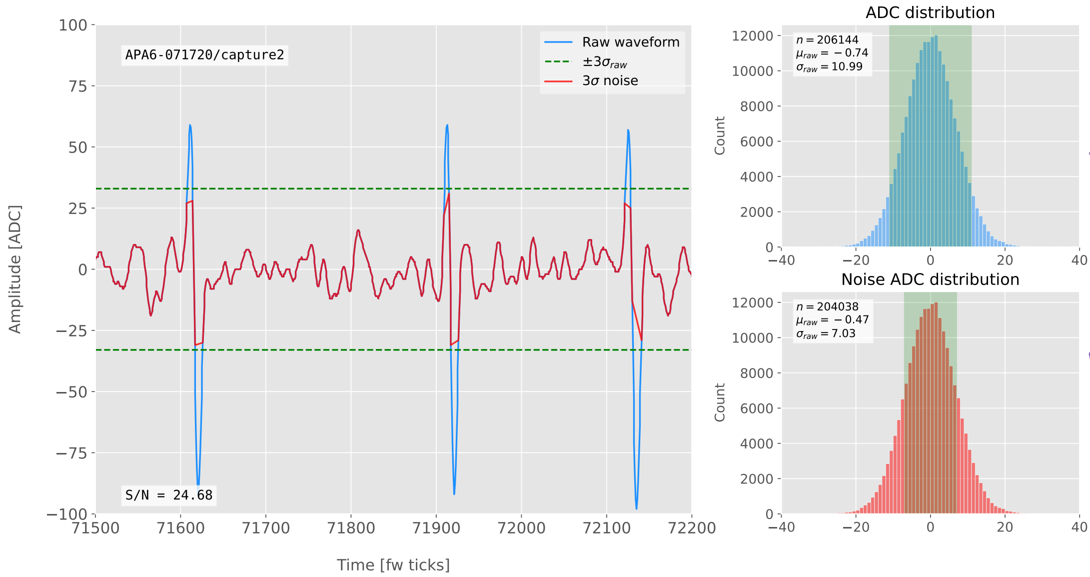
$$\mathrm{S/N} = \frac{\max\left[ADC\right] - \mu_{noise}}{\sigma_{noise}}, \tag{6.2}$$

where $\max\left[ADC\right]$ is simply the maximum ADC value found in the corresponding channel.

One can apply this definition of the S/N with a waveform from one of the channels of the data capture[1]. Fig. 6.2 shows a zoomed region of the waveform corresponding to channel 7840 (blue line), where one can clearly see three signal peaks and continuous additive noise (we actually see 6 peaks, 3 positive and 3 negative, but, because by design for induction channels the expected signal pulse shapes are bipolar, I treat them as a collection of 3 individual signal peaks). I estimated the standard deviation of this raw waveform to be $\sigma_{raw} = 7.84$ ADC, so I am able to define the noise waveform (red line) as the ADC values in the range ± 23.52 ADC. This way one obtains $\mu_{noise} = 0$ and $\sigma_{noise} = 5.58$ ADC, which gives S/N $= 30.84$.

We can repeat this calculation now for the corresponding filtered waveform (using the current firmware FIR filter). In Fig. 6.3 I plotted the same time window for the filtered waveform from channel 7840 (blue line). In this case, the standard deviation of the waveform is larger than before, giving $\sigma_{raw} = 10.99$ ADC. The resulting noise waveform (red line) results from selection the ADC values in the range ± 32.91 ADC, giving now $\mu_{noise} = -0.47$ ADC and $\sigma_{noise} = 7.03$ ADC. Finally, one obtains S/N $= 24.68$. Notice that the value of S/N decreases after the filtering. Clearly, one can see that the noise baseline has increased by a factor of 1.35 when we applied the FIR filter and at the same time the amplitude of the signal peaks has remained almost unchanged, leading to this poorer S/N value.

[1] All the original work was done within the `dtp-simulation` package [51], which offers a variety of tools to read raw data and emulate the TPG block (pedestal subtraction, filtering and hit finder). However, the results shown in this report were re-worked later using the C++ based `dtpemulator` package [52]. Its main purpose is the emulation of the TPG block and, in the same way as its predecessor, it has been cross-checked against the current firmware implementation.



SAME Q HERE

Figure 6.3: *Left panel: Zoomed filtered waveform corresponding to channel* 7840 *from the ProtoDUNE-SP raw data capture* `felix-2020-07-17-21:31:44` *(blue line). The filter used was the current implementation of the low-pass FIR filter in* `dtp-firmware`*. The green dashed lines mark the region* $\pm 3\sigma_{raw}$*. The resulting noise waveform is also shown (red line). Top right panel: ADC distribution for channel* 7840 *after filtering, where the green shaded region represents* $\pm\sigma_{raw}$*. Bottom right panel: noise ADC distribution for channel* 7840 *after filtering, where the green shaded region represents* $\pm\sigma_{noise}$

6.3 Low-pass FIR filter design

American spelling?

~~In general, when one uses a method~~ to optimize the frequency response of a digital filter, we can use such as the Parks-McClellan algorithm, where one finds a set of N real coefficients that give the best response for the specified pass-band and order of the filter [53].

In our case, as the sampling frequency is defined as 1 ticks^{-1}, the Nyquist frequency will simply be 1/2 ticks^{-1}. The current implementation of the filter seems to have as pass-band the range $[0, 0.1]$ ticks^{-1}. This can be seen in Fig. 6.4, where I show the power spectrum, in decibels, of such filter implementation (blue solid line). For instance, the Park-McClellan algorithm finds the optimal Chebyshev FIR filter taking as input the boundaries of the target pass-band and stop-band, which can be written in the form:

$$\begin{cases} [0, f_c] \\ [f_c + \delta f, f_N] \end{cases}, \qquad (6.3)$$

REP ?

where f_c is the cut-off frequency, δf is the transition width and f_N is the aforementioned Nyquist frequency. A similar behaviour to the one in the current filter can be obtained by setting $f_c = 0$ and $\delta f = 0.1$ ticks^{-1}. The response of the resulting filter is also shown in Fig. 6.4 (blue solid line). Notice that the suppression of the stop-band is enhanced for this optimal filter. For comparison I included the power response of the filter obtained by taking the integer part of the coefficients resulting from the Parks-McClellan method (red dashed line). One can see that it does not suppress that much the stop-band, in a similar way to the current implementation of the filter.

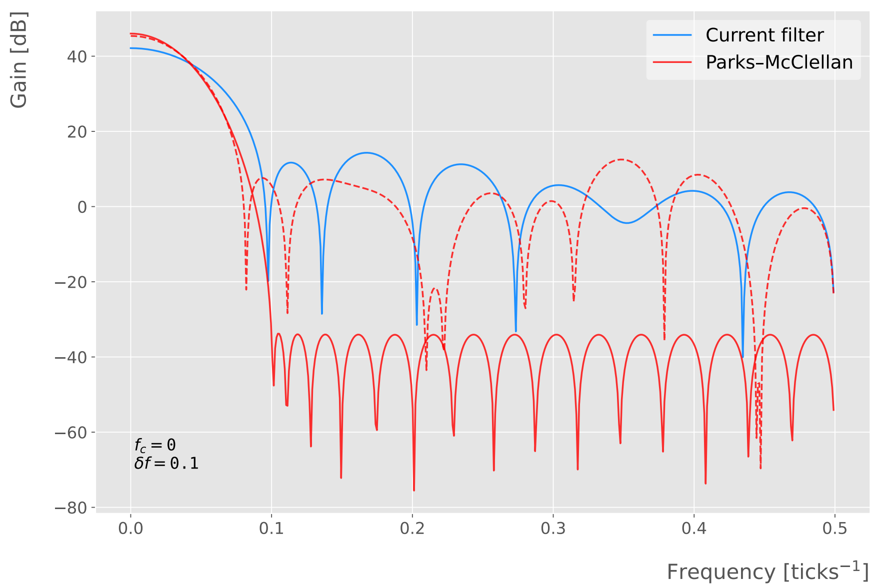


Figure 6.4: *Power spectrum in decibels for the current implementation of the low-pass FIR filter in* `dtp-firmware` *(blue line), compared to the response of an optimal filter obtained using the Parks-McClellan algorithm for the same pass-band (red line). Also for comparison I include the spectrum of the optimal filter when taking only the integer part of the coefficients (red dashed line).*

At this point, I tried to improve the performance of the FIR filter using the Park-McClellan method, i.e. maximize the overall S/N, using the available data captures. I did so by varying the values of the two quantities that parametrize the pass-band and stop-band, the cut-off frequency f_c and the transition width δf.

Fig. 6.5 shows the average relative change in the S/N (i.e. the ratio between the value of the S/N after and before the filtering) for capture `felix-2020-07-17-21:31:44`,

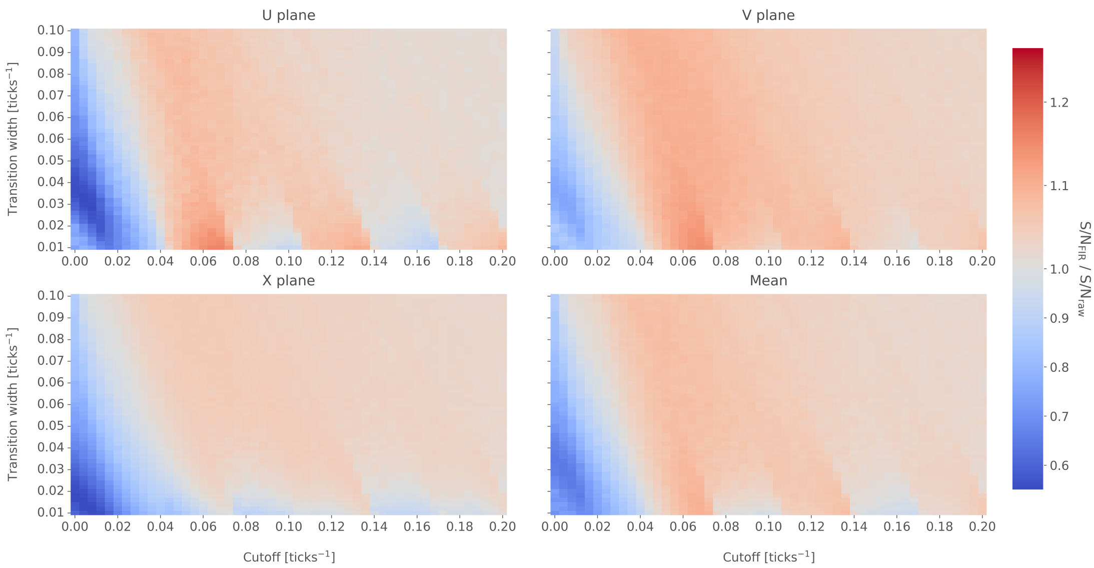
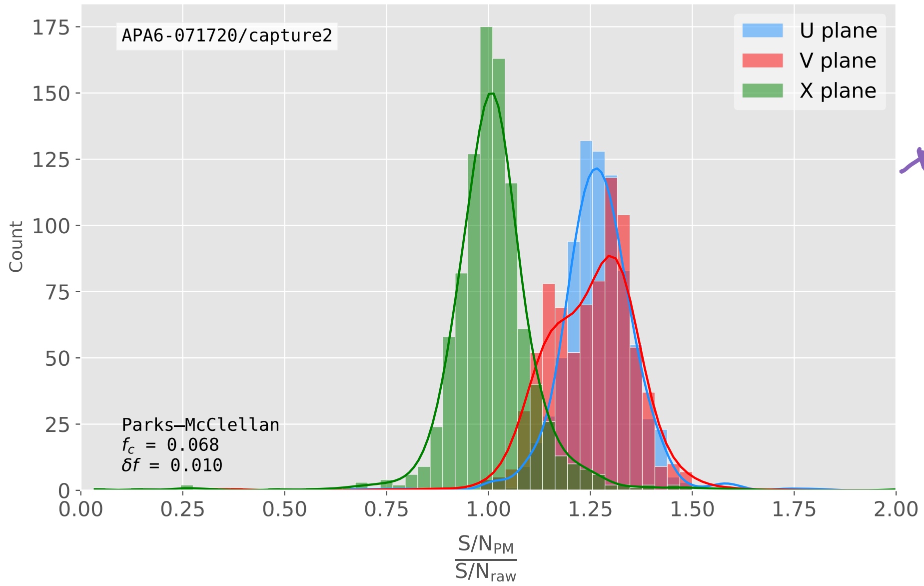


Figure 6.5: *Relative change in the S/N for the ProtoDUNE-SP raw data capture* `felix-2020-07-17-21:31:44`*, using different values of the cutoff frequency f_c and the transition width δf. The optimal Chebyshev filters were applied using just the integer part of the coefficients given by the Parks-McClellan algorithm.*

when using filters designed with the Parks-McClellan algorithm for the specified values of the cut-off frequency f_c and the transition width δf, restricted to integer values for the filter coefficients. One can clearly distinguish different regions where we get an improvement of up to a factor of 1.35 for the U plane. For large values of $f_c + \delta f$ the ratio tends to 1, as expected (in that limit the width of the stop-band goes to 0, meaning that no frequencies are filtered out and thus the waveform remains the same).

Using the configuration which gives the best mean performance for the three planes (see bottom right panel of Fig. 6.5), i.e. $f_c = 0.068$ ticks^{-1} and $\delta f = 0.010$ ticks^{-1}, we can see how such filter affects the different channels. Fig. 6.6 shows the distribution of the S/N improvement values for all the channels in the raw ADC capture `felix-2020-07-17-21:31:44`, separated by wire plane, after the optimal Chebyshev filter was applied. One can see that there is a clear improvement for both U and V induction wire planes, obtaining a mean change of 1.25 and 1.30 for them respectively. However, in the case of the collection plane X the mean of this distribution is roughly 1, meaning that a good fraction of channels in that plane get a slightly worse S/N after the



→ VIVA Q: DOUBLE PEAK IN V PLANE

Figure 6.6: *Distribution of the relative change of the S/N on the different wire planes from the ProtoDUNE-SP raw data capture* `felix-2020-07-17-21:31:44` *after the optimal Chebyshev filter was applied. The filter was computed with the Parks-McClellan algorithm using a cutoff of* $f_c = 0.068$ ticks^{-1} *and a transition width* $\delta f = 0.010$ ticks^{-1}.

I am not sure I agree w/ this statement. Let's talk about it at our next meeting

filter is applied. In any case, this is not a big issue as the S/N for collection channels is usually much higher than the one for induction channels.

The results I obtained optimising the low pass filter with the Parks-McClellan method are promising. Nonetheless, the improvement found is rather marginal so I wondered if there could be an alternative approach to the filtering problem which yields better outputs. At this point, I found a possible alternative in matched filters. By construction, this kind of filters offer the best improvement on the S/N.

6.4 Matched filters

In the context of signal processing, a matched filter is the optimal linear filter for maximising the signal-to-noise ratio (S/N) in the presence of additive noise, obtained by convolving a conjugated time-reversed known template with an unknown signal to detect the presence of the template in the signal [54].

Given a known signal sequence $s(t)$ and another (a priori unknown) noise sequence

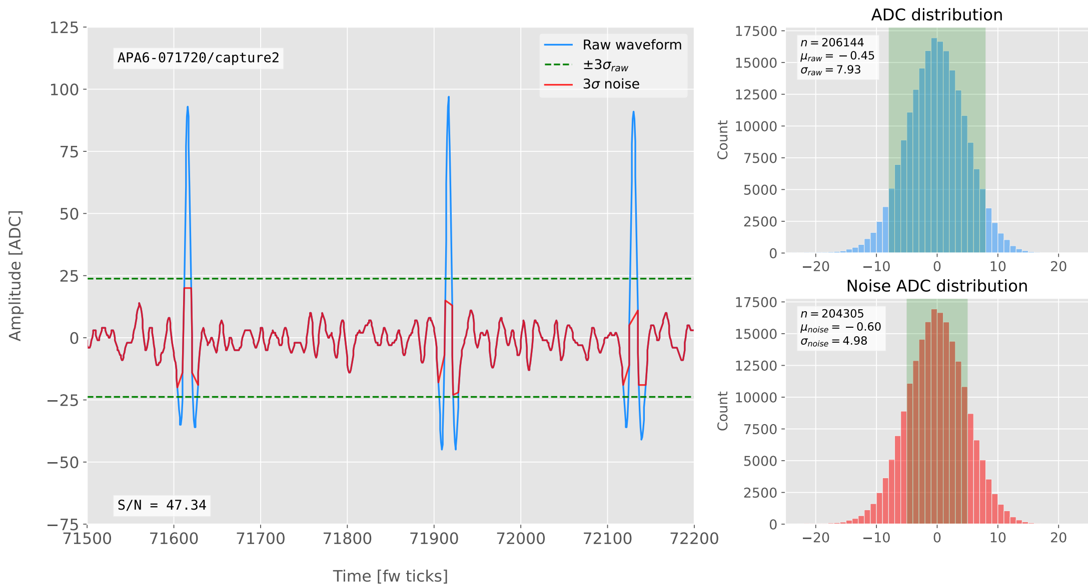


Figure 6.7: *Left panel: Zoomed match filtered waveform corresponding to channel* 7840 *from the ProtoDUNE-SP raw data capture* `felix-2020-07-17-21:31:44` *(blue line). The filter used was directly extracted from the data, being the* 32 *values around the first peak in the original waveform. The green dashed lines mark the region* $\pm 3\sigma_{raw}$. *The resulting noise waveform is also shown (red line). Top right panel: ADC distribution for channel* 7840 *after match filtering, where the green shaded region represents* $\pm\sigma_{raw}$. *Bottom right panel: noise ADC distribution for channel* 7840 *after match filtering, where the green shaded region represents* $\pm\sigma_{noise}$

$n(t)$, the input signal can be written as:

$$x(t) = s(t) + n(t). \tag{6.4}$$

Now, considering a linear time-invariant filter, whose impulse-response function I will refer to as $h(t)$, one can write the output signal as:

$$\begin{aligned} y(t) &= x(t) * h(t) \\ &= (s(t) + n(t)) * h(t) \\ &= y_s(t) + y_n(t), \end{aligned} \tag{6.5}$$

where $y_s(t)$ and $y_n(t)$ are simply the outputs of the filter due to the signal and the noise components respectively.

The goal of the matched filter is to detect the presence of the signal $s(t)$ in the input sample $x(t)$ at a certain time t_0, which effectively means we need to maximise the S/N. This way, what one wants is to have a filter which gives a much bigger output when the known signal is present than when it is not. Putting it in other words, the instantaneous power of the signal output $y_s(t)$ should be much larger than the average power of the noise output $y_n(t)$ at some time t_0.

For the case of the filtered signal, one can easily re-write it as an inverse Fourier transform:

$$y_s(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\omega\ H(\omega)S(\omega)\mathrm{e}^{i\omega t}, \tag{6.6}$$

where $H(\omega)$ and $S(\omega)$ are the Fourier transforms of the impulse-response function (i.e. the transfer function of the filter) and of the input signal, respectively.

Now focusing on the noise, we can use the Wiener-Khinchin theorem [55] to write the mean power of the noise after filtering as:

$$E|y_n(t)|^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\omega\ |H(\omega)|^2 S_n(\omega), \tag{6.7}$$

where $S_n(\omega)$ is the power spectral density of the noise.

Having these, one can write the instantaneous S/N at time t_0 as:

$$\begin{aligned}\left(\frac{S}{N}\right)_{t_0} &= \frac{|y_s|^2}{E|y_n(t)|^2} \\ &= \frac{1}{2\pi}\frac{\left|\int_{-\infty}^{\infty} \mathrm{d}\omega\ H(\omega)S(\omega)\mathrm{e}^{i\omega t_0}\right|^2}{\int_{-\infty}^{\infty} \mathrm{d}\omega\, |H(\omega)|^2\, S_n(\omega)}.\end{aligned} \tag{6.8}$$

Once we have this expression, we need to find the upper limit of it to determine what would be the optimal choice for the transfer function. One can use the Cauchy-Schwarz inequality, which in the present case takes the form:

$$\left|\int_{-\infty}^{\infty} \mathrm{d}x\ f(x)g(x)\right|^2 \leq \int_{-\infty}^{\infty} \mathrm{d}x\ \ |f(x)|^2 + \int_{-\infty}^{\infty} \mathrm{d}x\ \ |g(x)|^2\,, \tag{6.9}$$

for any two analytical functions $f(x)$ and $g(x)$. One can prove that making the choice:

$$\begin{aligned} f(x) &= H(\omega)\sqrt{S_n(\omega)}\mathrm{e}^{i\omega t_0}, \\ g(x) &= \frac{S(\omega)}{\sqrt{S_n(\omega)}}, \end{aligned} \tag{6.10}$$

leads to the following upper bound for the S/N:

$$\left(\frac{S}{N}\right)_{t_0} \leq \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{d}\omega \, \frac{|S(\omega)|^2}{S_n(\omega)}. \tag{6.11}$$

From Eqs. (6.8), (6.9) and (6.10) one can also derive the form of the transfer function such that the upper bound is exactly reached [56]:

$$H(\omega) \propto \frac{S^*(\omega)\mathrm{e}^{-i\omega t_0}}{S_n(\omega)}. \tag{6.12}$$

From this last expression we can clearly see the way the matched filter acts. As the transfer function is proportional to the Fourier transform of the signal it will try to only pick the frequencies present in the signal [57].

The matched filter transfer function can be greatly simplified if the input noise is Gaussian. In that case, the power spectral density of the noise is a constant, so it can be re-absorbed in the overall normalisation of the transfer function. Moreover, considering that the input signal is a real function, one can simply set $S^*(\omega) = S(-\omega)$, which gives:

$$H(\omega) \propto S(-\omega)\mathrm{e}^{-i\omega t_0}. \tag{6.13}$$

For a discrete signal, one can think of the input and impulse-response sequences as vectors of $\mathbb{R}^N$. Then, the matched filter tries to maximise the inner product of the signal and the filter while minimising the output due to the noise by choosing a filter vector orthogonal to the later. In the case of additive noise, that leads to the impulse-response vector:

$$h = \frac{1}{\sqrt{s^\dagger R_n^{-1} s}} \; R_n^{-1} s, \tag{6.14}$$

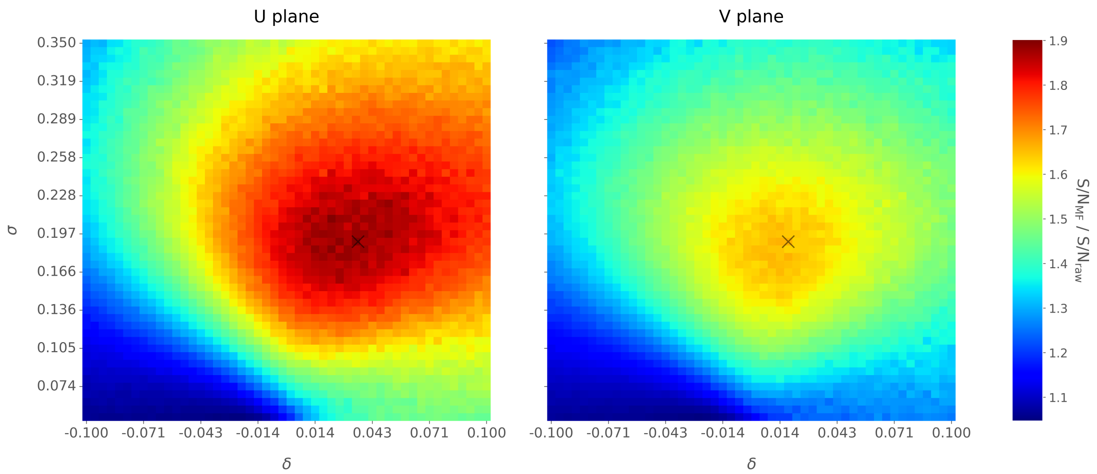


Figure 6.8: *Relative improvement in the S/N for the raw data capture* `felix-2020-07-17-21:31:44`*, using the matched filter following the parametrisation in Eq. (6.17). The black crosses in both panels denote the location of the maximum ratio value.*

where s is a reversed signal template sequence of length N equal to the order of the filter and R_n is the covariance matrix associated with the noise sequence n. For the Gaussian noise case, the covariance matrix is simply the unit matrix, so the above expression simplifies again to:

$$h = \frac{s}{|s|}. \tag{6.15}$$

isn't this the same as Eq. 6.2? just reference that

For this first stage of the study, I use a definition of the S/N per channel given by:

$$\text{S/N} = \frac{\max{[ADC]} - \mu_{noise}}{\sigma_{noise}}, \tag{6.16}$$

where the subscript *noise* refers to a subset of the data obtained by only taking into account waveform values within a $\pm 3\sigma$ range around the mean of the data and $\max{[ADC]}$ is the maximum of the original waveform. This definition is further discussed in App. 6.2, where I also show examples of its application to raw data and to a waveform filtered with the current low-pass FIR filter.

To test whether this choice of filter is appropriate one needs to choose a signal template. As an example of how a matched filter would affect our signal, I simply took the filter coefficients to be the 32 ADC values around a signal peak present in the data.

If you can, bring Fig 6.7 closer to here, so this is the first mention

In Fig. 6.7 (left panel) I plotted a zoomed region for channel 7840 in the raw data capture `felix-2020-07-17-21:31:44`, after applying the matched filter described before (blue line). When compared to the raw and FIR filtered case (see App. 6.2), after applying the match filter the standard deviation of the noise waveform (red line) decreases and at the same time the signal peaks are enhanced. This leads to an improvement of the S/N by a factor of 1.92 when compared to the raw waveform.

~~In order~~ T to obtain the matched filter that is more suitable for our data, I explored different configurations of signal templates. ~~In order to perform this exploration,~~ I parametrised the signal using the bipolar function:

$$f(x) = -A(x+\delta)\ \mathrm{e}^{-x^2/\sigma^2}, \tag{6.17}$$

where the parameter δ controls the asymmetry between the positive and negative peaks and σ controls their width. The amplitude parameter A is set such that it keeps the height of the biggest peak to be less than 200 ADC in absolute value.

As this parametrisation is only adequate for bipolar signals I will focus exclusively on the induction channels. Also, the optimal configurations I found for the U and V plane will be kept separate, i.e. I will have two sets of coefficients that will be applied to either the U and V planes of wires. I do so as I found this was the choice giving the best performance. Even so, as I will discuss, the differences are not very pronounced. In case it is not technically possible to separate channels in the firmware according to the wire plane they come from and use different sets of filter coefficients for them, we can just find a common unique set of coefficients. In such case, I do not expect our results to change dramatically.

In Fig. 6.8 I present the results of our parameter scan, for channels in the induction planes U (left panel) and V (right panel). For each configuration of σ and δ the resulting matched filter was applied to all channels in the corresponding plane within the data capture `felix-2020-07-17-21:31:44`, the S/N improvement was computed with respect to the raw waveforms and then the S/N mean value was kept as a score for such filter.

One can see that the improvement obtained for the U plane is in general higher than the one for the V plane. In any case, I got substantially higher ratios than the ones obtained for the low-pass FIR filters. For the optimal configurations I attained improvements up to a factor of 1.85 for the U plane and 1.65 for the V plane.

VIVA Q: WHY?

The sets of optimal matched filter coefficients were obtained for the parameters $\delta = 0.035$, $\sigma = 0.191$ for the U plane and $\delta = 0.018$, $\sigma = 0.191$ for the V plane. I show these two sets of coefficients in Fig. 6.9 (left panel). ~~Also~~ In Fig. 6.9 (right panel) I plot the distribution of the S/N improvement after the optimal match filters for the U and V were applied to the corresponding channels in the raw data capture `felix-2020-07-17-21:31:44`. As mentioned before, the mean improvement achieved for the U plane channels is slightly bigger than the one for the V channels. Note, however, that the spread of the distribution for the V plane is also smaller than the one for the U plane.

I also performed a similar scan for the case of a low-pass FIR filter using the Parks-McClellan algorithm. In that case, the parameters to check were the cutoff frequency and the transition width of the filter. A summary of the results is given in App. 6.3.

HOW DOES THIS DIFFER FROM THE PREVIOUS PASS?

Overall, one can see that the improvements on the S/N are much more significant in the case of the matched filter than it is for the low-pass FIR filters. The analysis of this and other raw data captures from ProtoDUNE-SP suggest that matched filters increase the S/ N of induction channels by a factor of 1.5 more than the optimal low-pass FIR filters.

Although these results are by themselves great points in favour of the matched filter, more studies are needed to completely assess the robustness of this approach. I proceeded then to test the matched filter with simulated data samples.

6.5 Using simulated samples

~~In order~~ to further test the matched filter, the next step was to generate and process data samples using *LArSoft* [58]. In this way, one can control the particle content of

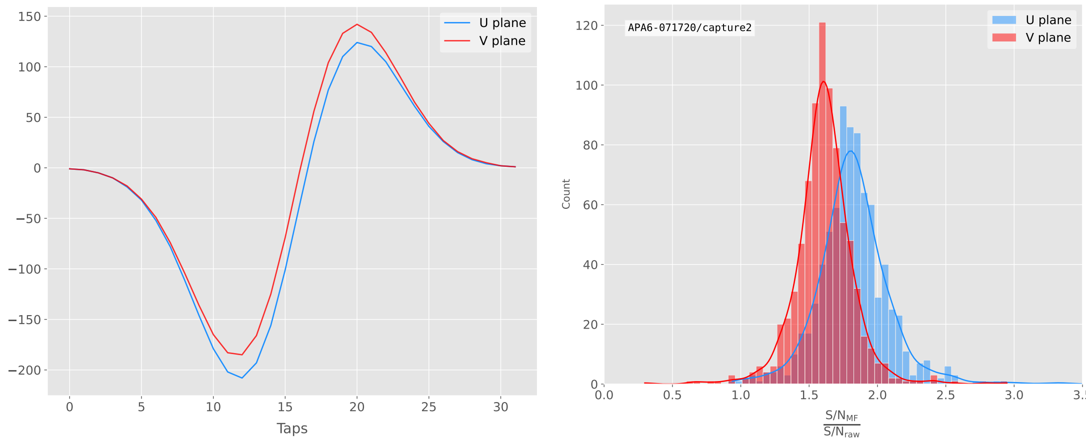


Figure 6.9: *Left panel: Optimal matched filter coefficients for the U (blue line) and V (red line) planes. The filters were computed with our parametrisation in Eq. (6.17) for the parameter values* $\delta = 0.035$*,* $\sigma = 0.191$ *and* $\delta = 0.018$*,* $\sigma = 0.191$ *respectively. Right panel: Distribution of the relative change of the S/N on the two induction wire planes from the ProtoDUNE-SP raw data capture* `felix-2020-07-17-21:31:44` *after their respective optimal matched filters were applied.*

the samples, the orientation of the tracks and their energy, and therefore see how the matched filter behaves in various situations.

To begin with, I prepared different monoenergetic and isotropic samples containing a single particle per event. Each sample contains a different particle species, namely electrons, muons, protons and neutral pions all with a kinetic energy of $E_k = 100$ MeV. I chose these because of the fairly different topologies they generate in the liquid argon, ranging from shower-like to track-like. The procedure I followed to generate the samples and process them is discussed in detail in App. ??.

VIVA Q: WHY 100 MeV?

These were generated with the single particle gun and the Geant4 stage of the *LArSoft* simulation [58] was performed with the standard configuration for the DUNE FD 10kt module.

For simplicity, I restricted the particles to start drifting in a single TPC volume (in this case TPC 0), so I can focus exclusively on the signals coming from one APA. The chosen kinetic energy for all the particles in my first trial is $E_k = 100$ MeV, so a necessary check is to see if all our tracks will be typically contained in one TPC volume. Fig. 6.10 (left panel) shows the distributions of the track lengths in the liquid argon

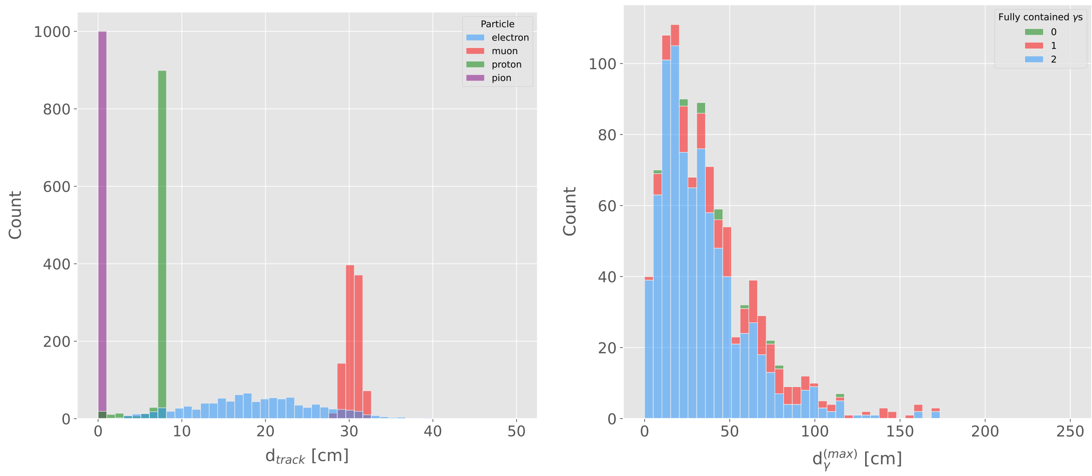


Figure 6.10: *Left panel: distributions of the particles track length in the liquid argon for the generated $E_k = 100$ MeV monoenergetic samples, electrons (blue), muons (red), protons (green) and neutral pions (purple). Right panel: distribution of the length of the longest photon in the neutral pion sample after the decay process $\pi^0 \rightarrow \gamma\gamma$.*

of all generated particles with $E_k = 100$ MeV. One can see that, in the case of the track-like particles (i.e. muons and protons), their length distributions are quite sharp and centered at relatively low distances (30 and 8 cm, respectively). For electrons, the distribution is quite broad but it does not extend past ~ 30 cm. The case of neutral pions can be misleading, as they decay promptly the track length associated with the true Monte Carlo particle is always < 1 cm. In Fig. 6.10 (right panel) I show the effective length distribution of the longest photon after the pion decays as $\pi^0 \rightarrow \gamma\gamma$, highlighting the number of fully contained photons in the TPC volume per event (either zero, one or both). One can see that the vast majority of events has both photons contained and that just a SMALL ~~negligible~~ number of them has none of them contained in the TPC volume. In any case, for the sake of caution, I will only keep the pion events with both photons contained.

The next step is to ~~Once I have prepared a sample at the Geant4 level, I need to~~ process the sample ~~it~~ trough the detector simulation. ~~In order~~ to make adequate estimations of the noise levels ~~and run the filtering and hit finder as I did with the ProtoDUNE data~~, one needs to turn off the default zero-suppression of the waveforms produced by the simulation. At this first stage I am only concerned with the waveforms with the noise added, so I keep the noise addition option as true in the configuration. However, for studies related to the hit finder performance one will also need to store the noiseless waveforms ~~in order~~ to retrieve the truth information of the hits. I will discuss this approach next.

REPETITIONS

After the detector simulation stage, one needs to extract the no zero-suppressed noisy waveforms, along with their offline channel numbers, and store them in a certain format to be analysed later. To reduce the amount of data that will go for processing, I used the information from the Geant4 step of the simulation to select only the active channels, i.e. the channels where some ionisation electrons arrive. Moreover, as said previously, I only extract the waveforms from APA 0 and exclusively the ones coming from induction channels. The resulting ROOT file contains a tree with two branches, one containing the waveforms for each event and channel and the other with the corresponding offline channel numbers.

you should decide how you want to write Geant & be consistent throughout the thesis

Finally, to extract the truth values for the orientation of the tracks and the energies of the particles I used a modified analysis module. This gives a ROOT file with a single tree, containing several branches with different information such as the components of the initial momentum of the particles, initial and final xyz location, track length, etc.

written by you?

For the analysis of the resulting waveforms and truth values I used a custom set of Python libraries (available at [???]). Among other functionalities, these enable the user to read the ROOT files, export the raw data as pandas objects, apply the filters and compute the S/N of both the raw and filtered signals. So far, the default configuration for the filtering uses the set of optimal matched filter coefficients that I found using the ProtoDUNE data samples.

Additionally, for the analysis of the samples it was necessary to use two different reference frames, to study separately the signals coming from the U and V induction wire planes. As I am focussing on a single APA, the U and V wires have a different orientation in the yz plane. In the case of U wires, these are tilted 35.7° clockwise from the vertical (y direction), whereas the V wires are at the same angle but in the counter clockwise direction. Because of this, the best option is to deal with two new coordinate systems rotated by ±35.7° along the x axis, so the new y' and y'' directions are aligned with the

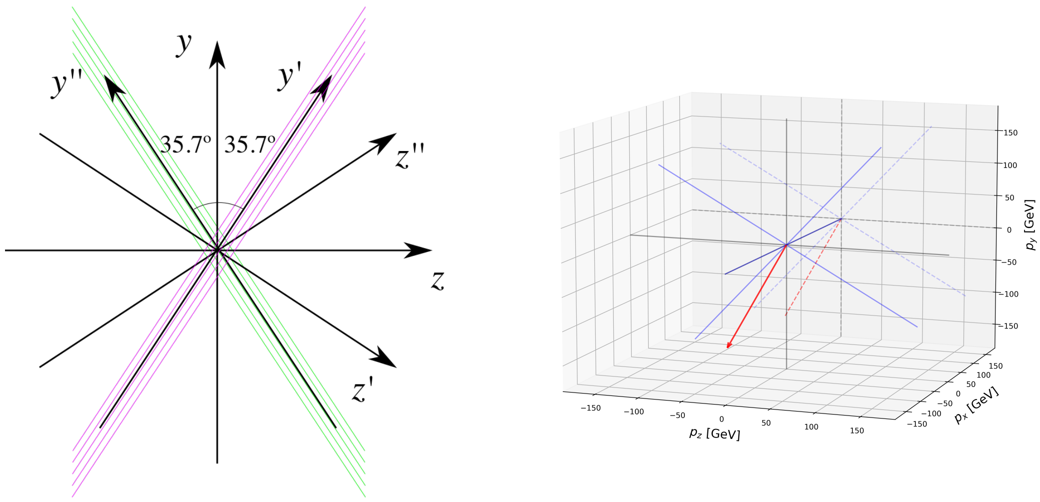


Figure 6.11: *Left panel: schematic representation of the two new rotated reference frames used in this analysis (denoted as prime and double prime), viewed from the yz plane. The magenta stack of lines represent the wires in the U plane, whereas the green lines correspond to the wires in the V plane. Right panel: 3D representation of the momentum of one of the generated monoenergetic muons (red arrow) in the original reference frame (black lines), along with the new reference frame used for the U plane waveforms (blue lines). In the yz plane I added the projection of these three.*

U and V induction wires. Fig. 6.11 (left panel) shows a schematic representation of the original reference frame together with the two rotated ones (denoted by primed and double primed). This way, one can easily understand how parallel was a track to the wires in the two induction planes. Fig. 6.11 (right panel) shows a 3D representation of the momentum of a track (red arrow) in the original reference frame (black lines), along with the new reference frame for U wires (blue lines). I added the projection in the yz plane of this three to show the usefulness of the new reference frame to tell whether a track is parallel or normal to the wires in the induction plane.

Fig. 6.12 shows the distribution of the average S/N improvement per event when one applies the optimal matched filters. I produced separate distributions for the channels in the U (red) and V (blue) induction wire planes. Notice that the S/N distributions for the track-like particles, i.e. muons (top left panel) and protons (bottom left panel), have significantly larger mean values than the distributions of the shower like particles, i.e. electrons (top right panel) and neutral pions (bottom right panel). An important

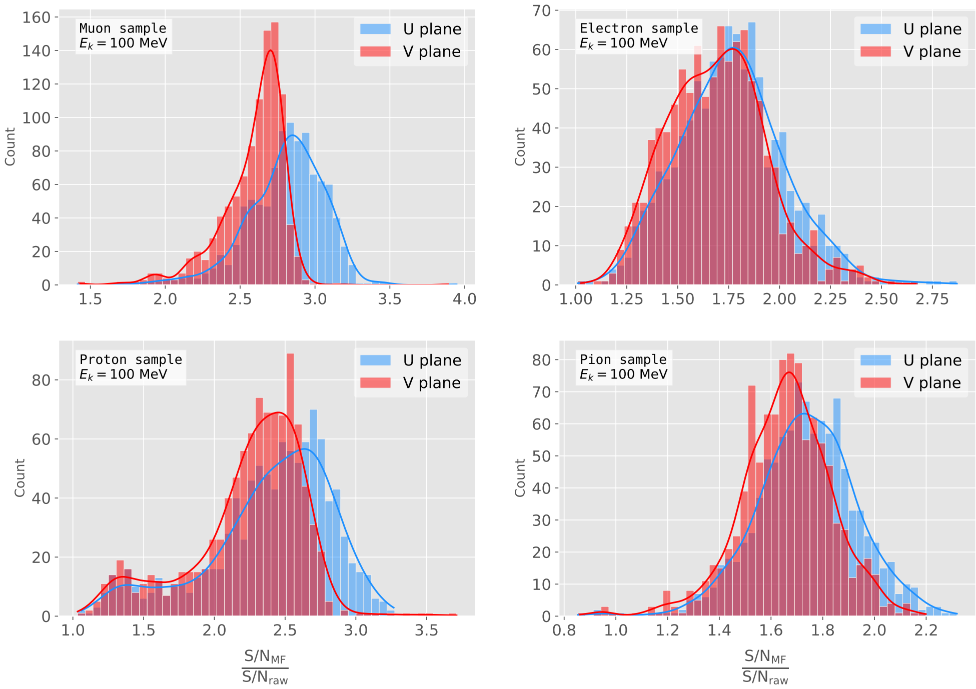


Figure 6.12: *Distributions of the mean S/N improvement per event for the corresponding sample after applying the matched filters. Here I separated the change in the U plane (blue)) and the V plane (red) channels. From top left to the right: muon, electron, proton and neutral pion. All the events have a fixed kinetic energy of* $E_k = 100$ MeV.

difference between these results and the ones seen before for ProtoDUNE data is that, overall, the improvements that I get for simulated data are bigger. This could be due either to the default noise model used in the *LArSoft* simulation or to the simulated hits having higher energy than the ones in the recorded data. Nonetheless, the concluding message is that the previously optimised matched filters give an overall significant improvement of the S/N for the different samples.

About the convention I followed for the plots and results, in the case of the raw and filtered S/N of each event in the sample I simply took the average of the quantities over all the active channels in the event. That is, if a certain event has N_{chan} active channels these two quantities are computed as:

$$\begin{aligned}(S/N_{fir})_{event} &= \frac{\sum_{i=0}^{N_{chan}} (S/N_{fir})_i}{N_{chan}}, \\ (S/N_{raw})_{event} &= \frac{\sum_{i=0}^{N_{chan}} (S/N_{raw})_i}{N_{chan}}.\end{aligned} \tag{6.18}$$

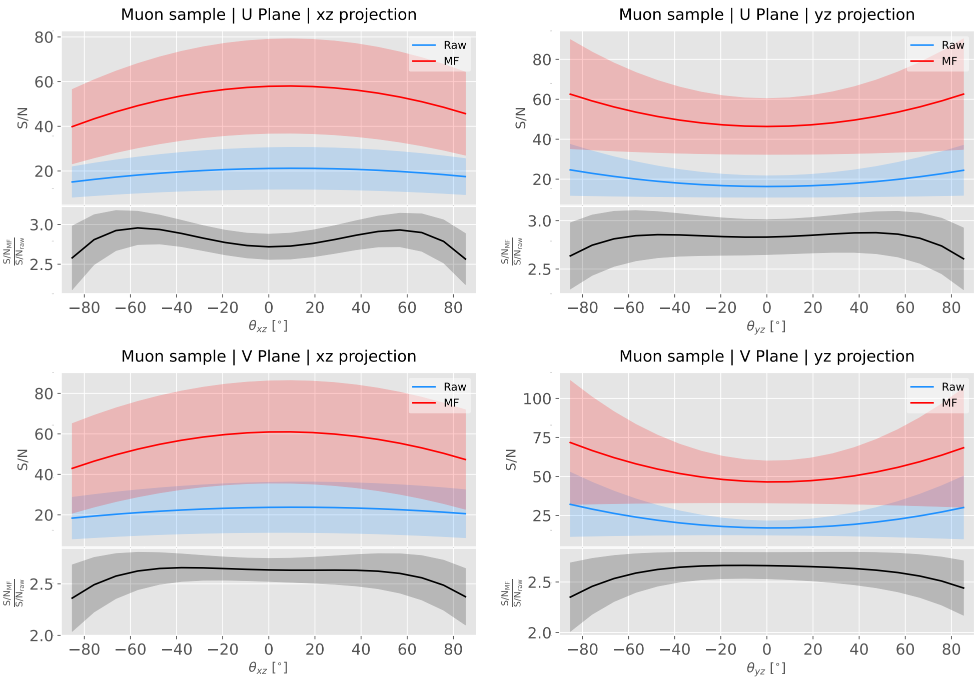


Figure 6.13: Angular dependence of the mean S/N and the S/N improvement, for the different monoenergetic samples considered (from top to bottom: electrons, muons, protons and neutral pions). The two columns on the left represent the values for the U plane waveforms. The top subplots show the mean S/N for raw (green) and filtered (red) waveforms whereas the bottom subplots depict the averaged S/N improvement (black).

However, for the ratio of the raw and filtered S/N (what I called the S/N improvement) per event I am not just taking the ratio of the previous two quantities but computing the average of the individual ratios per channel in the event:

$$\left(\frac{S/N_{fir}}{S/N_{raw}}\right)_{event} = \frac{\sum_{i=0}^{N_{chan}} \left(\frac{S/N_{fir}}{S/N_{raw}}\right)_i}{N_{chan}}, \tag{6.19}$$

and so:

$$\left(\frac{S/N_{fir}}{S/N_{raw}}\right)_{event} \neq \frac{(S/N_{fir})_{event}}{(S/N_{raw})_{event}}. \tag{6.20}$$

6.5.1 Angular dependence

Having these monoenergetic samples, one can also study the angular dependence of the performance of the matched filter. This is an important point, as it is a well established fact that for certain configurations (an extreme case configuration being signals normal

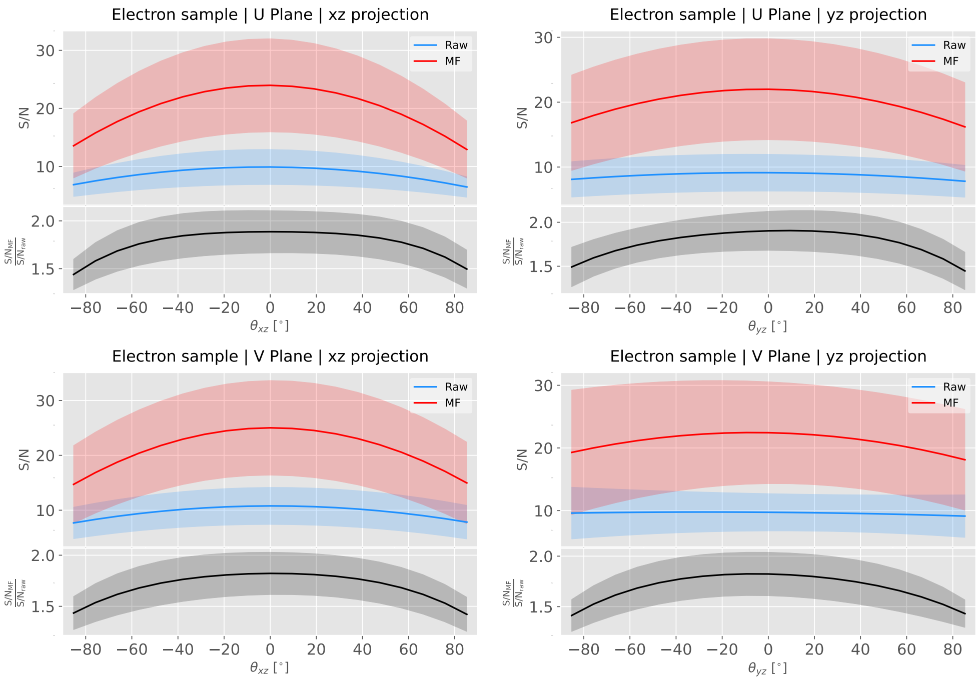


Figure 6.14: Angular dependence of the mean S/N and the S/N improvement, for the different monoenergetic samples considered (from top to bottom: electrons, muons, protons and neutral pions). The two columns on the left represent the values for the U plane waveforms. The top subplots show the mean S/N for raw (green) and filtered (red) waveforms whereas the bottom subplots depict the averaged S/N improvement (black).

to the wire plane and perpendicular to the induction wires at the same time) the S/N is much lower than average as the corresponding waveforms are severely distorted. In this sense, I am interested to see how the matched filter behaves for these cases and how the S/N improvement on those compare to the average.

Fig. 6.13 shows the angular dependence of the S/N for the monoenergetic $E_k =$ 100 MeV isotropic muons, for the different induction wire planes and projections. The angles for each event are given by the components of the initial value of the momentum of the particles, taking the angles of the projections on the xz and yz planes with respect to the z axis (more accurately, one needs to compute these angles twice for each event, a pair for the $xy'z'$ coordinate system and the other for the $xy''z''$). The top row shows the dependence on the angles corresponding to the U plane, i.e. $\theta_{xz'}$ and $\theta_{y'z'}$, whereas the bottom row shows the angular dependence viewed from the V plane, $\theta_{xz''}$ and $\theta_{y''z''}$. In each plot, the top subplot represents the mean values of the S/N for the raw (blue) and matched filtered (red) signals, and the bottom subplot the averaged S/N improvement (black). The solid lines represent the mean value obtained for the corresponding angular value, whereas the semitransparent bands represent one standard deviation around the mean at each point.

As expected, the S/N is in general higher when tracks are parallel to the APA (i.e. $\theta_{xz} \sim 0$) and lower when it is normal to the plane ($\theta_{xz} \sim \pm 90°$). In the same way, tracks parallel to the wires ($\theta_{yz} \sim \pm 90°$) tend to have higher S/N than those perpendicular to these ($\theta_{yz} \sim \pm 0$).

Fig. 6.14 shows the corresponding angular dependence information for the $E_k =$ 100 MeV electrons sample. Notice that, in this case, the S/N behaviour discussed above does not hold. A possible explanation can be that, because most hits in these events are produced by the secondary particles generated in the EM shower, the signal peaks whose S/N ratios were computed do not correspond to the directional information of the primary electron.

6.5.2 Distortion and peak asymmetry

As a ~~little~~ case ~~of~~ study, I selected two of the simulated $E_k = 100$ MeV monoenergetic muon events. With respect to the U induction plane, one is parallel to the APA (low $\theta_{xz'}$) and to the wires (high $\theta_{y'z'}$) and the other is normal to the APA plane (high $\theta_{xz'}$) and perpendicular to the wires (low $\theta_{y'z'}$). As expected from the results on the angular dependence discussed above, the former has a higher S/N (before and after the filtering) when compared to the latter. An interesting thing to notice about these two samples is that, even though one has a much bigger S/N than the other, it is the one with the smallest S/N the one that got the biggest averaged S/N improvement. In Table 6.1 I included all the relevant parameters of these two $E_k = 100$ MeV muon events ~~I am considering~~, namely, the angles with respect to the $xy'z'$ reference frame, the values of the S/N, the S/N improvement and also the so-called peak asymmetry Δ_{peak} that I will discuss next.

One can try to understand better what is going on with these two events by looking

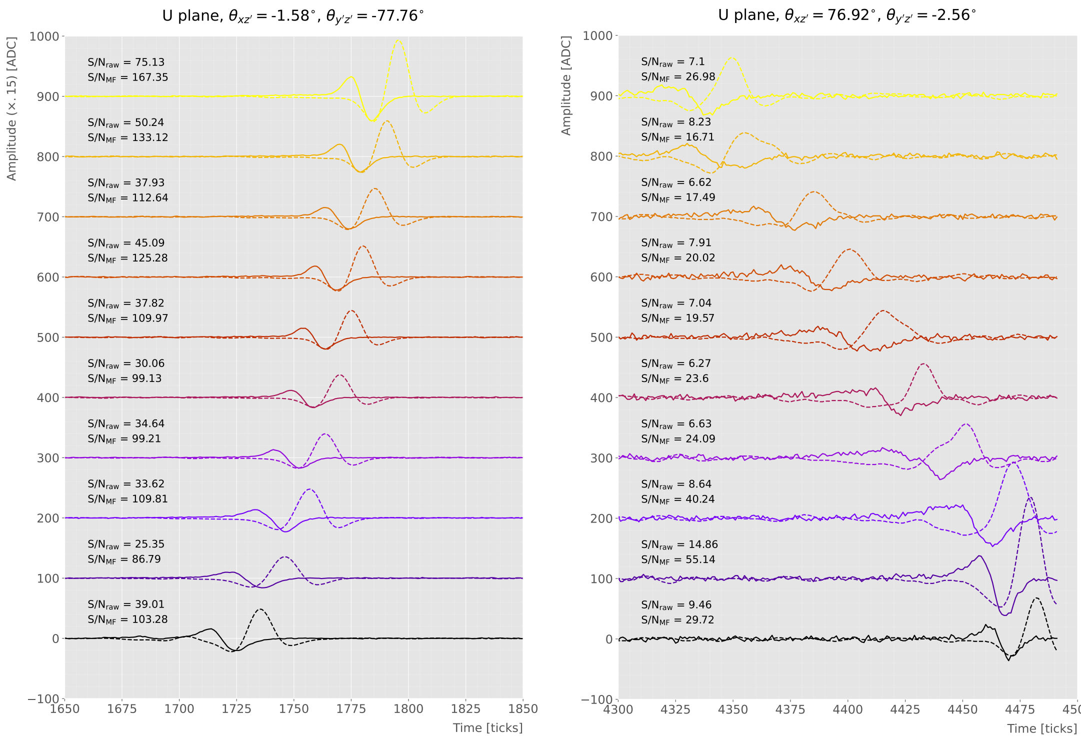


Figure 6.15: *Selected consecutive waveforms corresponding to two monoenergetic* $E_k = 100$ MeV *muon events, one is parallel to the APA and to the wires in the U plane (left panel) and the other is normal to the APA plane and perpendicular to the U plane wires (right panel). The solid lines represent the raw waveforms whereas the dashed lines correspond to the waveforms after the matched filter was applied. The waveforms on the left panel have been scaled by a factor of* 0.15 *to have similar amplitude to the ones on the right panel.*

Table 6.1: *Characteristic parameters of the two monoenergetic muon events selected, relative to the U plane: projected angles in the* xz' *and* $y'z'$ *planes, S/N values for the raw and filtered waveforms, mean improvement of the S/N and peak asymmetry.*

	$\theta_{xz'}$ (°)	$\theta_{y'z'}$ (°)	$\mathrm{S/N_{raw}}$	$\mathrm{S/N_{MF}}$	$\frac{\mathrm{S/N_{MF}}}{\mathrm{S/N_{raw}}}$	Δ_{peak} (ADC)
High ("parallel")	-1.58	-77.76	41.65	112.44	2.83	-35.73
Low ("normal")	76.92	-2.56	8.07	25.46	3.12	-10.38

at the raw and filtered data from some of their active channels. Fig. 6.15 shows a selection of consecutive raw and filtered U plane waveforms from the event with high S/N (left panel) and the one with low S/N (right panel). Notice that to show both collections of waveforms at a similar scale I had to apply a factor of 0.15 to the waveforms with high S/N. Additionally, next to each waveform I included the values of the raw and

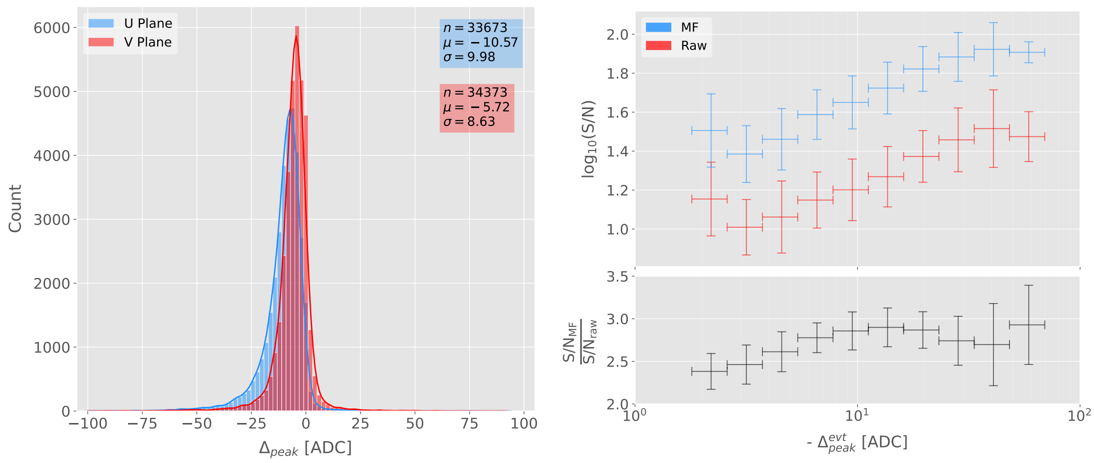


Figure 6.16: *Left panel: peak asymmetry distribution for the case of the monoenergetic* $E_k = 100$ MeV *muon sample. Each value corresponds to a single bipolar signal peak from a channel in any event. The blue distribution represents the peaks on U plane channels, whereas the red corresponds to signal peaks in V wires. Right panel: relation between the mean peak asymmetry per event with the S/N for U channel waveforms from the* $E_k = 100$ MeV *muon sample. The top subplot shows the decimal logarithm of the mean S/N for the raw (red) and the matched filtered (blue) waveforms. The bottom subplot contains the mean S/N improvement ratio after the matched filter was applied.*

matched filtered S/N for the corresponding channel. The first thing to notice in this plot is that the amplitude of the signal peaks from the normal track have a much smaller amplitude, and also appear quite distorted when compared to the others. On the other hand, although the matched filtered S/N is still smaller, the relative improvement is bigger than in the parallel case.

A way I found to quantify the difference between the shapes within these two events is their different peak asymmetry. One can define the peak asymmetry as the (signed) difference between the positive and the negative peaks of the bipolar shape, i.e.:

$$\Delta_{peak} \equiv h_+ - h_-, \tag{6.21}$$

where both heights h_+ and h_- are positive defined. Fig. 6.16 (left panel) shows the distribution of this peak asymmetry for all the waveforms corresponding to channels in the U (blue) and V (red) planes for the monoenergetic muon sample. One can see that these distributions are clearly shifted to negative values (with mean values

$\mu_{\Delta}^{\mathrm{U}} = -10.57$ ADC and $\mu_{\Delta}^{\mathrm{V}} = -5.72$ ADC respectively). It is interesting to notice that the peak asymmetry value of the sample with high S/N sits at the left tail of the distribution whereas the corresponding value of the sample with low S/N lies around the mean.

Now, one can try to correlate the peak asymmetry with the S/N and the S/N change per event. Fig. 6.16 (right panel) shows the result of comparing (minus) the mean peak asymmetry per event to the averaged raw (red) and matched filtered (blue) S/N per event (top subplot). The horizontal lines sit at the mean value obtained in the fit and represent the width of the $-\Delta_{peak}$ bins used, while the vertical lines indicate one standard deviation around that mean value. Notice that, when taking decimal logarithm on both, there is an approximate linear relation between these quantities, except for peak asymmetry values bigger than -5 ADC where the S/N remains constant.

Also, in the bottom subplot of Fig. 6.16 (right panel) I show the relation between the peak asymmetry and the mean S/N improvement. In this case, one see that there is a maximum at $\Delta_{peak} \sim -10$ ADC. As mentioned previously, this is also the value of the mean of the peak asymmetry distribution. In fact, it is expected that our filter favours the signal peaks with the most common values of the peak asymmetry, as this was one of the features I target in our filter coefficient optimisation through the parameter δ.

These results suggest that events with poorer values of the mean S/N, usually associated to non-favourable track orientations, tend to have smaller values of the mean peak asymmetry (in absolute value). Nonetheless, because our matched filters have been optimised to account for these asymmetries, the improvement on the S/N for these events is sizeable if not better than the one for events which already had a high S/N.

6.5.3 Hit sensitivity

One of the advantages of the matched filter, directly related to increasing the S/N, is the capability of picking hits that before fell below the threshold. For instance, Fig. 6.17 shows the raw ADC data from an example event (electron, $E_k = 100$ MeV) with the produced true hits superimposed (black boxes), together with the hits produced by the

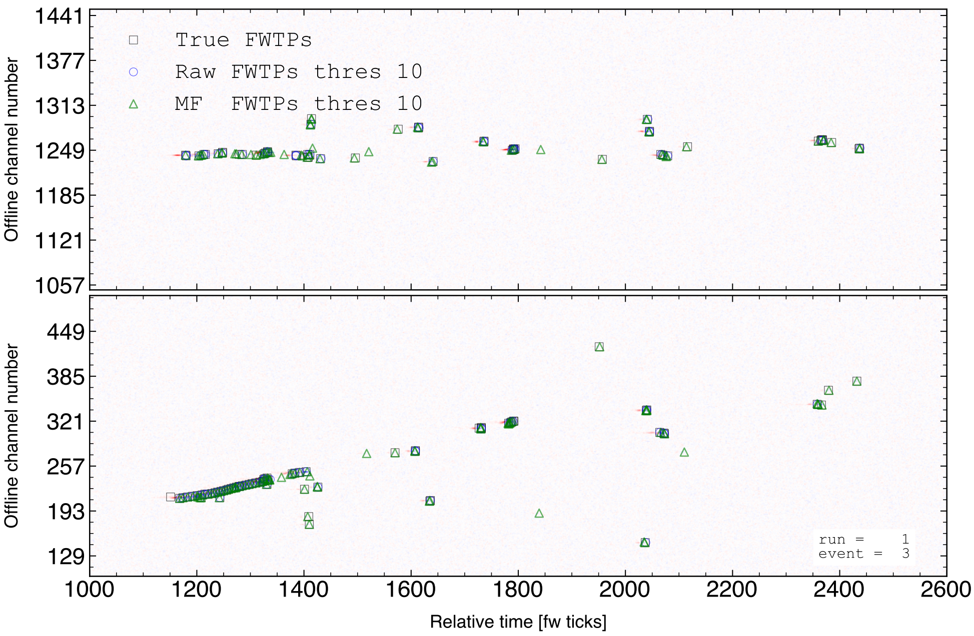


Figure 6.17: *Raw data display in the plane time (in firmware ticks) vs. offline channel number for an* $E_k = 100$ MeV *electron event. The produced true hits are superimposed (black boxes) as well as the hits comming from the standard hit finder chain (blue circles) and the hit finder using the matched filter (green triangles).*

standard hit finder chain (blue circles), i.e. using the current FIR filter, and the hits obtained using the matched filters (green triangles). Both the standard and the matched filter hit finders were run with a threshold of 10 ADC. Notice that the standard hits match well the true ones at the initial part of the event (where we have a track-like object), but they miss most of the hits produced by the EM shower at later times. On the other hand, the hits produced with the matched filter have a better agreement with the true hits even for the more diffuse shower activity.

Notwithstanding that ~~now I~~ the matched filter gets more hits ~~with this combination of matched filter and low threshold~~ as a results of the enhancement of the signal peaks relative to the noise level, it is also true that ~~I~~ it picks up some spurious hits not related to any real activity if one lowers the thresholds too much. Therefore, some optimisation of the threshold is needed, ~~Basically one will need to make~~ as there is a trade-off between precision and sensitivity.

Having this in mind, I tried to compare the produced hits one gets from the standard hit finder and the ones resulting from applying the matched filter with the true hits.

By running the hit finders on our samples with different values of the threshold one can understand, for instance, how low one can set the threshold without getting mostly spurious hits and then evaluate the gains obtained from this.

~~Because now I am also interested in seeing~~ To see how the hit sensitivity changes with the energy, I prepared new isotropic samples with the same types of particles as before (muons, electrons, protons and neutral pions) but with a flat kinetic energy distribution ranging from 5 to 100 MeV.

~~In order~~ to estimate the hit sensitivity, given a certain sample, one needs to recover the set of true hits to be able to compare these with the ones produced. To do so, a modification in the procedure I was using to extract the raw waveforms is needed. For this kind of study I run the detector simulation in two steps, first I produce the waveforms without noise and extract them in the same format I used for the raw data, then the noise is added and the noisy waveforms are then written to a file as well.

I am not sure how this is different from the true hits you used before?

To have a better comparison between the true hits and the ones produced from the raw waveforms after applying the two filters, I applied also the FIR filter and the matched filters to the noiseless waveforms and then I run the hit finder with a minimal threshold (in this case I used 1 ADC) on these noiseless filtered waveforms. In this way I generated two sets of true hits, I will refer to them as standard true hits (with the current/default FIR filter) and matched filter true hits respectively. This allows a more precise matching between the different groups of hits produced, as it will account for any delays and distortions introduced by the FIR and the matched filters.

In the case of the raw waveforms (with noise), I run the hit finder on them, with different values of the threshold, after applying either the FIR or the matched filters. I will name them simply standard hits and matched filter hits respectively. Then, I match the generated hits to the true hits (the standard hits with the standard true hits and the matched filter hits with the matched filter true hits). The matching is performed by comparing the channel number and the timestamp of the hits. To count as a match, I require that all hits with the same channel number and timestamp have overlapping hit windows, i.e. the time windows between their hit end and hit start times need to

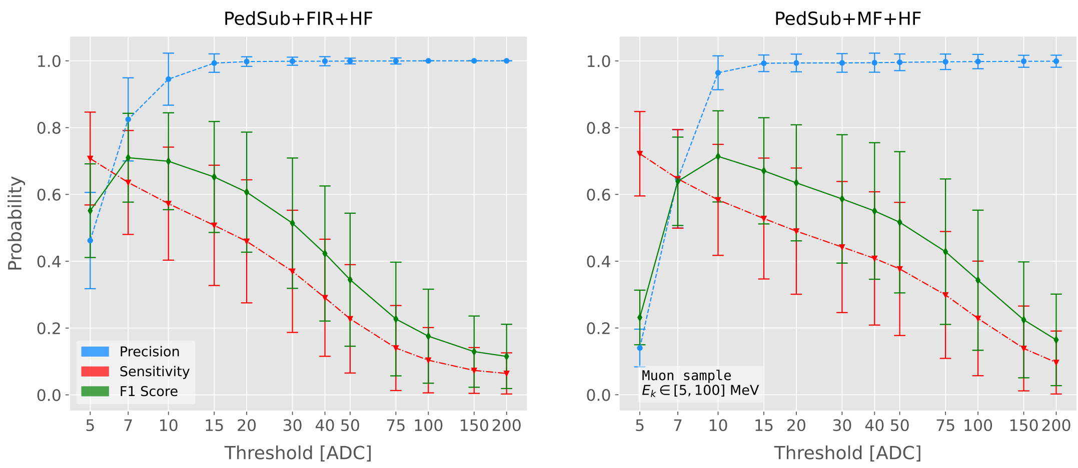


Figure 6.18: *Dependence of the precision (blue), sensitivity (red) and F_1 (green) scores on the threshold values used in the hit finder, for the FIR (left panel) and matched filter (right panel) cases. The results were obtained after matching the hits to the true hits in the case of the isotropic muon sample with kinetic energy in the range* 5 *to* 100 MeV*, taking only into account the induction plane channels. The points represent the mean value while the error bars indicate one standard deviation around that mean value.*

overlap. If more than one hit in one of the groups have hit overlap with the same hit in the other group I only count the hit with closer hit peak time value.

The generation of the samples, the procedure to produce the standard hits (with the default FIR filter) and matched filter hits and the matching of these with the true hits is described in detail in App. **??**.

To quantify the performance of the two hit finder approaches, I use a classical method from statistical classification known as confusion matrix [59]. ~~This is basically a way of sorting the outputs of a binary classifier, considering the true values of the classification and the predicted values.~~ It divides the outputs in four categories: true positive (TP, both true and predicted values are 1), false negative (FN, true value is 1 but predicted is 0), false positive (FP, true value is 0 but predicted is 1) and true negative (TN, both true and predicted values are 0)).

The contents of the confusion matrix allow us to compute other derived scores to judge the performance of our classifiers. In this study, I will make use of three of these metrics, namely the precision or positive predictive value:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{6.22}$$

the sensitivity or true positive rate:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{6.23}$$

and the F_1 score [60]:

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{6.24}$$

which is the harmonic mean of the precision and the sensitivity.

In our specific case I am not going to make use of the true negative value, as its definition in this context can be ambiguous because one does not have clear instances in the classification process. This way, I will only count the number of true positives as the total amount of hits I can match between true and raw populations, the number of false negatives will be the number of missing true hits and the false positive the number of hits which do not match any true hit.

In Fig. 6.18 I show the precision (blue), sensitivity (red) and F_1 (green) scores I obtained for different values of the threshold used in the hit finder for the case of the muon sample. Because the matched filters are only applied to induction channels, I only consider here hits coming from the U and V planes. The panel on the left corresponds to the scores I got when I ran the hit finder on the FIR filtered waveforms, whereas the right panel contains the scores for the matched filter case. The points are centered at the threshold value used and represent the mean value obtained for each score using all the generated events, while the error bars indicate one standard deviation around the mean value.

One can see that the precision for the matched filter case is lower when the thresholds are very low, as the noise baseline is slightly amplified, but then rises to high values quicker than for the FIR case. The other difference one can spot is that the sensitivity

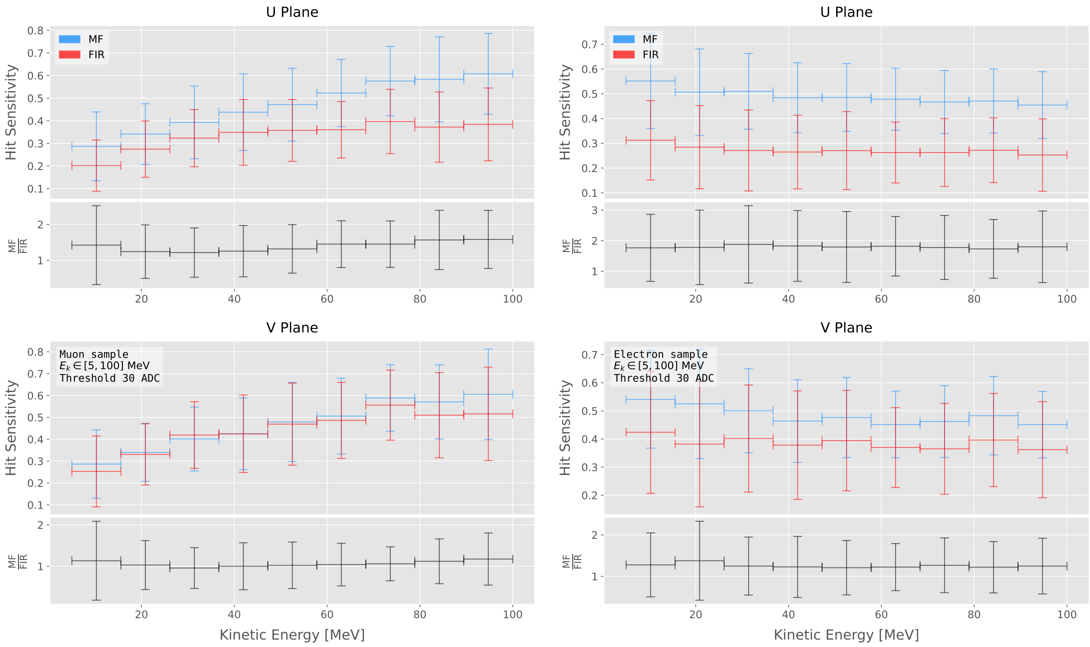


Figure 6.19: *Dependence of the averaged hit sensitivity on the kinetic energy of the events for the matched filter (blue) and standard (red) hits, for the case of the muon (left panel) and electron (right panel) samples, separated between U (top plots) and V (bottom plots) induction wire planes. The top subplots contain the hit sensitivities for the two hit finder alternatives, while the bottom subplots show the ratio between the two. The horizontal lines sit at the mean value and represent the size of the energy bins, while the vertical error bars indicate one standard deviation around that mean value.*

in the FIR case starts dropping faster at around the same threshold values where the precision stabilizes around 1, while in contrast for the matched filter this rapid decrease starts at higher threshold values. A similar scan for the same thresholds was performed for the electron sample in the same energy range, yielding similar results.

In Fig. 6.19 I show the averaged hit sensitivity versus the kinetic energy of the events, both for the matched filter hits (blue) and the standard hits (red). The left panel corresponds to the muon sample, whereas the one on the right corresponds to the electron sample, both with kinetic energies between 5 and 100 MeV. In each panel the top plot corresponds to hits in the U plane, while the bottom plot contains the same information for the V plane. Each plot contains two subplots, the one on the top shows the hit sensitivity values for the matched filter and standard hits separate, while the bottom subplot depicts the ratio between the matched filter and standard sensitivities.

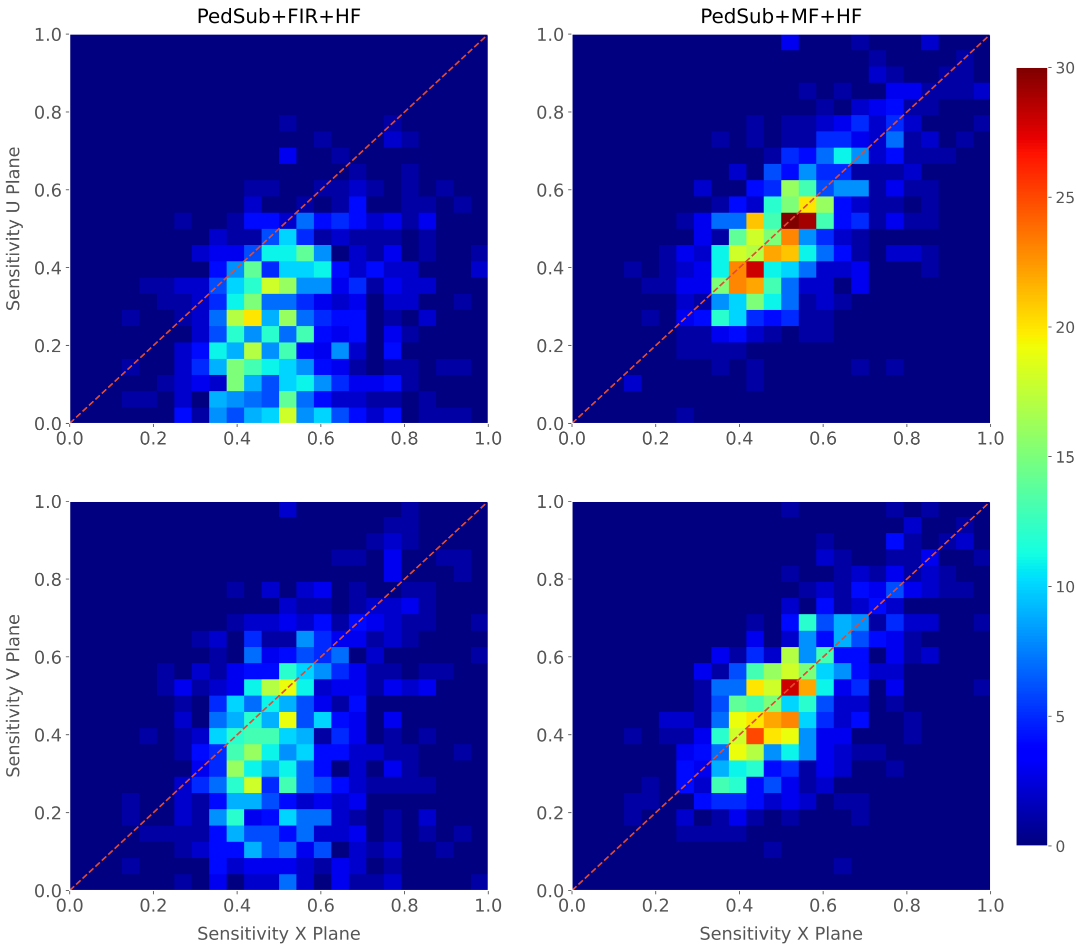


Figure 6.20: *Distributions of the hit sensitivity in the U (top panels) and V (bottom panels) planes versus the hit sensitivity in the X plane, both for the standard hits (left panels) and the matched filter hits (right panels), in the case of the electron sample and a threshold of* 30 ADC*.*

The horizontal lines are placed at the mean value obtained in the fit and represent the width of the E_k bins used, while the vertical error bars indicate one standard deviation around that mean value. In both cases the threshold used was 30 ADC, as I required the precision to be higher than 0.99 for both matched filter and standard cases.

One can see that, in general, the improvements are better for the U than for the V plane. While for the U channels I achieved a mean improvement of 50% and 80% for muons and electrons respectively, the improvement in the V plane is stalled at 10% and 25%. Nevertheless, if I look at the sensitivities for the matched filter hits in both planes one can see these have similar mean values for each energy bin, while on the contrary

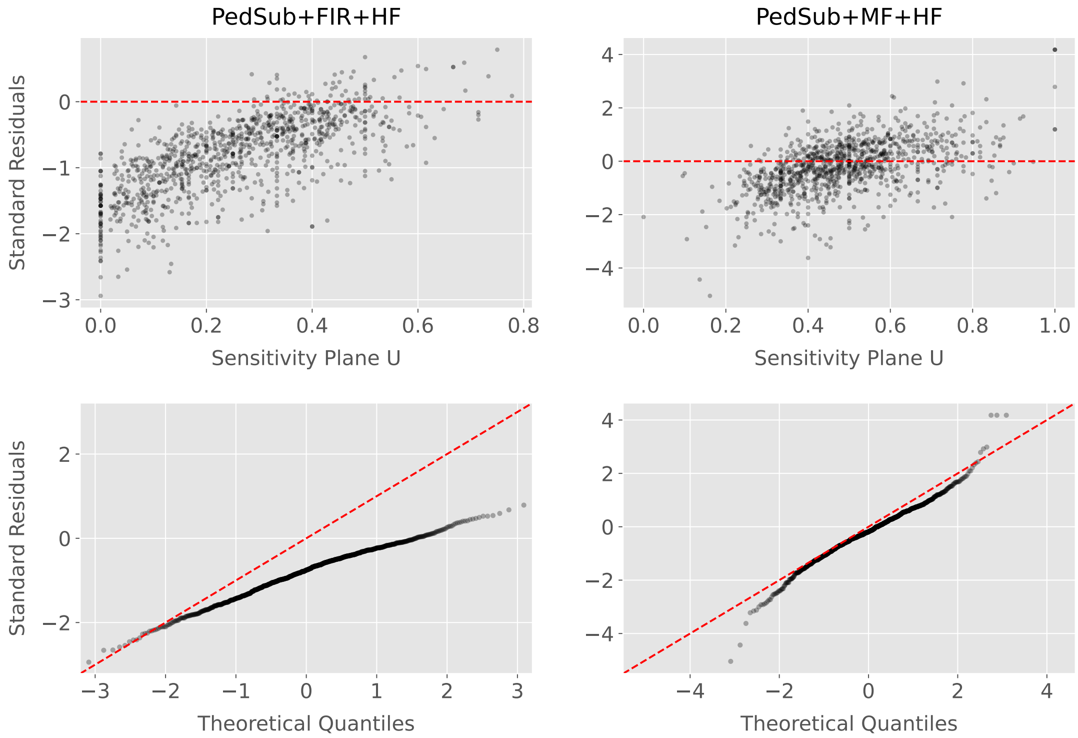


Figure 6.21: *Top panels: standard residual plots of the hit sensitivities between the X and U planes. Bottom panels: quantile-quantile plots of the hit sensitivity standard residuals between the X and U planes. In all cases, the left panel corresponds to the standard hits while the right panel represents the matched filter case, all from the electron sample with a* 30 ADC *threshold.*

for the standard hits the sensitivity remains relatively high for the V plane. This way, it looks there was a less significant gain because the hit sensitivity was already high.

Another interesting observation is the different behaviors for muons and electrons. While hit sensitivity for muons grows significantly with energy, in the case of electrons this slightly decreases the higher the kinetic energy of the event is. In any case, when it comes to the improvement on the sensitivities, this remains almost constant in all cases.

Furthermore, we can look at how the concurrence of hits between the different wire planes has changed. For any given event, I expect to have a similar number of hits in the three planes. As the ionisation electrons need to cross the U and V planes prior to reach the collection plane X they will induce current in those wire planes. A way to check the concurrence of hits across planes is looking at the relation between the hit sensitivities for each individual event. One cannot expect the sensitivities to be exactly equal across planes, but ideally they should be normally distributed around the diagonal.

Fig. 6.20 shows the hit sensitivity in the U (top panels) and V (bottom panels) planes versus the hit sensitivity in the X plane, for the case of the standard hits (left panels) and the matched filter hits (right panels). All plots were generated for the electron sample and a threshold of 30 ADC. From these one can see a clear trend, when I use the standard hit finder chain the sensitivities in the induction planes are systematically lower than the hit sensitivity in the X plane, i.e. most of the points sit below the diagonal (red dashed line). In contrast, when the matched filters are applied, the majority of the events are distributed around the diagonal. This points out that the concurrence of hits across planes has improved.

To exemplify the improvement I obtained, one can consider the residuals of the hit sensitivities for the X and U planes. Assuming the diagonal hypothesis, i.e. given a dataset of the form (x, y) for any x I take the predicted y value to be equal to the value of x, I can compute the standard residuals for the hit sensitivities in U given the ones for X. In Fig. 6.21 (top panels) I show these standard residuals against the corresponding values of the hit sensitivity in the U plane, for our electron sample with kinetic energy between 5 and 100 MeV. If I compare the scatter points in the case of the standard hits (left panel) and the matched filter hits (right panel), I see that the residuals of the standard hit finder case follow a certain pattern and their mean deviates from 0.

To see clearly if the residuals are normally distributed, in Fig. 6.21 (bottom panels) I plot the corresponding quantile-quantile plot for both the standard (left panel) and matched filter (right panel) standard residuals. One can clearly see that the points for the standard case follow a strongly non-linear pattern, suggesting that the residuals do not follow a normal distribution. In contrast, for the matched filter hits the points conform to a roughly linear path, implying that in this case the normality condition is fulfilled.

All these results hint at the fact that the concurrence of hits across the wire planes can be strengthened by applying the matched filters.

+ ADD PROTEOMICS DATA

In some figures you say FIR, in other "standard", I would choose one and be consistent

Chapter 7

DM searches with neutrinos from the Sun

7.1 Motivation

The idea of detecting neutrino signals coming from the Sun's core to probe DM is not new. The main focus of these searches has usually been high-energy neutrinos originated from DM annihilations into heavy particles [61–64], although recent studies have proposed to look at the low-energy neutrino flux arising from the decay of light mesons at rest in the Sun [65–68] previously thought undetectable.

In this chapter I try to demonstrate the capability of DUNE to constrain different DM scenarios. I used the neutrino fluxes arising from DM annihilations in the core of the Sun to compute the projected limits that DUNE would be able to set on the annihilation rates in the Sun and the DM scattering cross sections.

7.2 Gravitational capture of DM by the Sun

The Sun and the centre of the Earth are possible sources of DM annihilations, specially interesting because of their proximity. Their gravitational attraction ensured the capture of DM from the local halo through repeated scatterings of DM particles crossing them.

Only neutrinos produced from DM annihilations can escape the dense interior of these objects. Therefore, neutrino telescopes are the most useful experimental layouts to pursue DM searches from their cores.

The neutrino flux from DM annihilations inside the Sun depends on the DM capture rate, which is proportional to the DM scattering cross section, and the annihilation rate, which is proportional to the velocity-averaged DM annihilation cross-section. The total number of DM particles inside the Sun follows the Boltzmann equation [65]:

$$\frac{\mathrm{d}N_{DM}}{\mathrm{d}t} = C_\odot - A_\odot N_{DM}^2, \tag{7.1}$$

where $C_\odot$ and $A_\odot$ are the total Sun DM capture and annihilation rates respectively. In this expression I neglected the evaporation term, proportional to N_{DM}, which only contribute for $m_{DM} \lesssim 4$ GeV [69]. As the current threshold of neutrino telescopes is a few GeV, this region falls below the probed range but can be important in future low-energy projects.

This equation has an equilibrium solution:

$$N_{DM}^{eq} = \sqrt{\frac{C_\odot}{A_\odot}}, \tag{7.2}$$

which represents the amount of DM inside the Sun if the capture and annihilation have reached equilibrium. As the Sun is approximately 4.6 Gyr old, it is usually assumed that equilibrium has been achieved. Therefore, the anomalous neutrino flux from the Sun would only depend on the DM scattering cross section, enabling us to set limits on this quantity. If one does not assume equilibrium, some assumptions on the DM annihilation cross section are necessary to extract predictions from neutrino signals.

Here, I am going to consider three possible scenarios for the DM interactions: DM scattering off electrons, spin-dependent (SD) and spin-independent interactions off nuclei. For the case of these last two, the cross sections will be given in terms of the SD and SI elastic scattering DM cross section off protons (assuming that DM interactions off protons and neutrons are identical), $\sigma_p^{\rm SD}$ and $\sigma_p^{\rm SI}$, as [4, 65]:

$$\sigma_i^{\rm SD} = \left(\frac{\tilde{\mu}_{A_i}}{\tilde{\mu}_p}\right)^2 \frac{4(J_i+1)}{3J_i} |\langle S_{p,i}\rangle + \langle S_{n,i}\rangle|^2 \sigma_p^{\rm SD}, \tag{7.3}$$

$$\sigma_i^{\rm SI} = \left(\frac{\tilde{\mu}_{A_i}}{\tilde{\mu}_p}\right)^2 A_i^2 \sigma_p^{\rm SI}, \tag{7.4}$$

where $\tilde{\mu}_{A_i}$ is the reduced mass of the DM-nucleus i system, $\tilde{\mu}_p$ is the reduced mass of the DM-proton system, A_i and J_i the mass number and total angular momentum of nucleus i and $\langle S_{p,i}\rangle$ and $\langle S_{n,i}\rangle$ the expectation value of the spins of protons and neutrons averaged over all nucleons, respectively (see Ref. [70] for a review on spin expectation values).

Since the Sun is mainly composed of Hydrogen, the capture of DM from the halo is expected to occur mainly through spin-dependent scattering. However, since the spin-independent cross section is proportional to the square of the atomic mass, heavy elements can contribute to the capture rate (even though they constitute less than 2% of the mass of the Sun). Heavy elements can also contribute to the spin-dependent cross section if the DM has also momentum-dependent interactions.

DM particles can get captured by the Sun if after repeated scatterings off solar targets their final velocity is lower than the escape velocity of the Sun. In the limit of weak cross sections, this capture rate can be approximately written as [4]:

$$C_\odot^{\rm weak} = \sum_i \int_0^{R_\odot} {\rm d}r\ 4\pi r^2 \int_0^\infty {\rm d}u_\chi\ \frac{\rho_\chi}{m_\chi} \frac{f_{v_\odot}(u_\chi)}{u_\chi} \omega(r) \int_0^{v_e(r)} {\rm d}v\ R_i^-(\omega \to v) |F_i(q)|^2, \tag{7.5}$$

where the summation extends over all possible nuclear targets. In this expression, $R_\odot$ is the radius of the Sun, ρ_χ is the local DM density, m_χ the mass of the DM particle, $f_{v_\odot}(u_\chi)$ the DM velocity distribution seen from the Sun's reference frame, $R_i^-(\omega \to v)$ is the differential rate at which a DM particle with velocity v scatters a solar target of mass m_i to end up with a velocity ω and $|F_i(q)|$ is the nuclear form factor of target i.

The differential scattering rate takes a rather simple form when considering velocity-independent and isotropic cross sections. In that case, this quantity is given by [4, 71]:

$$R_i^-(\omega \to v) = \frac{2}{\sqrt{\pi}} \frac{\mu_{i,+}^2}{\mu_i} \frac{v}{\omega} n_i(r) \sigma_i \left[\chi(-\alpha_-, \alpha_+) + \chi(-\beta_-, \beta_+) \mathrm{e}^{\mu_i (\omega^2 - v^2)/u_i^2(r)} \right], \quad (7.6)$$

where μ_i is the ratio between the DM mass and the mass of target i, $\mu_{i,\pm}$ is defined as:

$$\mu_{i,\pm} \equiv \frac{\mu_i \pm 1}{2}, \quad (7.7)$$

$n_i(r)$ is the density profile of target i in the solar medium, $u_i(r)$ is the most probable velocity of target i given by:

$$u_i(r) = \sqrt{\frac{2T_\odot(r)}{m_i}}, \quad (7.8)$$

where $T_\odot(r)$ is the temperature of the Sun, the quantities $\alpha_\pm$ and $\beta_\pm$ are defined as:

$$\alpha_\pm \equiv \frac{\mu_{i,+} v \pm \mu_{i,-} \omega}{u_i(r)}, \quad (7.9)$$

$$\beta_\pm \equiv \frac{\mu_{i,-} v \pm \mu_{i,+} \omega}{u_i(r)}, \quad (7.10)$$

and the function $\chi(a, b)$ is a Gaussian integral of the form:

$$\chi(a, b) \equiv \int_a^b \mathrm{d}x \; \mathrm{e}^{-x^2}. \quad (7.11)$$

Finally, if one assumes the DM halo velocity distribution in the galactic rest frame to be a Maxwell-Boltzmann distribution, one can write the halo velocity distribution for an observer moving at the speed of the Sun with respect to the DM rest frame as:

$$f_{v_\odot}(u_\chi) = \sqrt{\frac{3}{2\pi}} \frac{u_\chi}{v_\odot v_d} \left(\mathrm{e}^{-\frac{3(u_\chi - v_\odot)^2}{2v_d^2}} - \mathrm{e}^{-\frac{3(u_\chi + v_\odot)^2}{2v_d^2}} \right), \quad (7.12)$$

where:

$$\omega^2 = u_\chi + v_e(r)^2, \quad (7.13)$$

is the DM velocity squared, $v_\odot$ the relative velocity of the Sun from the DM rest frame

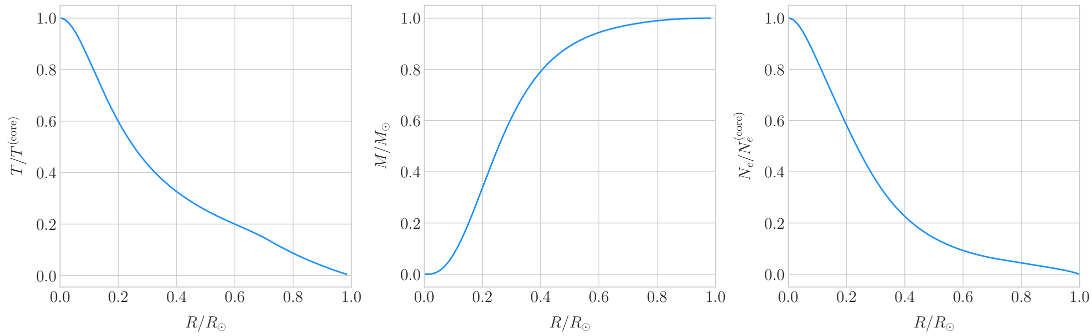


Figure 7.1: *Input solar parameters used in our capture rate computation as functions of the Sun's radius, from left to right: temperature (with respect to the temperature at the core), mass (in solar masses) and electron number density (with respect to the electron density at the core). All quantities shown correspond to the standard solar model BS2005-OP [3].*

and $v_d \simeq \sqrt{3/2} v_\odot$ the velocity dispersion.

For the case of strong scattering cross section, Eq. (7.5) ceases to be valid, as it escalates indefinitely with the cross section. In that limit, the capture rate saturates to the case where the probability of interaction is equal to one, which can be written as:

$$C_\odot^{\mathrm{geom}} = \pi R_\odot^2 \left(\frac{\rho_\chi}{m_\chi}\right) \langle v \rangle \left(1 + \frac{3}{2}\frac{v_e^2(R_\odot)}{v_d^2}\right) \xi(v_\odot, v_d), \tag{7.14}$$

where $v_d = \sqrt{8/3\pi} v_d$ is the mean velocity in the DM rest frame and the factor $\xi(v_\odot, v_d)$ accounts for the suppression due to the motion of the Sun:

$$\xi(v_\odot, v_d) = \frac{v_d^2 \mathrm{e}^{-\frac{3v_\odot^2}{2v_d^2}} + \sqrt{\frac{\pi}{6}}\frac{v_d}{v_\odot}\left(v_d^2 + 3v_e^2(R_\odot) + 3v_\odot^2\right) \mathrm{Erf}\left(\sqrt{\frac{3}{2}}\frac{v_\odot}{v_d}\right)}{2v_d^2 + 3v_e^2(R_\odot)}. \tag{7.15}$$

Having these into account, one can write the total capture rate as a combination of both contributions, allowing a smooth transition between the two, as:

$$C_\odot = C_\odot^{\mathrm{weak}} \left(1 - \mathrm{e}^{C_\odot^{\mathrm{geom}}/C_\odot^{\mathrm{weak}}}\right). \tag{7.16}$$

I computed the capture rate from Eq. (7.16) in the case of interactions with electrons. To do so, I used the standard solar model BS2005-OP [3]. Fig. 7.1 shows the

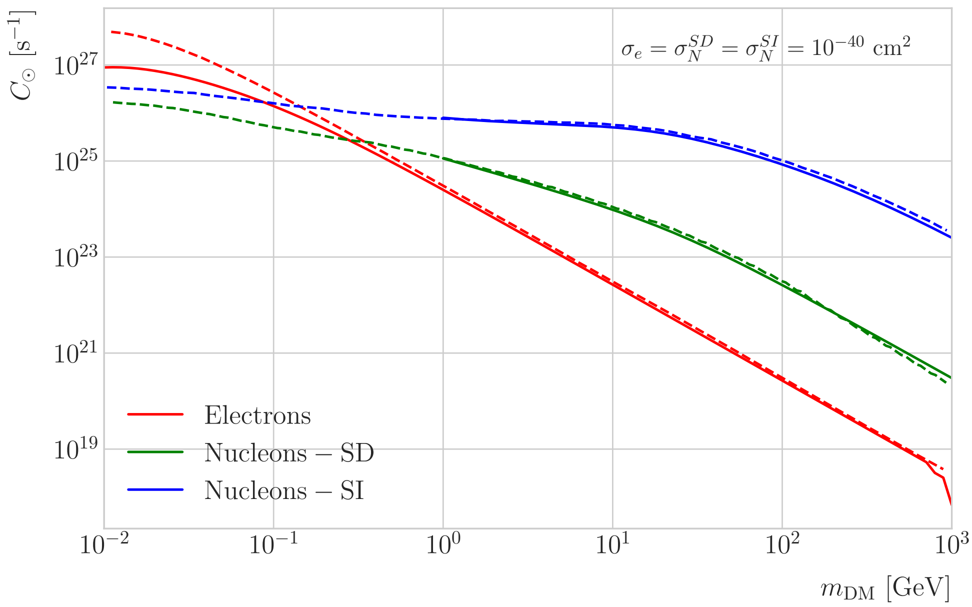


Figure 7.2: *Capture rates as a function of the DM mass for the DM-electron interactions (red lines), SD DM-nucleons interactions (green lines) and SI DM-nucleons interactions (blue lines). Solid lines represent the values computed in this work while the dashed lines are the one given in Ref. [4]. All the rates are shown for a choice of scattering cross section of* $\sigma_i = 10^{-40}$ cm^2.

three parameters from the solar model that are needed for the computation, the solar temperature (left panel), mass (central panel) and electron density (right panel) profiles.

For the case of the interactions off nuclei, the computations are more convoluted as one needs to add up the contributions of the different most abundant nuclei in the Sun. Also, in contrast to the electron scenario where the form factor is trivially $|F_e(q)|^2 = 1$, for any nucleus i one would need to consider some appropriate nuclear density distribution (either a Gaussian approximation, a Woods-Saxon distribution, etc) which would complicate the calculations even further.

That is the reason why, at this stage of our study, I decided to take an alternative approach to the computation of the DM-nucleus capture rates. I used the `DarkSUSY` software, that allows us to compute these quantities performing a full numerical integration over the momentum transfer of the form factors. The default standard solar model used by `DarkSUSY` is BP2000[1] [72].

[1]This is what they say in their manual, but I fear it is somewhat outdated. It appears to me this

In Fig. 7.2 I show the results I obtained for the capture rates, for the case of interactions off electrons (red solid line), SD (green solid line) and SI (blue solid line) interactions of nucleons. In all cases I used a value of the scattering cross sections of $\sigma_i = 10^{-40}$ cm^2. Note here one of the limitations of the `DarkSUSY` approach, one can not extend the computation below $m_{\rm DM} = 1$ GeV. Nevertheless, this is not something to worry about in this case, as I will discuss next. As a comparison, I added also the values computed in Ref. [4] (same color scheme, dashed lines). One can see there is good agreement between these and the `DarkSUSY` computation of the SD and SI interactions for $m_{\rm DM} \geq 1$ GeV. In this regime their computations also matches quite well our result for the electron capture rate. However, these start to differ significantly below $m_{\rm DM} = 1$ GeV, being their estimate up to a factor of 5 bigger than ours for low masses.

Let us comment briefly about the assumption I made before about not including an evaporation term in the Boltzmann equation. If I include this term in the equation (which will be proportional to the number of DM particles) the equilibrium solution takes the form:

$$N_{DM}^{eq} = \sqrt{\frac{C_\odot}{A_\odot}} \frac{1}{\kappa + \frac{1}{2} E_\odot \tau_{eq}}, \tag{7.17}$$

where $E_\odot$ is the total evaporation rate, τ_{eq} is the equilibrium time in the absence of evaporation:

$$\tau_{eq} = \frac{1}{\sqrt{C_\odot A_\odot}}, \tag{7.18}$$

and κ is defined as:

$$\kappa \equiv \sqrt{1 + \left(\frac{E_\odot \tau_{eq}}{2}\right)^2}. \tag{7.19}$$

Now, it is easy to proof that in case evaporation dominates $\kappa \gg 1$ and therefore:

$$N_{DM}^{eq} \simeq \frac{C_\odot}{E_\odot}. \tag{7.20}$$

In contrast, if evaporation is irrelevant $\kappa \simeq 1$ and one recovers Eq. (7.2).

model is relatively old and do not see why they are not using others like [3]. Maybe one can double-check in the code to make sure.

In this way, one can define the evaporation mass as the mass for which the number of DM particles in equilibrium approaches Eq. (7.20) at 10% level:

$$\left| N_{DM}^{eq}(m_{\text{evap}}) - \frac{C_{\odot}(m_{\text{evap}})}{E_{\odot}(m_{\text{evap}})} \right| = 0.1 N_{DM}^{eq}(m_{\text{evap}}). \tag{7.21}$$

This can be regarded as the minimum testable mass one can reach using the annihilation products of the DM in the Sun.

It was reported in Ref. [4] that, in the case of both SD and SI DM interactions off nuclei, this value ranges from 2 to 4 GeV depending on the specific scattering cross section value, compatible with the usual assumptions in the literature. What is interesting is the case of the electron capture. It was found that, when one applies a cutoff in the velocity distribution of the DM trapped in the Sun slightly below the escape velocity, the evaporation mass for the DM-electron interaction decreases remarkably. For a moderate choice of $v_c(r) = 0.9 v_e(r)$ one gets an evaporation mass of around 200 to 600 MeV. This possibility opens a region of the parameter space that could be tested with neutrino detectors.

7.3 Neutrino flux from DM annihilations

When WIMPs annihilate inside the Sun a flux of high-energy neutrinos is expected from heavy quarks, gauge bosons and $\tau^+\tau^-$ final states, which decay before losing energy in the dense solar medium, as they will produce a continuum spectra up to $E_\nu \sim m_\chi$ (in the case of direct annihilation to neutrinos one would have a line at $E_\nu = m_\chi$) [66]. This kind of signal has been extensively studied in the literature, allowing to put strong limits on the SD WIMP-proton cross section for large m_χ. However, the number of high-energy neutrinos per WIMP annihilation is small and the spectrum depends on the unknown final state. Moreover, background rejection is easier for large m_χ but neutrinos with $E_\nu \gtrsim 100$ GeV are significantly attenuated by interactions in the Sun.

Nevertheless, most WIMP annihilation final states eventually produce a low-energy neutrino spectrum. In this case one does not just consider the more massive final states but also annihilations into e^+e^-, $\mu^+\mu^-$ and light quarks [65]. In particular, light mesons would be produced and stopped in the dense medium, thus decaying at rest and producing a monoenergetic neutrino signal. The decay-at-rest of kaons will produce a $E_\nu = 236$ MeV ν_μ while in the case of pions one would have a $E_\nu = 29.8$ MeV ν_μ. In practice only K^+ and π^+ contribute to these signals, as K^- and π^- are usually Coulomb-captured in an atomic orbit and get absorbed by the nucleus. There is also a low-energy neutrino signal coming from muon decays, which are produced in kaon or pion decays, leptonic decays of other hadrons and heavy leptons or even directly from WIMP annihilations, which can decay at rest and contribute to the previous low-energy neutrino flux with a well known spectrum below 52.8 MeV.

These monoenergetic MeV neutrinos were previously considered undetectable but, due to the large yield, the known spectra and the modern advances in the detector technology, these low-energy neutrino flux can be a good probe of the SD WIMP-proton cross-section in standard solar WIMP capture scenario, as it is sensitive to low WIMP masses and insensitive to the particular final state. A good place to look for these signals are next-generation neutrino experiments such as DUNE.

7.4 Computing limits from solar neutrino fluxes

In order to use the neutrino fluxes from DM annihilations in the Sun, the first thing I need to do is to determine the expected number of atmospheric background events, for a given exposure, after directionality selection has been applied. I can write this number as:

$$N_B = \eta_B \int \mathrm{d}\Omega \int_{E_{min}}^{E_{max}} \mathrm{d}E_\nu \, \frac{\mathrm{d}^2\Phi^\mu_{atm}}{\mathrm{d}E_\nu \mathrm{d}\Omega} \times \left(A^{(\mu)}_{eff}(E_\nu) T \right), \tag{7.22}$$

where η_B is the background efficiency, E_{min} and E_{max} the minimum and maximum energies to integrate over, $\mathrm{d}^2\Phi^\mu_{atm}/\mathrm{d}E_\nu\mathrm{d}\Omega$ the differential flux of atmospheric muon neutrinos, $A^{(\mu)}_{eff}$ is the effective area of DUNE to muon neutrinos and T is the exposure time. The effective area can be expressed as the product of the neutrino-nucleus scattering cross section and the number of nuclei in the fiducial volume of the detector. This way

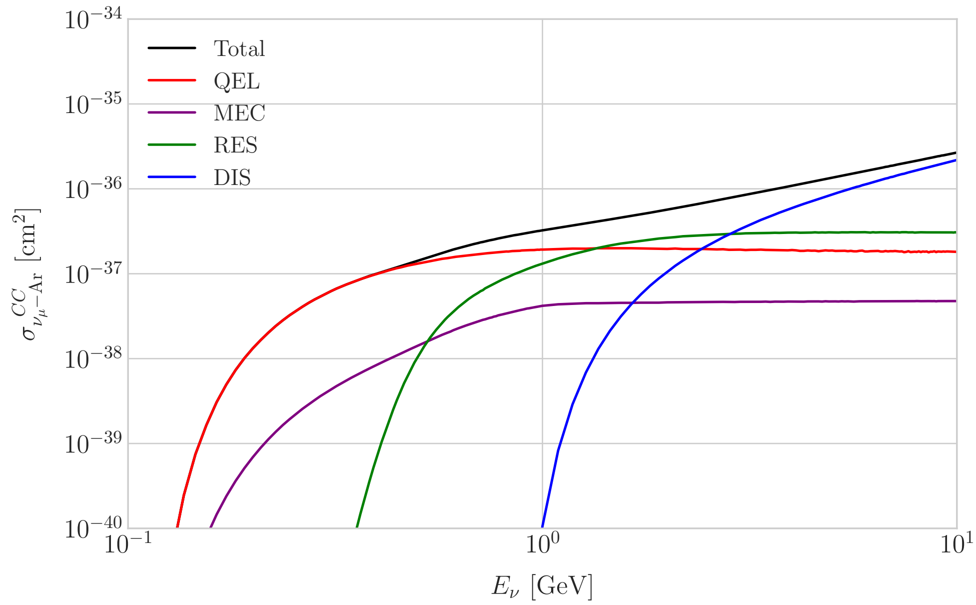


Figure 7.3: *`NuWro` computed $\nu_\mu - ^{40}\mathrm{Ar}$ charged-current scattering cross section as a function of the neutrino energy E_μ. The black line shows to the total cross section, whereas the others correspond to the different contributions (in red quasi-elastic scattering, in green resonant pion exchange, in blue deep inelastic scattering and in purple meson exchange current).*

for DUNE I can write:

$$A^{(\mu)}_{eff}(E_\nu) = (6.0 \times 10^{-10}\ \mathrm{m}^2)\left(\frac{\sigma^{(\mu)}_{\nu-\mathrm{Ar}}(E_\nu)}{10^{-38}\ \mathrm{cm}^2}\right)\left(\frac{M_{target}}{40\ \mathrm{kT}}\right), \tag{7.23}$$

where $\sigma^{(\mu)}_{\nu-\mathrm{Ar}}$ is the $\nu_\mu - ^{40}\mathrm{Ar}$ charged-current scattering cross section. In Fig. 7.3 I show the computed value of this cross section as a function of the neutrino energy E_ν, in the range of interest both for the atmospheric background and signal events. It was computed using the `NuWro` Monte Carlo neutrino event generator [73], including the charged-current contributions of the quasi-elastic scattering (red line), resonant pion exchange (green line), deep inelastic scattering (blue line) and meson exchange current (purple line).

The background rejection will depend on the resolution of the detector and the selection one applies on the events. A geometry argument can be used to estimate the maximum background rejection one can achieve in this case, considering one can efficiently discriminate all events coming from a direction different from that of the Sun. In that case, the optimal background efficiency will simply be the relative angular coverage of the Sun. Taking the angular diameter of the Sun as seen from the Earth to be 0.5°, I have:

$$\eta_B^{(opt)} \approx \frac{\pi \left(\frac{0.5}{2}\right)^2}{360 \times 180} \simeq 3.03 \times 10^{-6}. \tag{7.24}$$

This value will give a very optimistic estimate of the number of background events. However, it can be regarded as an lower limit, as it represents the best case scenario.

In Fig. 7.4 I show the fluxes of atmospheric neutrinos at the Homestake mine during solar minimum, taken from Ref. [5]. The values are averaged over the two angular directions. In blue I have the flux of muon neutrinos while in red I indicate the flux of electron neutrinos. Additionally, the dashed lines correspond to both antineutrino species.

Using these values for the muon neutrino and the corresponding total CC cross section, one can compute the number of expected background events by integrating over the given energy range (as in this case the angular integral is trivial). As for the energy range to integrate over, I choose the range for DUNE specified in [42], $E_{min} = 10^{-1}$ GeV and $E_{max} = 10$ GeV. Taking all these into account, I found the number of background events to be:

$$N_B \simeq \eta_B \times \left(3.827 \times 10^4\right) \times \left(\frac{\text{exposure}}{400 \text{ kT yr}}\right). \tag{7.25}$$

In order to estimate the sensitivity of DUNE to this kind signal, one can consider a hypothetical data set where the number of observed neutrinos is taken to be the expected number of background events rounded to the nearest integer, $N_{obs} = \text{round}(N_B)$ [74]. Now, if I assume that the number of signal and background events seen by DUNE are given by Poisson distributions with means equal to the expected number of signal and background events, N_S and N_B, one can denote by N_S^{90} to the number of expected signal events such that the probability of having an experimental run with a number of events greater than N_{obs} is 90%. This number can be obtained as the numerical solution

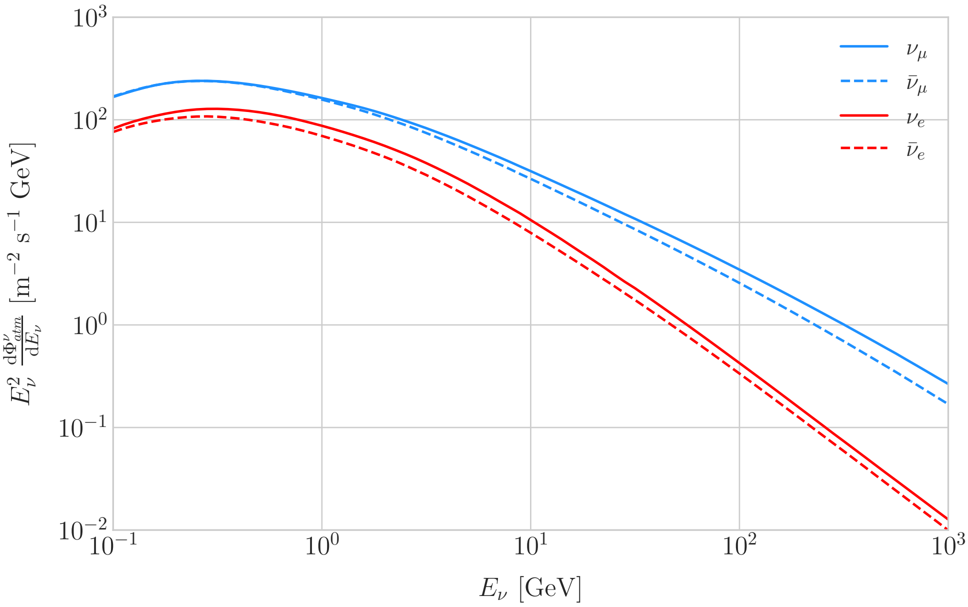


Figure 7.4: *Expected atmospheric neutrino flux as a function of the neutrino energy E_ν at Homestake at solar minimum, taken from Ref. [5]. The blue solid (dashed) line correspond to muon neutrinos (antineutrinos) and the red solid (dashed) line correspond to electron neutrinos (antineutrinos).*

to the equation:

$$1 - \frac{\Gamma\left(N_{obs}+1, N_S^{90}+N_B\right)}{N_{obs}!} = 0.9, \tag{7.26}$$

where $\Gamma(x, y)$ is the upper incomplete gamma function.

The number of signal events is related to the neutrino flux from DM annihilations in a similar way as the background events to the atmospheric neutrino flux. In this case I have:

$$N_S = \eta_S\, \Gamma_A^{eq} \int_{z_{min}}^{z_{max}} \mathrm{d}z\, \frac{\mathrm{d}N_\nu}{\mathrm{d}A\mathrm{d}N_A\mathrm{d}z} \times \left(A_{eff}^{\mu}(z) T\right), \tag{7.27}$$

where η_S is the signal efficiency, Γ_A^{eq} is the total annihilation rate of DM particles at equilibrium, $\Gamma_A^{eq} = A_\odot \left(N_{DM}^{eq}\right)^2$, z_{min} and z_{max} the minimum and maximum relative energies to integrate over (in such a way that $z_{min,max} \leq E_{min,max}/m_{\mathrm{DM}}$ for each m_{DM}) and $\mathrm{d}N_\nu/\mathrm{d}A\mathrm{d}N_A\mathrm{d}z$ the muon neutrino flux per DM annihilation in the Sun.

Knowing N_S^{90} one can use the relation in Eq. (7.27) to obtain $\Gamma_A^{eq,90}$ for different values of the DM mass. From there I can directly translate those values into the upper limits for DUNE on the DM scattering cross sections, for a given exposure. The relation between the annihilation rate and the DM-nucleon cross section comes from the equilibrium condition through the solar DM capture rate. The details of the evolution of the number of DM particles inside the Sun and the computation of the capture rates are discussed in App. 7.2.

7.5 Example: Kaluza-Klein Dark Matter

Even though there are plenty of BSM theories which provide viable dark matter candidates, Kaluza-Klein type of models [75, 76] within the universal extra dimensions (UED) paradigm naturally predict the existence of a massive, stable particle that can play the role of the dark matter. In the UED scenario all the SM fields can propagate in one or more compact extra dimensions [77], as opposed to the idea of brane worlds [78, 79], where just gravity can propagate in the bulk while SM particles live at fixed points.

Furthermore, in UED there is no violation of the translational invariance along the extra dimensions, thus leading to degenerate KK modes masses and also the conservation of the KK number in the effective four dimensional theory. At loop level, radiative corrections and boundary terms shift the masses of the KK modes and break KK number conservation into a KK parity. As a result, this theory only contains interactions between an even number of odd KK modes and therefore the lightest among the first KK excitations will be stable. This particle is usually denoted as the lightest Kaluza-Klein particle (LKP) and its mass is proportional to $1/R$, being R the size of the extra dimension.

A viable DM candidate needs to be electrically neutral and non-baryonic, therefore good candidates among the first Kaluza-Klein excitations would be the KK neutral gauge bosons and the KK neutrinos [80]. Another possible candidate is the first KK excitation of the graviton, which receives negligible radiate contributions and therefore has a mass almost equal to $1/R$, but it has been shown that the lightest eigenstate from

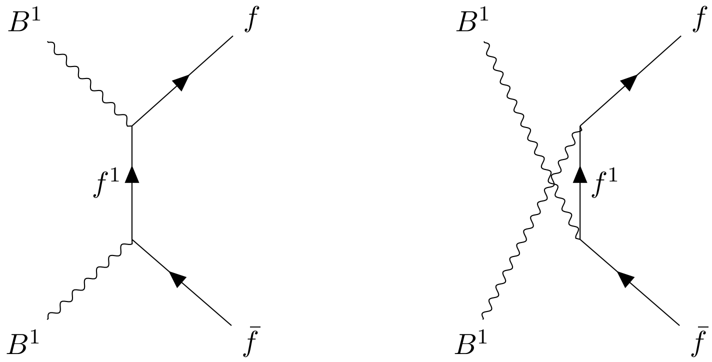


Figure 7.5: Feynman diagrams for B^1B^1 annihilation into SM fermions.

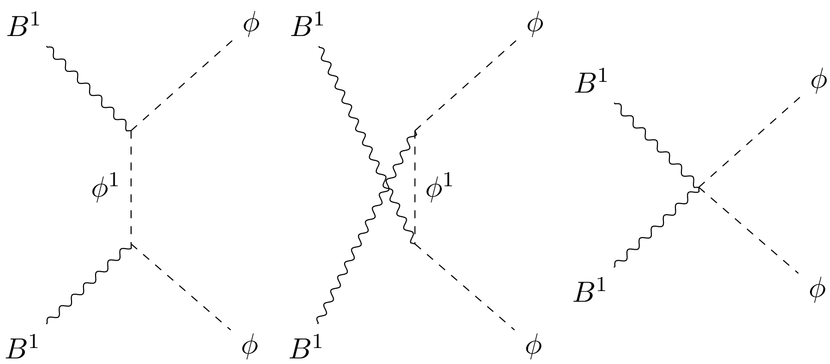


Figure 7.6: Feynman diagrams for B^1B^1 annihilation into a Higgs boson pair.

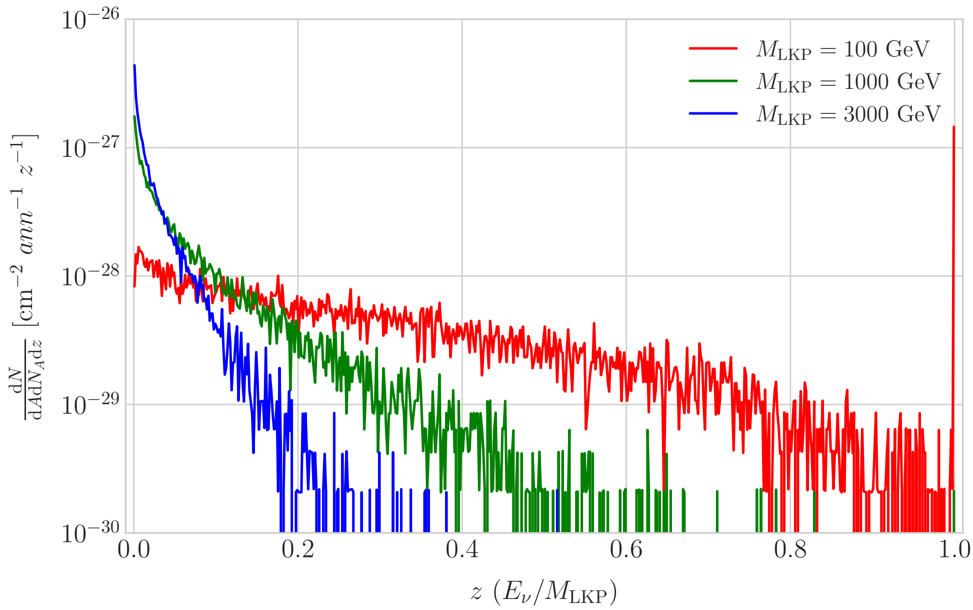


Figure 7.7: *Computed spectra of muon neutrinos at the DUNE FD site from B^1 annihilations in the Sun for three different values of $M_{\rm LKP}$, plotted in relative energy units for legibility.*

the mixing of the gauge mass states $\left(B^1, W_3^1\right)$ would be lighter, as B^1 and W_3^1 receive negative radiate corrections [81]. It is also understood that, when these corrections become sizeable, the eigenstates become approximately pure B^1 and W_3^1 states as the Weinberg mixing angle grows small with the KK number [81]. In that case, the LKP can be well-approximated as being entirely B^1.

I need to compute the neutrino flux produced by the annihilations of the LKP in the core of the Sun, taking into account their propagation in the solar medium, as well as neutrino oscillations. To this end I used `WimpSim` [82, 83] to generate one million annihilation events in the Sun over a time span of four years and propagate them to the DUNE FD location (44° 20' N, 103° 45' W), for different values of $M_{\rm LKP}$. In Fig. 7.7 I show the obtained muon neutrino spectra arriving to the detector from LKP annihilations in the Sun, per unit area and per annihilation, plotted in relative energy units for different values of the mass. As one could expect the spectra get steeper the higher is the mass, due to the absorption of high-energy neutrinos in the solar medium. Also, one can see the peak at $z = 1$ due to the direct annihilation into neutrinos $\chi\chi \to \nu\bar{\nu}$.

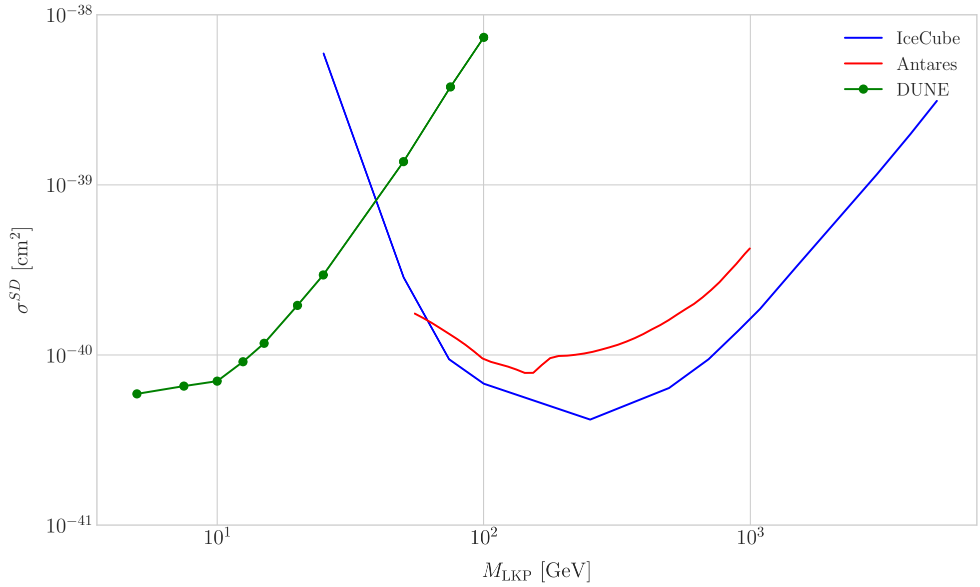


Figure 7.8: *Projected 90% confidence level upper limit for DUNE (400 kT yr) on the spin- dependent B^1-proton scattering cross section as a function of M_{LKP} (green dots). I also show the previous limits from IceCube [6] (blue line) and Antares [7] (red line) on the LKP cross section. The shaded area represents the disfavoured region (at 95% confidence level) on the mass of the LKP from LHC data [8].*

Now, one can estimate the sensitivity of DUNE to this particular model by using the methods I previously discussed. To begin with, I will use the optimistic estimation of the background efficiency in Eq. (7.24) to get our upper bound. Using it, one can directly compute the number of expected background events to be $N_B = 0.1101$ for an exposure of 400 kT yr. Then, Eq. (7.26) give us a value of $N_S^{90} = 2.20$ for the 90% exclusion number of expected signal events. By using the `NuWro` generated cross sections and the computed neutrino fluxes from B^1 annihilations in the Sun I can estimate the limits on the SD and SI DM-nucleus cross section using the relation in Eq. (7.2) and the capture rates I computed with `DarkSUSY`.

In Fig. 7.8 I show the projected sensitive for DUNE on the spin-dependent B^1-proton scattering cross section versus the mass of the DM particle, for a exposure of 400 kT yr (green dots). I also include the previous results from IceCube [6] (blue line) and Antares [7] (red line). The shaded area represents the disfavoured region from combined searches for UED by ATLAS and CMS [8].

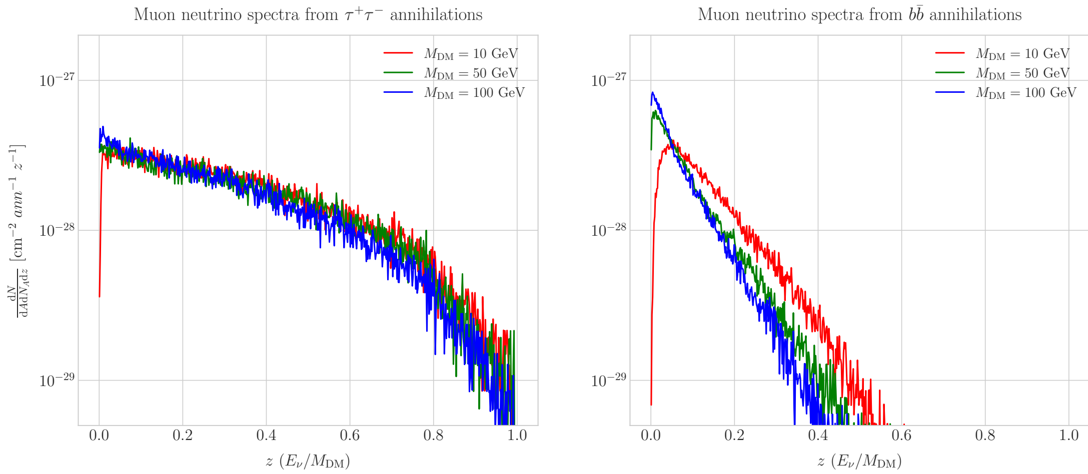


Figure 7.9: *Computed spectra of muon neutrinos at the DUNE FD site from* $\tau^+\tau^-$ *(left panel) and* $b\bar{b}$ *(right panel) annihilations in the Sun for the DM masses* $m_{\rm DM} = 10$ GeV *(red line),* 50 GeV *(green line) and* 100 GeV *(blue line), plotted in relative energy units.*

From the experimental point of view, this estimation lacked a detailed simulation of the detector response and thus this must be consider as a mere optimistic sensitivity computation. However, it shows the potential of DUNE to constrain this kind of exotic scenarios, showing the region where it will be in a position to compete with other neutrino telescopes. A more detailed analysis is needed if I am to make a realistic estimation. Even though the region of the parameter space where DUNE would be sensitive to this particular model is quite constrained by collider searches [8] and other rare decay measurements [84, 85], it still constitutes an alternative indirect probe.

7.6 High energy DM neutrino signals

To have better estimates on the capability of the DUNE FD to constrain the parameter space of DM using solar neutrino fluxes, I need to start accounting for the detector resolution effects and the topologies of the different signatures. As a starting point, I will focus on specific annihilation channels. For the case of DUNE, the relevant ones are mainly the hard channels $\tau^+\tau^-$ and $\nu\bar{\nu}$ and the soft channel $b\bar{b}$. These are the open annihilation channels for relatively low mass WIMPs that will actually give neutrino fluxes. Other channels, like W^+W^- and ZZ, are open for more massive WIMPs, but those will produce usually a higher energy neutrino flux that will be out of reach for DUNE (usually the maximum neutrino energy is taken to be $E_{max} = 10$ GeV).

In Fig. 7.9 I show the `WimpSim` [82, 83] generated muon neutrino spectra at the DUNE FD location (44° 20' N, 103° 45' W) from $\tau^+\tau^-$ (left panel) and $b\bar{b}$ (right panel) annihilations in the core of the Sun, for different DM masses. Here, one can clearly see the meaning of the previous distinction between hard and soft channels. For the same DM mass value, the muon neutrino spectrum from the $\tau^+\tau^-$ channel is more flat and reaches higher energies than the one from the $b\bar{b}$ channel, which drops faster.

In this case, I prepared two sets of files, one for $\tau^+\tau^-$ and the other for $b\bar{b}$, for DM masses in the range from 5 to 100 GeV (actually for $b\bar{b}$ the first mass point I took is 7.5 GeV, as a WIMP with $m_{\mathrm{DM}} = 5$ GeV can not kinematically self annihilate into $b\bar{b}$). Then, I prepared the `WimpSim` output fluxes in a specific way to use them as inputs to `NuWro`, which simulates the neutrino interaction with the argon.

Because `WimpSim` outputs an event list together with the fluxes, I can use the former to generate the events. The direction of these is given in terms of the azimuth and altitude angles viewed from the specified location, so first I need to convert these into the DUNE FD coordinates. Once I have done it, each event can be processed with `NuWro`. To increase the number of samples and optimise the computation time, I generate 100 interactions (i.e. `NuWro` events) for each `WimpSim` event[2]. I restrict the event generation to charged current interactions, but I allow all the different contributions to the CC cross section, i.e. quasielastic scattering (QEL), meson exchange current process (MEC), resonant pion production (RES) and deep inelastic scattering (DIS). I just take into account the CC contribution because I am only interested in final states with charged leptons, as we have better chances of reconstructing the kinematics of CC events.

For the atmospheric fluxes I follow a similar procedure, only that this time I do not have a set of events but the fluxes binned in azimuth and altitude angles. This way, I transform these to DUNE coordinates and process the fluxes for each bin separated with

[2] This also solves a problem related with the generation of the neutrino interactions in `NuWro`, as if you only produce one event each time you launch `NuWro` it will always produce an interaction of the dominant interaction type for that particular energy.

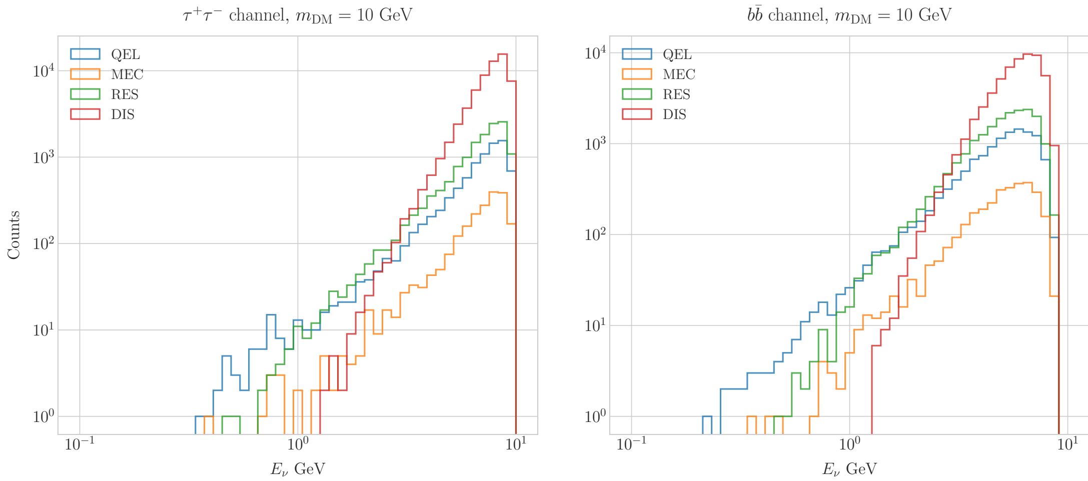


Figure 7.10: *Distribution of the muon neutrino energies from the $\tau^+\tau^-$ (left panel) and $b\bar{b}$ (right panel) annihilation channels, for $m_{\rm DM} = 10$ GeV, separated by CC interaction type: QEL (blue), MEC (orange), RES (green) and DIS (red).*

`NuWro`.

At this point, I have two sets of events with different energies and final states. In Fig. 7.10 one can see the distribution of the muon neutrino energies for the case $m_{\rm DM} = 10$ GeV, both for the $\tau^+\tau^-$ (left panel) and $b\bar{b}$ (right panel) channels, separated by interaction. One can clearly see that there are different energy regimes where the primary interaction type is different. This leads to a plurality of event topologies, therefore making it difficult to implement a general approach to the selection of events in detriment of the background. As a way to proceed, I decided to split our samples, based on the different interaction modes and contents of the final state, into a CC DIS sample and a single proton CC QEL sample.

7.6.1 DIS events

To begin with, I consider the high energy part of the spectrum. In this region DIS events dominate, i.e. interactions of the form $\nu_\mu + q_d(\bar{q}_u) \to \mu^- + q_u(\bar{q}_d)$. Therefore, our final estates will contain a muon and a hadronic jet from the fragmentation of the outgoing quark. As all these events have $E_\nu \gtrsim 1$ GeV the momentum transfer to the remnant nucleus is negligible, for this reason the neutrino energy can be effectively reconstructed

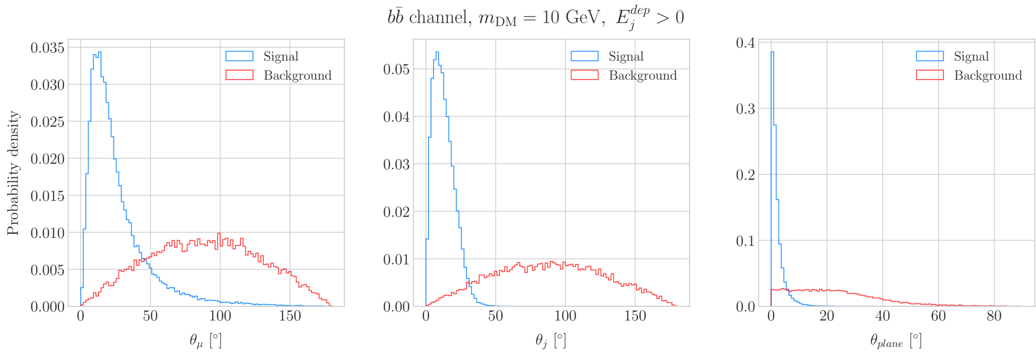


Figure 7.11: *Distributions of θ_μ (left panel), θ_j (central panel) and θ_{plane} (right panel) for the $b\bar{b}$ sample with $m_{\rm DM} = 10$ GeV (blue) and the atmospheric background (red).*

just taking into account the momenta of the muon and the jet. This technique was successfully used in Ref. [86] to select monoenergetic DM solar neutrino events from $\nu\bar{\nu}$ annihilation channels.

Using momentum conservation one sees that the plane generated by the momenta of the muon and the jet needs to also contain the momentum of the neutrino. As we are interested in neutrinos coming from the Sun, the momentum of the neutrino can be regarded as known beforehand. This will allow us to define the angle of the outgoing muon and jet with respect to the incoming neutrino. Moreover, one can also use that information to reject poorly reconstructed jets, checking for deviations of these from the momentum conservation plane.

To account for the limited angular resolution of the detector, I smeared the momenta of the muons and hadrons. In a liquid argon TPC muons are expected to be tracked with high precision, therefore I take the associated angular resolution to be 1°. In the case of jets, it is expected that for the hadrons dominating the cascade a detector like DUNE has an angular resolution between 1° to 5° [42], so I take the latter, more conservative, estimate.

As a first selection step, I will just take into account particles with kinetic energies above the detection threshold of DUNE. For muons and photons the specified threshold energy is 30 MeV, for charged pions 100 MeV and for other hadrons 50 MeV [42]. This way, if the outgoing muon in a certain event has an energy lower than the required

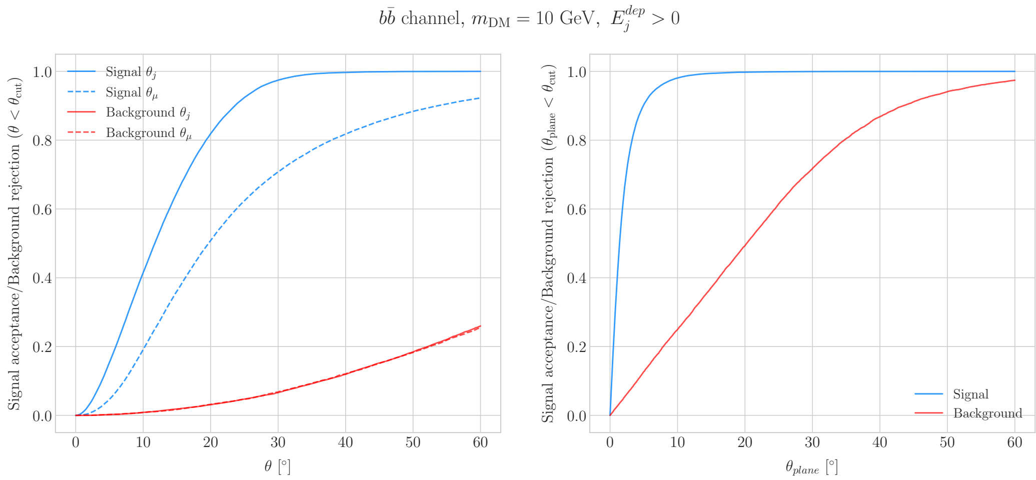


Figure 7.12: *Left panel: signal efficiencies (blue lines) and background rejections (red lines) for events passing the cuts $\theta < \theta_{cut}$ for the jet (solid lines) and muon (dashed lines) angles. Right panel: signal efficiency (blue line) and background rejection (red line) for events passing the cut $\theta_{plane} < \theta_{cut}$ for the momentum conservation plane deviation.*

threshold I will drop such event. For the case of hadrons and photons, I will only require to have at least one particle above the energy threshold, so then one can compute the jet momentum using the (smeared) momenta of the N particles above threshold as:

$$\vec{p}_j = \sum_{i=1}^{N} \vec{p}_i. \tag{7.28}$$

Additionally, I will also define an estimation of the deposited hadronic energy as:

$$E_j^{dep} = m_{^{39}\mathrm{Ar}} - m_{^{40}\mathrm{Ar}} + \sum_{i=1}^{N} \sqrt{|\vec{p}_i|^2 + m_i^2}. \tag{7.29}$$

This quantity is useful to select events with enough hadronic visible energy in the detector. For events where most of the hadronic energy is scattered across plenty of hadrons with individual energies below the detection threshold, this estimation will give $E_j^{dep} \leq 0$. In these cases it could be expected that the jet momentum is poorly reconstructed, and therefore I require events to pass the cut $E_j^{dep} > 0$.

For the events I can compute the angles for the muon and jet with respect to the incoming neutrino as:

$$\cos\,\theta_\mu = \hat{p}_\nu \cdot \hat{p}_\mu, \tag{7.30}$$

$$\cos\,\theta_j = \hat{p}_\nu \cdot \hat{p}_j, \tag{7.31}$$

and the deviation from the momentum conservation plane as:

$$\sin\,\theta_{plane} = \left| \frac{\hat{p}_\mu \times \hat{p}_\nu}{|\hat{p}_\mu \times \hat{p}_\nu|} \cdot \hat{p}_j \right|. \tag{7.32}$$

In Fig. 7.11 I show some distributions of these quantities for the case of the $b\bar{b}$ sample with $m_{\rm DM} = 10$ GeV (blue histograms) and for the atmospheric backgrounds (red). In order to select the atmospheric events I followed the same criteria as for the signal events. However, because in the signal case I used the true direction of the neutrino as input, as it should be that of the Sun at that time and therefore known, in the atmospheric case I used a set of solar positions as our ansatz for the neutrino direction. From the distributions, one can see that the muon and the jet for the signal events are predominantly forward and also that the deviations from the momentum conservation plane are peaked at zero, as one should expect.

Now, I can start applying cuts to maximise our signal selection efficiency while at the same time I try to minimise the amount of atmospheric background events passing the selection. To this end, I will need at to find some lower and upper cuts for θ_j and θ_μ and an upper bound for θ_{plane}. In Fig. 7.12 I show how upper bound cuts in the different angular variables affect the signal efficiency (blue lines) and the background rejection (red lines). Notice that the signal efficiency behaves in a quite different way when I apply cuts in the jet and the muon angles. On the contrary, the cuts on both variables have a similar effect on the background rejection.

In order to obtain the optimal set of cuts, I perform a multidimensional scan. I do this separately for the $\tau^+\tau^-$ and the $b\bar{b}$ samples. For each case, I scan the possible cuts for each mass point and then I take the mean value of the signal efficiency for

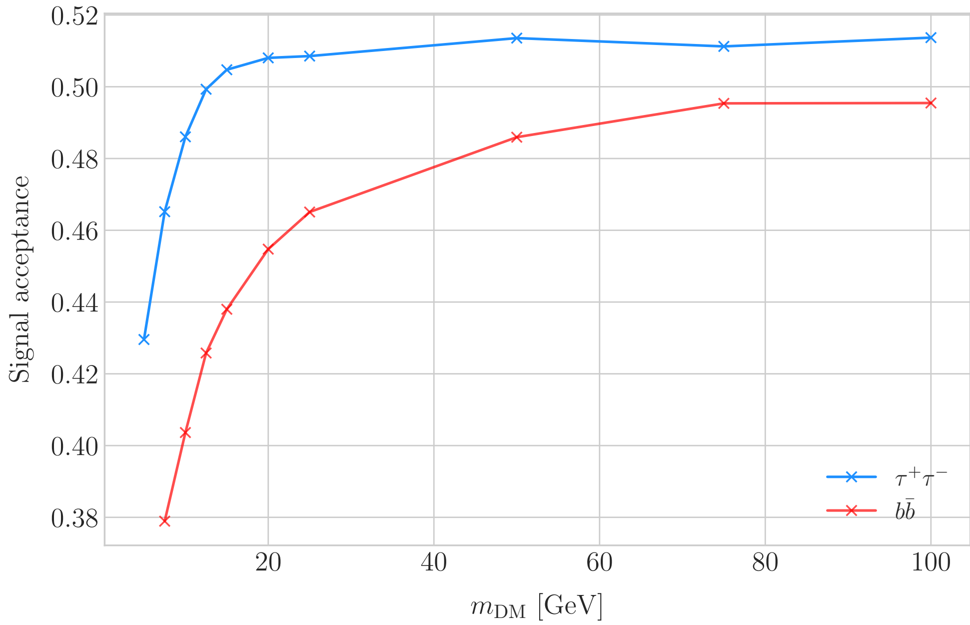


Figure 7.13: *Signal efficiencies for the $\tau^+\tau^-$ (blue line) and $b\bar{b}$ (red line) DIS samples as functions of the DM mass, m_{DM}, obtained by applying the optimal angular cuts $\theta_\mu < 27°$, $4° < \theta_j < 26°$ and $\theta_{plane} < 3.5°$.*

each configuration, to get the mean efficiency for each set of cuts. I do a similar scan for the atmospheric sample independently. Then, I take the sets of cuts such that the background rejection achieved is greater than 99.8% and search for the one which maximises the $\tau^+\tau^-$ and $b\bar{b}$ sample mean efficiencies. I found that with the cuts $\theta_\mu < 27°$, $4° < \theta_j < 26°$ and $\theta_{plane} < 3.5°$ I get a background rejection of 99.80% while achieving a 49.40% and 44.92% mean signal efficiencies for the $\tau^+\tau^-$ and $b\bar{b}$ signals respectively.

In Fig. 7.13 I show the signal efficiencies as a function of the DM mass for the $\tau^+\tau^-$ (blue line) and the $b\bar{b}$ (red line) DIS events, after applying the cuts discussed above, as well as the energy threshold and hadronic visible energy selections. One can see that the efficiency grows with the mass, as annihilations of more massive DM particles will produce a neutrino spectrum centered at higher energies, where DIS events dominate. Notice also that the efficiency is higher for the $\tau^+\tau^-$ case at every mass point, as in general this channel produces neutrinos at higher energies than the corresponding $b\bar{b}$ channel.

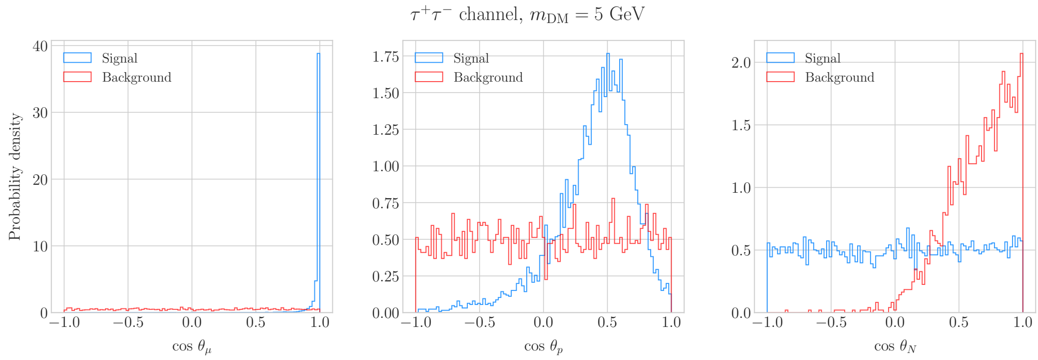


Figure 7.14: *Distributions of* $\cos\theta_\mu$ *(left panel),* $\cos\theta_p$ *(central panel) and* $\cos\theta_N$ *(right panel) for the* $\tau^+\tau^-$ *QEL sample with* $m_{\rm DM} = 5$ GeV *(blue) and the atmospheric background (red).*

7.6.2 Single proton QEL events

Now, one can try to explore the low energy tail of the neutrino energy distributions. This regime is dominated by the QEL interactions, i.e. events of the type $\nu_\mu + n \to \mu^- + p$. In this case, as the typical energies are $E_\nu \lesssim 1$ GeV, the momentum transfer to the remnant nucleus is sizeable. Therefore, I can not make the approximation I did before and assume that the momentum of the muon and the proton will give an adequate estimation of the reconstructed neutrino energy.

In any case, as before, I can take the direction of the incoming neutrino as known. That way, one can estimate the energy of the neutrino as:

$$E_\nu^{reco} = E_\mu + E_p + m_{^{39}\mathrm{Ar}} - m_{^{40}\mathrm{Ar}}, \tag{7.33}$$

and using momentum conservation I can write the momentum of the remnant nucleus as:

$$\vec{p}_N = \hat{p}_\nu \left(E_\mu + E_p + m_{^{39}\mathrm{Ar}} - m_{^{40}\mathrm{Ar}}\right) - \vec{p}_\mu - \vec{p}_p. \tag{7.34}$$

As in the previous case, I need to drop the events where the muon or the proton fall below the kinetic energy detection threshold [42]. Also, I again apply a smearing to the momenta of the particles, a 1% for muons and 5% for protons.

Having done that, one can compute the following angular variables for our selected

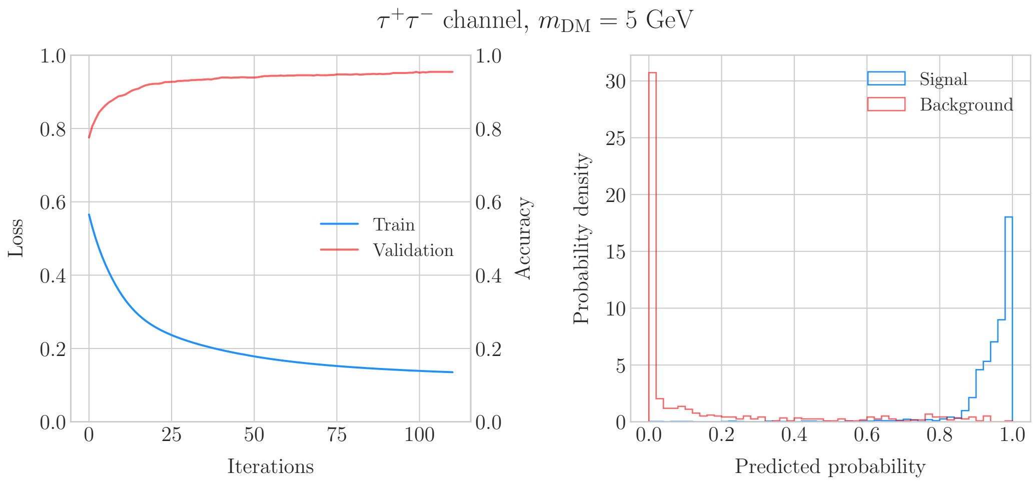


Figure 7.15: *Left panel: value of the loss function for the training sample (blue line) and accuracy for the validation sample (red line) versus the number of iterations for the MLP classifier training. Right panel: distributions of the predicted probabilities assigned by the MLP classifier to the test sample for the $\tau^+\tau^-$ QEL signal with $m_{\rm DM} = 5$ GeV (blue) and the atmospheric background (red).*

events:

$$\cos\,\theta_\mu = \hat{p}_\nu \cdot \hat{p}_\mu, \tag{7.35}$$

$$\cos\,\theta_p = \hat{p}_\nu \cdot \hat{p}_p, \tag{7.36}$$

$$\cos\,\theta_N = \hat{p}_\nu \cdot \hat{p}_N. \tag{7.37}$$

Fig. 7.14 shows the distributions of these angular variables for the $\tau^+\tau^-$ QEL sample with $m_{\rm DM} = 5$ GeV (blue) and the atmospheric background (red). Again, for the atmospheric events I used a random solar position as the ansatz for the incoming neutrino direction. Notice that now, opposed to the DIS case where the signal had very sharp distributions for the variables considered, the shapes of the angular distributions for signal and background are not that much different.

This effectively means that the usual approach of applying simple angular cuts would not work as well as in the previous situation. Therefore, as a possible solution, I tried to use a multilayer perceptron (MLP) classifier to separate between signal and background events. Thus, the power of the hypothesis test will serve as an estimate of the signal

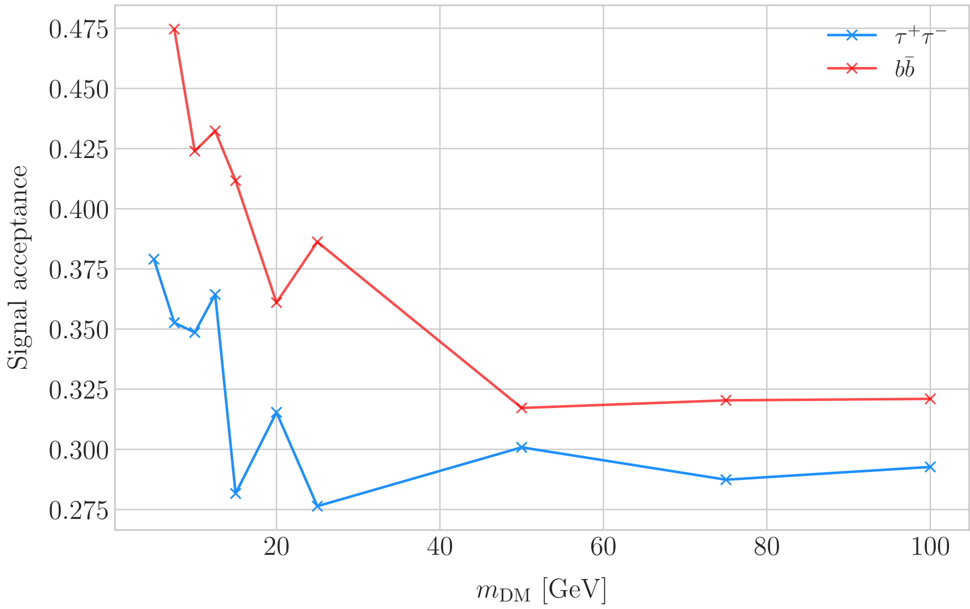


Figure 7.16: *Signal efficiencies for the $\tau^+\tau^-$ (blue line) and $b\bar{b}$ (red line) single proton QEL samples as functions of the DM mass, m_{DM}, obtained by requiring a minimum predicted probability from the MLP classifier of* 0.97 *in order to achieve a background rejection greater than* 99.8%.

efficiency, and in the same way one can take the size of the test to be our background rejection.

For each DM mass value and channel, as well as for the background sample, I divide our events into training, validation and test samples. The input variables for the classifier were the reconstructed neutrino energy from Eq. (7.33) and the angular variables defined in Eqs. (7.35 - 7.37). I used the MLP classifier implemented in `scikit-learn` [87], with a total of five hidden layers, the rectified linear unit activation function and adaptive learning rate. In order to account for fluctuations due to artifacts in the training process I repeated the training a thousand times for each sample, redefining each time the training, validation and test subsets, so one can take as our signal efficiency and background rejection the mean values of the powers and sizes of the tests.

The results of one of these training processes for the $\tau^+\tau^-$ QEL signal with $m_{\mathrm{DM}} =$ 5 GeV is shown in Fig. 7.15. On the left panel I show the loss function values (blue) and accuracy (red) at each iteration for the training and the validation samples respectively.

The training stops either when the maximum number of iterations is reached (1000 in this case) or when the accuracy for the validation sample reaches a certain tolerance (I chose 10^{-4} as our tolerance). On the right panel I have the distributions for the predicted probability by the model, separated in true signal (blue) and background (red) events, for the test sample. One can see that both populations are well separated, obtaining a power of 44.97% and a size of 0.17% when I require a predicted probability greater than 0.97.

Applying this criteria for each sample, I obtain the mean signal efficiencies shown in Fig. 7.16. Notice that the efficiencies for the channel $\tau^+\tau^-$ (blue line) are consistently lower than the ones for the $b\bar{b}$ channel (red line). This can be due to the fact that, for each DM mass point, the neutrino spectrum coming from the $b\bar{b}$ annihilation channel is centered at lower energies when compared to the $\tau^+\tau^-$ spectrum. This directly translates into more low energy neutrinos undergoing QEL interactions, which give signals that can be easily separated from the atmospheric background. This explanation also help us understand why in both cases the signal acceptance drops when the DM mass increases. In all cases, the background rejection took values between 99.8% to 99.9%. I will assume a 99.8% background rejection value in all cases to keep our estimation conservative.

7.6.3 Results

In order to estimate the DM-nucleon cross section sensitivities in the present case I need again to compute the expected number of background events. As I am now separating events by interaction type Eq. (7.25) does not hold anymore, as in that case I integrated over the total neutrino-argon cross section. In this instance, the expected background events for DIS events is approximately given by:

$$N_B^{DIS} \simeq \eta_B^{DIS} \times \left(4.655 \times 10^3\right) \times \left(\frac{\text{exposure}}{400 \text{ kT yr}}\right), \tag{7.38}$$

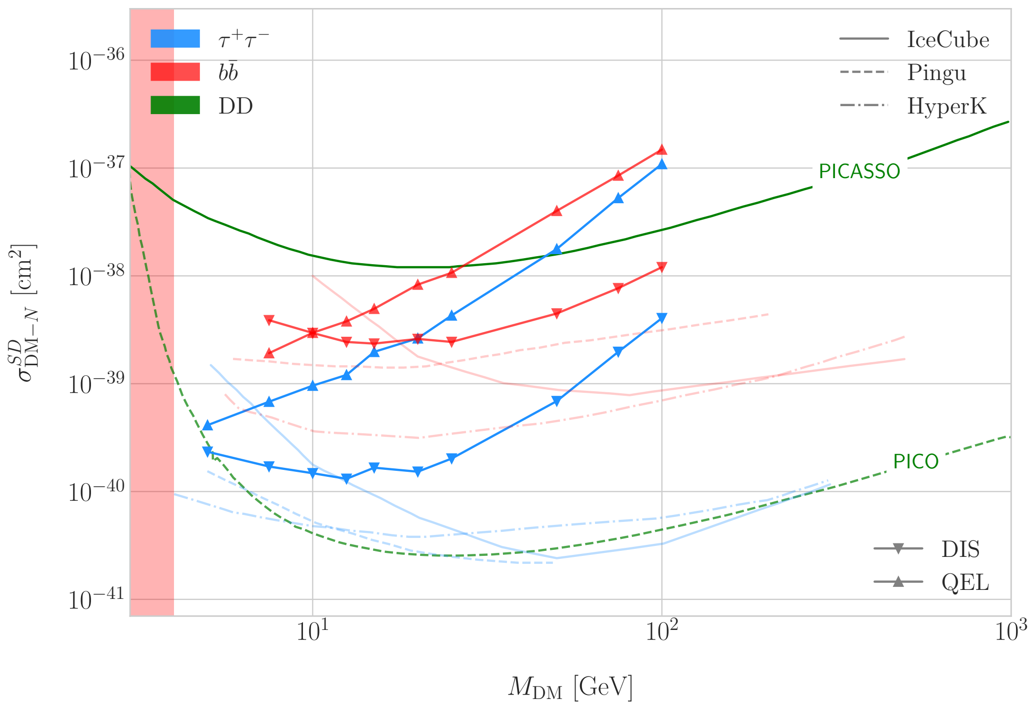


Figure 7.17: *Projected 90% confidence level upper limit for DUNE (400 kT yr) on the spin-dependent DM-nucleon scattering cross section as a function of* m_{DM}*, for the annihilation channels* $\tau^+\tau^-$ *(blue) and* $b\bar{b}$ *(red) separated by interaction type (up triangles denote DIS interactions whereas down triangles represent QEL interactions). I also show the previous limits from IceCube [9] (solid lines) and the projected sensitivities for Pingu [10] (dashed lines) and Hyper-Kamiokande [11] (dash-dotted lines), as well as the direct detection limits from PICASSO [12] (solid green line) and PICO-60* C_3F_8 *[13] (dashed green line).*

whereas for QEL events we have:

$$N_B^{QEL} \simeq \eta_B^{QEL} \times \left(2.248 \times 10^4\right) \times \left(\frac{\text{exposure}}{400 \text{ kT yr}}\right). \tag{7.39}$$

Now, using these together with Eqs. (7.26) and (7.27) one can obtain the 90% C.L. upper limit on the total annihilation rate at equilibrium for both kind of events. Then, applying the computed DM-nucleons capture rates I can translate these into limits on the DM-nucleon cross section by means of Eqs. (7.2), (7.5) and (7.6).

Fig. 7.17 shows the obtained limits on the SD DM-nucleon cross section for DUNE, using the DIS (up triangles) and QEL (down triangles) events both for the $\tau^+\tau^-$ (blue) and the $b\bar{b}$ (red) samples, for an exposure of 400 kT yr. I also include the corresponding current limits from IceCube [9] (solid lines), as well as the projected sensitivities of Pingu [10] (dashed lines) and Hyper-Kamiokande [11] (dash-dotted lines). For comparison, I also show the reported direct detection limits from PICASSO [12] (solid green line) and PICO-60 C_3F_8 [13] (dashed green line).

Notice that, for most of the mass range, the limits one can set by using the DIS events are stronger than those of the QEL interactions, except for the low mass part of both the $\tau^+\tau^-$ and the $b\bar{b}$ curves where the QEL events dominate. In general, the expected sensitivity of DUNE for DM masses $\lesssim 25$ GeV surpasses the stronger current indirect limits. However, experiments like Hyper-Kamiokande are foreseen to have an overall better sensitivity in this kind of searches, as they have a bigger active volume and accept a broader energy range.

A pending question is what happens when we add the RES and MEC charged-current interaction contributions. In that case it would probably be more convenient to split the samples by final state interaction topologies. Also, another necessary improvement would be adding a full detector simulation and reconstructions. This will also require considering the effect of poorly reconstructed events or final states containing neutral particles such that they mimic the desired topology at the reconstruction level.

7.7 Example: Leptophilic Dark Matter

In general, the capture rate of DM particles by the Sun via interactions with electrons is several orders of magnitude smaller than the capture via DM-nucleus scattering. Thus, it would be sub-leading even when nucleon capture is loop suppressed. As I showed in Fig. 7.2, the capture rate via scattering off electrons only surpasses the capture rates via DM-nucleons interactions for DM masses $\lesssim 100 - 500$ MeV.

However, if one considers a model where DM-nucleon interactions are forbidden even at loop level, then electron interactions will be the sole contributor to DM capture in the Sun. One can describe such scenario where the DM particles couple to leptons but not to the quark sector using effective operators.

In general, assuming that the DM particle is a Dirac fermion, the dimension six operators describing the interaction between two DM particles and two leptons can be written as:

$$\mathcal{L}_{eff} = G \sum_i \left(\bar{\chi} \Gamma^i_\chi \chi \right) \left(\bar{\ell} \Gamma^i_\ell \ell \right) , \tag{7.40}$$

where $G = 1/\Lambda^2$ is the effective coupling strength, Λ the cut-off of the effective field theory and ℓ denotes any lepton. In principle, one should consider all the possible Lorentz structures Γ^i_f in order to have a complete set of effective operators.

However, some combinations will induce interactions with nucleons at loop level. As we are specifically interested in interactions which forbid any communication with the quark sector, I will not consider those [88]. In addition, some of the effective operators give rise to velocity-suppressed scattering cross sections between DM particles and leptons. I will also neglect those, as the suppression goes with the square of the DM halo velocity which in units of the speed of light is $\sim 10^{-6}$.

This way, the only Lorentz tensor structure that do not induce interactions with quarks at loop level and gives a contribution to the scattering cross section that is not velocity suppress is the axial-axial interaction. The effective Lagrangian is then given by:

$$\mathcal{L}_{eff} = \frac{c^\chi_A c^\ell_A}{\Lambda^2} \left(\bar{\chi} \gamma^\mu \gamma^5 \chi \right) \left(\bar{\ell} \gamma_\mu \gamma^5 \ell \right) , \tag{7.41}$$

where c^χ_A and c^ℓ_A are the couplings for the different species. As the DM coupling appears as a common factor for any lepton choice, I will redefine the corresponding coupling c^ℓ_A to absorb c^χ_A. Also, for simplicity, I will assume that the couplings between the DM particles and the leptons are flavour independent, i.e. I have just two couplings, c^e_A for charged leptons and c^ν_A for neutrinos.

In the case of a scalar DM particle, the lowest order effective interaction with leptons happens through a dimension five operator, generating scalar and pseudoscalar interactions. However, the former induces interactions with quarks at two loop level whereas the latter gives a velocity suppressed scattering cross section.

From the effective Lagrangian in Eq. (7.41) it can be shown that the axial-axial contribution to the scattering cross section for the fermionic DM and a charged lepton is given by:

$$\sigma^{AA}_{\mathrm{DM}-e} = 3\,(c^e_A)^2\,\frac{m_e^2}{\pi\Lambda^4}. \tag{7.42}$$

If the DM interacts exclusively with fermions, then the only annihilation channels that will give us a measurable neutrino flux coming out of the Sun are $\tau^+\tau^-$ and $\nu\bar{\nu}$. The former channel, already explored previously in the more mainstream scenario of the DM capture via scattering off nucleons, is open only for $m_{\mathrm{DM}} > m_\tau \simeq 1776.86 \pm 0.12$ MeV [89], a mass region where the solar DM capture by electrons is at least one order of magnitude smaller than the capture via interactions with nucleons. On the contrary, the latter allows us to explore a region where the capture rate via scattering off electrons dominates over the rest.

One downside of focusing in such low mass range is that it falls bellow the usual limit of $m_{\mathrm{evap}} \sim 4$ GeV usually explored in the literature. The pretext to explore this region is the result discussed previously reported in Ref. [4], where DM evaporation in the Sun for the case of capture via electron scattering could be negligible for masses as low as $m_{\mathrm{evap}} \sim 200$ MeV. This result is quite sensitive to the high velocity tail of the DM velocity distribution in equilibrium inside the Sun, and therefore full numerical simulations would be needed to asses the impact of this effect. However, this falls out of the scope of our work.

In this case, as I have an specific realisation of the interaction between the DM and leptons, one can estimate the relic density of our DM for different values of the couplings and the effective field theory scale Λ. The first step to do so is compute the self-annihilation cross section. Because I consider cold relics, at the freeze-out time our DM particles were non-relativistic and so one can expand the annihilation cross section in terms of the relative velocity v between two annihilating DM particles as [90]:

$$\sigma^{AA}_{ann}|v| \approx \frac{1}{2\pi\Lambda^4}\sum_{\ell}\left(c^{\ell}_A\right)^2 m_\chi^2\sqrt{1-\frac{m_\ell^2}{m_\chi^2}}\left[\frac{m_\ell^2}{m_\chi^2}+\frac{1}{12}\left(2-\frac{m_\ell^2}{m_\chi^2}\right)v^2\right], \tag{7.43}$$

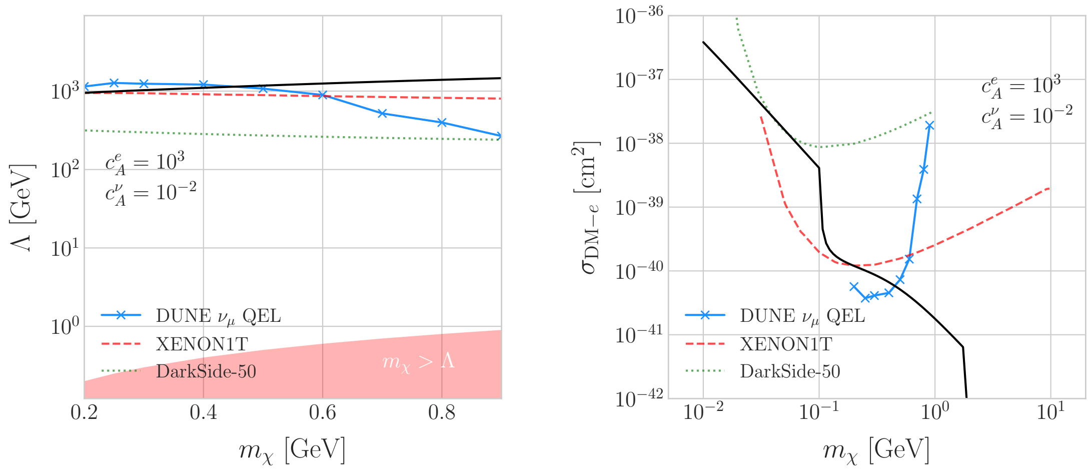


Figure 7.18: *Left panel: Projected 90% confidence level sensitivity of DUNE (400 kT yr) to the scale Λ of an EFT containing only leptophilic DM axial-axial interactions (blue line). Right panel: . In both cases the corresponding limits from DarkSide-50 [14] (dotted green line) and XENON1T [15] (dashed red line) are also shown, together with the configurations for which the correct relic density is achieved (black line), all for the coupling values $c_A^e = 10^3$ and $c_A^\nu = 10^{-2}$.*

where the sum includes all the possible lepton final states with mass m_ℓ.

Solving the Boltzmann equation for the evolution of the DM density gives as a solution a relic density of:

$$\Omega_\chi h^2 \approx \frac{(1.04 \times 10^9) x_F}{M_{Pl}\sqrt{g_*}(a + 3b/x_F)}, \tag{7.44}$$

where $x_F = m_\chi/T_F$ being T_F the freeze-out temperature, g_* the number of relativistic degrees of freedom at freeze-out and a and b the terms in the annihilation cross section expansion $\sigma_{ann}|v| \approx a + bv^2 + \mathcal{O}(v^4)$. Using the current best fit for the relic DM density $\Omega_\chi h^2 = 0.1198 \pm 0.0012$ [91] one can use these relations to compute the required effective theory scale Λ at which the correct density is achieved for any combinations of m_χ and c_A^ℓ.

As discussed before, in the low DM mass region QEL interactions dominate. Moreover, if I focus on direct annihilation to neutrinos, the energy of the muon neutrino flux is known as it must be equal to the mass of the DM particle, $E_\nu = m_\chi$. That way, now I do not need to use Eq. (7.33) in order to estimate the momentum transfer to the remnant nucleus, I can simply take:

$$\vec{p}_N = \hat{p}_\nu m_\chi - \vec{p}_\mu - \vec{p}_p. \tag{7.45}$$

To estimate the signal efficiency and background rejection for this case I used again the MLP classifier from `scikit-learn`, using the same specifications as before. The only difference now is that I add also the reconstructed neutrino energy as one of the features to train the classifier with, because the characteristic monoenergetic flux for each m_χ value will help to distinguish between signal and background events.

In this case, for masses below ~ 500 MeV I obtain a signal efficiency close to unity while keeping a background rejection of 99.9%. For bigger values of the mass, the signal efficiency drops significantly if I require to keep the background acceptance under 0.01%. However, because this kind of search is dominated by the background, sacrificing the signal acceptance to keep the background rejection to a minimum enhances the reach of the analysis. This way, for DM masses of the order of $m_\chi \sim 1$ GeV I end up with efficiencies as low as 1%.

Now, estimating the number of background events using Eq. (7.39) one can go on and apply Eqs. (7.26) and (7.27) together with Eq. (7.42) to derive the sensitivity of DUNE to this kind of model. Fig. 7.18 (left panel) shows the potential reach of DUNE to constrain the EFT scale Λ this model containing only leptophilic DM axial-axial interactions (blue line), for a choice of couplings $c_A^e = 10^3$ and $c_A^\nu = 10^{-2}$. I also included the current limits on the DM-electron scattering cross section from DarkSide-50 [14] (dotted green line) and XENON1T [15] (dashed red line), reworked with Eq. (7.42) to show their implications for the EFT scale. The values of Λ for which the correct DM relic density value is achieved for each mass are also shown (black line). This tells us that, for that specific choice of couplings, DUNE would be sensitive to DM configurations allowed by the relic density constraint up to a mass of $m_\chi \sim 400$ MeV.

In Fig. 7.18 (right panel) I show the same upper limits but for the DM-electron scattering cross section. From this view one can see that DUNE would be able to offer complementary information to the low energy DM-electron interaction searches performed by direct detection experiments, in a slightly higher mass range.

With the present example, although it focuses on a very specific realisation of the DM interactions, I show the potential of DUNE to constrain exotic DM scenarios. Thanks to its low backgrounds and superb angular resolution DUNE will be able to help with the systematic searches for dark sectors physics.

Chapter 8

Particle ID in GArSoft

ND-GAr is a magnetised, high-pressure gaseous argon TPC (HPgTPC), surrounded by an electromagnetic calorimeter (ECal) and a muon detector (commonly refer to as μID). A detailed discussion on the requirements, design, performance and physics of ND-GAr can be found in the DUNE ND CDR [48] and the ND-GAr whitepaper (cite).

In DUNE Phase II ND-GAr will fulfill the role of TMS, measuring the momentum and sign of the charged particles exiting ND-LAr. Additionally, it will be able to measure neutrino interactions inside the HPgTPC, achieving lower energy thresholds than those of the ND and FD LArTPCs. By doing so ND-GAr will allow to constrain the relevant systematic uncertainties for the LBL analysis even further.

The goal of the present chapter is to review the requirements that the physics program of DUNE impose on ND-GAr, present the current status of its design and describe the GArSoft package, its simulation and reconstruction software.

As decided during the DUNE Phase II workshop in June 2023 [reference], we want to build ND-GAr physics case by showing:

- That ND-GAr can constrain systematic uncertainties that ND-LAr might miss.
- The impact on the neutrino oscillation results if such systematic uncertainties are missed.
- That ND-GAr is necessary to reach DUNE's main physics goals.

This way, the design of ND-GAr will be physics driven.

In order to study the effects of final state interactions (FSI) in CC interactions, ND-GAr should be able to measure the spectrum of protons and charged pions at low energies. ND-GAr also needs to be able to measure the pion multiplicity, specially for energies above 100 MeV as at these energies the pions shower in the LAr, to inform the pion mass correction in the ND and FD LArTPCs.

In order to correctly identify electrons, muons, pions, kaons and protons ND-GAr can use a combination of: $\mathrm{d}E/\mathrm{d}x$ measurements in the HPgTPC, E_{ECAL}/p using the ECAL total energy and the momentum obtained from magnetic spectroscopy in the HPgTPC and penetration information through the ECAL and muon tagger.

8.1 dE/dx measurement in the TPC

Among the parameters extracted from the track fitting, ionisation is particularly useful for particle identification, as it is a function of the particle velocity. Although for the case of relativistic particles this dependence is not very strong, measuring the track on a large number of points may allow us to estimate the amount of ionisation accuratel. This, paired with a measurement of the momentum, may allow us to identify the particle type.

The first calculation of the energy loss per unit length of relativistic particles using a quantum-mechanical treatment is due to Bethe [?]. Using this approach, the mean ionisation rate of a charged particle traveling through a material medium is (using natural units $G = \hbar = c = 1$):

$$\left\langle \frac{\mathrm{d}E}{\mathrm{d}x} \right\rangle = \frac{4\pi N e^4}{m_e \beta^2} z^2 \left(\log \frac{2 m_e \beta^2 \gamma^2}{I} - \beta^2 \right), \tag{8.1}$$

where N is the number density of electrons in the medium, e the elementary charge, m_e is the electron mass, z the charge of the particle in units of e, β is the velocity of the particle, $\gamma = (1 - \beta^2)^{-1}$ and I denotes the effective ionisation potential averaged over all electrons. This relation is known as the Bethe-Bloch formula.

From Eq. (8.1) one can see that the ionisation loss does not depend explicitly on the mass of the charged particle, that for non-relativistic velocities it falls as β^{-2}, then goes through a minimum and increases as the logarithm of γ. This behaviour at high velocities is commonly known as the relativistic rise. The physical origin of this effect is partly due to the fact that the transverse electromagnetic field of the particle is proportional to γ, therefore as it increases so does the cross section.

It was later understood that the relativistic rise could not grow indefinitely with γ. A way to add this feature in the Bethe-Bloch formula is by introducing the so-called density effect term. It accounts for the polarisation effect of the atoms in the medium, which effectively shield the electromagnetic field of the charged particle halting any further increase of the energy loss [?]. Denoting the correction as $\delta(\beta)$, one can rewrite Eq. (8.1) as:

$$\left\langle \frac{\mathrm{d}E}{\mathrm{d}x} \right\rangle = \frac{4\pi N e^4}{m_e \beta^2} z^2 \left(\log \frac{2 m_e \beta^2 \gamma^2}{I} - \beta^2 - \frac{\delta(\beta)}{2} \right). \tag{8.2}$$

In general, the form of $\delta(\beta)$ depends on the medium and its state of aggregation, involving the usage of tabulated parameters and implicit relations [?].

Another standard method to compute the amount of ionisation a charged particle produces is the so-called photo-absorption ionisation (PAI) model proposed by Allison and Cobb [?]. Within their approach, the mean ionisation is evaluated using a semiclassical calculation in which one characterises the continuum material medium by means of a complex dielectric constant $\varepsilon(k, \omega)$. However, in order to model the dielectric constant they rely on the quantum-mechanical picture of photon absorption and collision. Therefore, in the PAI model the computation of the ionisation loss involves a numerical integration of the measured photo-absorption cross-section for the relevant material.

In a particle physics experiment, the typical way of determining the energy loss per unit length as a function of the particle velocity is studying identified particles over a range of momenta. Once we have established this relation we can use it for other, unknown particles. In this sense, it makes sense to have a regular mathematical expression for this relation that one can use.

It happens that neither the Bethe-Bloch theory nor the PAI model from Allison and Cobb offer a close mathematical form for the ionisation curve. This is the reason why a full parametrisation of the ionisation curves can be useful. A parametrisation originally proposed for the ALEPH TPC [?] and later used by the ALICE TPC [?] group that manages to capture the features of the ionisation energy loss is:

$$f(\beta\gamma) = \frac{P_1}{\beta^{P_4}} \left(P_2 - \beta^{P_4} - \log\left[P_3 + \frac{1}{(\beta\gamma)^{P_5}} \right] \right), \tag{8.3}$$

where P_i are five free parameters. Hereafter, we will refer to Eq. (8.3) as the ALEPH $\mathrm{d}E/\mathrm{d}x$ parametrisation.

8.1.1 Energy calibration

In order to obtain the amount of energy loss by a charged particle due to ionisation in our TPC we need to determine the conversion between the charge deposited in our readout planes and the actual energy depositions. This procedure is known as energy calibration.

In a general, the first step of the calibration involves a non-uniformity correction, to make sure that the detector response is uniform throughout the TPC. These are typically divided into three categories, non-uniformities in the transverse YZ plane, non-uniformities along the drift direction X and variations of the detector response over time (would not apply to us as the detector is not built yet). These would correct for effects such as electron diffusion and attenuation, space charge effects or channel misconfiguration. However, because at the moment I am only interested in making sure we recover a sensible result from our simulation, I will not apply uniformity corrections to our charge deposits.

Other effects, like electron-ion recombination or ADC saturation, lead to a non-linear relation between the observed charge and the deposited energy in the detector, with the observed readout charge saturating at high ionisation energies. In this case, because we are dealing with gaseous argon and therefore recombination is not as important as in

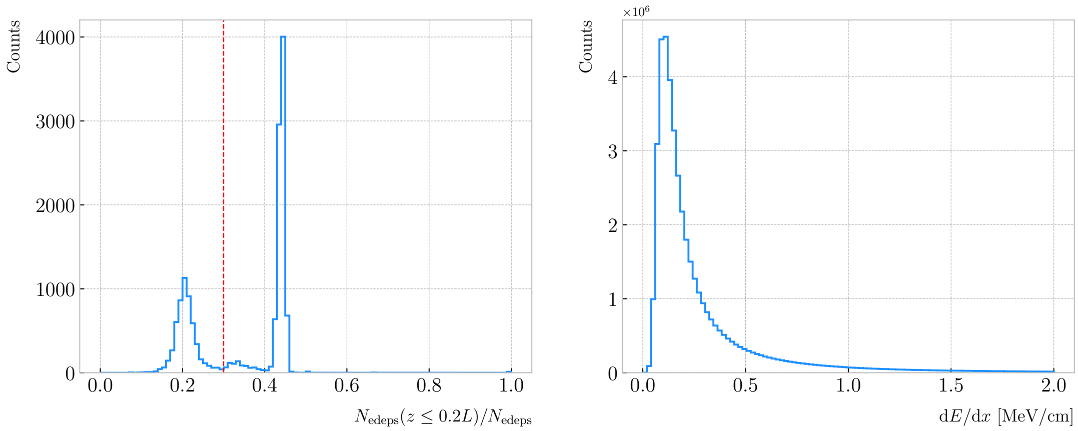


Figure 8.1: Left panel: distribution of the fraction of Geant4-level energy deposits per track with residual range less than 20% of the total track length, for the isotropic proton sample. Right panel: distribution of the ionisation per unit length of the energy deposits in the proton sample after removing the tracks with less than 30% of their energy deposits in the last 20% of the track.

liquid, we do not simulate recombination effects in the TPC. Even so, the simulation of the electronic response will still introduce charge saturation, and one needs to correct for it in order to obtain the exact amount of energy loss due to ionisation.

By default, the track fitting algorithm in GArSoft provides a `TrackIonization` object associated to each reconstructed track. It contains two collections of charge deposits, one for each fitting direction, consisting on pairs of charge values (dQ, in ADC) and step sizes (dx, in cm).

In order to estimate the ionisation loss in the ND-GAr TPC, I have used an MC sample consisting of single, isotropic protons propagating in the TPC. The starting points of the protons were sampled inside a $50 \times 50 \times 25$ cm box centered at $(100, -150, 1250)$, and their momenta are uniformly distributed in the range $0.25 - 1.75$ GeV. I ran the simulated sample through GArSoft's default detector simulation and reconstruction, and then a custom analyser module that extracts the ionisation data together with other reconstructed track information from the Kalman fit.

For studying the energy loss of the protons I select the reconstructed tracks that range out (i.e. slow down to rest) inside the TPC. A characteristic feature of the energy loss profile of any stopping ionising particle is the so-called Bragg peak, a pronounced

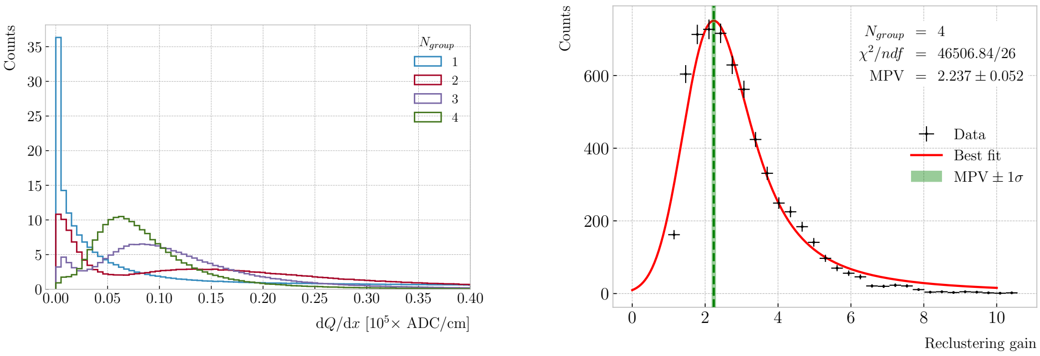


Figure 8.2: Left panel: distribution of the reconstructed ionisation charge per unit length for our MC stopping proton sample. The different colors indicate how many consecutive dQ/dx pairs were grouped together. Right panel: distribution of the median change in dQ/dx per track after $N_{group} = 4$ clusters were reclustered together.

peak that occurs immediately before the particle comes to rest. From Eq. (8.1) we can see that this behaviour is expected, as the energy loss for non-relativistic particles is inversely proportional to β^2. In data, a way of identifying the Bragg peak, and thus select the stopping particles, is checking the number of energy deposits towards the end of the track. In this case, I count the fraction of the Geant4 simulated energy deposits with a residual range value (the distance from a given energy deposit to the last deposit in the track trajectory) less than a 20% of the corresponding track length[1]. The distribution of this fraction of energy deposits for our proton sample is shown in Fig. 8.1 (left panel). We can clearly see two well separated peaks in this distribution, one centered at 0.2 and another, narrower, one centered at a higher value. The first one corresponds to non-stopping protons, as in that case the number of energy deposits towards the end of the track is uniformly distributed due to the absence of the Bragg peak. In that way, I apply a cut in this distribution, requiring that at least 30% of the simulated energy deposits sit in the last 20% of the tracks, to ensure that the Bragg peak is present.

Figure 8.1 (right panel) shows the distribution of the energy loss per unit length for the Geant4 simulated energy deposits of the selected stopping protons. We can see that

[1]As we are applying this selection at the Geant4 level we could have simply selected the stopping protons using the `EndProcess` labels from the simulation. However, the Bragg peak identification method displayed here could serve as a starting point for a selection of stopping protons in real data.

it follows the expected shape of a Landau distribution, which describes the fluctuations of the ionisation energy losses [?]. This distribution has a characteristic asymmetric PDF, with a long right tail that translates into a high probability for high-energy ionisation losses. The origin of these fluctuations is mainly the possibility of transferring a high enough energy to an electron, so it becomes a ionising particle itself.

Now, from the point of view of the reconstruction, the objects that we have available to extract the ionisation information for the different reconstructed tracks are the collections of dQ and dx pairs, as stated before. The dQ values come from adding up the amplitude of all the reconstructed hits in a cluster, which is the input object to the Kalman fit.

Figure 8.2 (left panel) shows the distribution of the ionisation charge deposits per unit length for the track in the stopping proton sample (blue line). As one can notice, this distribution does not resemble the expected shape of the Landau PDF. This distribution peaks sharply at 0 and has a heavy tailed behaviour. Notice, however, how the distribution changes its shape as we group together N_{group} consecutive charge deposit pairs (red, purple and green lines). The distribution in the $N_{group} = 4$ case already has a shape which resembles that of the Geant4-level ionisation per unit length, so I will proceed using this amount of reclustering for the reconstruction-level depositions.

An extra factor I need to account for, when reclustering is applied, is how the overall dQ/dx per track changes. To do so, we can look at the ratio between the median dQ/dx after and before the reclustering. Figure 8.2 (right panel) shows the median enhancement in dQ/dx per track for the stopping proton sample in the case $N_{group} = 4$. Fitting a Landau distribution convolved with a Gaussian[2], I estimate the most probable value of this ratio to be $G_{group} = 2.24 \pm 0.05$.

At this point, I am left with determining the conversion between the charge deposits per unit length dQ/dx and the energy deposits per unit length dE/dx. To this end, we need a way of comparing the two. I can use the residual range z to get a prediction of

[2] In the literature, this distribution is often referred to as Landau+Gaussian or langau. In the following, I will use LanGauss to refer to such PDF.

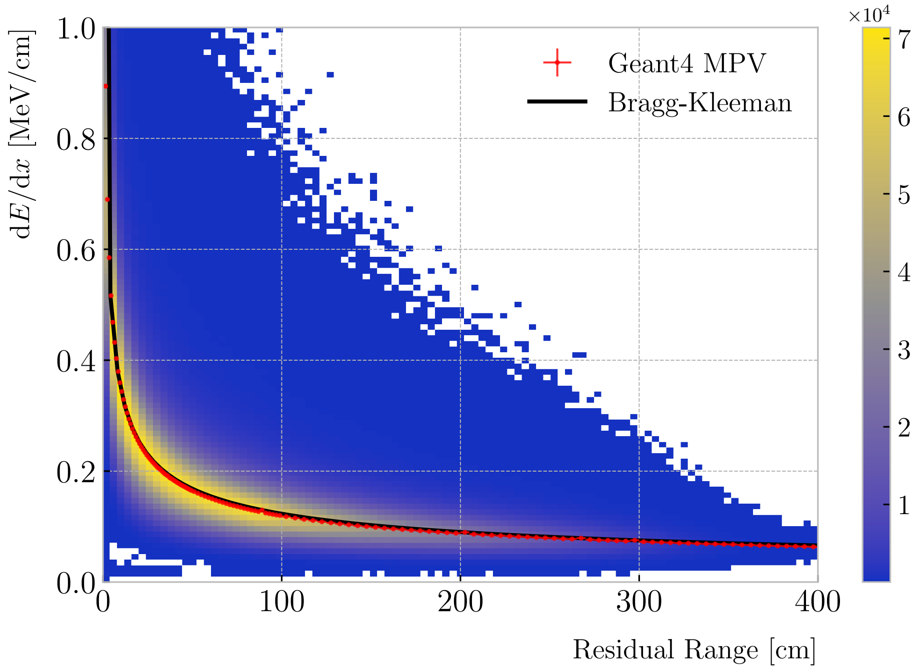


Figure 8.3: Distribution of the Geant4-simulated energy losses per unit length versus residual range for the stopping proton sample. The overlaid points represent the fitted most probable value of the $\mathrm{d}E/\mathrm{d}x$ distribution in each residual range bin, whereas the curve is their best fit to the Bragg-Kleeman formula from Eq. (8.4).

the most probable $\mathrm{d}E/\mathrm{d}x$ by using the following empirical parametrisation [?]:

$$\frac{\mathrm{d}E}{\mathrm{d}x}(z) = \frac{z^{\frac{1}{p}-1}}{p\Lambda^{\frac{1}{p}}}, \tag{8.4}$$

which is quoted in the literature as the Bragg-Kleeman formula. In order to obtain the p and Λ parameters I perform a fit using the energy losses and the residual ranges given by the Geant4 stage of our proton sample.

Within our simulation, the residual range is sampled with a maximum size of 5 mm. Therefore, to perform the fit to the Bragg-Kleeman formula, we can use a fine-grained residual range binning. For each of the residual range bins I extract the $\mathrm{d}E/\mathrm{d}x$ distribution and fit it to a LanGauss distribution, to obtain the value of the most probable $\mathrm{d}E/\mathrm{d}x$ in the bin together with a statistical uncertainty. I then fit Eq. (8.4) to these most probable values and the centres of the residual range bins. This

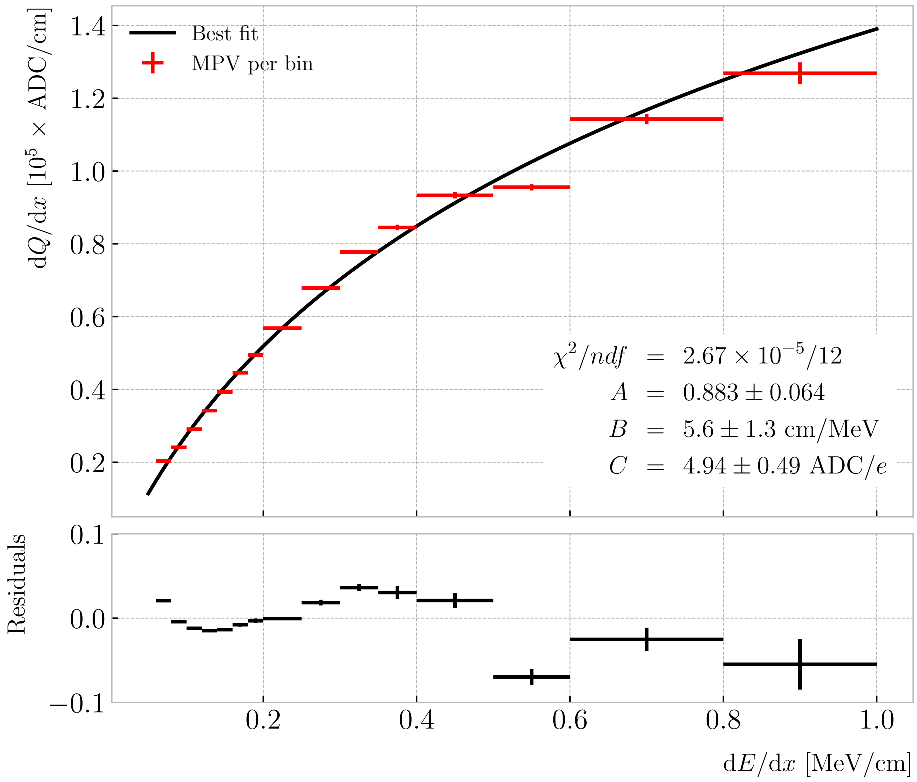


Figure 8.4: Fitted most probable dQ/dx values for each dE/dx bin (red points), obtained from the stopping proton sample. The overlaid curve (black line) represents the best fit to the logarithmic calibration function from Eq. (8.5).

procedure is depicted in Fig. 8.3, where I show the distribution of the energy loss per unit length versus the residual range, together with the most probable dE/dx values and their uncertainty in each bin (red points) and the curve with the best fit of the Bragg-Kleeman relation to those values (black line). The best fit is obtained for the parameter values $p = 1.8192 \pm 0.0005$ and $\Lambda = 0.3497 \pm 0.0008$ cm/MeV^{p}[3].

Having an analytical expression that relates the residual range to dE/dx, I can take our reconstruction-level residual ranges from the stopping proton sample and compute the most probable energy loss associated.

In order to parametrise the charge saturation, we can use the following logarithmic

[3]These strange units for Λ come from dimensional analysis, just to keep the Bragg-Kleeman formula (8.4) consistent.

function inspired by the modified box model for recombination:

$$\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{\mathrm{e}^{\frac{\mathrm{d}Q}{\mathrm{d}x} B \frac{W_{ion}}{G_{group} C}} - A}{B}, \tag{8.5}$$

where A and B are the calibration parameters we need to determine, W_{ion} is the average energy to produce an electron-ion pair, G_{group} is the gain from the reclustering discussed above and C is the calibration constant to convert number of electrons to ADC counts, commonly refer to as gain (also to be obtained in the fit). In this case, I use a value for the electron-ion production energy of $W_{ion} = 26.4$ eV [?]. This value, used in our simulation as well, was measured for gaseous argon in normal conditions, and therefore should be checked in the future to describe correctly the high-pressure argon-CH_4 mixture of ND-GAr.

For the calibration fit I follow a procedure similar to the previous one for Eq. (8.4). Binning the $\mathrm{d}E/\mathrm{d}x$ range, I fit a LanGauss distribution to the corresponding $\mathrm{d}Q/\mathrm{d}x$ distribution to obtain the most probable value. The resulting data points (red bars) are shown in Fig. 8.4 (top panel), the horizontal error bars depict the width of the $\mathrm{d}E/\mathrm{d}x$ bin whereas the vertical bars represent the error associated to the most probable value estimation. A fit to the logarithmic function in Eq. (8.5) is also shown (black line). For this I weighted the data points using the inverse of their relative error, obtaining a reduced chi-square value of $\chi^2/ndf = 2.22 \times 10^{-6}$. The best fit parameters I found from this fit are $A = 0.883 \pm 0.064$, $B = 5.6 \pm 1.3$ cm/MeV and $C = 4.94 \pm 0.49$ ADC/e. Figure 8.4 (bottom panel) shows the residuals between the data points and the fit.

The value for the gain I obtained from the fit is in reasonable agreement with our expectation. This value is set in GArSoft to 5 ADC/e by default.

One interesting thing to check is what induces this non-linear relation between charge and energy. The only effects that modify the amount of electrons reaching the readout planes in the simulation are the transverse diffusion and the finite electron lifetime. Once the electrons reach the readout chambers, the pad response functions are applied, together with an electrons-to-ADC conversion and the ADC saturation limit.

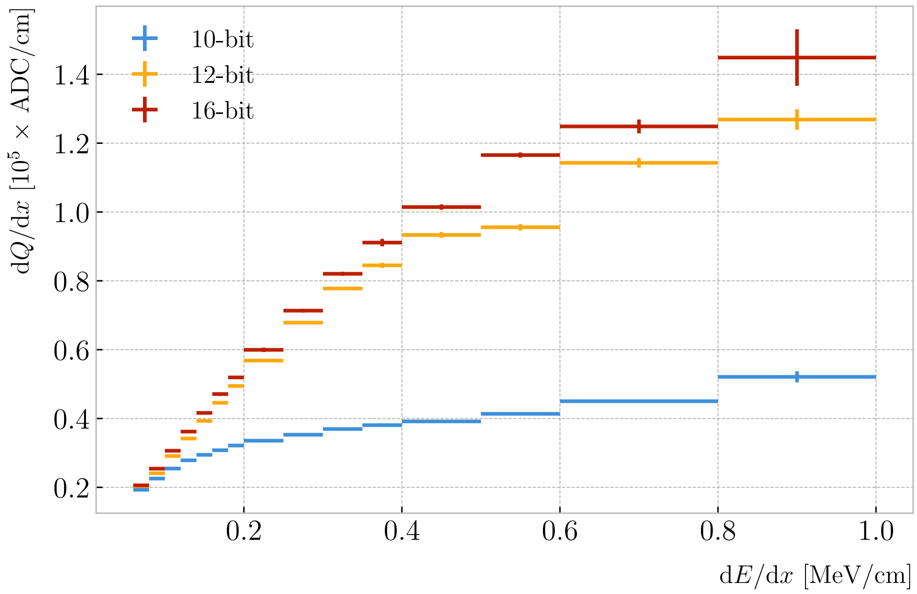


Figure 8.5: Fitted most probable dQ/dx values for each dE/dx bin for three different ADC bit limits, 10 (blue points), 12 (default, yellow points) and 16-bit (red points).

Table 8.1: Calibration parameters obtained from the fit of the ND-GAr simulated stopping proton sample to the calibration function from Eq. (8.5). The fits were performed for the 10, 12, and 16-bit ADC limits.

	χ^2/ndf	Best fit $\pm 1\sigma$		
		A	B (cm/MeV)	C (ADC/e)
10-bit	$1.83 \times 10^{-6}/12$	-9.3 ± 3.9	270 ± 69	27.1 ± 5.4
12-bit	$2.67 \times 10^{-5}/12$	0.883 ± 0.064	5.6 ± 1.3	4.94 ± 0.49
16-bit	$1.44 \times 10^{-5}/12$	0.949 ± 0.024	3.53 ± 0.58	4.52 ± 0.29

By default, GArSot applies a 12-bit ADC limit, which can be changed in the simulation configuration. However, it can only be increased up to 16-bit, as we represent the ADC collection as a `std::vector<short>`. This way, I tried to change the saturation parameter to see how it affects the relation between reconstructed charge and energy. Figure 8.5 shows a comparison between the most probable dQ/dx for 10, 12 and 16-bit ADC limits. As expected, the lower the limit is the sooner the charge saturates. For higher ADC limits the relation between energy and charge remains linear up to higher dE/dx values, but even for the 16-bit limit the saturation is noticeable for values $\gtrsim 0.5$ MeV/cm.

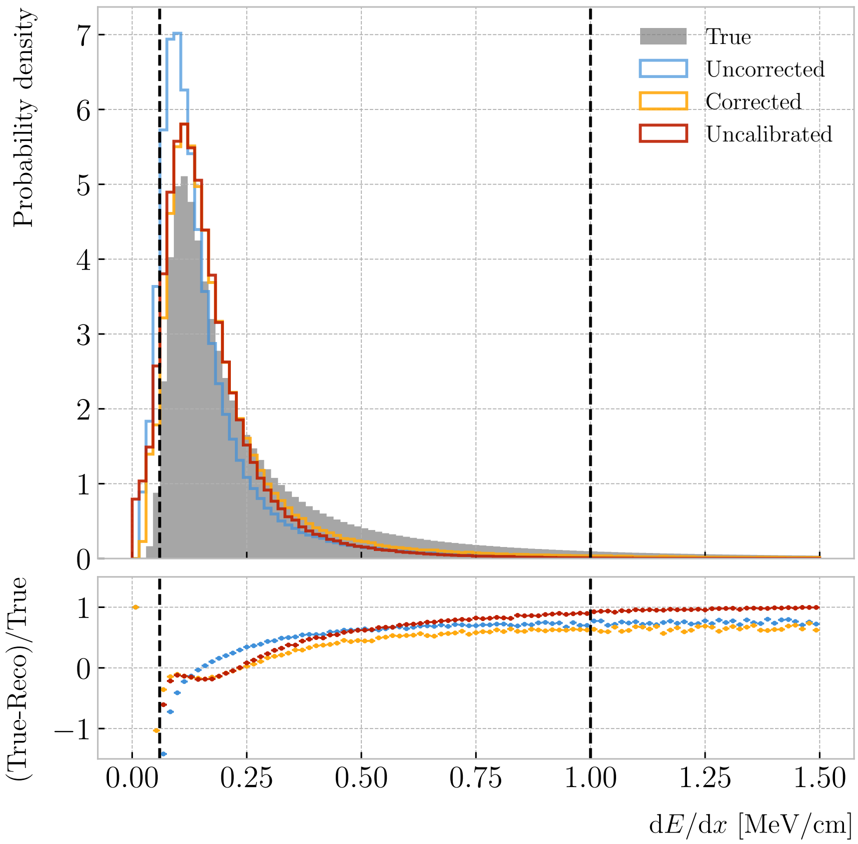


Figure 8.6: Top panel: area normalised dE/dx distributions for the true (solid grey) and the reconstructed energy deposits in the stopping proton sample, both after applying the calibration (blue) and the calibration and the normalisation correction (yellow). Also shown is the distribution obtained by applying a correction factor to the dQ/dx values but not the calibration (red). Bottom panel: fractional residuals for the uncorrected (blue), corrected (yellow) and uncalibrated (red) samples.

Table 8.1 shows the results of fitting the samples with 10 and 16-bits ADC limits to the calibration function from Eq. (8.5), using the weights based on their relative error as described previously. One interesting feature to notice is how different the best fit points look for the 10-bit ADC saturation when compared to the other two, which are consistent with each other.

At this point we can compare the dE/dx distribution one gets from Geant4, i.e. the true energy loss distribution, and the distribution I found by applying the calibration function to our collection of reconstructed dQ/dx values. Figure 8.6 (top panel) shows the true (solid grey) and reconstructed (blue, labeled as uncorrected) distributions together. The dashed vertical lines indicate the region of validity of the calibration fit, i.e. the left and right edges of the first and last dE/dx bin respectively. Notice that these histograms are area-normalised, as the total number of true energy deposits is much higher than the number of reconstructed charge deposits. This is due to a combination of effects, like the finite spatial resolution of the detector, the hit clustering used in the track fitting and the reclustering we have applied here.

The two distributions are significantly different. That can be seen clearly when looking at the fractional residuals, shown in Fig. 8.6 (bottom panel). In particular, the position of the peak is off, which could bias the mean energy loss predictions. It seems like the difference between these may be due to an overall scaling factor. One possibility is to scale the most probable value of the reconstructed distribution to the most probable value predicted by Geant4. I do this by fitting both distributions using a LanGauss function, obtaining $\mathrm{d}E/\mathrm{d}x_{\mathrm{MPV},\ true} = 0.1145 \pm 0.0005$ MeV/cm and $\mathrm{d}E/\mathrm{d}x_{\mathrm{MPV},\ reco} = 0.0928 \pm 0.0005$ MeV/cm for the true and reconstructed most probable values respectively. These can be translated into an scaling factor $S = 0.579 \pm 0.006$.

The result of applying the scaling correction can be seen in Fig. 8.6 (top panel). The corrected dE/dx distribution (yellow, labeled as corrected) peaks around the same value the true distribution does, as expected. Moreover, the high energy region is also slightly better described. For low ionisations, below the lower limit of the calibration fit, the differences between true and reconstructed are still significant. This low energy excess may be migration of some events from the peak region. The overall effect of the correction can be seen in the fractional residual plot in Fig. 8.6 (bottom panel).

One can also check what happens if instead of applying the logarithmic calibration we simply scale the dQ/dx distribution (post reclustering) to have the same most probable value as the true dE/dx distribution. In this case, following an analogous procedure to the one described earlier, I found the scaling factor $S_{uncalibrated} = 0.414 \pm 0.002$ MeV/ADC[4]. The resulting distribution (red, labeled as uncalibrated) is also shown in in Fig. 8.6 (top panel). The behaviour of the new distribution is similar to the corrected case at low

[4]Notice that now the scaling factor is not dimensionless, as it acts more like a conversion factor here.

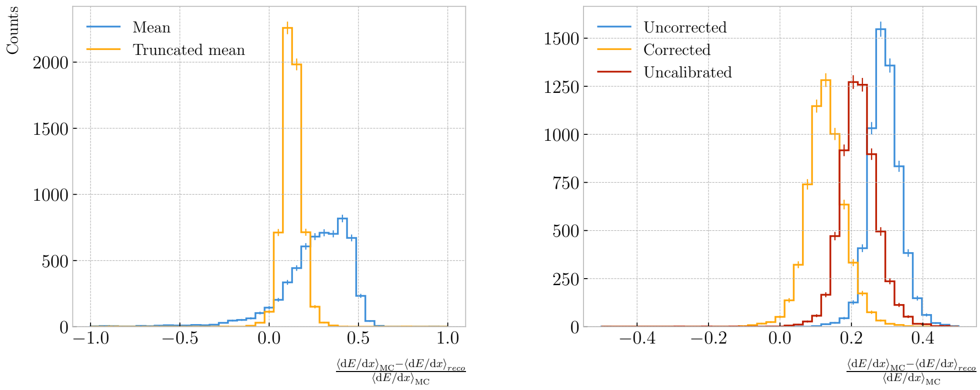


Figure 8.7: Left panel: fractional residuals between the true and the corrected dE/dx means (blue) and the 60% truncated means (yellow), for each event in the stopping proton sample. Right panel: fractional residuals between the true and the uncorrected (blue), corrected (yellow) and uncalibrated (red) dE/dx 60% truncated means, for each event in the stopping proton sample.

energy losses, around the peak of the true distribution, but it is worse at describing the high energy tail. This is expected, it is in the high ionisation regime where saturation effects apply and therefore calibration is needed.

8.1.2 Truncated dE/dx mean

Once we have a collection of dE/dx values for each reconstructed track, we can compute the corresponding most probable ionisation loss per unit length of the particle. This is the value predicted by the Bethe-Bloch or the PAI models, and together with a measurement of the momentum it allows for particle identification.

However, estimating the most probable dE/dx value for each reconstructed track is not a trivial task. As mentioned before, the dE/dx distributions follow Landau-like distributions. Therefore, one should perform e.g. a LanGauss fit to correctly estimate the most probable values. Automating this kind of fits is often problematic, as they usually incur in convergence problems. Moreover, the reconstructed dE/dx distributions we obtain tend to have relatively small statistics, which may also produce poor fits. In practice, doing these unsupervised fits may degrade our performance, and a more robust method is preferred.

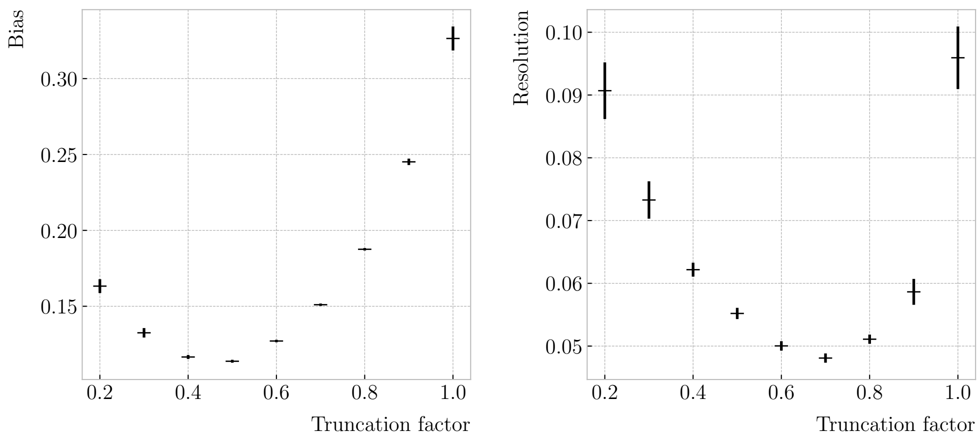


Figure 8.8: Estimated values of the mean dE/dx bias (left panel) and resolution (right panel) obtained using the corrected data from the stopping proton sample, for different values of the truncation factor.

A possibility could be taking the mean of the reconstructed dE/dx distribution for each particle. The problem with this approach is that the high energy Landau tail, combined with our limited statistics, can induce large fluctuations in the computation of the mean. Imagine you have two protons with the same kinetic energy, but due to reconstruction problems in one case you did not get as many charge deposits reconstructed in its high ionisation loss region. If you do not remove the tails the computed dE/dx means will be significantly different.

In order to avoid those fluctuations, one can compute the mean of a truncated dE/dx distribution instead. By keeping only a given fraction of the lowest energy deposits we obtain an estimate of the mean energy loss that is more resilient to reconstruction inefficiencies and statistical effects. Figure 8.7 (left panel) shows a comparison between the $\langle dE/dx \rangle$ computed by taking the mean of the full distribution (blue line) and the 60% lowest energy clusters (yellow line), for the stopping proton sample. The fractional residuals are computed for each proton, taking the corresponding means using their collections of true and reconstructed energy deposits. One can see that using the simple mean translates into a high bias and uncertainty in the $\langle dE/dx \rangle$ estimation, whereas applying the truncation reduces both significantly.

Additionally, I performed a comparison between the 60% truncated mean dE/dx

obtained using the different calibration methods discussed earlier, namely the uncorrected (blue), corrected (yellow) and uncalibrated (red) distributions. The results are shown in Fig. 8.7 (right panel). While the widths of these distributions are similar, the bias obtained for the corrected sample, i.e. calibration function and correction factor applied, is a factor of ~ 2 lower than in the uncalibrated case and almost three times smaller than for the uncorrected sample.

The next step is to optimise the level of truncation we are going to apply to our data. To do so, I used different truncation factors, i.e. the percentage of energy-ordered reconstructed energy deposits we keep to compute the mean, on the corrected $\mathrm{d}E/\mathrm{d}x$ sample of the stopping protons. Then, following the same procedure of computing the fractional residuals as before, I fitted the resulting histograms using a double Gaussian function. This is simply the sum of two Gaussian functions of the type:

$$g(x;\ \mu, \sigma, A) = A\ \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{8.6}$$

I do not add the classical normalisation factor of the Gaussian, $1/\sqrt{2\pi}\sigma$, therefore the amplitude A simply represents the maximum of the function. One of the two Gaussian functions describes the core part of the distribution, while the other captures the behaviour of the tails.

For each truncation factor, I look at the bias and the resolution I obtain. I define these as the weighted means of the corresponding parameters in the fits:

$$\bar{x} = \frac{A_{core}\ x_{core} + A_{tail}\ x_{tail}}{A_{core} + A_{tail}}, \tag{8.7}$$

where A_{core} and A_{tail} are the amplitudes of the core and tail distributions respectively and x is either the mean μ or the width σ of said distributions.

Figure 8.8 shows the bias (left panel) and the resolution (right panel) I obtained for the stopping proton sample, using different values of the truncation. From these, it can be seen that a truncation factor of 50% minimises the bias in the estimation, while

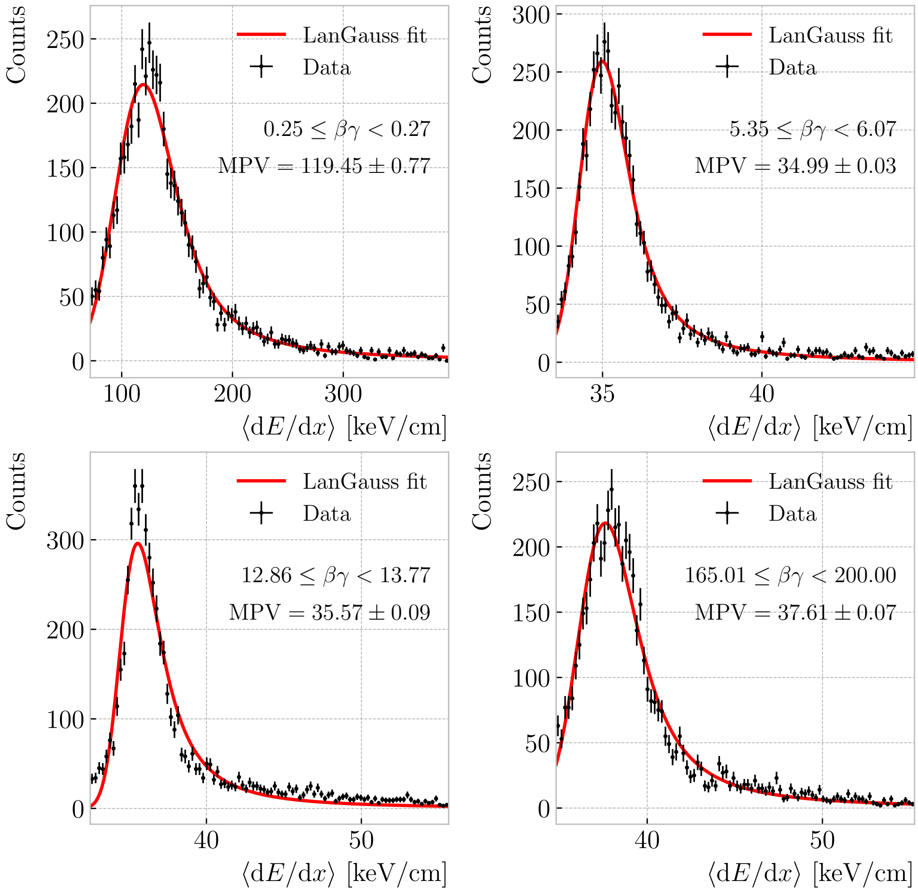


Figure 8.9: Examples of the truncated mean dE/dx LanGauss fits for various $\beta\gamma$ bins, from a simulated FHC neutrino sample.

70% gives the best resolution. That way, I settled on the intermediate value of 60% truncation, which yields a $\langle \mathrm{d}E/\mathrm{d}x \rangle$ resolution of 5.00 ± 0.08 % for stopping protons.

8.1.3 Mean dE/dx parametrisation

Now that we have a way to estimate the mean energy loss of a particle in the HPgTPC, we can determine the value of the free parameters in the ALEPH formula, Eq. (8.3). For this, I used a sample of 10^5 reconstructed FHC neutrino events inside ND-GAr. In this case I cannot use the stopping proton sample, as we need to cover the full kinematic range of interest for the neutrino interactions in our detector.

The original data does not contain an estimation of the velocity of the tracks, instead the tracks have a value for the reconstructed momentum and the associated PDG code

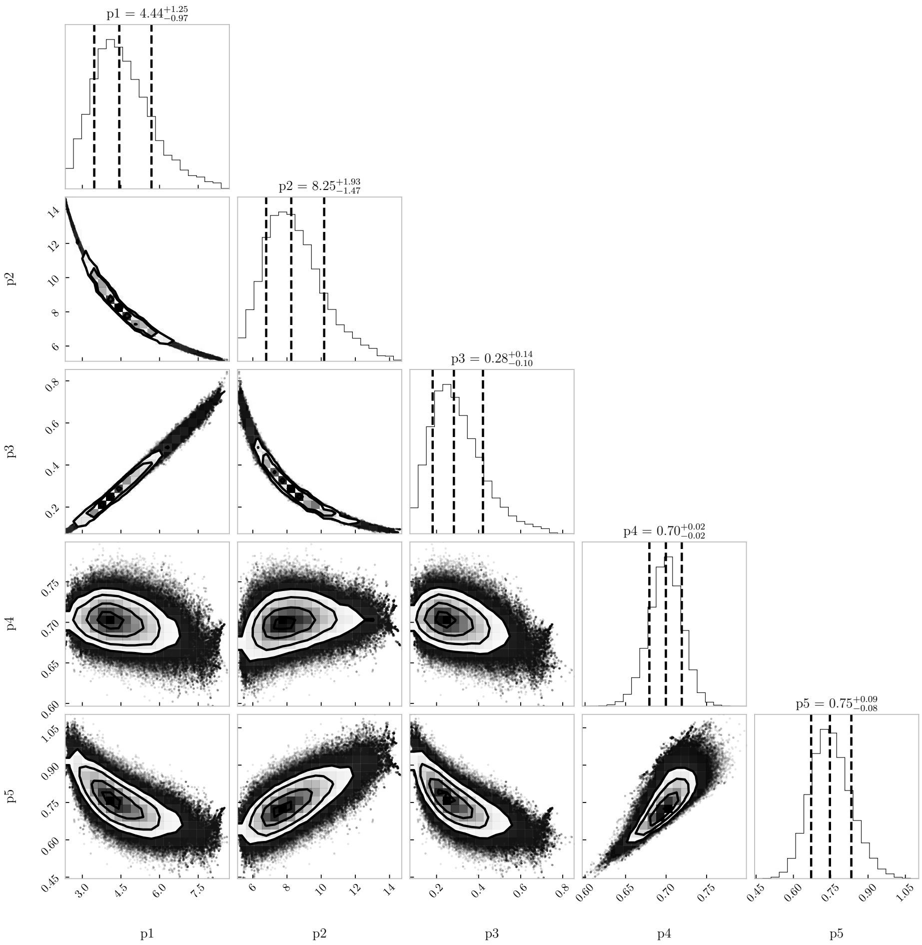


Figure 8.10: Resulting one and two dimensional projections of the posterior probability distributions of the ALEPH $\langle \mathrm{d}E/\mathrm{d}x \rangle$ parameters obtained by fitting the 60% truncated mean $\mathrm{d}E/\mathrm{d}x$ values from a FHC neutrino sample in ND-GAr. The vertical dashed lines in the 1D distributions represent the 16th, 50th and 84th percentiles.

of the Geant4-level particle that created the track. Therefore, one can select some of the particles in the data, in this case I selected electrons, muons, pions and protons, and compute β and γ using the reconstructed momentum and their mass. In terms of $\beta\gamma$ the mean $\mathrm{d}E/\mathrm{d}x$ does not depend on the particle species, so one can consider all the dataset as a whole. For this fit, I will express β in terms of the $\beta\gamma$ product as:

$$\beta = \frac{\beta\gamma}{\sqrt{1 + (\beta\gamma)^2}}, \tag{8.8}$$

which can be easily proven from the definition of γ.

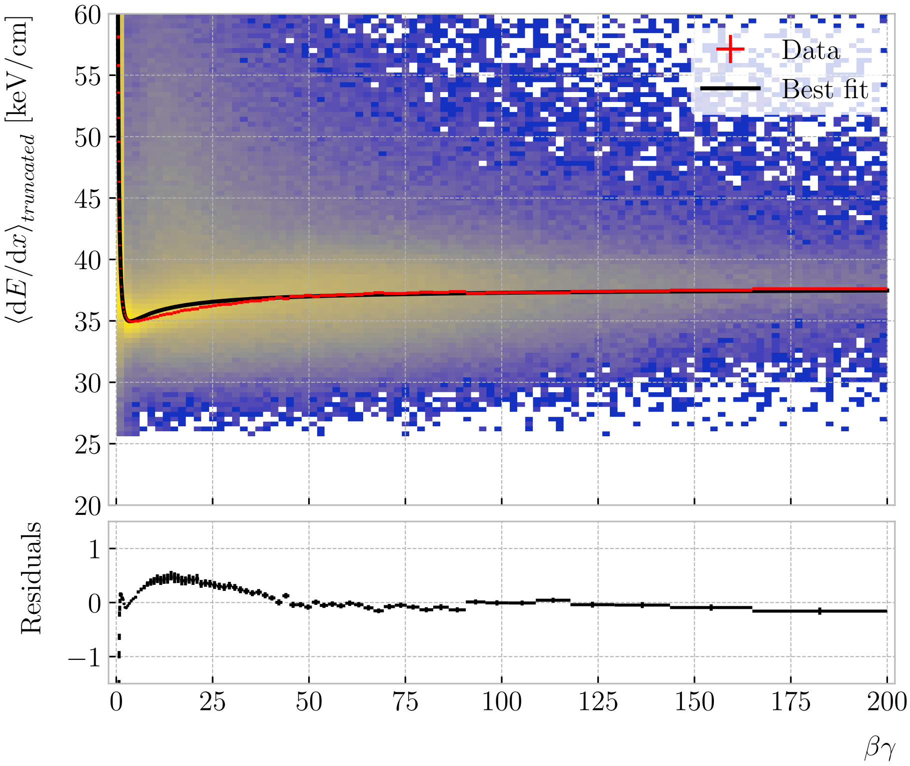


Figure 8.11: Truncated mean dE/dx obtained for the FHC neutrino sample as a function of the $\beta\gamma$ product (upper panel). Also shown are the fitted most probable values for each $\beta\gamma$ bin (red points) and the best fit obtained using the ALEPH parametrisation (black line). The residuals resulting from the fit are shown in the lower panel.

Next, I bin the data in $\beta\gamma$. I chose a fine binning so as to capture the different features of the ionisation curve. Instead of fixing the bin width, I select them so each one has approximately the same statistics. Then, for each $\beta\gamma$ slice, I compute the median and the interquartile range (IQR) of the $\langle dE/dx \rangle$ distribution. Using these, I make a histogram in the range [median − IQR, median + 5 IQR), which I fit to a LanGauss function in order to extract the MPV. Using this range accounts for the asymmetric nature of the distributions, while also helps avoiding a second, lower maximum present at low $\beta\gamma$, probably a result of reconstruction failures.

A few examples of these fits are shown in Fig. 8.9. The chosen values of $\beta\gamma$ sit in very distinct points along the $\langle dE/dx \rangle$ curve, going from the high ionisation region at low velocities (top left panel), to the minimum point (top right panel), the beginning of

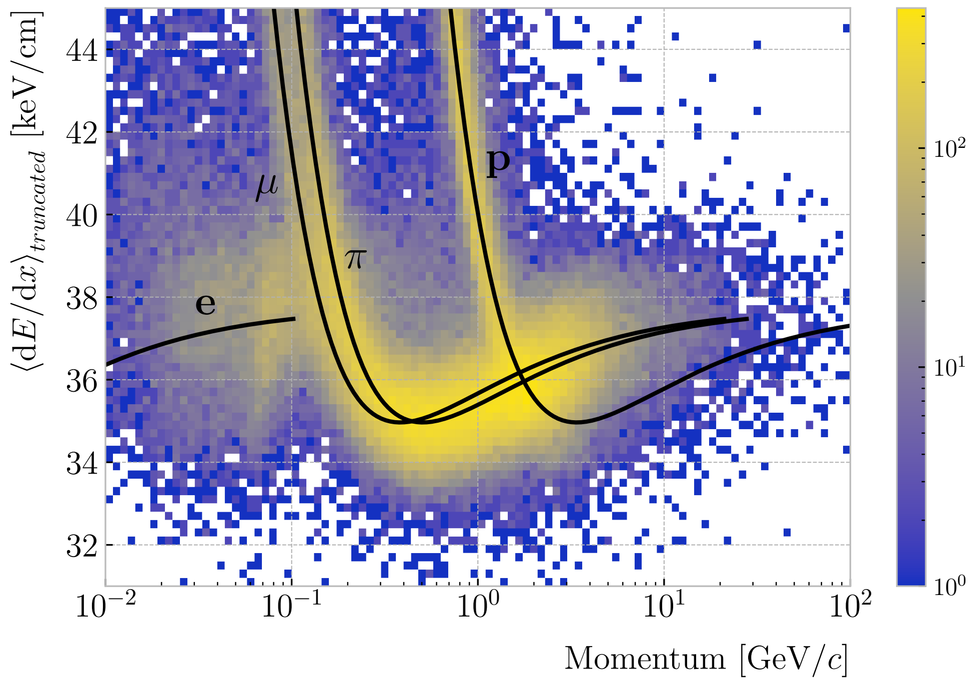


Figure 8.12: Distribution of the 60% truncated mean dE/dx versus reconstructed momentum for the FHC neutrino sample. The black lines indicate the predictions of the ALEPH parametrisation for electrons, muons, charged pions and protons.

the relativistic rise (bottom left panel), and the plateau produced by the density effect (bottom right panel).

I used the resulting most probable $\langle dE/dx \rangle$ values and the centres of the $\beta\gamma$ bins as the points to fit to the ALEPH formula. For this particular fit I used the least-squares method to get a first estimation of the ALEPH parameters. Applying some uniform priors, I then used these values as the starting point of a 100000 steps MCMC. Figure 8.10 shows the posterior probability distributions I obtain for each parameter. The reported best fit points are based on the 16th, 50th, and 84th percentiles in the marginalised distributions.

The resulting fit (black line), compared to the data points (red points) and the underlying distribution is shown in Fig. 8.11 (top panel). The overall fit is good, with a reduced chi-squared of $\chi^2/ndf = 1.02$. However, there are some regions where the fit does not describe the data correctly, like the very low $\beta\gamma$ regime, where the fit severely underestimates for energy losses $\gtrsim 50$ keV/cm, and the start of the relativistic raise,

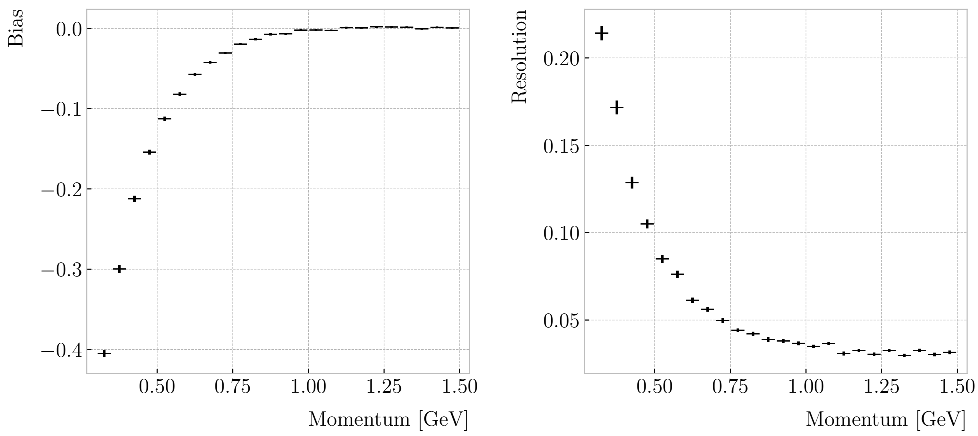


Figure 8.13: Estimated values of the mean $\mathrm{d}E/\mathrm{d}x$ bias (left panel) and resolution (right panel) obtained for the true protons in a FHC neutrino sample.

where we have a slight overestimation. This is a result of those points having a larger uncertainty when compared to the ones around the dip or the plateau areas. These differences can be better seen in the residual plot, Fig. 8.11 (bottom panel).

8.1.4 Particle identification

8.2 Muon and pion separation in the ECal and MuID

As it could be seen from Fig. 8.12, it is not possible to separate muons and charged pions in the HPgTPC using $\mathrm{d}E/\mathrm{d}x$ for momenta $\gtrsim 300$ MeV/c. In ND-GAr, approximately 70% of the interactions in FHC mode will be ν_μ CC (compared to the 47% of $\bar{\nu}_\mu$ CC interactions when operating in RHC mode), while 24% are neutral currents. Out of these, around 53% and 47% of them will produce at leat one charged pion in the final state, respectively. Figure 8.14 shows a comparison between the spectra of the primary muons and the charged pions for ν_μ CC interactions in ND-GAr producing one or more charged pions. From this, one can see that (i) the majority of muons and charged pions are not going to be distinguishable with a $\langle \mathrm{d}E/\mathrm{d}x \rangle$ measurement, and that (ii) particle identification in necessary both to classify correctly the ν_μ CC events and identify the primary muon within them.

ND-GAr features two other subdetectors which can provide additional information

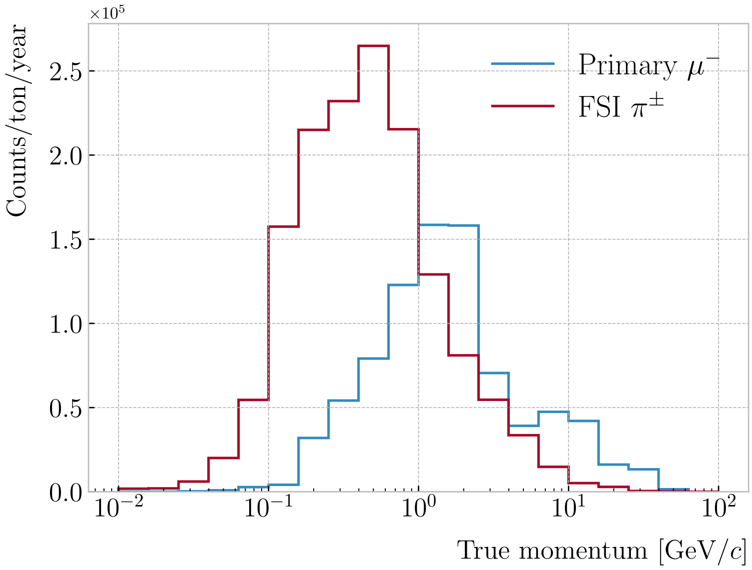


Figure 8.14: True momentum distribution for the primary muon in ν_μ CC $N\pi^\pm$ interactions inside the fiducial volume of ND-GAr (blue line), compared to the post FSI charged pion spectrum (red line).

for this task, namely the ECal and MuID. The current ECal design, described in (ref section), consists of 42 layers, made of 5 mm of Pb, 7 mm of plastic scintillator and a 1 mm PCB board. The total thickness of this calorimeter is 1.66 nuclear interaction lengths or 1.39 pion interaction lengths. The MuID design is in a more conceptual stage, however it is envisioned to feature layers with 10 cm of Fe and 2 cm of plastic scintillator[5]. With its three layers, it will have a thickness of 1.87 or 1.53 nuclear or pion interaction lengths, respectively.

Because pion showers are dominated by inelastic nuclear interactions, the signatures of these particles in the calorimeter will look significantly different from those of muons. Although our ECal is not thick enough to fully contain the hadronic showers of the charged pions at their typical energies in FHC neutrino interactions, they can still be used to understand whether the original particle was more hadron-like or MIP-like. In Fig. 8.15 I show two examples of energy distributions created by a muon (left panel) and a charged pion (right panel) of similar momenta interacting in the ECal. These

[5]It is not mentioned anywhere, but I assume that there should also be another layer of PCB board of 1 mm. However, in this case its contribution to the total thickness of the sampling calorimeter would be negligible.

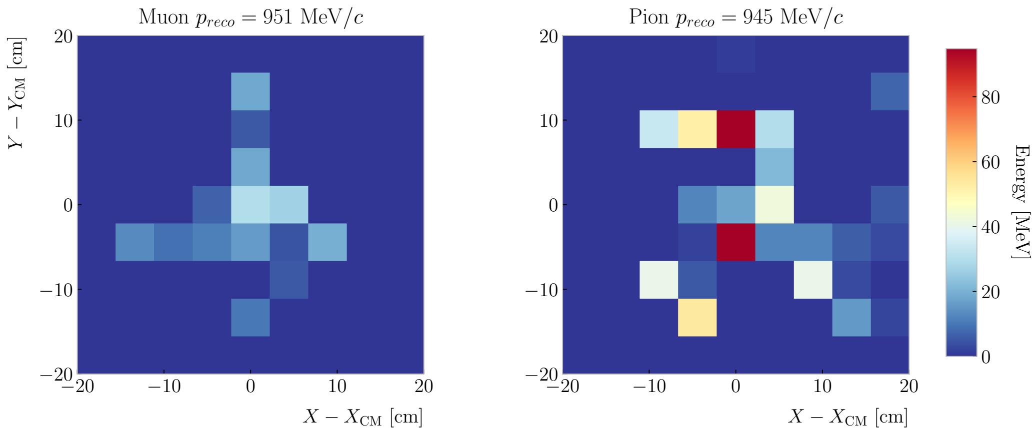


Figure 8.15: Distributions of energy deposits in the ECal for a muon (left panel) and a charged pion (right panel) with similar momenta. The energy is projected onto the plane perpendicular to the principal component of the hits, and the positions are relative to the center of the interaction.

figures represent the transverse development of the interactions. For each of them, I computed the principal component and centre of mass of the interaction, projecting the position of the hits onto the plane perpendicular to that direction, and taking the distances relative to the centre. It can be seen that the muon follows an almost MIP-like behaviour, being the central bin in the histogram the one with the highest deposited energy. On the other hand, the pion not only deposits more energy overall, but also this energy is more spread-out among the different hits. It is this kind of information that would allow us to tell apart muons from pions.

This way, I identify three main action points that need to be addressed if one wants to use these detectors to distinguish between muons and charged pions. These are:

1. the way we make the associations between tracks in the HPgTPC to the activities (what in GArSoft we call clusters) in the ECal and the MuID,

2. what variables or features one can extract from the calorimeters that encapsulate the information we are interested about,

3. and how to carry out the classification problem.

8.2.1 Track-ECal matching

One of the main players in the muon and pion separation is the way we associate clusters in the ECal to reconstructed tracks in the TPC. Missing some associations or making wrong ones can bias the ECal quantities that we can use for classifying particles. The current algorithm in GArSoft provides precise associations, i.e. most of the associations that it produces are correct, but it appears to miss an important number of associations (at least when using the default configuration).

The current TPC track-ECal cluster association algorithm is divided in four parts. It first checks whether the track end point fulfils certain conditions to be extrapolated. There are two cut values in this step, one for the drift direction and other radial.

If the point can be extrapolated, the code computes the coordinates of the centre of curvature using the Kalman fit estimates at the track end $(y,\ z,\ 1/R,\ \phi,\ \tan\lambda)$. It then compares the distance between this and the cluster in the (z, y) plane with R. This introduces another cut in the perpendicular direction.

The next step is different for clusters is in the barrel or in one of the end caps. If it is a barrel cluster the algorithm extrapolates the track up to the radial distance of the cluster. There are three possible outcomes, the extrapolated helix can cut the cylinder of radius r_{clus} two, one or zero times. I get the cut point that is closer to the cluster and check that it is either in the barrel or the end caps. Computing the difference between the x coordinates of the cluster and the extrapolated point, the module checks that this is not greater than a certain cut. If the cluster is in an end cap, I propagate the track up to the x position of the cluster. Then, the algorithm computes the angle in the (z, y) plane between the centre of curvature and the cluster, α, and the centre of curvature and the propagated point, α'. A cut is applied to the quantity $(\alpha - \alpha')R$.

If the cluster contains more than a certain number N of hits, I apply an extra cut to the dot product of the direction of the track at the propagated x value and the cluster direction.

The code makes sure to only associate one end of the track (if any) to a cluster.

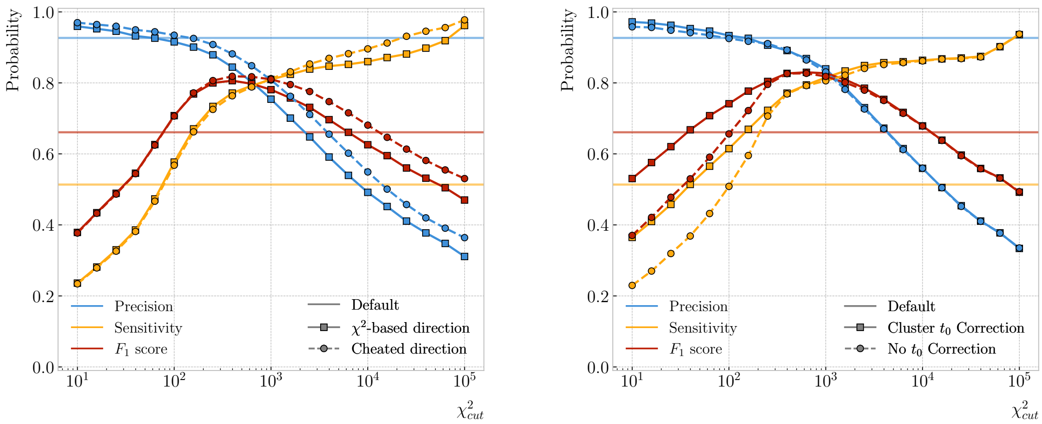


Figure 8.16: Left panel: comparison between the precision (blue), sensitivity (yellow) and F_1 score (red) obtained for the default (horizontal lines) and new algorithms, both with the χ^2-based direction estimator (squares) and cheating the directions (circles), for different values of the chi^2 cut. Right panel: comparison of the performance of the new algorithm when applying the cluster t_0 correction (squares) and when (circles).

However, it can associate more than one track to the same cluster. This makes sense, as different particles can contribute to the same cluster in the ECal, but it makes it difficult to quantify the relative contributions of the tracks to a certain cluster.

As a way of comparing the performance of this algorithm, a new, simpler association module was written. The goal was to have a simple and robust algorithm, which depends on as few parameters as possible and that can produce a one-to-one matching between tracks and ECal clusters.

For each reconstructed track, the new algorithms applies the same procedure to the forward and the backward fits irrespective of their end point positions. It first gets the Kalman fit parameters at the corresponding end point together with the X position, x_0, $(y_0,\ z_0,\ 1/R,\ \phi_0,\ \tan\lambda)$.

For each ECal cluster, I compute the radial distance to the centre of the TPC and find the ϕ value in the range $[\phi_0,\ \phi_0 + \text{sign}(R)\phi_{max})$ that makes the propagated helix intersect with the circle defined with such radius. The (x, y, z) position of the helix for the ϕ value found (if any) is then computed. In case there are two intersections, I keep the one that minimises the distance between (y, z) and (y_c, z_c).

I then calculate χ^2 value based on the Euclidean distance between the propagated point and the cluster:

$$\chi^2/ndf = \frac{\sum_{n=0}^{2}\left(x^{(n)} - x_c^{(n)}\right)^2}{3}. \tag{8.9}$$

If there was no intersection I store a -1 instead. In the end, for each reconstructed track in the event one ends up with two collections of χ^2 values, one for each ECal cluster and fit directions.

The current code only supports having ECal clusters associated to one end of each track. We have two options to decide what track end to keep. The first one tries to cheat the selection, looking at the distance between the two track ends and the true start position of the associated MC particle. The second one keeps the track end with more χ^2 entries below the cut.

This feature of only considering one track end limits the algorithm, making it not suitable for reconstructing events with particles originating outside the TPC. However, as for the moment the main concern of the group is the study of neutrino interactions off the gaseous argon, this is an acceptable assumption.

In order to associate a cluster to a track, I take all clusters with a χ^2 value in the range $[0, \chi^2_{cut})$. If a cluster has been assigned to more than one track we leave it with the one with the lowest χ^2.

This default behaviour of the algorithm can be modified to associate more than one track to each cluster. Not only that, but the χ^2 values can be used to assign relative weights to the different contributions.

To evaluate the performance of the association method, I use a binary classification approach. In this case, I check the leading MC Track IDs associated to the reconstructed tracks and ECal clusters. I count an association as true positive (TP) if both Track IDs coincide. An association is considered false positive (FP) when the Track IDs are different. If a cluster has not been associated to any track but it shares the Track ID with a reconstructed track it is counted as a false negative (FN).

For the testing, I used a sample of 10000 FHC neutrino events inside the HPgTPC. Figure 8.16 (left panel) shows the precision (blue line), sensitivity (yellow line), and F_1

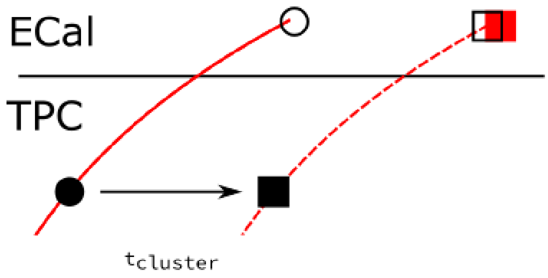


Figure 8.17: Schematics of a possible option to deal with track-ECal associations in non-zero t_0 neutrino interaction events, trying to correct for the drift direction uncertainty in a cluster-by-cluster basis using the cluster time, $t_{cluster}$.

score (red line) I obtained for different values of χ^2_{cut}. For comparison, the same metrics computed for the default algorithm with the current configuration are also shown (dashed lines). In the case of the new algorithm, I used both the χ^2-based method to estimate the track direction described earlier (square markers) and the cheated direction from the Geant-level information (circle markers). For either of these we achieve similar values of the precision compared to the old code, while having a considerably higher sensitivity. It can be seen that cheating the direction of the tracks only makes a difference at high χ^2_{cut}, past the optimal value of the cut around the F_1 score maximum. Therefore, I set the χ^2 method as the default.

One of the possible weak points of this approach is that it relies on the position along the drift direction to make the decisions. Within the current ND-GAr design implemented in GArSoft, the timing information is provided by the ECal. That effectively means that prior to make the track-ECal associations the reconstructed x positions of the track trajectories differ from the simulated ones by an amount:

$$x^{(n)}_{reco} - x^{(n)}_{sim} = v_{drift}\, t_0, \tag{8.10}$$

where v_{drift} is the mean drift velocity in our medium and the initial time is in the range $t_0 \in [0, t_{spill})$ where t_{spill} is the spill length. For a 10 μs spill this translates into a maximum 30 cm uncertainty on the drift direction position.

The current default in GArSoft sets $t_0 = 0$, but the functionality to randomly sample this within the spill time is in place. Therefore, we need to understand what is the impact of a non-zero t_0 on the associations algorithm and foresee possible ways of minimising a loss in performance.

Figure 8.17 represents a possible option to tackle the association problem when having events with a non-zero initial time t_0. The black and white circles represent the original points, whereas the squares indicate the corrected positions. The end points of the track and the propagated points up to the cluster radius are indicated using filled and unfilled markers respectively. The red square represents the position of the cluster.

Here I try to correct for the drift coordinate position using the time associated to the cluster. Assuming that the drift time is much larger than the propagation time, $t_{cluster}$ could be used as a good estimation of the t_0. An alternative can be using the earliest time associated to a hit in said cluster. Doing this for each cluster before computing the χ^2 value could be used as an alternative to knowing the specific value of the t_0, as when the association is correct this will provide the right correction but its impact is small enough to not change the position significantly in the case the cluster does not correspond to a given track.

I tested the effect of this correction again using a sample of 10000 FHC neutrino events. Figure 8.16 (right panel) shows the precision (blue line), sensitivity (yellow line), and F_1 score (red line) for the case the cluster t_0 correction is applied (square markers) and for the no correction case (circle markers), as a function of χ^2_{cut}. In this case, the differences are particularly notorious at low values of the cut. It makes sense, as the t_0 effect becomes subdominant when the distance we consider grows large. Overall, the correction increases the sensitivity while keeping the precision almost unchanged. As a result, I apply the t_0 correction to the generated samples as the default.

8.2.2 Classification strategy

The problem of the muon and charged pion separation has to be viewed in the broader context of the particle identification in our detector. Focusing on the beam neutrino interactions, it is clear that we are going to have muons and pions spanning a broad

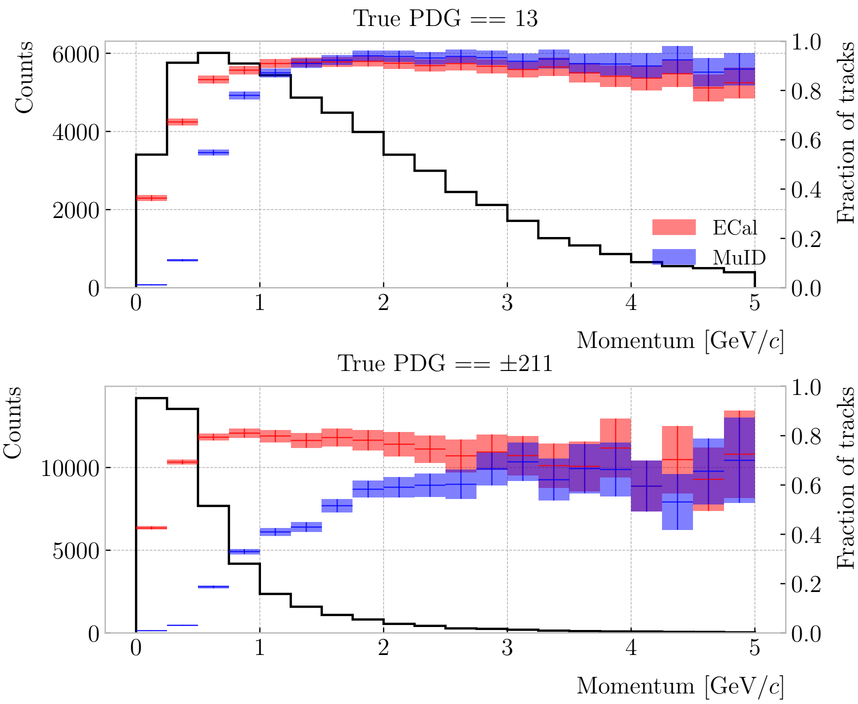


Figure 8.18: Momentum distribution for the reconstructed muons (top panel) and charged pions (bottom panel) in a FHC neutrino sample, together with the fraction of them reaching the ECal (red) and MuID (blue). Each entry corresponds to a reconstructed track, backtracked to a true muon or pion which has not produced any other reconstructed track.

momentum range. Not only that, but we will also have other particles with similar characteristics that will make the classification even more challenging. Therefore, we are presented with a task that will depend heavily on the kinematic range we are looking at each time, as both the available information and the possible impurities of other particle species vary.

For instance, distinguishing muons from pions could be difficult at low momenta, as a great number of them do not reach the ECal. Therefore, we could think of tailoring a version of the classification for that particular case, which could be complemented with a dE/dx measurement. Likewise, for momenta $\gtrsim 1$ GeV muons and pions reach the calorimeters efficiently, but so do protons. Because of this, one can try to train another classifier for this energy range, and rely on other methods to remove as many of the protons as possible.

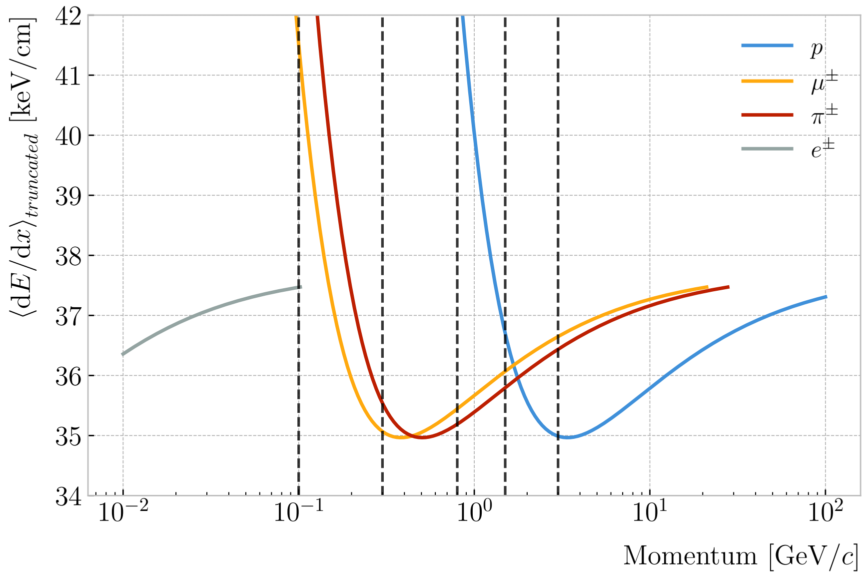


Figure 8.19: Predicted truncated mean $\mathrm{d}E/\mathrm{d}x$ versus momentum, for electrons, muons, charged pions and protons, obtained using the ALEPH parametrisation. The vertical dashed lines represent the boundaries of the six regions used for the muon and pion classification training.

Figure 8.18 shows the momentum distribution of the reconstructed muons (top) and pions (bottom) in a FHC sample. It also contains the fraction of particles reaching the ECal (red) and MuID (blue), for the different momentum bins. In Fig. 8.19 I show the mean $\mathrm{d}E/\mathrm{d}x$ of different particles as a function of the momentum, computed using the ALEPH parametrisation with the best fit parameters found in Subsec. 8.1.2.

Using these two figures as references, I decided to approach the classification by dividing the problem into six different momentum regions. A summary of these can be found in Tab. 8.2. The basic idea is to exploit all the information that is available in each region and . For the problem at hand, I prepared separated samples of isotropic single muons and pions, with momenta uniformly distributed along the corresponding momentum range. Each sample contains 50000 events of the corresponding particle species. I did not generate samples for the first region, as it is assumed that the separation can be achieved using $\mathrm{d}E/\mathrm{d}x$ only. For the last region, I generated particles up to a momentum of 10 GeV/c, as that is well above the typical energies of muons and pions from FHC neutrino interactions in ND-GAr.

Table 8.2: Momentum ranges and description of the PID approach assumed for the muon and pion classification task.

Momentum range	Description
< 0.1 GeV/c	All tracks can be separated with dE/dx
$[0.1,\ 0.3)$ GeV/c	Use ECal for reaching muons and pions, dE/dx for the rest
$[0.3,\ 0.8)$ GeV/c	Use ECal for muons and pions, dE/dx for protons
$[0.8,\ 1.5)$ GeV/c	Use ECal and MuID for muons and pions, dE/dx for protons
$[1.5,\ 3.0)$ GeV/c	Use ECal and MuID for muons and pions, ToF for protons
≥ 3.0 GeV/c	Use ECal and MuID for muons and pions, dE/dx and ToF for protons

Additionally, I prepared another sample of 100000 FHC neutrino events. For each interaction, I select the reconstructed particles which were backtracked to true muons or charged pions. I use this dataset to perform validation checks, to see how the models trained with the single particle data generalise to a more realistic scenario.

To tackle this classification problem, I make use of Boosted Decision Trees (BDT). A decision tree uses a flowchart-like structure to make decisions based on some input data. It starts from a root node, which represents the complete dataset, and then it splits this based on the variable or feature which gives the best separation between classes, creating two new nodes. The process repeats for each node until it reaches a certain limit, like a maximum number of splits or some tolerance criteria. The last set of nodes are often called leave nodes, and represent the final prediction of the classifier.

Boosting refers to a family of methods to combine the predictions from multiple classifiers, following a sequential approach where each new model learns from the errors of the previous one. The process starts with a simple decision tree, which is used to make predictions on the training data. Then, the data points misclassified by the first model are assigned higher weights, and another decision tree is trained on the data with adjusted weights. The predictions of the two trees are then combined, and the cycle repeats for a predefined number of iterations. Gradient boosting uses the direction of the steepest error descent to guide the learning process and improve the accuracy with each iteration.

8.2.3 Feature selection and importance

Using the reconstructed tracks as a starting point, I compute a number of ECal and MuID variables for each of them. As there can be more than one cluster associated to a track, what I do is collect all associated clusters and compute these variables from the complete collection of associated hits. For the MuID, because it only features three layers and typically there will be less hits, I also allow single hits to be associated with tracks[6]. I can roughly divide the variables in three types: energy-related, geometry-related and statistical. In the following, I briefly describe the variables related exclusively to the ECal:

- **Energy-related ECal**
  - ECal total energy (ClusterTotalEnergy): sum of the energy of all the ECal hits.
  - Mean ECal hit energy (HitMeanEnergy): mean of the hit energy distribution.
  - Standard deviation ECal hit energy (HitStdEnergy): standard deviation of the hit energy distribution.
  - Maximum ECal hit energy (HitMaxEnergy): maximum of the hit energy distribution.
- **Geometry-related ECal**
  - Mean distance hit-to-cluster (DistHitClusterMean): mean of the distance distribution between the hits and the corresponding cluster's main axis.
  - RMS distance hit-to-cluster (DistHitClusterRMS): root mean square of the distance distribution between the hits and the corresponding cluster's main axis.
  - Maximum distance hit-to-centre (DistHitCenterMax): maximum of the distance distribution between the hits and the centre of the TPC.

[6] At the reconstruction level what happens is that non-clustered hits are put into single hit clusters, instead of being thrown away. This is necessary to keep the consistency of the track-cluster association code.

  - Time-of-Flight velocity (TOFVelocity): slope obtained when fitting a straight line to the hit time versus hit distance to the centre (i.e. $d = v \times t$).

- **Energy and geometry ECal**

  - Radius 90% energy (Radius90E): distance in the hit-to-cluster distribution for which 90% of the total energy is contained in the hits that are closer to the axis (i.e. radius that contains 90% of the energy).

- **Statistical ECal**

  - Number of hits (NHits): total number of hits associated to the track.
  - Number of layers with hits (NLayers): not really a count of all layers with hits but the difference between the last and the first layer with hits.

Figure 8.20 shows the distributions of three different ECal variables, separating true muons (blue) and charged pions (red), for the five momentum ranges considered. I chose to show one feature from each category, namely the mean energy per hit (left column), the mean distance between the hits and the centre of the cluster (middle column), and the number of ECal layers with hits (right column). These give an idea of the separating power of the different features, and how it changes considerably with the energy. In the number of layers with hits distributions, the peak at 6 is due to the fact that the first six ECal layers sit inside the pressure vessel[7]. Therefore, some of the particles get stopped crossing it, never making it to the seventh layer.

In the case of the MuID, because at low momenta a significant fraction of the particles do not make it past the ECal, I only consider the information coming from this detector for momenta ≥ 0.8 GeV/c, i.e. for the last three momentum regions. The variables I extract from it are the following:

- **Energy-related MuID**

[7] Note to self: check this. I thought the ECal barrel had 8 layers of tiles, and that all of them were inside the pressure vessel.

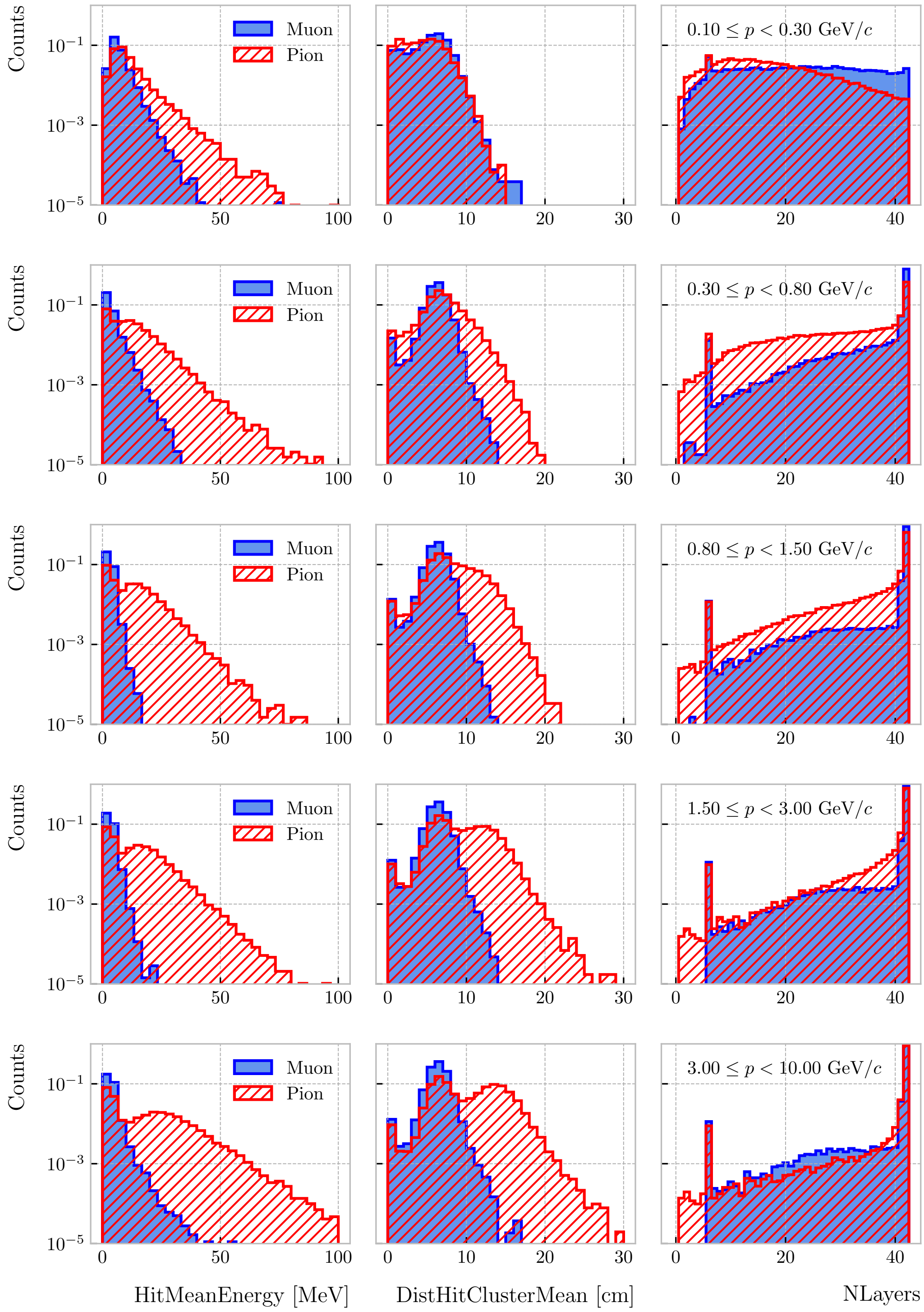


Figure 8.20: Example ECal feature distributions for muons (blue) and charged pions (red) in the five different momentum ranges considered (from top to bottom, in ascending momentum order). From left to right: mean hit energy, mean distance hit-to-cluster, and number of layers with hits.

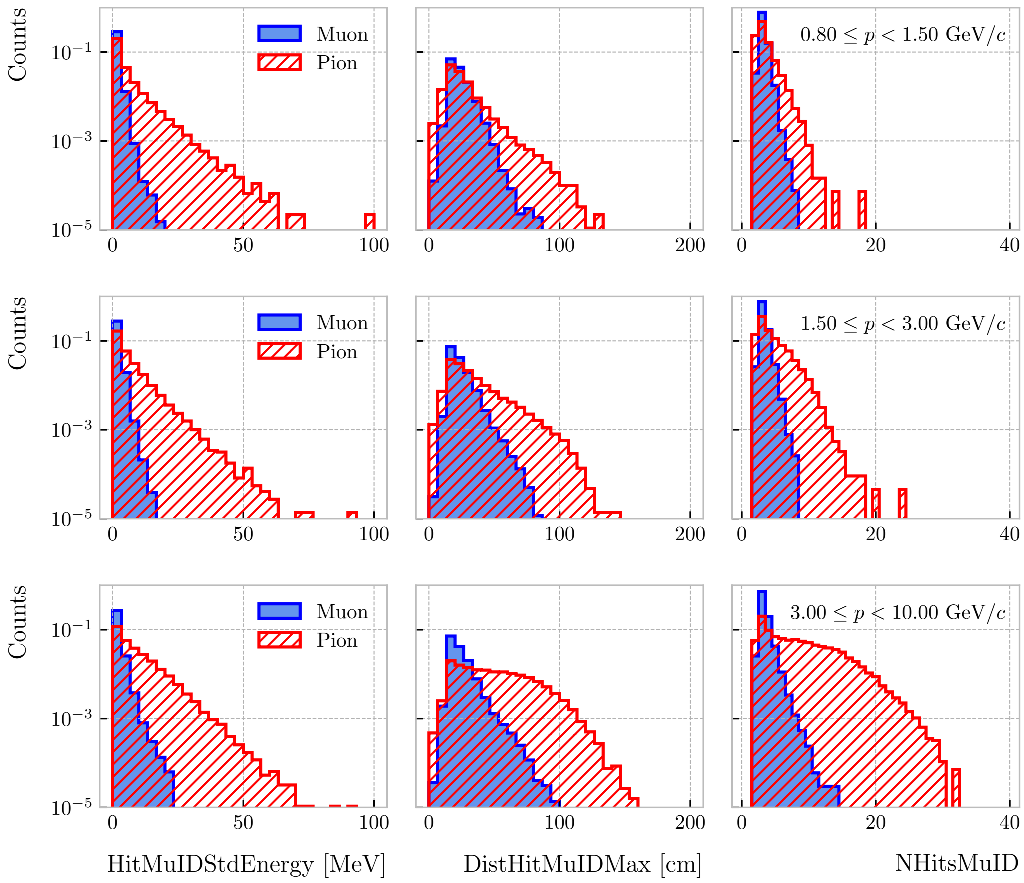


Figure 8.21: Example MuID feature distributions for muons (blue) and charged pions (red) in the three different momentum ranges considered (from top to bottom, in ascending momentum order). From left to right: standard deviation hit energy, maximum distance hit-to-hit, and number of hits.

- MuID total energy (ClusterMuIDTotalEnergy): sum of the energy of all the MuID hits.
- Mean MuID hit energy (HitMuIDMeanEnergy): mean of the MuID hit energy distribution.
- Standard deviation MuID hit energy (HitMuIDStdEnergy): standard deviation of the MuID hit energy distribution.
- Maximum MuID hit energy (HitMuIDMaxEnergy): maximum of the MuID hit energy distribution.

• **Geometry-related MuID**

- Maximum distance MuID hit-to-hit (DistHitMuIDMax): maximum distance between pairs of MuID hits (not sure this is a good variable, distribution looks nuts).
- Maximum distance MuID hit-to-centre (DistHitCenterMuIDMax): maximum of the distance distribution between the MuID hits and the centre of the TPC.

- **Statistical MuID**
  - Number of hits (NHitsMuID): total number of MuID hits associated to the track.
  - Number of layers with hits (NLayersMuID): not really a count of all layers with MuID hits but the difference between the last and the first layer with MuIDhits.

Figure 8.21 shows the distributions of three different MuID variables, separating true muons (blue) and charged pions (red), for the three momentum ranges which use the muon tagger information. In this case I decided to standard deviation of the MuID hit energy distribution (left column), the maximum distance between the MuID hit pairs (middle column), and the number of MuID hits (right column). These variables are used together with the ECal features at high momenta, providing additional disambiguation power.

Once our features have been defined, one can do some exploratory analysis to understand how well the variables describe the target class, and avoid the black-box approach by what features are most relevant for the learning process. This way, I performed a feature analysis for each of the momentum ranges I divided this classification problem into. It follows three steps: first a principal component analysis (PCA), followed by a feature importance study using Gini and Shapley values, and finally a feature permutation importance analysis.

The PCA is useful to understand the variance of the feature space. It is an unsupervised machine learning technique that allows the user to perform a dimensionality

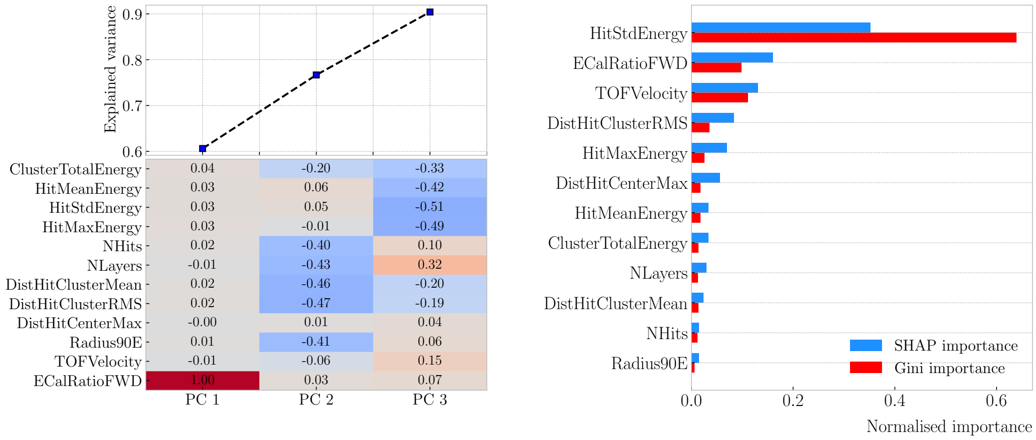


Figure 8.22: Left panel: cumulative explained variance for the first three principal components (top panel) and contribution of the different features to the principal axes in feature space (bottom panel). Right panel: Shapley (blue) and Gini (red) feature importances for the different input features. Both figures correspond to the samples in the momentum range $0.3 \leq p < 0.8$ GeV/c.

reduction. It uses a singular value decomposition of the input features to project them into a lower dimensional space. The idea is to find the matrix $\mathbf{C}_m$, whose columns are the first m orthonormal eigenvectors of the input covariance matrix. Consider the $n \times p$ real matrix of input data $\mathbf{X}$, where n is the number of samples and p the number of features. If $\mathbf{X}$ is centred, i.e. the means of its columns are equal to zero, we can write the covariance matrix of $\mathbf{X}$ as $\mathbf{C} = \mathbf{X}^\intercal\mathbf{X}/(n-1)$. This matrix can be diagonalised, yielding:

$$\mathbf{C} = \mathbf{V}\mathbf{L}\mathbf{V}^\intercal, \tag{8.11}$$

where $\mathbf{V}$ is a matrix of eigenvectors and $\mathbf{L}$ a diagonal matrix with eigenvalues λ_i. Then, performing SVD on $\mathbf{X}$ gives us:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{W}^\intercal, \tag{8.12}$$

where $\mathbf{U}$ is a unitary matrix, whose columns are called left singular vectors, $\mathbf{S}$ is a diagonal matrix of single values s_i, and $\mathbf{W}$ is another unitary matrix, its columns known as right singular vectors. This way, we can write:

$$\mathbf{C} = \mathbf{W}\mathbf{S}\mathbf{U}^\intercal\mathbf{U}\mathbf{S}\mathbf{W}^\intercal/(n-1) = \mathbf{W}\frac{\mathbf{S}^2}{n-1}\mathbf{W}^\intercal. \tag{8.13}$$

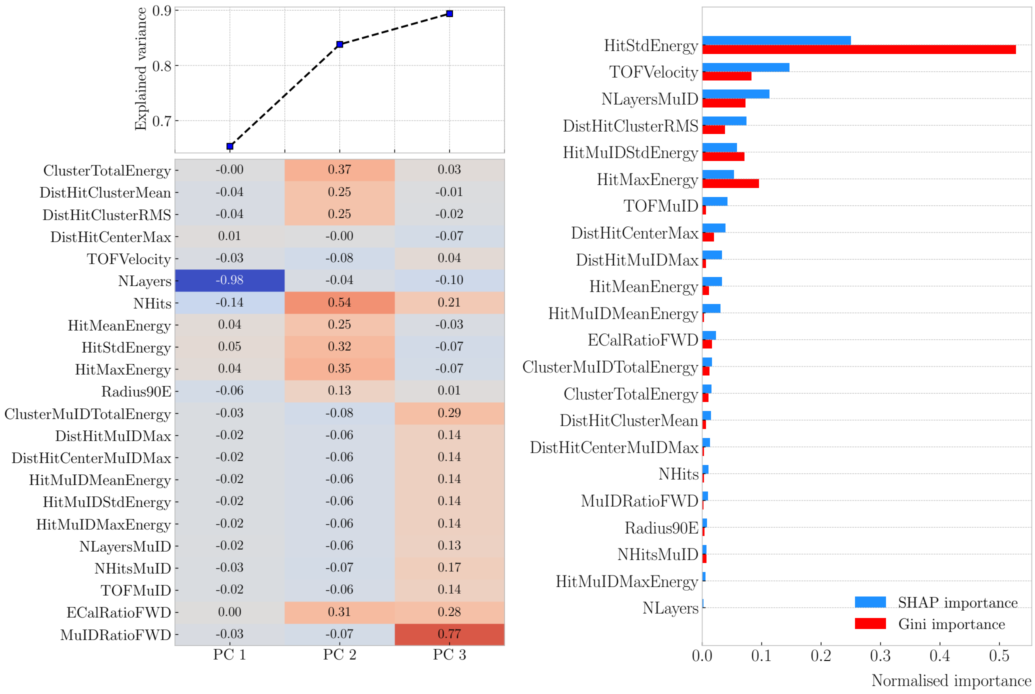


Figure 8.23: Left panel: cumulative explained variance for the first three principal components (top panel) and contribution of the different features to the principal axes in feature space (bottom panel). Right panel: Shapley (blue) and Gini (red) feature importances for the different input features. Both figures correspond to the samples in the momentum range $0.8 \leq p < 1.5$ GeV/c.

meaning that the right singular vectors are also the eigenvectors of the covariance matrix. The SVD can be computed numerically following an iterative approach.

This way, taking an input data vector $X \in \mathbb{R}^n$, the resulting feature vector $Y \in \mathbb{R}^m$ is given by:

$$Y = \mathbf{C}_m^{\intercal} X. \tag{8.14}$$

The new features capture most of the variance of the original sample, while being lower dimensional, as $m < n$.

Before applying the PCA reduction one needs to centre and scale the input data. Centring is necessary when using SVD to obtain the eigenvectors of the covariance matrix, as only in that case we can do the identification with the right singular vectors from the input data. Scaling is needed when variables are on different scales, as some can then dominate the PCA procedure.

I used the PCA module of `scikit-learn`, together with the `RobustScaler`, which centres the data and scales it based on the interquartile range. In Fig. 8.22 (left panel) and Fig. 8.23 (left panel) I show the results I obtained from the PCA for the momentum ranges $0.3 \leq p < 0.8$ GeV/c and $0.8 \leq p < 1.5$ GeV/c, respectively. Notice that in the second case the number of features increases considerably, as this is the first region which uses the MuID variables. I found that, in all the cases, adding a fourth PC does not add additional information. As it can be seen in the top panels of the figures, the cumulative explained variance is already over 80% with three PCs.

The bottom panels show the contribution of the variables to the principal axes. For the two first momentum regions, I observe a tendency of the energy-related and the geometry-related ECal variables to be clustered together. For the other ranges, when I include the MuID variables, there seems to be a division between ECal and MuID variables. For these, it seems like the number of ECal layers with hits also plays an important role.

The next step in the analysis is to quantify the importance of the features based on two additional metrics, namely the Gini and the Shapley values. The Gini importance, often called mean decrease impurity, is based on how much a feature contributes to the purity improvement at the splits in each decision tree. The purity is measured in terms of the Gini impurity index, defined as:

$$I_G = 1 - \sum_i f_i, \tag{8.15}$$

where f_i is the fractional abundance of the i-th class. Then, for each split one can compute the weighted decrease in impurity as:

$$\Delta_G = \frac{N_t}{N}\left(I_G - \frac{N_t^R}{N_t} I_G^R - \frac{N_t^L}{N_t} I_G^L\right), \tag{8.16}$$

where N represents the total number of samples, N_t the number of samples at the current node, N_t^R and N_t^L the number of samples in the right and left children respectively,

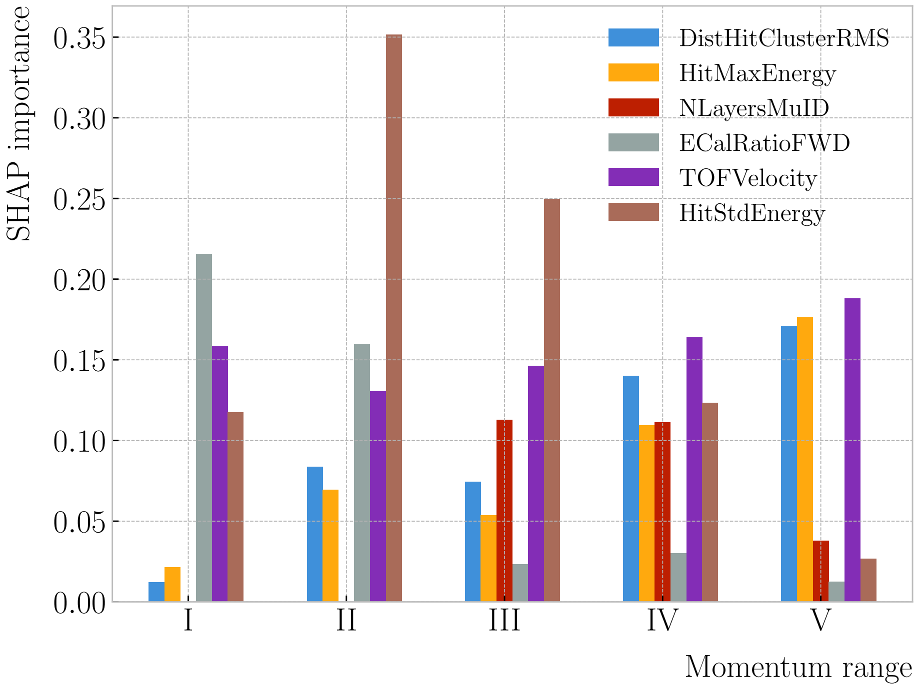


Figure 8.24: Evolution of the SHAP importance for the top six most important features across all five momentum ranges.

I_G is the Gini impurity at the current node, and I_G^R and I_G^L the Gini impurities of the resulting right and left children.

For each decision tree, one will have a normalised vector with the accumulated decrease in Gini impurity for each feature. In the case of a BDT, the feature importances are simply the mean for all the estimators in the ensemble[8].

The concept of Shapley values originated in the context of game theory, and it measures the marginal contribution of a feature in enhancing the accuracy of a classifier. Take F to be the set of all features in a problem, and $S \subseteq F$ the subset of features. To compute the Shapley value of the i-th feature, one has to train a model with that feature present, $f_{S\cup\{i\}}$, and another model trained without it, f_S. This has to be repeated for all possible combinations of subsets $S \subset F \setminus \{i\}$, and evaluating the models predictions

[8]Note to self: this appears not to be the case. If you get the `feature_importance` for each tree in the BDT and take the average, the result is not the same to be one reported in the `feature_importance` attribute of the BDT.

on the appropriate sets of data x_S. This way, the Shapley value results:

$$\varphi_i = \sum_{S \subset F \setminus \{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S) \right]. \tag{8.17}$$

I trained the `GradientBoostingClassifier` from `scikit-learn` with the default configuration in order to evaluate both the Gini and Shapley importances. The Gini scores are automatically computed by `scikit-learn`, using the training data. For the Shapley importance, I used the implementation from the `SHAP` package, computing it using the test sample. The results can be seen in Fig. 8.22 (right panel) and Fig. 8.23 (right panel), again for the momentum ranges $0.3 \leq p < 0.8$ GeV/c and $0.8 \leq p < 1.5$ GeV/c. The length of the bars denote either the SHAP (blue) or the Gini (red) importance of the feature. One interesting thing to notice is that, when looking at the Gini importance, there is always one feature that dominates over the rest. This is not the case for the SHAP importance, where importances tend to be more balanced.

Across all momentum ranges, I observe that the most important features are. For the five momentum ranges considered, only six variables sit in the top five at least once. Figure 8.24 shows the evolution of the SHAP importance of these six features. It is interesting to see that the time-of-flight variable keeps its importance almost unchanged for all momenta. Also, it looks like the ECal energy ratio gets less relevant the higher the momentum is, but the RMS of the hit-to-cluster distance distribution and the maximum ECal hit energy become more important in the last momentum ranges.

The last step in the feature selection analysis is the feature permutation. This technique measures the contribution of each feature to the performance of a model by randomly shuffling its values and checking how some scores degrade. For the present case, I am interested in the precision or purity, and the sensitivity or efficiency, as these two are the most relevant metrics from a physics point of view. The `scikit-learn` module provides the user with a method to perform the permutation scans.

The results of these are shown in Fig. 8.25. For the different momentum ranges I show the permutation importances for the ten most important features. For each

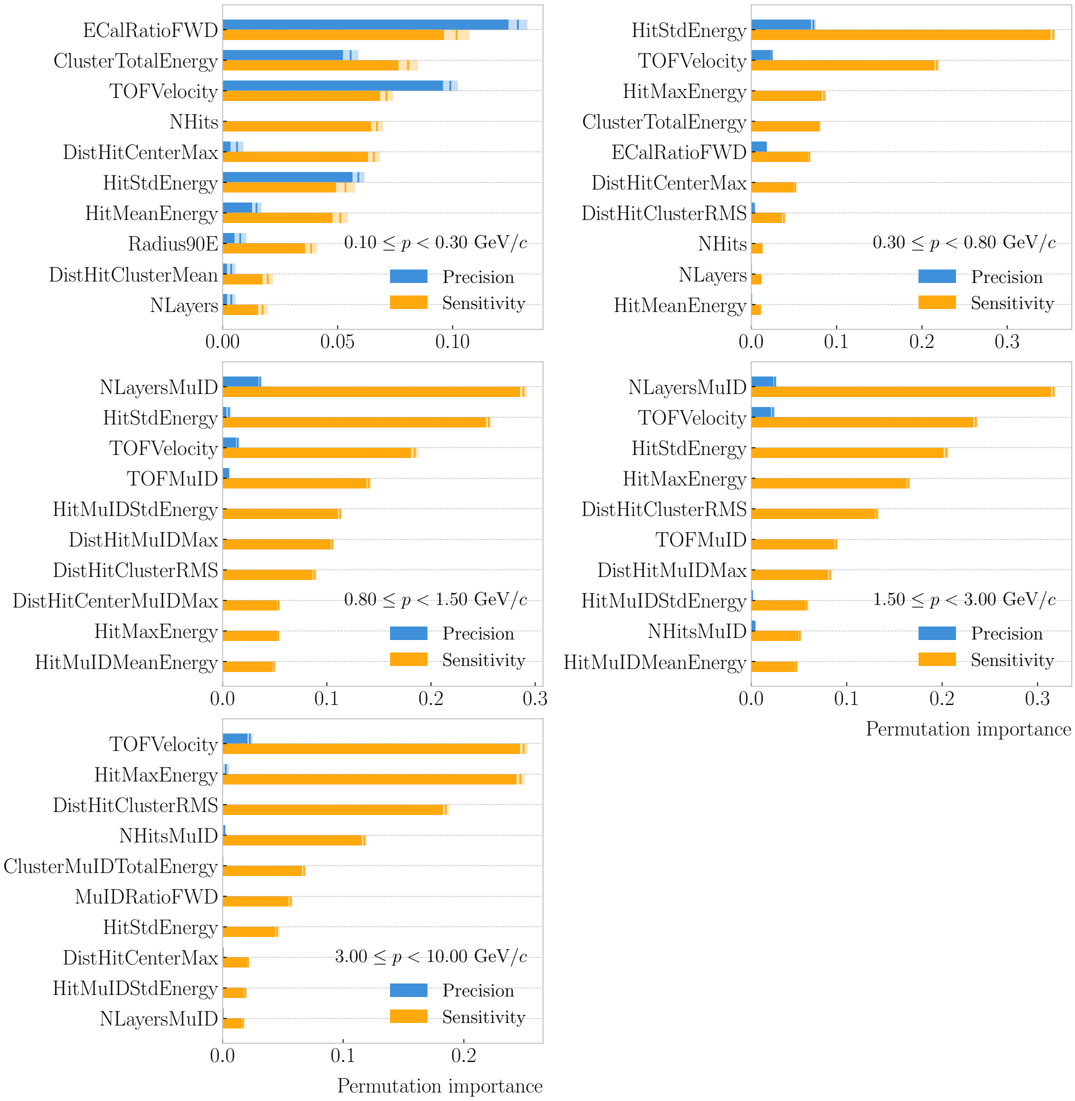


Figure 8.25: Permutation importances for the ten most important features in the different momentum ranges (from left to right, top to bottom, in increasing momentum order). The bars indicate the effect that permutations of each feature have on the purity (blue) and the sensitivity (yellow), the translucent regions representing one standard deviation around the central value.

of the variables I report the effect the permutations have on the precision (blue) and sensitivity (yellow) of the models. The bars indicate the importance value, with the lighter part representing one standard deviation around the mean (hinted as an additional vertical line). Something to notice is that, in the first momentum region, the feature permutations have an effect on both the precision and the sensitivity. However, for the rest the precision is almost unaffected, while the sensitivity changes are considerably larger.

It is also interesting to see that most of the variables identified as important here are the same I found when looking at the Shapley values. The behaviour of these across the momentum ranges is also similar, with the same patterns of some features being important at low momenta and then dropping in importance for the high momentum ranges.

Wit this, I conclude the study of the features. I have prepared the training and testing datasets and understood what features are likely to have the largest impact on the performance of the classifiers.

8.2.4 Hyperparameter optimisation

Any BDT requires the user to specify a number of parameters that will dictate its behaviour. They can be divided into two categories: (i) tree-specific parameters, which affect each individual tree in the model, and (ii) boosting parameters, which control the boosting operation in the model. The value of these so-called hyperparameters affect the performance and predictive power of the models. Therefore, one needs to carefully select their optimal values in order to extract as much information as possible from the data.

From all the parameters used to define a tree in the `scikit-learn` implementation of the BDT classifier, I only consider a subset of them. This is due to the fact that some are mutually exclusive, but also because I noticed that others have little effect on the problem at hand. Therefore, the parameters I investigate are the following:

- `min_samples_split`: defines the minimum number of samples required in a node to be considered for splitting. Higher values prevent a model from learning relations which might be highly specific to the particular sample, but may led to under-fitting if the value is too low.

- `min_samples_leaf`: defines the minimum samples required in a leaf node. For imbalanced problems it should take a low value, as there will not be many cases where the minority class dominates.

- `max_depth`: maximum depth of a tree. Useful to prevent over-fitting, as higher depth will allow a model to learn relations specific to the training sample.

In the case of the boosting parameters, the ones I look at are:

- `learning_rate`: determines the impact of each tree on the final outcome. Low values make the model robust to the specific characteristics of a tree, and thus allow it to generalise well. However, that usually requires a large number of trees to model the data properly.

- `n_estimators`: number of sequential trees to be trained. In general, BDTs are fairly robust at higher number of trees but it can still overfit at a point.

- `subsample`: fraction of observations to be selected for each tree. Values slightly less than 1 make the model robust by reducing the variance.

In general, hyperparameters depend on each other. Thus, it is not possible to optimise them independently. In the literature, we find two main strategies to explore the hyperparameter space. We could use a grid search, in which one discretises a portion of the space of hyperparameters and evaluates the model at each point. Another approach is the randomised search, where a certain number of random configurations of hyperparameters are explored.

In this case, I used the random search to scan the hyperparameter space. Also, because it is not guaranteed that a set of hyperparameters can be efficiently applied across different datasets, I perform the optimisation for each of the momentum ranges considered. Table 8.3 shows the list of hyperparameters considered, and the range within

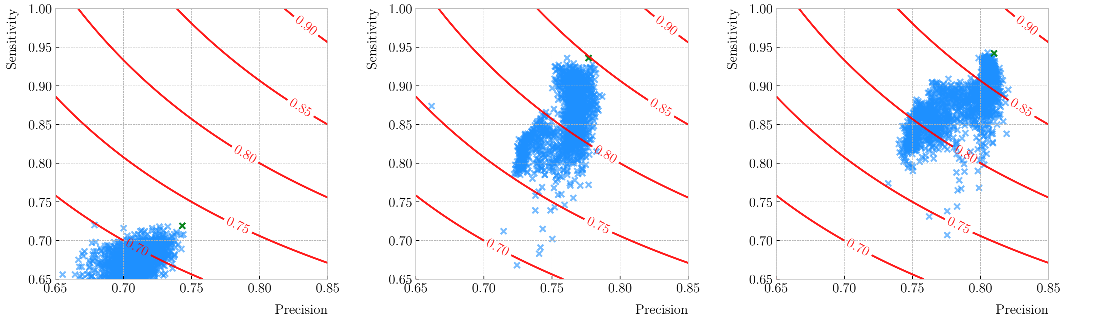


Figure 8.26: Values of the precision and sensitivity obtained for 10000 BDT hyperparameter configurations, for the momentum regions I, III and V. The red contours indicate the curves of equal F_1-score, while the green crosses are the selected configurations.

which I let them vary. I decided to fix the number of estimators to 400 in all cases, as its value is correlated with that of the learning rate.

I evaluate 10000 different hyperparameter configurations for each momentum range. For the hyperparameter tunning, I used subsamples containing 10% of the full datasets, keeping the original proportions between classes, in order to reduce the computational load. The performance of the models was assessed using a stratified 3-fold cross-validation with replacement. Cross-validation involves dividing the data in a number of subsets, training the model using some of them, and testing it with the rest. In our case, I divide the data in 3 equal-sized subsets, maintaining the class proportions of the original dataset. Then, for 3 consecutive iterations, I train the models using 2 of the subsets while I compute the precision and sensitivity scores with the other. This approach provides a more robust estimate of the performance on unseen data.

Figure 8.26 shows the results in the precision versus sensitivity plane, for the momentum regions I, III and V (from left to right). The contours represent the curves of equal F_1-score, i.e. the harmonic mean of the precision and the sensitivity. In order to select the optimal configurations (indicated in the plots with a green cross), I chose the point with the highest F_1-score.

The results for the different momentum ranges are summarised in Tab. 8.3. One can see some consistency in hyperparameter choices, with models generally preferring

Table 8.3: Optimal values of the hyperparameters used by the BDT, for each momentum range.

Hyperparameter	Range	Best value				
		I	II	III	IV	V
`min_samples_split`	$[0.001,\ 1]$	0.10	0.06	0.05	0.27	0.06
`min_samples_leaf`	$[0.001,\ 1]$	0.02	0.03	0.01	0.02	0.07
`max_depth`	$\{2,\ 3,\ ...,\ 8\}$	8	2	4	2	7
`learning_rate`	$[0.05,\ 1]$	0.10	0.23	0.07	0.13	0.09
`subsample`	$[0.01,\ 1]$	0.75	0.65	0.79	0.86	0.95

Table 8.4: Performance metrics of the BDTs with optimal hyperparameters, for the different momentum ranges.

Metric	Value $\pm$ 1σ				
	I	II	III	IV	V
Accuracy	0.779 ± 0.003	0.812 ± 0.003	0.846 ± 0.002	0.861 ± 0.003	0.874 ± 0.002
Precision	0.769 ± 0.003	0.752 ± 0.005	0.788 ± 0.002	0.805 ± 0.003	0.815 ± 0.003
Sensitivity	0.745 ± 0.009	0.921 ± 0.006	0.965 ± 0.002	0.967 ± 0.002	0.976 ± 0.001
F_1-score	0.757 ± 0.004	0.828 ± 0.003	0.867 ± 0.002	0.879 ± 0.002	0.889 ± 0.002
ROC AUC	0.868 ± 0.003	0.865 ± 0.003	0.899 ± 0.002	0.902 ± 0.002	0.911 ± 0.001

small values for the tree-specific parameters, small learning rate, and relatively large subsample sizes.

Now that I have obtained the optimal values of the hyperparameters, I can train the different BDTs. In this case I use the complete datasets, keeping 20% of the data for testing. Table 8.4 shows the values of the different performance metrics obtained using the selected hyperparameters and 5-fold cross-validation. The last row indicates the value of the area under the receiver operating characteristic (ROC) curve. This represents the sensitivity of a model as a function of the false positive rate. I have included it here as it is a classic model metric used in the machine learning community. Overall, there is a clear trend of models performing better at higher momentum.

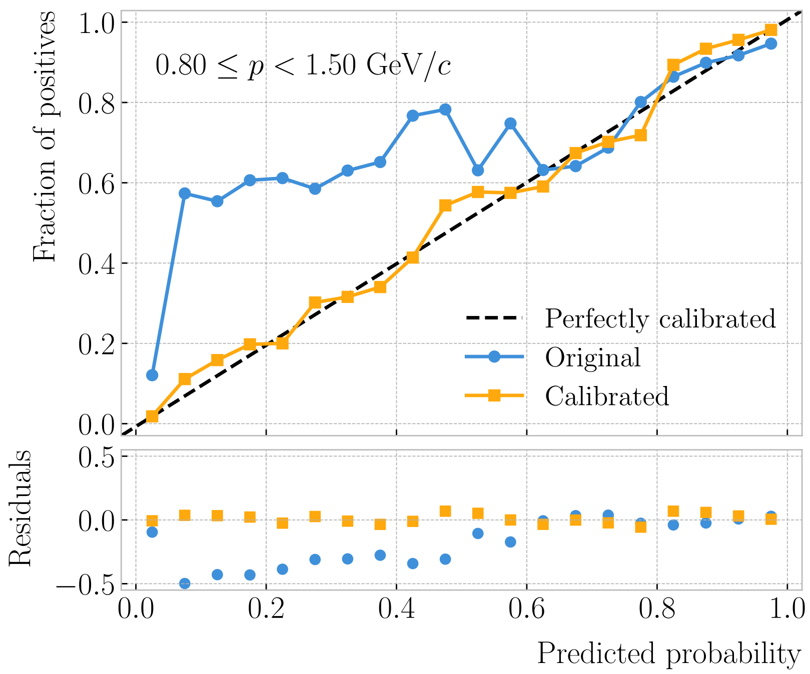


Figure 8.27: Reliability diagrams for the BDT classifier used in the momentum range $0.3 \leq p < 0.8$ GeV/c, both for the original (blue circles) and calibrated (yellow squares) responses. For reference, the response of a perfectly calibrated classifier is also shown (black dashed line).

3118
8.2.5 Probability calibration

3119 So far, the trained BDTs are able to provide predictions of the class labels. Ideally,
3120 one would like the output of a classifier to give a confidence level about the prediction.
3121 However, it is not straightforward to interpret the outputs of our BDTs in terms of
3122 probabilities.

3123 A way to visualise how well the predictions of a classifier are calibrated is using
3124 reliability diagrams [92]. They represent the probability of the positive label versus the
3125 probability predicted by the classifier. These can be obtained by binning the predicted
3126 probabilities, and then compute the conditional probability $P(y_{true} = 1 | y_i \leq y_{pred} <$
3127 $y_{i+1})$ by checking the fraction of true positive instances in each bin. The reliability
3128 diagram of a perfectly calibrated classifier would be a diagonal line.

3129 In this case, I try to correct the raw response of the classifiers by applying a sigmoid

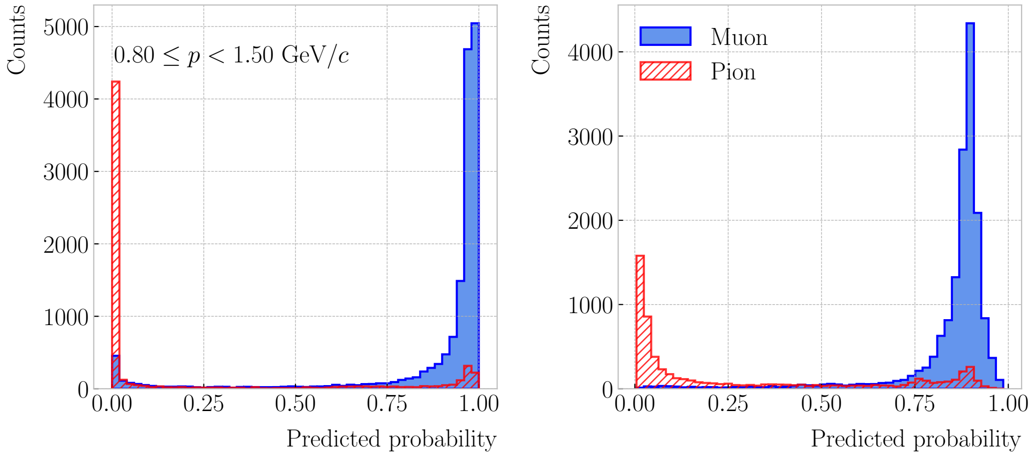


Figure 8.28: Uncalibrated (left panel) and calibrated (right panel) predicted probabilities assigned by the BDT classifiers for true muons (blue) and charged pions (red) in the momentum range $0.3 \leq p < 0.8$ GeV/c.

function:

$$\sigma(x;\ A, B) = \frac{1}{1 + e^{Ax+B}}, \tag{8.18}$$

where the parameters A and B are real numbers determined using the method of least squares.

For each classifier, I perform a grid search to obtain the optimal values of A and B. For any pair, I compute the predicted probabilities as $y_{pred} = \sigma(y_{raw};\ A, B)$, where y_{raw} are the raw predictions of the classifier[9]. Then, I calculate the corresponding reliability curve, and take the sum of the squared residuals between it and the response of the perfectly calibrated classifier.

Figure 8.27 shows the reliability diagrams for the original (blue) and calibrated (yellow) probability predictions of the classifier for the III momentum range, $0.3 \leq p < 0.8$ GeV/c. The original response of the classifier is given by $y_{pred} = \sigma(y_{raw};\ -2, 0)$, which is the transformation applied by `scikit-learn` to produce the probability estimate. Notice how the calibrated prediction matches the ideal response much better than the original, across all the probability range.

One can also compare the responses of the uncalibrated and calibrated classifiers

[9]In `scikit-learn` these correspond to the outputs of the `decision_function` method.

broken down by true particle type, as shown in Fig. 8.28. It can be seen that the distributions for both muons (blue) and charged pions (red) smoothen after calibration, but still the separating power of the classifier remains unchanged.

At this point, having the trained classifiers and the probability calibration parameters, I am able to assess the performance of the classification strategy in a physics-relevant case.

8.2.6 Performance

8.3 ECal time-of-flight

8.3.1 Arrival time estimations

8.3.2 Proton and pion separation

8.4 Charged pion decay in flight

As discussed previously, in GArSoft the TPC tracks are formed after a pattern recognition algorithm and a Kalman filter are applied to the TPC clusters. These two steps can find discontinuities in the track candidates (e.g. due to a particle decay) when these so-called breakpoints are large enough. However, for some, more subtle, cases they may miss them and form a single reconstructed track. It has been noted in the literature that Kalman filters offer, as a by-product, additional information to form test statistics to identify these breakpoints [?, ?].

Considering the mean life of the charged pion, $\tau = (2.6033 \pm 0.0005) \times 10^{-8}$ s, one can estimate that about 12% of the pions with momentum $p \sim \mathcal{O}(500\text{ MeV})$ (roughly the peak of the pion momentum distribution in ν_μ CC interactions off argon) decay inside the TPC. Figure 8.29 (left panel) shows the amount of charged pions decaying in the full TPC and fiducial volumes from an isotropic, monoenergetic sample of 10^5 negatively charged pions with $p = 500$ MeV. We see that about 10% of those decayed, with more than half of them decaying inside the TPC fiducial volume.

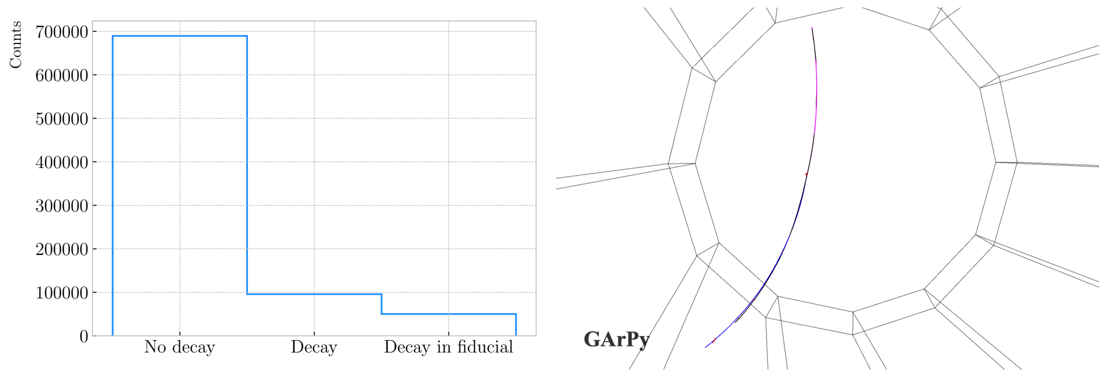


Figure 8.29: Left panel: number of non-decaying, decaying and decaying in the fiducial volume pions for a MC sample of 10^5, $p = 500$ MeV isotropic positively charged pions inside the TPC. Right panel: event display for a positive pion decaying inside the fiducial volume, with a single reconstructed track for the pion and muon system.

Figure 8.29 (right panel) shows an example event display of a charged pion (magenta line) decays in flight inside the TPC, but because the angle of the muon (blue line) is small both were reconstructed as one single track (black line). In this case, the composite track reaches the ECal, where it undergoes a muon-like interaction, thus being classified as a muon.

A way to understand what decaying pion tracks were totally or partially reconstructed together with the daughter muon is looking at the relative energy contributions to the reconstructed track. In order to select a sample of such events, I require that a minimum 50% of the total energy comes from the pion and at least 20% from the muon.

To identify potential decays we can use the information we obtain from the Kalman filter at each step of the fitted track. The simplest test we can think about is computing the χ^2 of the mismatch between all the parameters in the forward and the backward fits:

$$\chi_k^{2\ (FB)} = (\hat{\mathrm{x}}_k^B - \hat{\mathrm{x}}_k^F)^T [V^{(\hat{\mathrm{x}}_k,B)} + V^{(\hat{\mathrm{x}}_k,F)}]^{-1} (\hat{\mathrm{x}}_k^B - \hat{\mathrm{x}}_k^F), \tag{8.19}$$

where $\hat{\mathrm{x}}_k^F$, $\hat{\mathrm{x}}_k^B$ are the Kalman filter state vector estimates at step k in the forward and backward fits and $V^{(\hat{\mathrm{x}}_k,F)}$, $V^{(\hat{\mathrm{x}}_k,B)}$ the covariance matrices of $\hat{\mathrm{x}}_k^F$ and $\hat{\mathrm{x}}_k^B$ respectively. Using the values of the χ^2 at measurement k for the forward and backward fits we can

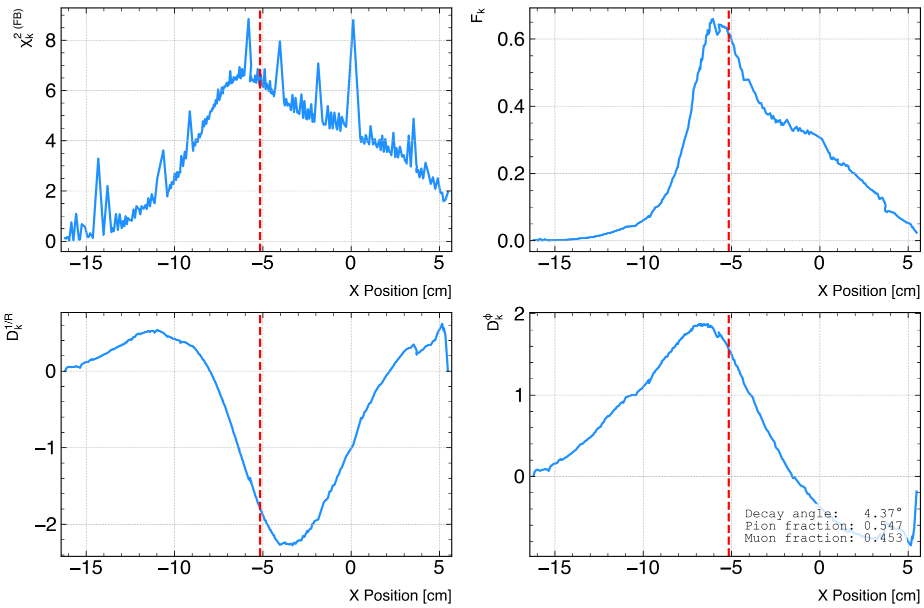


Figure 8.30: Values of $\chi_k^{2\ (FB)}$ (top left panel), F_k (top right panel), $D_k^{1/R}$ (bottom left panel) and D_k^{ϕ} (bottom right panel) versus position along the drift direction for a reconstructed track in a positive pion decay event. The vertical red dashed line indicates the true location of the decay point.

compute another χ^2 value that characterises the overall track fit:

$$\chi^2_{track} = \chi_k^{2\ (F)} + \chi_k^{2\ (B)} + \chi_k^{2\ (FB)}, \tag{8.20}$$

which remains approximately constant for all k.

An alternative approach proposed in the context of the NOMAD experiment was using a fit with a more elaborate breakpoint hypothesis, so we can perform a comparison of the χ^2 with and without breakpoints. This can be achieved by using some alternative parametrisation with extra parameters, which allows some of the track parameters to be discontinuous at certain points. A decay changes the momentum magnitude and direction, so we can use the new state vector:

$$\alpha = \Big(y, \quad z, \quad 1/R_F, \quad 1/R_B, \quad \phi_F, \quad \phi_B, \quad \tan\lambda_F, \quad \tan\lambda_B \Big)^T. \tag{8.21}$$

As we already have the estimates from the standard Kalman filter and their covariance matrices at each point, we do not need to repeat the Kalman fit for the new parametrisation. Instead, I can compute the values of α at each point k that minimise the χ^2 resulting from comparing them to $\{\hat{\mathrm{x}}_k^B, \hat{\mathrm{x}}_k^F\}$. Introducing the two 5×8 matrices:

$$H^F = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \; H^B = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{8.22}$$

we can write this as:

$$\begin{aligned} \chi_k^{2\;(FB)}(\alpha) = (\hat{\mathrm{x}}_k^F - H^F\alpha)^T \left[V^{(\hat{\mathrm{x}}_k,F)}\right]^{-1} (\hat{\mathrm{x}}_k^F - H^F\alpha) \\ + (\hat{\mathrm{x}}_k^B - H^B\alpha)^T \left[V^{(\hat{\mathrm{x}}_k,B)}\right]^{-1} (\hat{\mathrm{x}}_k^B - H^B\alpha). \end{aligned} \tag{8.23}$$

The minimum of $\chi_k^{2\;(FB)}(\alpha)$ is found when the measured new state vector takes the value:

$$\hat{\alpha}_k = V^{(\hat{\alpha}_k)} H^T (V^{(\hat{\mathrm{x}}_k)})^{-1} \hat{\mathrm{X}}, \tag{8.24}$$

where $\hat{\mathrm{X}} = \{\hat{\mathrm{x}}_k^B, \hat{\mathrm{x}}_k^F\}$, $V^{(\hat{\mathrm{x}}_k)}$ is the block diagonal matrix formed by $V^{(\hat{\mathrm{x}}_k,F)}$ and $V^{(\hat{\mathrm{x}}_k,B)}$ and $V^{(\hat{\alpha}_k)}$ is the covariance matrix of $\hat{\alpha}_k$, given by:

$$V^{(\hat{\alpha}_k)} = \left(H^T (V^{(\hat{\mathrm{x}}_k)})^{-1} H\right)^{-1}. \tag{8.25}$$

From these new fit estimates we can compute the F statistic, which tells us whether the model with breakpoint provides a statistically significant better fit:

$$F_k = \left(\frac{\chi^2_{track,k} - \chi^2_{full,k}}{8-5}\right) / \left(\frac{\chi^2_{full,k}}{N-8}\right). \tag{8.26}$$

One can also compute the signed difference of the duplicated variables divided by their standard deviation at each point. These represent how significant the discontinuity

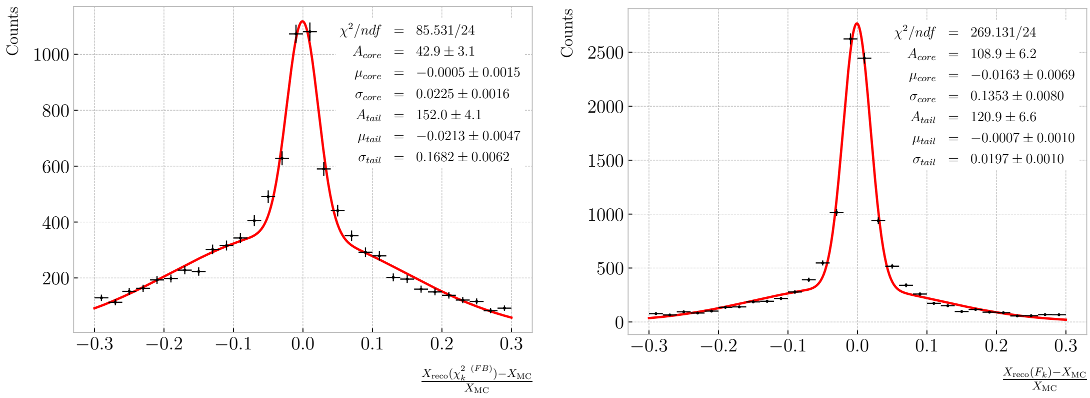


Figure 8.31: Fractional residual distributions of the true and reconstructed decay position along the drift coordinate, using the position of the maximum of $\chi_k^{2\,(FB)}$ (left panel) and F_k (right panel) as estimates of the decay position. Also shown are double Gaussian fits to these points (red lines).

in each variable is. For any variable η we can write it as:

$$D_k^{\eta} = \frac{\hat{\eta}_k^B - \hat{\eta}_k^F}{\sqrt{\mathrm{Var}[\hat{\eta}_k^F] + \mathrm{Var}[\hat{\eta}_k^B] - 2\mathrm{Cov}[\hat{\eta}_k^F, \hat{\eta}_k^B]}}. \tag{8.27}$$

In our case, the relevant ones to look at are $D_k^{1/R}$ and D_k^{ϕ}.

Figure 8.30 shows the values of $\chi_k^{2\,(FB)}$, F_k, $D_k^{1/R}$ and D_k^{ϕ} as functions of the position along the drift direction, for an example reconstructed track with 55.5% of the energy coming from the charged pion and 45.5% from the daughter muon. The true position of the decay is indicated (dashed red lines). Notice how $\chi_k^{2\,(FB)}$ and F_k, $D_k^{1/R}$ reach their maxima near the decay point. In the former case this indicates a large forward-backward difference in the track fit. In the later it represents that the extended state vector improves the fit particularly around that point.

I can estimate the decay position finding resolution by computing the difference between the X position of the maxima of $\chi_k^{2\,(FB)}$ and F_k and the X position of the true decay. Figure 8.31 represent the the fractional residual distributions for both cases, from the sample of tracks containing pion decays. Fitting a double Gaussian to the distributions (red lines) I find a resolution of 13.62% and 7.45% respectively.

In principle, the F-statistic should follow a Fisher distribution with $(8-5)$ and

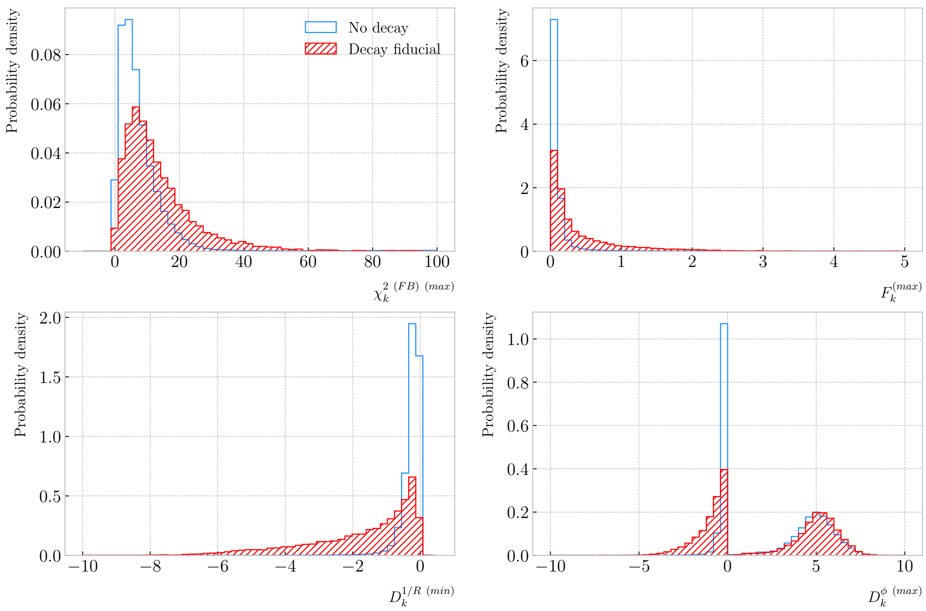


Figure 8.32: Distributions of the extreme values of $\chi_k^{2\ (FB)}$ (top left panel), F_k (top right panel), $D_k^{1/R}$ (bottom left panel) and D_k^{ϕ} (bottom right panel) for non-decaying reconstructed pion tracks (blue) and tracks which include the decay inside the fiducial volume (red).

$(N-8)$ degrees of freedom under the null hypothesis. In most of our cases $N \sim \mathcal{O}(100)$, so the probability density functions will look very similar. In this case, it is safe to take the limit $N \to \infty$ in the Fisher PDF:

$$\begin{aligned} \tilde{f}(x; a-b) &= \lim_{N\to\infty} f(x; a-b, N-a) \\ &= \frac{2^{-\frac{a-b}{2}}}{\Gamma\left(\frac{a-b}{2}\right)} (a-b)^{\frac{a-b}{2}}\, x^{\frac{a-b}{2}-1} \mathrm{e}^{-\frac{a-b}{2}x}. \end{aligned} \tag{8.28}$$

In our case $a-b = 8-5 = 3$, so we would obtain a p-value of 0.05 at $x = 2.60$.

Figure 8.32 contains the distributions of the maxima of $\chi_k^{2\ (FB)}$, F_k and D_k^{ϕ} and the minima of $D_k^{1/R}$ for a sample of non-decaying pion tracks (blue) and another sample of reconstructed tracks containing part of the pion and the daughter muon from a decay inside the fiducial volume (red). Notice that, even though the values of $F_k^{(max)}$ for the decay sample are typically larger than for the non-decaying one, just a small fraction of the events go beyond the aforementioned value of $F = 2.60$. Therefore, from a practical

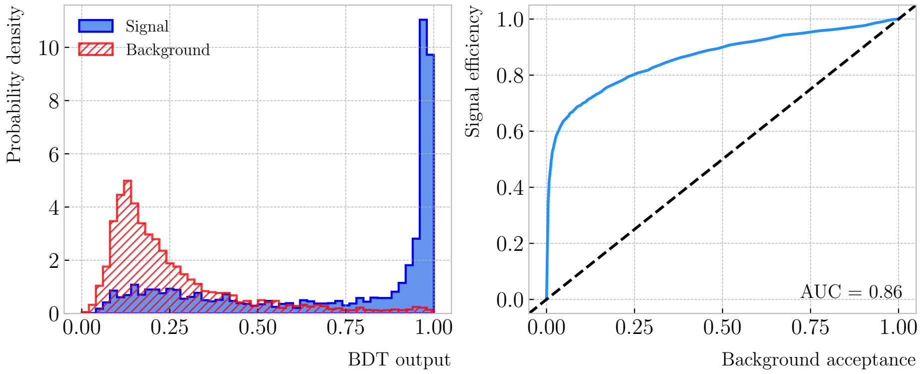


Figure 8.33: Left panel: distributions of the predicted probabilities assigned by the BDT classifier to a test sample of decaying pion+muon tracks (blue) and non-decaying pion tracks (red). Left: signal efficiency versus background acceptance (ROC curve) obtained from the BDT for the test sample.

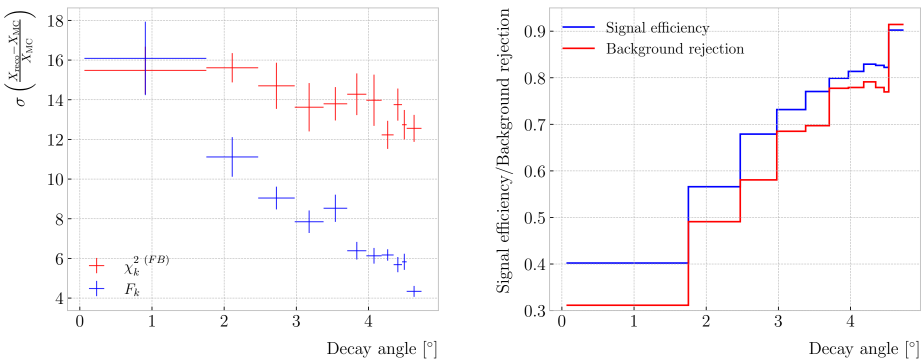


Figure 8.34: Left panel: dependence of the decay position finding resolution on the true value of the decay angle for the $\chi_k^{2\ (FB)}$ (red) and F_k (blue) methods. Right panel: signal efficiency (blue line) and background rejection (red line) from the BDT classifier versus true decay angle.

point of view, it is not the most efficient variable to use for selecting the decay events.

However, looking at the $D_k^{1/R\ (min)}$ distribution we can see there is a big difference between non-decaying and decaying events in this variable. One can use a combination of these four variables to distinguish between the pion decay events (signal) and the non-decaying pions (background).

An approach to this classification could be using a boosted decision tree (BDT). One of the advantages of BDTs is that they are easy to interpret and identify the relative importance of the different input variables. Training a BDT with 400 estimators and a maximum depth of 4 I can obtain an efficient classification without overtraining. Figure

8.33 (left panel) shows the distribution of probabilities predicted by the BDT for a test sample. The signal efficiency as a function of background acceptance, the so-called ROC curve, is shown in Fig. 8.33 (right panel). With a relative importance of 0.83, the most important variable turned out to be $D_k^{1/R\ (min)}$.

One thing we can check is how the resolution to the decay and the signal efficiency in the classification changes with the true decay angle. Using an equal-frequency binning for the decay angles, we can repeat the previous steps for each bin.

Figure 8.34 (left panel) shows the dependence on the decay angle of the decay finding resolution. We can see that for the $\chi_k^{2\ (FB)}$ maximum location method the resolution consistently lies between 12 to 16%. However, the $F_k^{(max)}$ approach gives a significantly better resolution for high angle values, reaching the $4-6\%$ range for decay angles $\geq 4°$.

For the classification dependence on the angle, I use the same classifier I trained before but evaluating the test sample for each individual angular bin. I compute the signal efficiency in each bin for a fixed value of the background rejection, in this case 90%. Similarly, for the background rejection estimation I use a fixed signal efficiency value of 90%. Figure 8.34 (right panel) represents the change in signal efficiency (blue) and background rejection (red) with the value of the true decay angles.

8.4.1 Track breakpoints

8.5 Neutral particle identification

8.5.1 ECal clustering

Another important reconstruction item is the clustering algorithm of ECal hits in GArSoft. The default module features a NN algorithm that treats all hits in the same way, independently of the layer each hit comes from. However, the current ECal design of ND-GAr has two very different types of scintillator layers. The inner layers are made out of tiles, which provide excellent angular and timing resolutions. On the other hand, the outer layers are cross scintillator strips. That way, an algorithm that treats hits from both kinds of layers differently may be able to improve the current performance.

Inspired by the reconstruction of T2K's ND280 downstream ECal [**?**], the idea was to put together a clustering module that first builds clusters for the different ECal views (tiles, strips segmented in the X direction and strips segmented in Y direction), and then tries to match them together to form the final clusters.

Working on a module-by-module basis, the algorithm first separates the hits depending on the layer type they come from. Then, it performs a NN clustering for the 3 sets of hits separately. For the tile hits it clusters together all the hits which are in nearest-neighbouring tiles and nearest-neighbouring layers, for strip hits it looks at nearest-neighbouring strips and next-to-nearest-neighbouring layers (as the layers with strips along the two directions are alternated). For strip clusters an additional cut in the direction along the strip length is needed.

After this first clustering I then apply a recursive re-clustering for each collection of strip clusters based on a PCA method. In each case, we loop over the clusters with $N_{hits} \geq 2$, computing the centre of mass and three principal components. Propagating these axes up to the layers of the rest of the clusters, we check if the propagated point and the centre of mass of the second cluster are within next-to-nearest-neighbouring strips. An additional cut in the direction along the strip length is also needed. Moreover, I require that the two closest hits across the two clusters are at most in next-to-nearest-neighbouring strips. I merge the clusters if these three conditions are satisfied. The re-clustering is repeated until no more cluster pairs pass the cuts.

The clusters in each strip view are combined if their centres of mass are close enough and they point in the same direction. An alternative approach for the strip cluster merging could be to compute the overlap between the ellipsoids defined by the principal axes of the clusters, and then merge the pair if the overlap exceeds some threshold. Further study is needed to understand if this change would have an impact in the overall clustering performance.

To merge the tile clusters to the combined strip clusters I propagate the principal axis of the strip cluster towards the inner layers, up to the centre of mass layer of the tile cluster. I merge the clusters if the distance between the propagated point and the

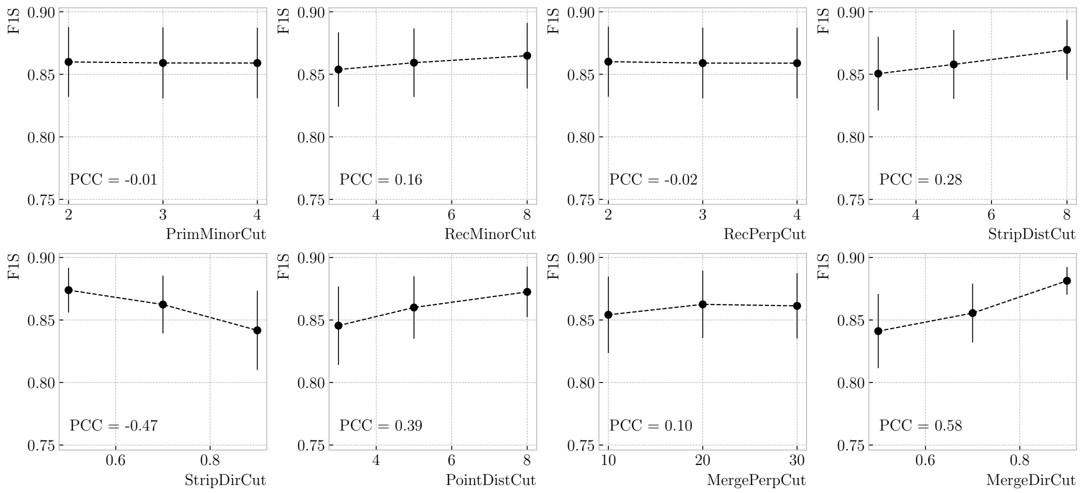


Figure 8.35: Mean values of the F_1-score marginal distributions for the different free parameters of the new clustering algorithm, with the error bars representing one standard deviation around the mean. The F_1-score values were computed for the 6561 possible parameter configurations using 1000 ν_μ CC interaction events.

centre of mass is bellow a certain cut.

The last step is to check if clusters in neighbouring modules should be merged together, both across two barrel modules, across end cap modules and between barrel end cap modules. I check the distance between the two closest hits in the pair of clusters and merge them if it passes this and an additional direction cut.

Figure **??** presents an example of the clustering steps relevant for strip layer hits, from the input hits (top left panel) to the NN clustering (top right panel) and re-clustering (bottom left panel) for each strip view and the final merging strip clusters (bottom right panel). It shows the hits from a single ECal barrel module in a ν_μ CC interaction event with a neutral pion and a proton in the final state. The two clusters on the left correspond to the photon pair from the π^0 decay and the one on the upper right corner is associated to the proton.

This algorithm has a total number of eight free parameters that need to be optimised. I used a sample of 1000 ν_μ CC interactions in order to obtain the optimal configuration of clustering parameters. This sample was generated up to the default ECal hit clustering level, so then I could run the new clustering algorithm each time with a different

Table 8.5: Summary of parameters and sampled values used in the optimisation of the clustering algorithm.

Name	Units	Sampled values	Description
PrimMinorCut	strips	2, 3, 4	Distance along strip length in NN clustering
RecMinorCut	strips	3, 5, 8	Distance between propagated point and CM along strip length in re-clustering
RecPerpCut	strips	2, 3, 4	Closest hit pair distance in re-clustering
StripDistCut	strips	3, 5, 8	Distance between CMs in strip cluster merging
StripDirCut	cos	0.5, 0.7, 0.9	Main axes direction cut in strip cluster merging
PointDistCut	tiles	3, 5, 8	Distance between propagated point and CM in strip-tile matching
MergePerpCut	cm	10, 20, 30	Closest hit pair distance in module merging
MergeDirCut	cos	0.5, 0.7, 0.9	Main axes direction cut in module merging

configuration of parameters. As the number of parameters is relatively large, I only performed a coarse-grained scan of the parameter space. Sampling each of the eight parameters at three different points each I obtain 6561 different configurations. These parameters, together with the used values, are summarised in Tab. 8.5.

In order to measure the performance of the clustering, I use a binary classification approach. For each formed cluster, I identify the Geant4 Track ID of the matching MC particle and the energy fraction of each hit. Then, I assign to each cluster the Track ID with the highest total energy fraction. For each of the different Track IDs associated to the clusters, I select the cluster with the highest energy (only from the hits with the same Track ID). I identify such a cluster as the main cluster for that Track ID. I count as true positives (TPs) the hits with the correct Track ID in each main cluster. False positives (FPs) are the hits with the incorrect Track ID for the cluster they are in, not only main clusters. The false negatives (FNs) are the hits with the correct Track ID in clusters other than the main.

Figure 8.35 shows the computed F_1-score values for the different cuts. In each case, the central value represents the mean of the F_1-score distribution for the specified value of the corresponding variable and the vertical error bar represents one standard deviation

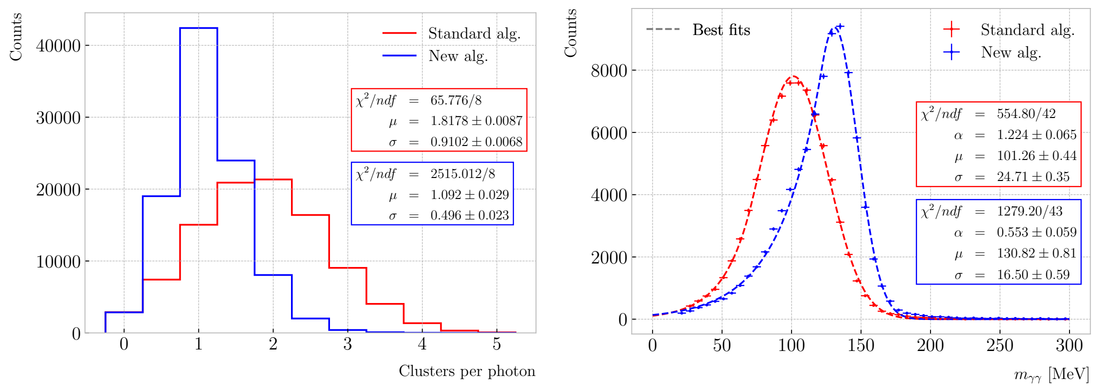


Figure 8.36: Left panel: distributions of the number of ECal clusters per photon from π^0 decays for the standard (red) and new (blue) clustering algorithms. Right panel: reconstructed invariant mass distributions for photon pairs from single π^0 events using the standard (red) and new (blue) ECal clustering algorithms.

around the mean. Also shown are the Pearson correlation coefficients of these central values. We can see that five of the variables have a sizeable effect on the F_1-score, with an absolute difference between the last and first values as big as 4%.

The working configuration is obtained as follows. I first select all configurations with purity $\geq 90\%$. Among those, I choose the combinations that yield the maximum F_1-score. If more than one configuration remains I select the one with the highest sensitivity. Doing so, I end up with a parameter configuration with an efficiency of 88% and a 90% purity. Compared with the default algorithm, which gives an efficiency of 76% and a purity of 91% for the same sample, I have managed to improve the efficiency by a factor of 1.16.

8.5.2 π^0 reconstruction

One of the potential applications of the new ECal hit clustering is the reconstruction of neutral particles, in particular pions. Neutral pions decay promptly after being produced, through the $\pi^0 \to \gamma\gamma$ channel $(98.823 \pm 0.034)\%$ of the time. The photon pair does not leave any traces in the HPgTPC (unless one or both of them converts into an electron-positron pair), but each of them will produced an electromagnetic shower in the ECal.

To test the potential impact of the new algorithm in π^0 reconstruction, I generated a MC sample of single, isotropic neutral pions inside the HPgTPC. All pions were generated with a momentum $p = 500$ MeV and their initial positions were uniformly sampled inside a $2 \times 2 \times 2$ m box aligned with the centre of the TPC. I ran both the default and the new clustering algorithms, using for the latter the optimised configuration discussed above.

The first thing to notice is that the number of clusters produced per photon has decreased. Figure 8.36 (left panel) shows these distributions for the default (red) and new (blue) algorithms. Using a simple Gaussian fit, we see that the mean number of ECal clusters per photon went from 1.82 ± 0.01 to 1.09 ± 0.03. This effectively means that with the new algorithm the ECal activity of one true particle is typically reconstructed as a single object. From the reconstruction point of view this can be an advantage. As now most of the photon energy ends up in a single ECal cluster, I can simply use cluster pairs to identify the π^0 decay.

In general, one calculates the invariant mass of the photon pair as:

$$m_{\gamma\gamma} = \sqrt{2E_1E_2(1 - \cos\,\theta)}, \tag{8.29}$$

where E_i are the energies of the photons and θ the opening angle between them. In this case I can use the energies deposited in the ECal and their incident directions. This quantity is computed for all possible pairs of clusters, using their position together with the true decay point. In a more realistic scenario, e.g. ν_μ CC interaction, one could use the position of the reconstructed primary vertex instead. I also tried to use the principal direction of the clusters, but that approach gave considerably worse results. For each event I only keep the pair with an invariant mass closer to the true π^0 mass value.

Figure 8.36 (right panel) shows the invariant mass distributions for the photon pairs we get using the default (red) and the new (blue) ECal clustering algorithms. For the fit I used a modified version of the Crystal Ball function [?], obtained by taking the limit where the parameter controlling the power-law tail goes to infinity:

$$f(x; N, \mu, \sigma, \alpha) = N \cdot \begin{cases} \mathrm{e}^{\frac{\alpha(2x-2\mu+\alpha\sigma)}{2\sigma}}; & x \leq \mu - \alpha\sigma, \\ \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}; & x > \mu - \alpha\sigma. \end{cases} \tag{8.30}$$

Comparing the fitted mean and standard deviation values for the Gaussian cores, we see that the distribution for the new algorithm is a 67% narrower and also peaks much closer to the true m_{π^0} value, going from 101.3 ± 0.4 MeV to 130.8 ± 0.6 MeV.

Chapter 9

Event selection in ND-GAr

9.1 CAFs and CAFAna

9.2 Event selection

9.2.1 ν_μ CC selection

9.2.2 Charged pion multiplicity

Chapter 10

Conclusions

Appendix A

An appendix

Bibliography

[1] DUNE collaboration, *Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume I Introduction to DUNE*, *JINST* **15** (2020) T08008 [2002.02967]. 15, 45, 46, 47, 50, 51, 52, 58, 59, 60, 67, 68

[2] DUNE collaboration, *Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume IV: Far Detector Single-phase Technology*, *JINST* **15** (2020) T08010 [2002.03010]. 15, 55

[3] J.N. Bahcall, A.M. Serenelli and S. Basu, *New solar opacities, abundances, helioseismology, and neutrino fluxes*, *Astrophys. J. Lett.* **621** (2005) L85 [astro-ph/0412440]. 21, 115, 117

[4] R. Garani and S. Palomares-Ruiz, *Dark matter in the Sun: scattering off electrons vs nucleons*, *JCAP* **05** (2017) 007 [1702.02768]. 21, 113, 114, 116, 117, 118, 141

[5] M. Honda, M. Sajjad Athar, T. Kajita, K. Kasahara and S. Midorikawa, *Atmospheric neutrino flux calculation using the NRLMSISE-00 atmospheric model*, *Phys. Rev. D* **92** (2015) 023004 [1502.03916]. 22, 121, 122

[6] M. Colom i Bernadich and C. Pérez de los Heros, *Limits on Kaluza-Klein dark matter annihilation in the Sun from recent IceCube results*, *Eur. Phys. J. C, 80 2 (2020) 129* **80** (2019) [1912.04585]. 22, 126

[7] ANTARES collaboration, *Search for Dark Matter in the Sun with the ANTARES Neutrino Telescope in the CMSSM and mUED frameworks*, *Nucl. Instrum. Meth. A* **725** (2013) 76 [1204.5290]. 22, 126

[8] N. Deutschmann, T. Flacke and J.S. Kim, *Current LHC Constraints on Minimal Universal Extra Dimensions*, *Phys. Lett. B* **771** (2017) 515 [1702.00410]. 22, 126, 127

[9] IceCube collaboration, *Search for GeV-scale dark matter annihilation in the Sun with IceCube DeepCore*, *Phys. Rev. D* **105** (2022) 062004 [2111.09970]. 24, 138, 139

[10] C.-S. Chen, F.-F. Lee, G.-L. Lin and Y.-H. Lin, *Probing Dark Matter Self-Interaction in the Sun with IceCube-PINGU*, *JCAP* **10** (2014) 049 [1408.5471]. 24, 138, 139

[11] N.F. Bell, M.J. Dolan and S. Robles, *Searching for dark matter in the Sun using Hyper-Kamiokande*, *JCAP* **11** (2021) 004 [2107.04216]. 24, 138, 139

[12] E. Behnke et al., *Final Results of the PICASSO Dark Matter Search Experiment*, *Astropart. Phys.* **90** (2017) 85 [1611.01499]. 24, 138, 139

[13] PICO collaboration, *Dark Matter Search Results from the Complete Exposure of the PICO-60 C_3F_8 Bubble Chamber*, *Phys. Rev. D* **100** (2019) 022001 [1902.04031]. 24, 138, 139

[14] DarkSide collaboration, *Constraints on Sub-GeV Dark-Matter–Electron Scattering from the DarkSide-50 Experiment*, *Phys. Rev. Lett.* **121** (2018) 111303 [1802.06998]. 24, 142, 143

[15] XENON collaboration, *Light Dark Matter Search with Ionization Signals in XENON1T*, *Phys. Rev. Lett.* **123** (2019) 251801 [1907.11485]. 24, 142, 143

[16] P.F. de Salas, D.V. Forero, S. Gariazzo, P. Martínez-Miravé, O. Mena, C.A. Ternes et al., *2020 global reassessment of the neutrino oscillation picture*, *JHEP* **02** (2021) 071 [2006.11237]. 31, 42, 47

[17] W. Pauli, *Dear radioactive ladies and gentlemen*, *Phys. Today* **31N9** (1978) 27. 37

[18] F. Reines and C.L. Cowan, *Detection of the free neutrino*, *Phys. Rev.* **92** (1953) 830. 37

[19] ALEPH, DELPHI, L3, OPAL, SLD, LEP Electroweak Working Group, SLD Electroweak Group, SLD Heavy Flavour Group collaboration, *Precision electroweak measurements on the Z resonance*, *Phys. Rept.* **427** (2006) 257 [hep-ex/0509008]. 38

[20] Super-Kamiokande collaboration, *Evidence for oscillation of atmospheric neutrinos*, *Phys. Rev. Lett.* **81** (1998) 1562 [hep-ex/9807003]. 38

[21] B. Pontecorvo, *Inverse beta processes and nonconservation of lepton charge*, *Zh. Eksp. Teor. Fiz.* **34** (1957) 247. 38

[22] Z. Maki, M. Nakagawa and S. Sakata, *Remarks on the unified model of elementary particles*, *Prog. Theor. Phys.* **28** (1962) 870. 38

[23] L. Wolfenstein, *Neutrino Oscillations in Matter*, *Phys. Rev. D* **17** (1978) 2369. 40

[24] B.T. Cleveland, T. Daily, R. Davis, Jr., J.R. Distel, K. Lande, C.K. Lee et al., *Measurement of the solar electron neutrino flux with the Homestake chlorine detector*, *Astrophys. J.* **496** (1998) 505. 41

[25] F. Kaether, W. Hampel, G. Heusser, J. Kiko and T. Kirsten, *Reanalysis of the GALLEX solar neutrino flux and source experiments*, *Phys. Lett. B* **685** (2010) 47 [1001.2731]. 41

[26] SAGE collaboration, *Measurement of the solar neutrino capture rate with gallium metal. III: Results for the 2002–2007 data-taking period*, *Phys. Rev. C* **80** (2009) 015807 [0901.2200]. 41

[27] G. Bellini et al., *Precision measurement of the 7Be solar neutrino interaction rate in Borexino*, *Phys. Rev. Lett.* **107** (2011) 141302 [1104.1816]. 41

[28] Super-Kamiokande collaboration, *Solar neutrino measurements in super-Kamiokande-I*, *Phys. Rev. D* **73** (2006) 112001 [hep-ex/0508053]. 41

[29] SNO collaboration, *Combined Analysis of all Three Phases of Solar Neutrino Data from the Sudbury Neutrino Observatory*, *Phys. Rev. C* **88** (2013) 025501 [1109.0763]. 41

[30] Super-Kamiokande collaboration, *Atmospheric neutrino oscillation analysis with external constraints in Super-Kamiokande I-IV*, *Phys. Rev. D* **97** (2018) 072001 [1710.09126]. 41

[31] IceCube collaboration, *Measurement of Atmospheric Neutrino Oscillations at 6–56 GeV with IceCube DeepCore*, *Phys. Rev. Lett.* **120** (2018) 071801 [1707.07081]. 41

[32] KamLAND collaboration, *Reactor On-Off Antineutrino Measurement with KamLAND*, *Phys. Rev. D* **88** (2013) 033001 [1303.4667]. 41

[33] RENO collaboration, *Measurement of Reactor Antineutrino Oscillation Amplitude and Frequency at RENO*, *Phys. Rev. Lett.* **121** (2018) 201801 [1806.00248]. 41

[34] Daya Bay collaboration, *Measurement of the Electron Antineutrino Oscillation with 1958 Days of Operation at Daya Bay*, *Phys. Rev. Lett.* **121** (2018) 241805 [1809.02261]. 41

[35] K.J. Kelly, P.A.N. Machado, S.J. Parke, Y.F. Perez-Gonzalez and R.Z. Funchal, *Neutrino mass ordering in light of recent data*, *Phys. Rev. D* **103** (2021) 013004. 42

[36] P. Dunne, "Latest Neutrino Oscillation Results from T2K." Jul, 2020. 42

[37] MINOS collaboration, *Combined analysis of ν_μ disappearance and $\nu_\mu \to \nu_e$ appearance in MINOS using accelerator and atmospheric neutrinos*, *Phys. Rev. Lett.* **112** (2014) 191801 [1403.0867]. 42

[38] K2K collaboration, *Measurement of Neutrino Oscillation by the K2K Experiment*, *Phys. Rev. D* **74** (2006) 072003 [hep-ex/0606032]. 42

[39] DUNE collaboration, *Long-baseline neutrino oscillation physics potential of the DUNE experiment*, *Eur. Phys. J. C* **80** (2020) 978 [2006.16043]. 42

[40] Hyper-Kamiokande collaboration, *Physics potential of Hyper-Kamiokande for neutrino oscillation measurements*, *PoS* **NuFact2019** (2019) 040. 42

[41] DUNE collaboration, *Snowmass Neutrino Frontier: DUNE Physics Summary*, 2203.06100. 47, 48

[42] DUNE collaboration, *Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume II: DUNE Physics*, 2002.03005. 47, 49, 63, 65, 121, 130, 134

[43] Super-Kamiokande collaboration, *Search for Proton Decay via $p \to e^+\pi_0$ and $p \to \mu^+\pi_0$ in a Large Water Cherenkov Detector*, *Phys. Rev. Lett.* **102** (2009) 141801 [0903.0676]. 48

[44] S. Raby, *Grand Unified Theories*, in *2nd World Summit: Physics Beyond the Standard Model*, 8, 2006 [hep-ph/0608183]. 48

[45] Kamiokande-II collaboration, *Observation of a Neutrino Burst from the Supernova SN 1987a*, *Phys. Rev. Lett.* **58** (1987) 1490. 49

[46] R.M. Bionta et al., *Observation of a Neutrino Burst in Coincidence with Supernova SN 1987a in the Large Magellanic Cloud*, *Phys. Rev. Lett.* **58** (1987) 1494. 49

[47] DUNE collaboration, *The DUNE Far Detector Vertical Drift Technology, Technical Design Report*, 2312.03130. 53, 54

[48] DUNE collaboration, *Deep Underground Neutrino Experiment (DUNE) Near Detector Conceptual Design Report*, *Instruments* **5** (2021) 31 [2103.13910]. 57, 58, 61, 65, 145

[49] J. Strait, E. McCluskey, T. Lundin, J. Willhite, T. Hamernik, V. Papadimitriou et al., *Long-baseline neutrino facility (lbnf) and deep underground neutrino experiment (dune) conceptual design report volume 3: Long-baseline neutrino facility for dune june 24, 2015*, `1601.05823`. 62

[50] DUNE DAQ, "dtp-firmware." `https://gitlab.cern.ch/dune-daq/readout/dtp-firmware`, 2020. 77

[51] DUNE DAQ, "dtp-simulation." `https://gitlab.cern.ch/dune-daq/readout/dtp-simulation`, 2020. 80

[52] DUNE DAQ, "dtpemulator." `https://github.com/DUNE-DAQ/dtpemulator/tree/fmlopez/filter_ana`, 2022. 80

[53] J. McClellan and T. Parks, *A personal history of the Parks-McClellan algorithm*, *IEEE Signal Processing Magazine* **22** (2005) 82. 81

[54] G. Turin, *An introduction to matched filters*, *IEEE Transactions on Information Theory* **6** (1960) 311. 84

[55] J.W. Goodman, *Statistical Optics*, Wiley (1985). 86

[56] B. Dwork, *Detection of a pulse superimposed on fluctuation noise*, *Proceedings of the IRE* **38** (1950) 771. 87

[57] L. Wainstein and V. Zubakov, *Extraction of Signals from Noise*, Prentice-Hall (1962). 87

[58] E.D. Church, *Larsoft: A software package for liquid argon time projection drift chambers*, `1311.6774`. 90, 91

[59] S.V. Stehman, *Selecting and interpreting measures of thematic classification accuracy*, *Remote Sensing of Environment* **62** (1997) 77. 104

[60] A.A. Taha and A. Hanbury, *Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool*, *BMC Medical Imaging* **15** (2015) . 105

[61] J. Silk, K.A. Olive and M. Srednicki, *The Photino, the Sun and High-Energy Neutrinos*, *Phys. Rev. Lett.* **55** (1985) 257. 111

[62] M. Srednicki, K.A. Olive and J. Silk, *High-Energy Neutrinos from the Sun and Cold Dark Matter*, *Nucl. Phys. B* **279** (1987) 804. 111

[63] J.S. Hagelin, K.W. Ng and K.A. Olive, *A High-energy Neutrino Signature From Supersymmetric Relics*, *Phys. Lett. B* **180** (1986) 375. 111

[64] T.K. Gaisser, G. Steigman and S. Tilav, *Limits on Cold Dark Matter Candidates from Deep Underground Detectors*, *Phys. Rev. D* **34** (1986) 2206. 111

[65] N. Bernal, J. Martín-Albo and S. Palomares-Ruiz, *A novel way of constraining WIMPs annihilations in the Sun: MeV neutrinos*, *JCAP* **08** (2013) 011 [1208.0834]. 111, 112, 113, 119

[66] C. Rott, J. Siegal-Gaskins and J.F. Beacom, *New Sensitivity to Solar WIMP Annihilation using Low-Energy Neutrinos*, *Phys. Rev. D* **88** (2013) 055005 [1208.0827]. 111, 118

[67] C. Rott, S. In, J. Kumar and D. Yaylali, *Dark Matter Searches for Monoenergetic Neutrinos Arising from Stopped Meson Decay in the Sun*, *JCAP* **11** (2015) 039 [1510.00170]. 111

[68] DUNE collaboration, *Searching for solar KDAR with DUNE*, *JCAP* **10** (2021) 065 [2107.09109]. 111

[69] G. Busoni, A. De Simone and W.-C. Huang, *On the Minimum Dark Matter Mass Testable by Neutrinos from the Sun*, *JCAP* **07** (2013) 010 [1305.1817]. 112

[70] V.A. Bednyakov and F. Simkovic, *Nuclear spin structure in dark matter search: The Zero momentum transfer limit*, *Phys. Part. Nucl.* **36** (2005) 131 [hep-ph/0406218]. 113

[71] A. Gould, *WIMP Distribution in and Evaporation From the Sun*, *Astrophys. J.* **321** (1987) 560. 114

[72] J.N. Bahcall, M.H. Pinsonneault and S. Basu, *Solar models: Current epoch and time dependences, neutrinos, and helioseismological properties*, *Astrophys. J.* **555** (2001) 990 [astro-ph/0010346]. 116

[73] T. Golan, J.T. Sobczyk and J. Zmuda, *NuWro: the Wroclaw Monte Carlo Generator of Neutrino Interactions*, *Nucl. Phys. B Proc. Suppl.* **229-232** (2012) 499. 120

[74] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, *Eur. Phys. J. C* **71** (2011) 1554 [1007.1727]. 121

[75] T. Kaluza, *Zum Unitätsproblem der Physik*, *Sitzungsber. Preuss. Akad. Wiss. Berlin (Math. Phys.)* **1921** (1921) 966 [1803.08616]. 123

[76] O. Klein, *Quantum Theory and Five-Dimensional Theory of Relativity. (In German and English)*, *Z. Phys.* **37** (1926) 895. 123

[77] T. Appelquist, H.-C. Cheng and B.A. Dobrescu, *Bounds on universal extra dimensions*, *Phys. Rev. D* **64** (2001) 035002 [hep-ph/0012100]. 123

[78] N. Arkani-Hamed, S. Dimopoulos and G.R. Dvali, *The Hierarchy problem and new dimensions at a millimeter*, *Phys. Lett. B* **429** (1998) 263 [hep-ph/9803315]. 123

[79] L. Randall and R. Sundrum, *A Large mass hierarchy from a small extra dimension*, *Phys. Rev. Lett.* **83** (1999) 3370 [hep-ph/9905221]. 123

[80] G. Servant and T.M.P. Tait, *Is the lightest Kaluza-Klein particle a viable dark matter candidate?*, *Nucl. Phys. B* **650** (2003) 391 [hep-ph/0206071]. 123

[81] H.-C. Cheng, K.T. Matchev and M. Schmaltz, *Radiative corrections to Kaluza-Klein masses*, *Phys. Rev. D* **66** (2002) 036005 [hep-ph/0204342]. 125

[82] M. Blennow, J. Edsjö and T. Ohlsson, *Neutrinos from WIMP annihilations using a full three-flavor Monte Carlo*, *JCAP* **01** (2008) 021 [0709.3898]. 125, 128

[83] J. Edsjö, J. Elevant and C. Niblaeus, *WimpSim Neutrino Monte Carlo*. 125, 128

[84] U. Haisch and A. Weiler, *Bound on minimal universal extra dimensions from anti-B —> X(s)gamma*, *Phys. Rev. D* **76** (2007) 034014 [hep-ph/0703064]. 127

[85] A. Freitas and U. Haisch, *Anti-B —> X(s) gamma in two universal extra dimensions*, *Phys. Rev. D* **77** (2008) 093008 [0801.4346]. 127

[86] C. Rott, D. Jeong, J. Kumar and D. Yaylali, *Neutrino Topology Reconstruction at DUNE and Applications to Searches for Dark Matter Annihilation in the Sun*, *JCAP* **07** (2019) 006 [1903.04175]. 130

[87] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al., *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* **12** (2011) 2825. 136

[88] J. Kopp, V. Niro, T. Schwetz and J. Zupan, *DAMA/LIBRA and leptonically interacting Dark Matter*, *Phys. Rev. D* **80** (2009) 083502 [0907.3159]. 140

[89] Particle Data Group collaboration, *Review of Particle Physics*, *PTEP* **2020** (2020) 083C01. 141

[90] M. Beltran, D. Hooper, E.W. Kolb and Z.C. Krusberg, *Deducing the nature of dark matter from direct and indirect detection experiments in the absence of collider signatures of new physics*, *Phys. Rev. D* **80** (2009) 043509 [0808.3384]. 141

[91] Planck collaboration, *Planck 2018 results. VI. Cosmological parameters*, *Astron. Astrophys.* **641** (2020) A6 [1807.06209]. 142

[92] D.S. Wilks, *Statistical Methods in the Atmospheric Sciences*, Academic Press. 191