Improving event classification at KM3NeT with OrcaNet

A thesis submitted in fulfillment of the requirements for
Master in Experimental Physics
at
Universiteit Utrecht

Author

Enrique Huesca Santiago

July, 2019

Abstract

Neutrino telescopes such as KM3NeT are being built to detect these tiny, elusive particles. In the case of KM3NeT ORCA the aim is to determine the currently unknown neutrino mass hierarchy, which has far reaching implications for scientific research. Distinguishing between the different types of neutrino flavour interactions seen in the detector is critical for this goal. In order to achieve this, Deep-Learning algorithms such as the OrcaNet framework for KM3NeT are being developed and tested. This work consists of an exploration of the performance of this tool for the concrete case of event identification in KM3NeT, and its implications for determining the neutrino mass hierarchy. Here, clear evidence is presented that there is potential for event classification and identification beyond the current binary track-shower scheme, including up to 40% separation for electron neutrino charged current events.

Para mamá y papá, los mejores científicos que nunca han sido.

Student number: 6310141
Thesis Research Project: 60 ECTS
Contact: e.huescasantiago@uu.nl
Supervisor: Paul de Jong
First Examiner: Raimond Snellings
Second Examiner: Alessandro Grelli
Word Count: 24966

Contents

1 Introduction 6

1.1 Motivation . . . 6

1.2 Neutrino classification . . . 10

1.3 Cherenkov Radiation . . . 11

2 Neutrino Physics 12

2.1 Neutrino Discovery . . . 12

2.2 Neutrino Oscillations . . . 14

2.2.1 Neutrino Mixing Formalism . . . 14

2.2.2 The Solar Neutrino Problem . . . 16

2.2.3 3 flavour neutrino oscillations . . . 19

2.3 Neutrino Astronomy . . . 21

3 KM3NeT 23

3.1 KM3NeT in a nutshell . . . 23

3.2 ARCA . . . 25

3.3 ORCA . . . 26

3.4 Software Infrastructure . . . 28

3.4.1 Triggering . . . 28

3.4.2 Reconstruction and Particle Identification . . . 28

3.5 Mass ordering fit . . . 30

4 Methods 32

4.1 OrcaSong . . . 32

4.2 OrcaNet . . . 35

4.3 Statistical Classification . . . 38

4.3.1 Threshold . . . 38

4.3.2 Estimators . . . 40

4.3.3 The receiver operating characteristic . . . 42

5 Results 43

5.1 Prediction probability distributions . . . 43

5.1.1 Track probability distribution . . . 43

5.1.2 New probability distributions . . . 50

5.2 Visible Energy dependency . . . 56

5.2.1 Threshold effect . . . 57

5.3 Categories comparison . . . 58

5.3.1 Hadronic and Tau shower . . . 60

5.4 Separation . . . 62

5.4.1 2 categories: e-CC vs e-NC 62
5.4.2 2 categories: Track-shower 64
5.4.3 3 categories: track, EM shower, tau + NC 65
5.4.4 4 categories 65
5.5 ROC curves 67

6 Conclusions 72
6.1 Research Objectives 72
6.2 Practical Implications 76
6.3 Future Research 77

Bibliography 79

Online References 81

A Introduction to Machine Learning 82
A.1 Neural Network 84
A.2 Model's Architecture 85

B Hidden Feature Check 88

C Technical Details 92
C.1 The Pipeline 92
C.2 OrcaSong 2.0 94

Note on the content

The following work is structured as follows: Chapter 1 explains the present work within neutrino telescopes research. Chapter 2 is a complete overview of the neutrino history from its discovery to the present. Chapter 3 is dedicated to describe the KM3NeT neutrino telescope. Chapter 4 explains explains the concrete tools and methods used in this thesis. Chapter 5 contains and comments the different types of results. Finally, Chapter 6 is dedicated to the different conclusions and what to do next. Chapters 1,2,3,4 are heavily based on bibliography. Two appendices are included to aid with the Machine Learning model and other useful information.

1 Introduction

1.1 Motivation

Neutrinos are neutrally charged fermions (spin 1/2) suggested by Pauli and introduced by Fermi to correctly balance the beta decay and other nuclear interactions. Fermi's description, now fully developed as the weak interaction, has been extended for the three families of the Standard Model of elementary particles between the leptons ($l = e, \mu, \tau$) and their associated neutrinos (ν_e, ν_μ, ν_τ).

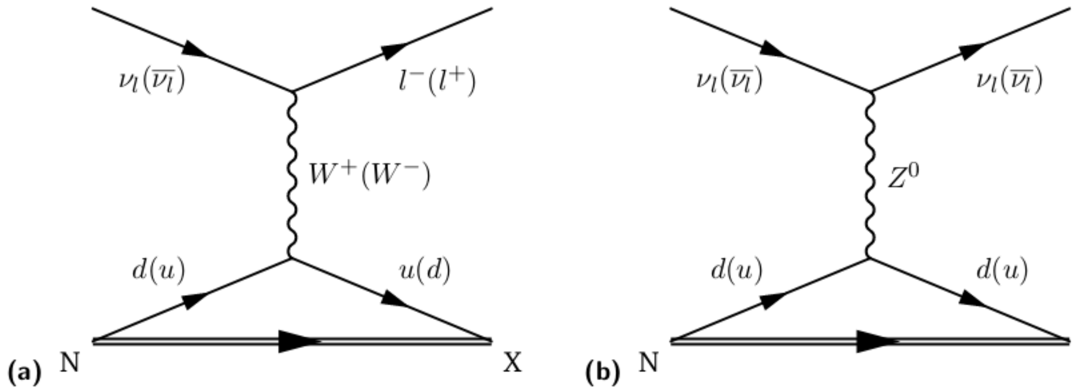

$\nu_l(\overline{\nu_l})$
$l^-(l^+)$
$W^+(W^-)$
$d(u)$
$u(d)$
(a)
N
X
$\nu_l(\overline{\nu_l})$
$\nu_l(\overline{\nu_l})$
Z^0
$d(u)$
$d(u)$
(b)
N
X

Figure 1.1: *The Feynman diagram of weak neutrino interactions with a nucleon. a) shows a charged current (CC) interaction, mediated by a weak boson $W^\pm$. b) shows a neutral current (NC) interaction, scattering via a Z^0 boson. Source [38].*

Neutrinos have the lowest cross sections known of all elementary particles, which makes their direct detection and a study of their properties almost impossible. What makes them interesting is that our limited knowledge about neutrinos includes plenty of puzzling phenomena that cannot be explained within the Standard Model of particle physics. At the same time, their ghost-like nature is also the reason why neutrinos originating from extreme astrophysical phenomena, such as the core of a supernova or deep within a black hole, are able to travel unimpeded in the universe, bringing that information to Earth.

Because of the low cross sections, neutrinos were assumed to be massless and were placed as a footnote of the physics revolution of the 20th century. Today, thanks to the revolutionary breakthroughs by the experiments of the Kamiokande collaboration and many others, neutrino physics has become an area of active and important research. The elusive particles now seem to hold some clues about the nature of the biggest and smallest length and energy scales at the same time, lying in the intersection between the disciplines of particle physics and astrophysics alike[1]. For this reason, today we have the bizarre concept of neutrino telescopes, experiments that are actively tackling particle physics questions alongside astrophysical research, such is the case of the ongoing experiments ANTARES/KM3NeT and IceCube.

A neutrino telescope is essentially a very large array of photodetectors placed deep under water (KM3NeT) or ice (IceCube) that targets atmospheric and cosmic neutrinos (see 1.2). The surrounding medium (water or ice) is used for both shielding against other interactions and Cherenkov light radiation (see 1.3). This simple idea does not capture the fact that neutrino telescopes need to be (and indeed are) the biggest experiments ever designed for two different reasons:

First, the interactions in the water/ice with energies of a GeV and beyond create a shower of particles (with their associated Cherenkov light) that spans hundreds of meters. Recording the complete light profile is necessary to fully understand this events. Second, neutrinos with extragalactic origin have an extremely low flux (seen in 1.4). Other neutrino experiments, like the accelerator experiments or nuclear reactor experiments (like the one who discovered the neutrino, see section 2.1), have a very clear source of neutrinos with a controlled direction and position. On the other hand, neutrino telescopes observe the neutrinos that come from the most remote and singular sources in the universe. IceCube, for example, has seen only 82 cosmic events between 60 TeV and 10 PeV in six years of operation [39]. This simply means that in order to have a non-negligible interaction rate (and study enough events to draw statistically-relevant conclusions) the instrumentalized volume needs to be as big as possible. This alone makes the minimum size for a viable telescope, just enough to detect as little as 10 events in the PeV scale per year, around 1 cubic kilometer, containing water or ice in the order of gigatons.

Capturing cosmic neutrinos is promising goal for potential knowledge that we can get in return, but is indeed a very difficult task which requires tremendous human and material effort. Due to sheer material and economical reasons, it is impossible to fully cover a cubic kilometer with photodetectors. This is why neutrino detectors have a sparse configuration. The sparsity means that most of the light produced by an event will be lost, either absorbed or scattered by the medium. Thus, the signature of the event, the event we can reconstruct from the light distribution of the shower of particles, is only a poor image of the actual interaction. Neutrino telescopes have a very low resolution, especially in the short range (meters). This means that, out of the many different particles produced in the showers, we can only record a handful of signatures based on the original neutrino interaction, shown in Fig 1.2 below. Even then, the large overlap between the light profiles of the three of the showers (electromagnetic, hadronic and tau) makes them effectively indistinguishable. The single signature that is clearly identifiable with our current resolution is the muon neutrino charged-current interaction ($\mu - CC$), because the outgoing muon is long-lived and travels in a straight line (emitting light along its track). Hence, these type of events are called "tracks", against the the rest of the "showers".

[1] Two groups usually so separated that they only meet at the University's cafeteria.

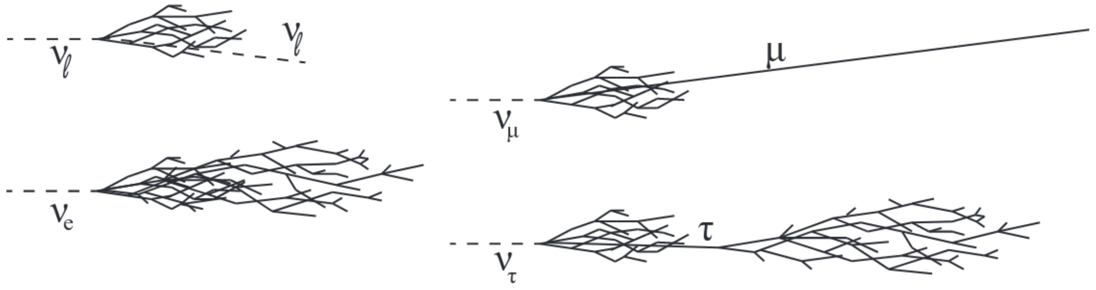


Figure 1.2: *Different neutrino interactions with water nuclei and their respective signatures. Top left is a hadronic shower, a NC scattering which produces a single particle shower + an outgoing invisible neutrino. Top right is a muon neutrino charged current event ($\mu - CC$), with an outgoing muon. Bottom left is a electromagnetic (EM) shower, an electron neutrino charged current ($e - CC$) where the outgoing charged electron scatters and produces a secondary shower. Bottom right is a tau shower ($\tau - CC$), where the tau travels for a short range before decaying and producing a secondary shower. The event's "double bang" name comes from the spatial and temporal separation. Source: Karel Melis [30].*

Track/shower classification is the common standard in particle identification (PID) in neutrino telescopes, but it is a deeply flawed classification. For a start, it's an imbalanced classification, since you are separating 1 out the 4 interactions in water. To make things worse, track classification does not fully identify the muon neutrino events. This is, distinguishing between a track and a shower event does not tell you whether it was a muon or a electron/tau/neutral event. The hadronic showers result from neutral interactions which do not distinguish flavours, so some muon neutrinos do produce showers. Looking at charged interactions only, the produced tau in a τ-CC event will decay into a muon with a branching ratio of 17% [44], which can be confused with a pure track event. Lastly, this classification does not take into account that the image is not perfect, so a track signature can be confused for a shower at lower energy, or because of any other external factors (coincidental events, backgrounds, secondary particle decay like the tau double bang, etc).

The thesis statement of this work is that **the particle identification scheme of track - shower in ORCA can be improved upon by using better state-of-the-art Deep Learning tools**. Deep Learning, based on the Convolutional Neural Network, is the method behind the current revolution in computer vision. This work, a first effort of this kind, tries to apply this tools to search for an improvement in the PID performance in the ORCA detector. This improvement can come within the track/shower scheme or overcoming it altogether. The work is based on Monte Carlo (MC) neutrino events of a full 115 lines ORCA detector, used to prepare 3 different implementations of ORCA's Deep Learning framework OrcaNet. The models have different output categories for classification: 2 categories a.k.a. track-shower classifier, 4 categories based on the 4 signatures and one model in between with 3 categories. The preparation and quantification of the performance of these models have been made under the same conditions, so a direct comparison between them can be drawn.

The starting goal of this thesis was to explore how the Deep Learning (DL) methodologies, which are starting to become popular within the high-energy physics community and KM3NeT in particular, can be applied to improve the mass hierarchy fit. The mass hierarchy or ordering problem essentially concerns the ordering of the neutrino masses, which can take one of two configurations (normal or inverted). The main purpose of the ORCA detector is to find from the neutrino appearance distribution, which ordering is followed. The origin of the mass hierarchy is described in section 2.2.3, and ORCA's hierarchy analysis is outlined in section 3.5.

The hierarchy analysis (a multidimensional parameter fit) is an already established study with its own fixed tools, and so an indirect approach is used here: The straightforward approach is testing this method for an improvement in track-shower classification, which implies an increase in the event's quality, which in turn translates to an increase in the statistics available for the posterior hierarchy fit, and a reduced data-taking time. The secondary approach aims at a more realistic classification on the flavour components. This is a far more ambitious goal, due to the sheer resolution of the detector (seen in 1.3). However, even an imperfect success opens the possibility of tackling further neutrino studies that require the three flavour components, like a unitary test of the PMNS matrix.

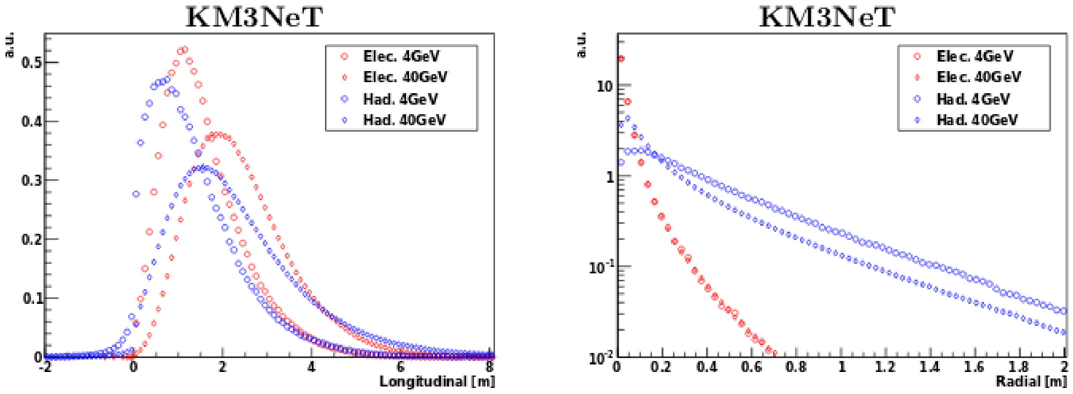


Figure 1.3: *Light emission profiles for different signatures and energies in ORCA. In spite of the large overlap, the detector is able to resolve to some degree the differences between shower events. A good event reconstruction algorithm should be able to pick these differences whenever possible. From [32], also [36].*

The current metric for the expected performance of the ORCA detector is called *sensitivity*, is it is defined defined as the data-taking time needed to perform the mass ordering analysis and separate between the normal and inverted orderings enough for evidence (3σ signal) or discovery (5σ). The current sensitivity estimate for ORCA is 3 years for the 3σ signal under all of the different possible scenarios. All the improvements in performance can be defined as improvements on sensitivity. This thesis does not include sensitivity calculations due to technological and material constraints. However, this is not to say that a sensitivity study based on the improved performance shown here would be out of the scope of interest of this of work, and might be a crucial step for future work along this line.

1.2 Neutrino classification

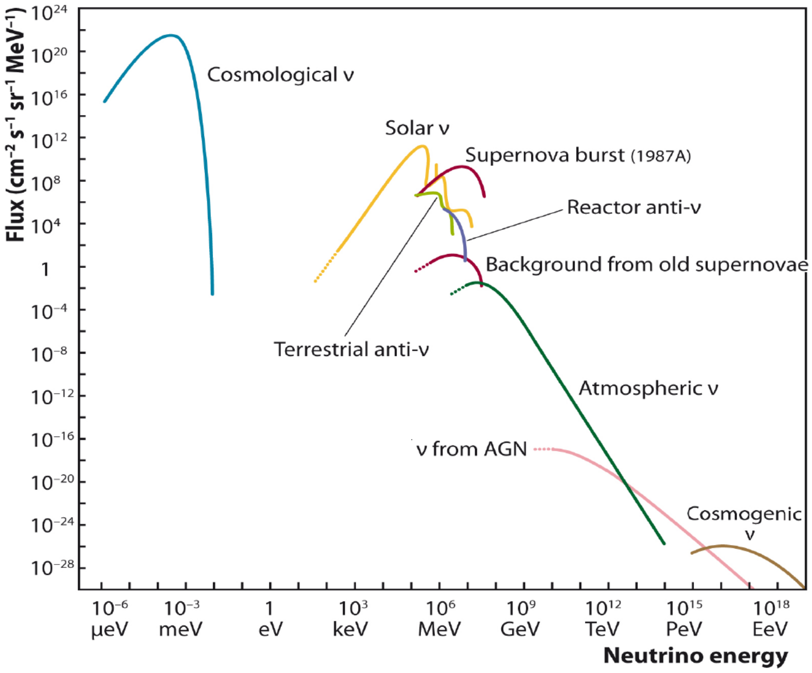


Figure 1.4: *The neutrino spectrum, classified according to the source. The production mechanism usually determines the observed energy and flux observed. For this work, we will call high-energy neutrinos to those above 1 GeV. Source: Uli Katz [29]*

- **Cosmological Neutrinos**: Cosmological or relic neutrinos are the ones produced in the earliest nuclear reactions in the universe, 1 second after the Big Bang. Prior to the CMB, they formed the CνB (cosmic neutrino background) which has been redshifted down to 1.9 K, beyond experimental detection.

- **Terrestrial (Anti-)Neutrinos**: Neutrinos produced in the Earth from heavy element decay. To distinguish the natural phenomena from the man-made in a nuclear plant, these last ones are called **reactor neutrinos**.

- **Solar neutrinos**: Produced by the nuclear reactions in the Sun. Every element in the sun constitutes a different channel, but they are mainly produced by the low-energy pp chain. The neutrinos produced by a solar supernova have a whole different nature, with higher energy and flux.

- **Atmospheric neutrinos**: Produced by the decay of the products of a cosmic-ray interaction in the upper atmosphere. The cosmic rays are accelerated outside of the Solar System, and probably outside of the galaxy, so the neutrinos (a small byproduct) can reach energies up to PeV (10^{15} eV).

- **Cosmic Neutrinos**: Cosmic neutrinos are the byproduct of the interactions outside of the Galaxy. There are only two known sources, active galactic nuclei (AGN), which refer to supermassive black holes; and cosmogenic neutrinos, supposed to be coming from the decay of an GKZ interaction [15] [16] between a ultra-high energy proton and a CMB photon.

1.3 Cherenkov Radiation

Cherenkov radiation is the name of the light emitted by the Cherenkov effect, and is one of the main physics mechanisms behind particle detection in high energy physics and the main mechanism in astrophysics.

The Cherenkov effect is the following: when a relativistic charged particle enters in a dielectric material, the so-called-radiator medium, their speed falls above the relative speed of the light in the medium $v = c/n$, where n is the medium's refractive index[2]. The speed of the particle simply causes a coherent interference of its own electric field, usually evanescent, into a conical wavefront that tails the particle, as illustrated in 1.5 below.

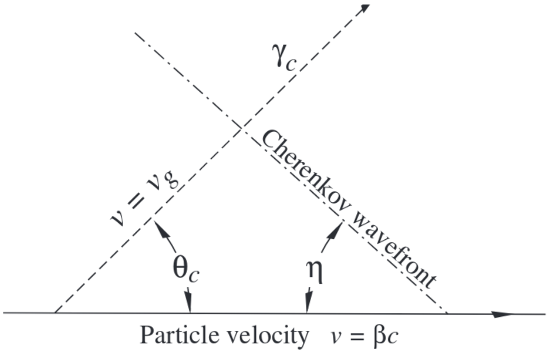


Figure 1.5: *Cherenkov effect schematic depiction. The particle's position is in the lower arrow, with the line indicating its direction. The posterior cone of light (of which the top half is presented here) can be represented in two equivalent frames, of which θ or θ_{Ch} is commonly referred as the Cherenkov angle. γ_c represents the direction of the light propagation, normal to the wavefront. Source: [44], Passage of particles through matter review, pg 33.*

The particle's speed β (and thus, its energy) is related to the angle of the cone of light,

$$\theta_{Ch} = arcos\left(\frac{1}{\beta n}\right), \tag{1.1}$$

and the number of photons N it produces per unit distance x and wavelength λ,

$$\frac{d^2N}{dxd\lambda} = \frac{2\pi\alpha z^2}{\lambda^2}(1 - (\beta n(\lambda)^{-2})), \tag{1.2}$$

with z the charge (in e units) and α the fine structure constant. For the KM3NeT experiment, the number of photons is favored against the angle to make an energy estimation.

Cherenkov radiation is a macroscopic phenomena, where the particle sees the radiation medium as a continuum. This means that Cherenkov radiation has more in common with the shockwave of a supersonic plane that other radiation phenomena like *Bremsstrahlung* or braking radiation, where the charged particle interacts with the individual atoms in the medium.

[2] In seawater, $n \sim 1.35$, although this depends on the wavelength, temperature, salinity,etc.

2 Neutrino Physics

2.1 Neutrino Discovery

The idea of the particle that we know now as neutrino was proposed by Wolfgang Pauli the 4th of December of 1930, in an open letter to a physics meeting in Tübingen, Germany [3]. In the letter, Pauli explains how the neutrino would solve the continuous beta spectrum problem, and tries to outline some of its properties. The continuous shape of the beta particle energy spectrum represented a serious challenge for the law of conservation of energy. Since the discovery of beta radiation by Rutherford in 1899 and its characterization as electrons by Henri Bequerel in 1900 (they shared the same mass-to-charge ratio), the radioactivity components were well understood, more so after the radioactive displacement law of Soddy[2] and Fajans[1] in 1913. If the source of the decay and the product are the same, and the emitted particle is always an electron, why do not all the emitted electrons have the same kinetic energy when emitted?.

While some physicists, like Bohr, argued that the law of conservation of energy was only obeyed "in a statistical average sense", and not for every interaction; Pauli argued that presence of a "neutron", of electrically neutral fermion "in any case no larger than 0.01 proton mass", could carry the missing energy (and spin) of the electron, so the that total is constant. At the same time (1931-1932), James Chadwick discovers the particle that we know today as neutron as massive, uncharged radiation [4]. In order to distinguish Pauli's "neutron" from Chadwick's neutron, in a conversation between Edoardo Amaldi and Enrico Fermi in Rome the "neutrino" name (little neutron in Italian) was used, in reference to its lower mass [21]. Fermi replicated the name in following scientific conferences and it was quickly adopted.

Fermi was the one who made sense of the particle puzzle when he published his theory of beta decay in 1933-1934, where he correctly described a 4-point interaction of a neutron decaying in a proton, an electron and a neutrino (later corrected to an antineutrino)[1]. This theory was only superseded by the correct formulation in 1968 of the electroweak force[17] and the direct discovery of the W, Z bosons in 1983 [20]. However, Fermi's 4 interaction picture is still used at energies below the electroweak scale of $\sim$ 100 GeV as an effective field theory, see 2.1.

[1] It is a remarkable comment that this theory was initially rejected by the *Nature* journal, which later openly regretted the mistake. An English version of the paper would have to wait until 1968. Fermi, troubled from this rejection, switched to experimental physics, shortly discovering the radioactive activation of nuclei with neutrons.

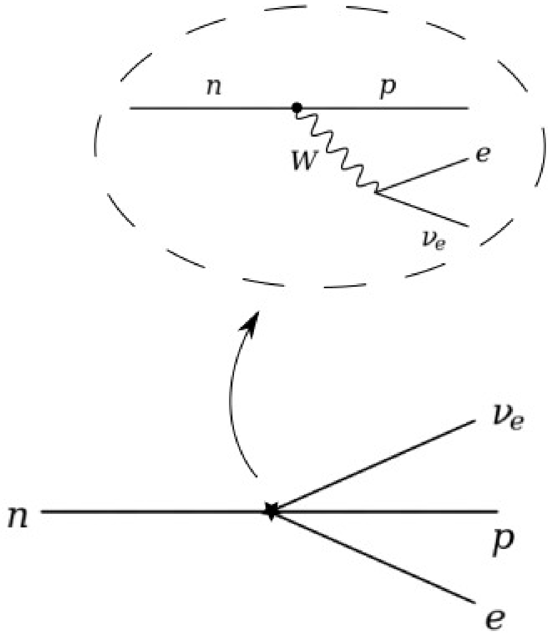


Figure 2.1: *Fermi's 4 point interaction picture: Read left to right, it explains a beta decay, while read right to left, it shows and inverse beta decay via electron capture. The zoomed area explains the short-range weak interaction. Figure by Gonzalo Contreras Aso.*

The only thing left to prove the existence of neutrinos was the direct experimental detection. In 1934 already, Bethe and Peierls [5] already assume the existence of a neutrino annihilation process (now known as beta capture), and correctly estimate that its cross section is $\sigma < 10^{-44}$ cm^2: "corresponding to a penetrating power of 10^{16} km in solid matter". Thus, neutrinos could cross astronomical distances (Our galaxy's radius is $5 \cdot 10^{17}$ km) through matter unfazed. The authors concluded: "besides creation and annihilation [...] there is not practically possible way of observing the neutrino".

A formal proposal of detecting the neutrinos from a nuclear reactor would come after the Second World War. Cowan and Reines, scientists at Los Alamos, started Project Poltergeist in 1951 with the goal of directly detecting the neutrino through inverse beta decay [24]. They built a detector consisting of two 100-liter water tanks doped with Cadmium and surrounded by scintillating layers containing photomultiplier tubes (PMTs). This detector was placed 11 m away from the Savannah River nuclear reactor (South Carolina) in order to make use of its large neutrino flux. Within the detector, the following interactions were taking place:

$$\begin{gathered}\bar{\nu}_e + H(= p^+ + e^-) \rightarrow e^+ + n \\ e^+ + e^- \rightarrow 2\gamma \\ n + Cd^{108} \rightarrow (Cd^{109})^* \rightarrow Cd^{109} + \gamma\end{gathered} \tag{2.1}$$

The two bursts of light observed (e^+ annihilation $+$ n absorption milliseconds later) was the perfect double signature that proved the existence of the neutrinos. The measured cross section, published in 1956 [7], was of $6.3 * 10^{-44}$ cm^2. Pauli, 25 years after his initial idea, commented: "Everything comes to him who knows how to wait".

2.2 Neutrino Oscillations

In 1956, an equally important discovery was made. A team, leaded by the Chinese American physicist Chien-Shing Wu, established, thanks to studying the Cobalt-60 beta decays, that the weak interactions do not conserve parity, or P-symmetry[2] [33]. This means that the neutrinos involved in this decay, as in all other weak interactions, always have the "left" state only (And the antineutrinos always have the "right" parity.). This result states that nature does not obey a fundamental symmetry, and was as deeply unexpected as troubling at the time (Pauli, less nonchalantly this time, supposedly claimed "That's total nonsense" [19]). Since a massless particle does not have more than fixed 1 parity state, this result, with the small mass and cross section measurements, cemented the idea that the neutrino was massless.

2.2.1 Neutrino Mixing Formalism

The Italian physicist Bruno Pontecorvo, already in 1957 posed the idea that neutrinos could oscillate between its different configurations [8]. Inspired by the Gell-Mann and Pais idea on CP-violating neutral kaon oscillations (who, in turn, developed the idea after Wu's result), its first concept actually involved neutrino-antineutrino oscillations. He soon followed with this "2-component neutrinos" in 1960, claiming that muon neutrinos were different than electron neutrinos. When the muon neutrino was discovered in 962 at Brookhaven [11], the neutrino mixing formalism by Maki, Nakagwa and Sakata quickly followed. [12]

This mixing formalism states that the neutrino eigenstates that form the basis for weak interactions $\{\nu_e, \nu_\mu, \nu_\tau\}$ are not the same as the "free particle" mass eigenstates $\{\nu_1, \nu_2, \nu_3\}$, but rotated. In other words, the mass interactions (that know we know are with the Higgs boson) and the weak interactions (with the W,Z bosons) do not see the same side of the neutrinos, but their respective pictures are connected by the PMNS matrix (Pontecorvo-Maki-Nakagwa-Sakata), usually called U. Using the current picture with 3 neutrino flavours and the standard notation

$$\begin{bmatrix}\nu_e\\ \nu_\mu\\ \nu_\tau\end{bmatrix} = \begin{bmatrix}U_{e1} & U_{e2} & U_{e3}\\ U_{\mu 1} & U_{\mu 2} & U_{\mu 3}\\ U_{\tau 1} & U_{\tau 2} & U_{\tau 3}\end{bmatrix}\begin{bmatrix}\nu_1\\ \nu_2\\ \nu_3\end{bmatrix}. \tag{2.2}$$

The PMNS mixing formalism offered, however, is the "virtual transmutation" between this two flavours ($\nu_e \leftrightarrow \nu_\mu$), a phenomena now called neutrino oscillation or neutrino probability oscillation.

To understand the concept of probability oscillation, we can go through classic and simple example 2 flavours (electron and muon, e.g.) and 2 mass states in vacuum. The following formulation, reproduced from [35], is not the most precised but is used for its clarity.

[2] Parity symmetry is the symmetry present in a mirror, or why your left hand seen in a mirror is the same as your and right hand. This is why the possible configurations of this symmetry are called left and right handed.

Let's start with the (simplified) mixing matrix, which can be described as a rotation matrix between two orthonormal states bases. The mixing matrix is parametrized with a single angle (mixing angle θ),

$$\begin{bmatrix} \nu_e \\ \nu_\mu \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}. \tag{2.3}$$

This means that any electron neutrino is produced as a combination of the two mass states

$$|\nu_e(t=0)\rangle = cos\theta|\nu_1\rangle + sin\theta|\nu_2\rangle. \tag{2.4}$$

Applying the Schrödinger equation of the time evolution for a free relativistic particle to the neutrino we obtain

$$|\nu_e(t=0)\rangle = e^{-ip_1^\alpha x_\alpha} cos\theta|\nu_1\rangle + e^{-ip_2^\alpha x_\alpha} sin\theta|\nu_2\rangle, \tag{2.5}$$

with $\hbar = c = 1$, p^α the 4-momentum of the particle and the x_α the 4 positions. For a distance L travelled,

$$p_i^\alpha x_\alpha = E_i t - \bar{p}\bar{x} = L(\sqrt{p^2 + m_i^2} - p) = Lp(\sqrt{1 + (m_i/p)^2} - 1). \tag{2.6}$$

Now we recall that the neutrino is relativistic, almost massless ($m_i << p$, $p \simeq E$), so we can apply the Taylor expansion above and

$$Lp(\sqrt{1 + (m_i/p)^2} - 1) \simeq L\frac{m_i^2}{2p} \simeq \frac{m_i^2 L}{2E}. \tag{2.7}$$

The the probability of the oscillation at time t is:

$$P(\nu_e \to \nu_\mu) = |\langle \nu_\mu | \nu_e(t) \rangle|^2 = sin^2\theta cos^2\theta (e^{-ip_1^\alpha x_\alpha} - e^{-ip_2^\alpha x_\alpha})^2. \tag{2.8}$$

Plugging the expanded approximation and rearranging,

$$P(\nu_e \to \nu_\mu) = \frac{1}{2} sin(2\theta)(e^{\frac{iLm_1^2}{2E}} - e^{\frac{iLm_2^2}{2E}})^2. \tag{2.9}$$

If we define the mass difference Δm as $\Delta m_{21}^2 = m_2^2 - m_1^2$, then

$$P(\nu_e \to \nu_\mu) = \frac{1}{2} sin(2\theta)(e^{\frac{iL\Delta m^2}{2E}} - 1)^2 = sin(2\theta) sin^2\left(\frac{(\Delta m_{21}^2)L}{2E}\right). \tag{2.10}$$

This are but the broad lines of the actual formulation. However, the neutrino mixing theory relies on the fact that neutrinos have mass, something that, while it does not fully contradict Wu's result on CP violation, is clearly against the most simple interpretations of their result. For this reason, the neutrino mixing and the oscillations were widely disregarded by the physics community until it provided the solution for the solar neutrino problem.

2.2.2 The Solar Neutrino Problem

Pontecorvo, who was assistant of Fermi, had the idea of using the inverse beta decay of chlorine to capture neutrinos, and record the posterior radioactive decay:

$$\nu + Cl^{37} = e^{-} + Ar^{37}; \quad Ar^{37} \xrightarrow{\tau(Ar^{37}) = \ 34 \ days} e^{-} + Ar^{36} \tag{2.11}$$

His first attempt at detecting reactor neutrinos was not successful because reactors emit antineutrinos. However, this technique would later be used by Ray Davis Jr in the 1970s to detect solar neutrinos. In the Homestake experiment, a 380 cubic meter full of perchloroethylene (C_2Cl_4), a common cleaning fluid, was placed in the Homestake Gold mine in South Dakota, 1500 m underground for cosmic shielding. The amount of Argon-37 production in the tank gives a measure of the neutrino flux, so the experiment only had to recover the argon in the tank and count its decays. John N. Bachall performed the calculation of the expected flux in the detector based on the Standard Solar Model, and the experimental results presented a deficit by a factor of 3. This discrepancy was known as the Solar Neutrino Problem, and lasted for about 4 decades.

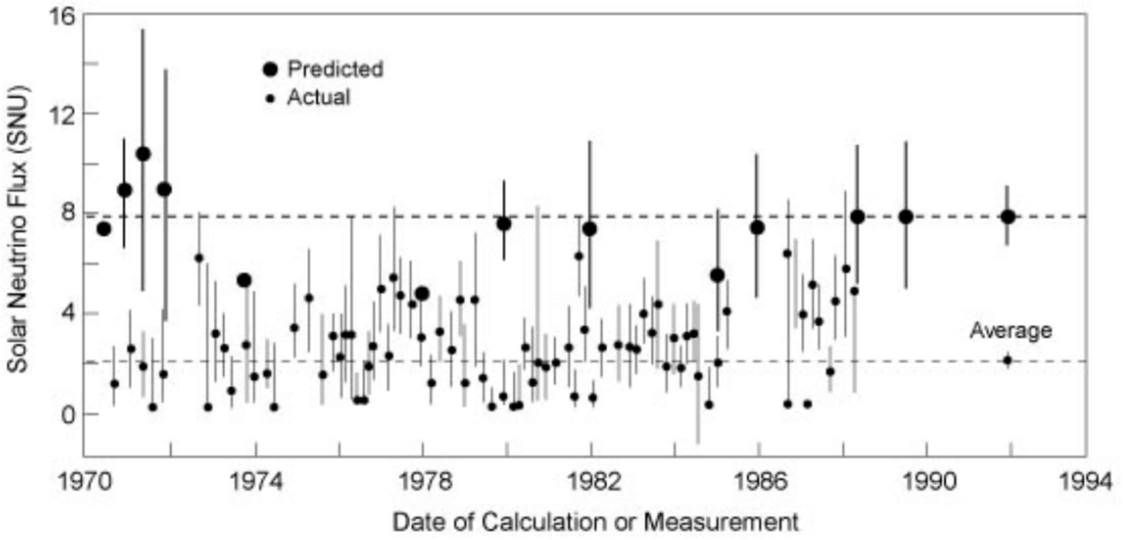


Figure 2.2: *One of the final results from the Homestake experiment, after 20 years of data taking. 1 SNU is 1 interaction per* 10^{36} *chlorine atoms per second. From [23]*

The discrepancy was assumed to be a mistake in the experiment, but, after verification, the problem persisted. Improved theoretical solar models (Helioseismology) revealed the inner structure of the sun in the 1990s and validated Bachall's calculations. Extended experimental data came from Homestake itself (see 2.2) and newer, more precise, Gallium based experiments, GALLEX (Italian) and SAGE (Russian), replicating the discrepancy. The 1989 Kamiokande experiment, a huge underground water detector designed to test nucleon decay, was the last validation of the Solar Neutrino Problem.

Immediately after the original discovery, in 1969, Pontecorvo and Gribov already suggested the idea that the neutrino oscillations were the responsible mechanism for the neutrino deficit [18]. In this theory, the electron neutrinos emitted from the sun can oscillate to other flavours during their travel, making them "blind" to the chlorine and gallium detectors. This theory, which contradicted the established idea of massless neutrinos, was not considered until kamiokande's successor, Super-Kamiokande, founded in 1998 the first proof of the oscillation mechanism.

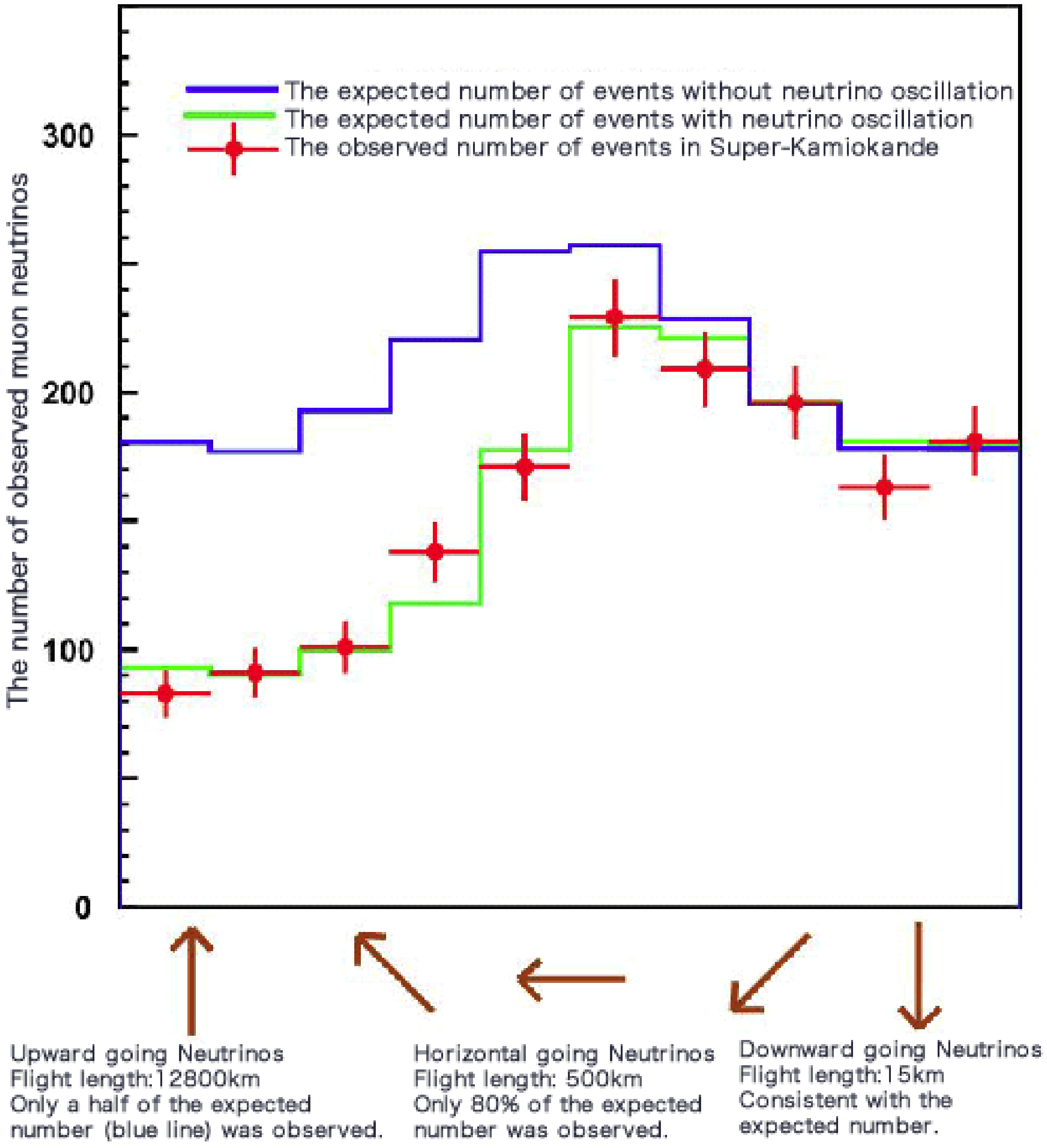


Figure 2.3: The atmospheric muon neutrinos coming into Super-Kamiokande. When the neutrinos move through the Earth, the MSW effect changes the neutrino oscillations and the expected number in accordance. From below the detector (the inside of the Earth) have a higher chance to oscillate than the ones coming from above (the atmosphere), so their final number is less. From [25].

The final proof would come from the Sudbury Neutrino Observatory (SNO), in Canada. SNO had the chance to build a detector based on heavy water or deuterium (D_2O). Deuterium is sensitive charged and the neutral current interactions, so both types could be observed, at the same time:

$$\begin{aligned} CC &: v_e + D \rightarrow e^- + 2p \\ NC &: v_x + D \rightarrow v_x - +p + n \end{aligned} \tag{2.12}$$

While the charged current interactions can only be carried out by an electron neutrino, all flavours participate equally in the neutral current interactions. The chance to measure both ratios at the same time is all that was needed. Indeed, in 2001, only 2 years after its start, the SNO collaboration announced a 2-1 ratio between the non-active (μ, τ) and active (electron) interactions in the detector.

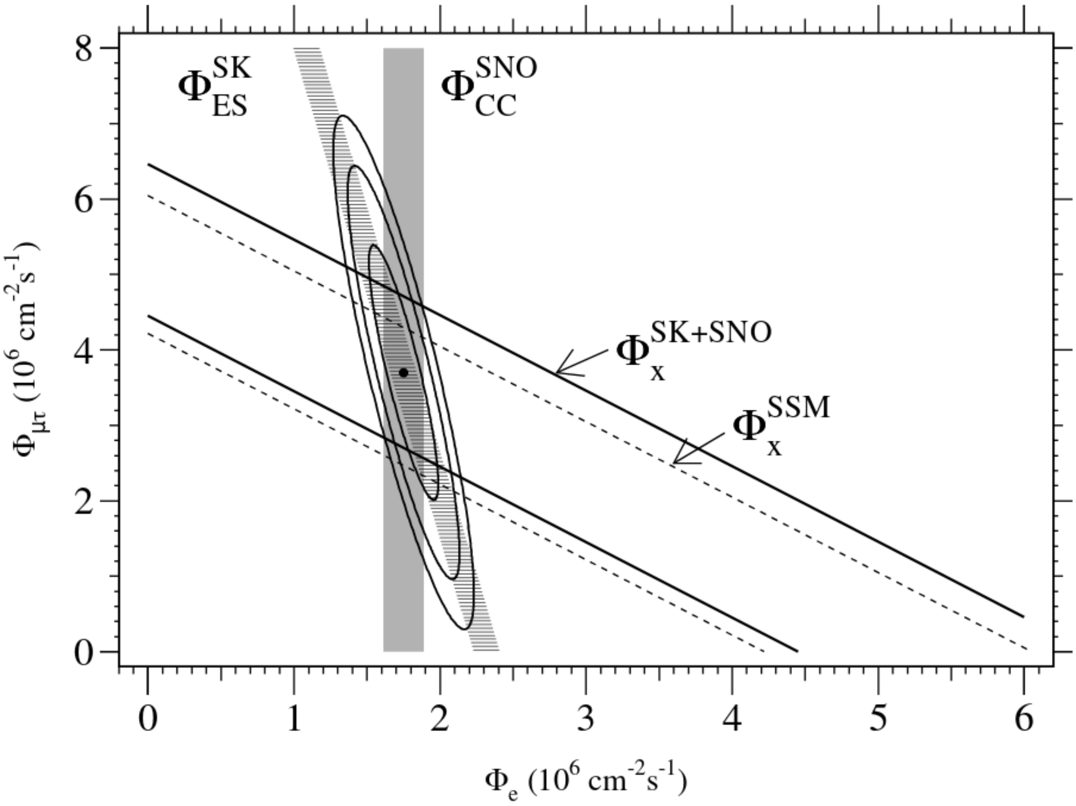


Figure 2.4: *SNO's final result: The electron neutrino flux (x axis) vs the "non-electron flavour active" μ, τ (y axis). It includes their result (SNO) with Super-Kamiokande (SK) their combined prediction with the Standard Solar Model (SSM). From [26].*

Neutrinos have three mass states, and they oscillate between them.

2.2.3 3 flavour neutrino oscillations

The full oscillation formalism is outlined below. This is just a more detailed version of the same calculation as in 2.2 above. This complete, precise description of the oscillations can be found at [42].

First, there are three flavours that participate in neutrino oscillations. This means that there is not one mixing angle and one mass difference, but 3 of each: $\theta_{12}, \theta_{23}, \theta_{13}, \Delta^2 m_{12}, \Delta^2 m_{23}$, $\Delta^2 m_{13}$. The mixing matrix, ignoring non-contributing terms (Majorana phases) to the oscillation, now looks like this:

$$U_{PMNS} = \begin{bmatrix} c_{12}c_{13} & s_{12}c_{13} & s_{13}e^{-i\delta_{CP}} \\ -s_{12}c_{23} - c_{12}s_{23}s_{13}e^{-i\delta_{CP}} & c_{12}c_{23} - s_{12}s_{23}s_{13}e^{i\delta_{CP}} & s_{23}c_{13} \\ -s_{12}s_{23} - c_{12}c_{23}s_{13}e^{-i\delta_{CP}} & -c_{12}s_{23} - s_{12}c_{23}s_{13}e^{-i\delta_{CP}}c_{12}c_{13} & c_{23}c_{13} \end{bmatrix}. \tag{2.13}$$

The oscillation probabilities are indeed changed too, but not very different:

$$P_{3\nu}(\nu_\mu \to \nu_e) \simeq sin^2\theta_{23} sin^2 2\theta_{13} sin^2\left(\frac{\Delta m^2_{31} L}{4E_\nu}\right). \tag{2.14}$$

The 3 dimensional PMNS matrix also gains a global phase, δ_{CP}, that is the source of the CP violation effects in the neutrino sector (exactly like the quark counterpart). Notice that identifying ν_1 as the lightest state and ν_3 as the heaviest state, $\Delta^2 m_{13}, = \Delta^2 m_{12} + \Delta^2 m_{23}$. So we are finally with 6 "free" parameters for the neutrino sector. The best inferred knowledge about this parameters, where all the results of all the neutrino experiments worldwide are combined, can be seen in nu-fit.org [47].

Second, we need to account for the "matter effects", formally known as MSW effect, from Mikheyev, Smirnov and Wolfestein [22]. Simply speaking, the matter effects come from the presence of the electrons along the neutrino path. The electron density increases the chance of a CC interaction for ν_e, adding an effective potential A and raising its effective mass in the oscillation scheme. The updated, more precise probability reads:

$$P^m_{3\nu}(\nu_\mu \to \nu_e) \simeq sin^2\theta_{23} sin^2 2\theta^m_{13} sin^2\left(\frac{\Delta^m m^2 L}{4E_\nu}\right), \tag{2.15}$$

where

$$\begin{gathered} sin^2(2\theta^m_{13}) = sin^2 2\theta_{13}\left(\frac{\Delta m^2_{31}}{\Delta^m m^2}\right)^2 \\ \Delta^m m^2 = \Delta m_{31}\sqrt{(cos2\theta_{13} - 2\frac{E_\nu A}{\Delta m^2_{31}})^2 + (sin2\theta_{13})^2}, \\ A = \pm\sqrt{2}G_F N_e. \end{gathered} \tag{2.16}$$

The potential A has the Fermi constant G_F, the electron density N_e and a different sign for ν_e (+) and $\bar{\nu}_e$ (-), since only the antineutrino can have a CC interaction with the electron. Other NC effects are averaged out between the protons and the electrons for the three flavours.

Third, the variable in the oscillation probability is $\Delta^2 m$, not Δm. When an oscillation probability experiment is carried, and after detailed computation with carefully controlled L and E, we only know the value of the mass difference squared. This means that there are two options for the mass difference, $\pm\Delta m$. This means that we know the mass differences but not the exact mass distribution of the mass states. This is called the mass hierarchy or mass ordering problem, and two mass differences gives in return two possible configurations, normal or inverted hierarchy, illustrated below 2.5. This problem is a roadblock for fundamental physics and cosmology research, due to its big influence on plenty of models and predictions 2.6.

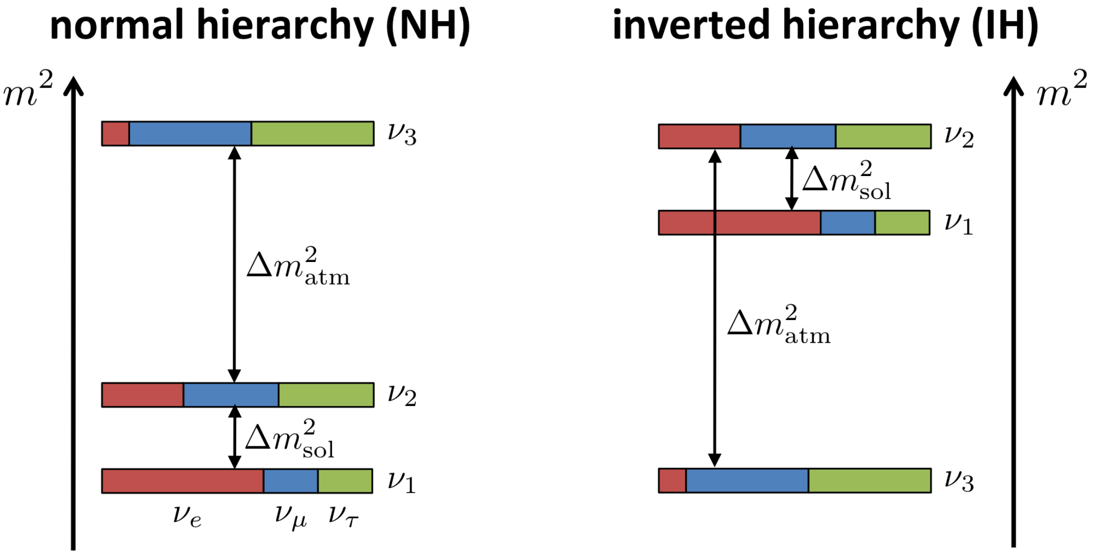


Figure 2.5: *The two possible solutions of the mass hierarchy. The mass differences take their name from the experiments that are sensitive to them. Notice that the colors are the different flavour contributions to the mass eigenstates. Source: JUNO experiment.*

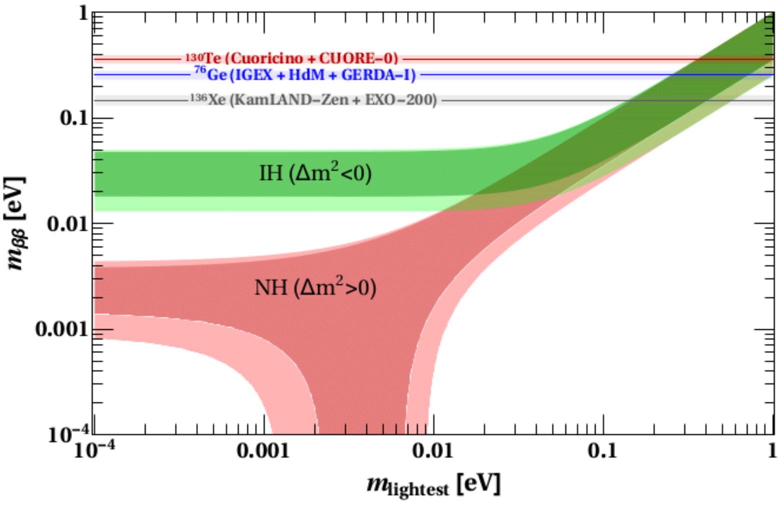


Figure 2.6: *Effect of the hierarchy on the combined neutrino masses $< m_{\beta\beta} >$ for neutrinoless double beta decay. Normal hierarchy is the one that assumes the lowest total mass. Figure from the KAMLAND experiment [43].*

2.3 Neutrino Astronomy

Werner Heisemberg, in 1936, already had the idea that neutrinos might be a substantial part of the radiation in the cosmic showers [6], but he overestimated the cross section, since he assumed that the neutrinos would cause secondary showers. The idea of neutrino astronomy was officially born in 1960 in two different publications: *Cosmic Ray Showers*, by Kenneth Greisen [9] and *On high energy neutrino physics in cosmic rays*, by Moisei Markov.[10] Greisen proposed a 15 m underground, 3000 ton detector based in Cherenkov-light counters. Despite calculating a very low ratio of events, he asserts that "within the next decade, cosmic ray neutrino detection will become one of the tools of both physics and astronomy". Markov, on the other hand, proposed the idea of installing detectors deep underwater (a lake or a sea) to determine the direction of charged particles with the help of Cherenkov radiation, in order to reach a detector volume bigger than 10 kton. This is the idea that has been developed in the modern neutrino telescopes. Although established since, this claims were not trivial for the era, since F. Reines (the same that discovered the neutrino) was claiming in 1960 that "cosmic neutrino flux cannot be predicted" while "cosmic ray neutrinos [atmospheric neutrinos] are predictable and of less interest".

However, it is a team lead by Reines, the CWI/SAND experiment, who detected the first atmospheric neutrino [14], the 23 of February of 1965, in a 3.5 km deep mine in South Africa, although this discovery was done almost simultaneously by the KGF team in India [13]. Exactly 22 years later, the first neutrinos emitted by a supernova (1987A, from the Large Magellanic Cloud), were recorded by Kamiokande (and others). This detection was the final proof of the relevance and feasibility of the study of the atmospheric neutrinos, formally establishing the field and triggering all the experiments that we have today.

Due to advances in atmospheric neutrino detection, Markov's idea of an underwater neutrino telescope started to became feasible. Reines himself, with the collaboration of other physicists (such as A.Roberts, J.Learned and S.Miyake) pioneered the first works in 1975 for a Deep Underwater Muon and Neutrino Detector, the DUMAND experiment. DUMAND was the project where all our actual, standard ideas and methods were proposed and developed: A cubic kilometer array of photomultiplier tubes (PMTs), housed in transparent, pressure resistant spheres would be placed deep underwater and record up-travelling muons that can only be generated by neutrinos undergoing charged current interactions in the water, since no other particle can move freely through the Earth. The large water depth was immediately recognized as a main factor to suppress down-moving muons and all other charged particles. This left the PMTs to record the Cherenkov light from the atmospheric and cosmic neutrinos. DUMAND, who was meant to be built in the coast of Hawaii, did pioneering technological developments and overcome different setbacks over 20 years until, due to lack of funding and technical difficulties, was cancelled in 1995. However, the DUMMAND experiment ideas and methods have been inherited by other 3 different experiments worldwide. The successors of those original experiments have increased their size and scale, and have become the very first Very Large Volume Neutrino Telescopes (VLVNT), today together, they form the Global Neutrino Network:

- **Lake Baikal**: The political relations during the Cold War forced the Russian contribution to the DUMMAND project to retire in 1980 and pursue their own version of the project. This would turn to be the Baikal Deep Underwater Neutrino Telescope (BDUNT), which worked from 1990 - 2015. BDUNT was the first experiment to have 3 detector strings deployed in 1993, allowing for the spatial reconstruction of the events. The BNUDT has become, since 2015, Baikal-GVD (Gigaton Volume Detector). The GVD telescope is currently operating and is expected to reach its maximum final size of 2 cubic km around 2021 [46]. GVD is designed to search for neutrinos in the high TeV - PeV range, the highest energy probe of the 3 telescopes.

- **South Pole**: An idea suggested by Francis Halzen and J. Leanerd (from DUMMAND) in 1988. In this case, the ice cap would substitute the water as the blocking and radiation material, offering a unique medium free from most secondary backgrounds, although the different properties of ice vs water, have proven to hinder the resolution of the instrument. The Antarctic Muon And Neutrino Detector, AMANDA, was built in 1996 at the Amundsen-Scott South Pole station, digging ~ 2 km deep holes to house the detectors. In 2005, it was decommissioned in favor of its successor, the IceCube Neutrino Observatory, completed in 2010. With 5040 optical modules, a cubic kilometer of instrumented volume, and 9 years of data-taking, IceCube is the actual standard in astronomical neutrino detection. Last July 2018, IceCube announced that they have recorded the first high-energy neutrino event that has been able to be reconstructed to its source, the blazar TXS 0506 +056 [41]. This work was possible thanks to the alert system, that allowed other optical telescopes to look at the source after the event was detected, and found the blazar in a flaring state. This joint work at detecting rare events is one of the first examples of the so-called multi-messenger astronomy.

- **Mediterranean Sea**: The Mediterranean Sea is a natural candidate for the underwater telescope efforts of the European countries, and a perfect complement for the South Pole observatory, since a single down-facing telescope cannot cover the whole sky. The earliest efforts were made in 1990/1991 by a Russian-Greek collaboration in the NESTOR project. Although their location in the coast of Pylos was privileged, their 2003/2004 prototype was short-lived and was terminated in 2005. The French and Italian teams in NESTOR left in 1997 and 1998 and founded the ANTARES (Astronomy with a Neutrino Telescope and Abyss environmental RESearch) and NEMO (the NEutrino Mediterranean Observatory) [3] experiments respectively. The NEMO team, together with teams from the Netherlands and Germany they joined the French idea and founded the ANTARES collaboration in 1999. The ANTARES telescope resides near the coast of Toulon, France, was completed in 2008 and is still in operation. Already in 2006, the ANTARES collaboration members from the different teams proposed the more ambitious project of building a Cubic Kilometer Neutrino Telescope (KM^{3}NeT, or KM3NeT).

[3] Indeed, the naming conventions for experiments have long gotten out of hand.

3 KM3NeT

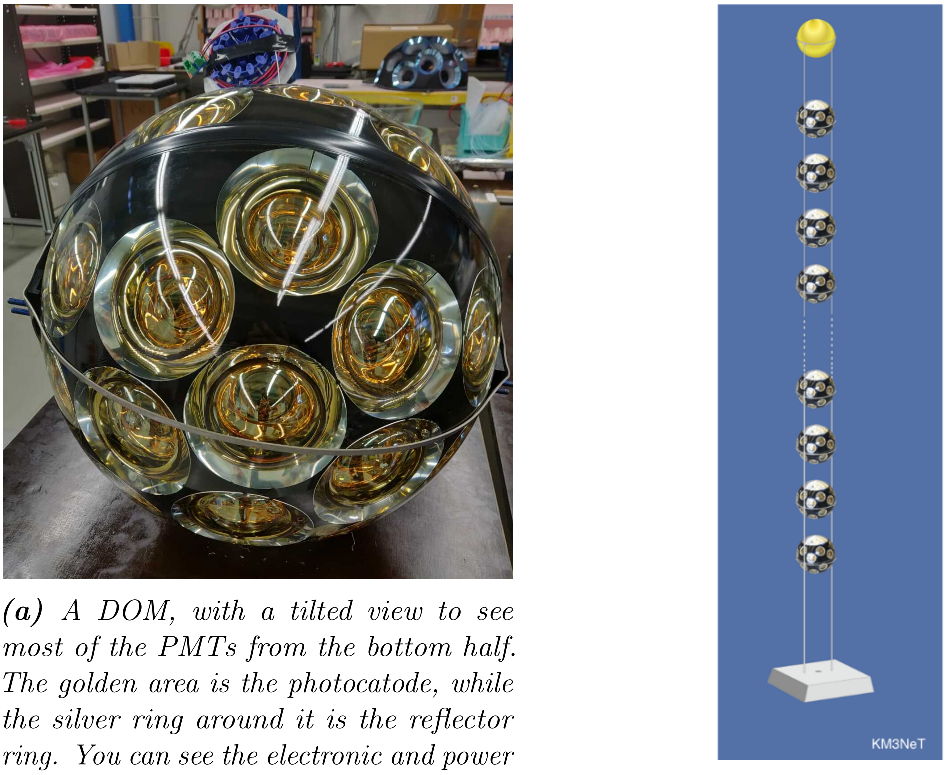
3.1 KM3NeT in a nutshell

KM3NeT, the Cubic Kilometer Neutrino Telescope, is a European-led, global collaboration that has integrated the efforts and ideas of is predecessors (ANTARES, NEMO and NESTOR) towards the task of building a network of deep sea neutrino detectors in the Mediterranean Sea[1]. As stated before, neutrino physics research involves very different fields and energy scales. For this reason, the KM3NeT collaboration is developing two different experiments: ARCA, near Italy, for astrophysics research and ORCA, in France (where ANTARES is still located), for particle physics research. How the collaboration is capable of running two experiments is one of the key ideas of this new neutrino telescope: A modular design of the seawater detectors allows for freedom to choose the energy range that will be probed by controlling the geometry (the components' density). The energy range is defined by the the effective volume and total amount of detectors that you have, as you can see in Fig 3.2.

The fundamental detection unit in the experiments is a Digital Optical Module (DOM, see Fig 3.1a), a pressure-resistant glass sphere that houses 31 PMTs (Photomultiplier Tubes), 5 horizontal rings of 6 PMTs that provide uniform angular coverage plus 1 PMT pointing vertically downwards, to increase coverage of the signals of upwards-moving neutrinos. Upward moving neutrinos are the only unequivocal neutrino signals because they come from the interior of the Earth, while downward neutrinos can be confused with an atmospheric muons. This system has an enlarged photo-cathode area of 3 to 4 times over a single large PMT (IceCube, ANTARES), depending on the presence of additional reflector rings. The DOM houses all necessary electronics: A FPGA-based digital readout board, Gb Ethernet with optical fibre to shore and calibration systems such as a LED beacon, a compass and a acoustic system for time, tilt and position calibration, respectively.

18 DOMs are joined together with vertical strong cables forming a detector unit (DU, Fig 3.1b), also called a line or string. To be deployed underwater, the strings are first curled in a 2m-diameter aluminum structure, transported to the corresponding site and anchored to the sea floor. Then the string, aided by a buoy at the top, slowly uncurls until gets its vertical shape. The aluminum structure floats to the top and is recovered, while the buoy remains fully emerged to help reduce drag and sideways motion.

[1] Besides the detectors and the support electronics, a few systems for long-term measurements of the deep-sea environment will be installed, which can be useful for Earth and Sea science projects ([32], pg 3).

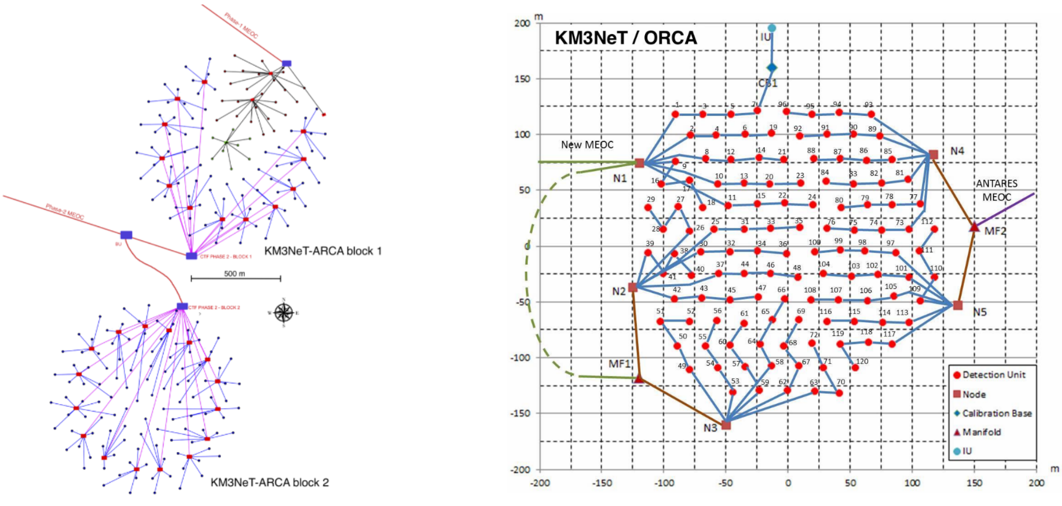


(a) *A DOM, with a tilted view to see most of the PMTs from the bottom half. The golden area is the photocatode, while the silver ring around it is the reflector ring. You can see the electronic and power connectors on top.*
Credit: Brian Ó Fearraigh.

(b) *Artistic rendering of a detector unit. Source: [32], pg. 12*

Figure 3.1: *KM3NeT modular components.*

115 strings, arranged with an hexagonal-triangular pattern, forms a building block, a fully-operational module for the detectors.



(a) *ARCA layout*

(b) *ORCA layout.*

Figure 3.2: *ARCA and ORCA detector unit footprints, 1 dot = 1 DU. The modular design allows the building blocks to have greatly different spatial scales. Source: [32], pgs 4,5.*

The expected resolution of a building block is astonishing: ~ 1 ns in time, $\lesssim 10$ cm in position and $\lesssim 10$ degrees in angular resolution, depending in the neutrino energy ([32], pgs 10, 103).

3.2 ARCA

ARCA stands for Astrophysics Research in the Abyss, and its main goal is the detection of high energy-neutrinos of cosmic origin and characterization of its sources. ARCA is the proper successor of ANTARES as a neutrino telescope, while ORCA, that will replace ANTARES in France, is a neutrino detector without astrophysical goals. Placed 3.5 km underwater, at around 100 km away from the coast of Sicily, ARCA will cover 87% of the sky. This field of view comes from using the Earth for shielding (a self-imposed restraint rather than an actual obstruction) combined with Earth's movement (not all parts of the sky will be available all the time). Luckily, its Northern Hemisphere position includes a privileged, full-year view of the Southern Hemisphere sky (unlike IceCube), where most Galactic sources are (as already seen by ANTARES, see the skymap in Fig 3.3 below).

Looking at figure 1.4, we see that the energy threshold of neutrinos of cosmic origin lies in the high-energy tail of the atmospheric neutrinos distribution, in the TeV region. Thus, ARCA will ideally cover the range between 100 GeV and 100 PeV (10^{11} to 10^{17} eV), but for source identification the lower bound might be placed at 10 TeV, where the resolution is improved and the atmospheric flux contamination is more reduced.

The associated flux with these energies is incredibly small, in the order of 10^{15} times the one for the companion ORCA detector. However, a rough estimate of the expected number of neutrino events per day in each detector (without precise knowledge of the sensibility) is 2-3 for ARCA and in the order of 200 - 300 for ORCA. ARCA will be able to compensate for the low flux with superior reconstruction and a greater effective detector volume. This is achieved by 2 adjacent building blocks of 115 strings (230 strings total), with an average spacing of 90 m horizontally between the strings, and 36 m vertically between DOMs in a string. This geometry provides an instrumented volume of 0.48 km^3 per DOM and over 1 cubic kilometer overall.

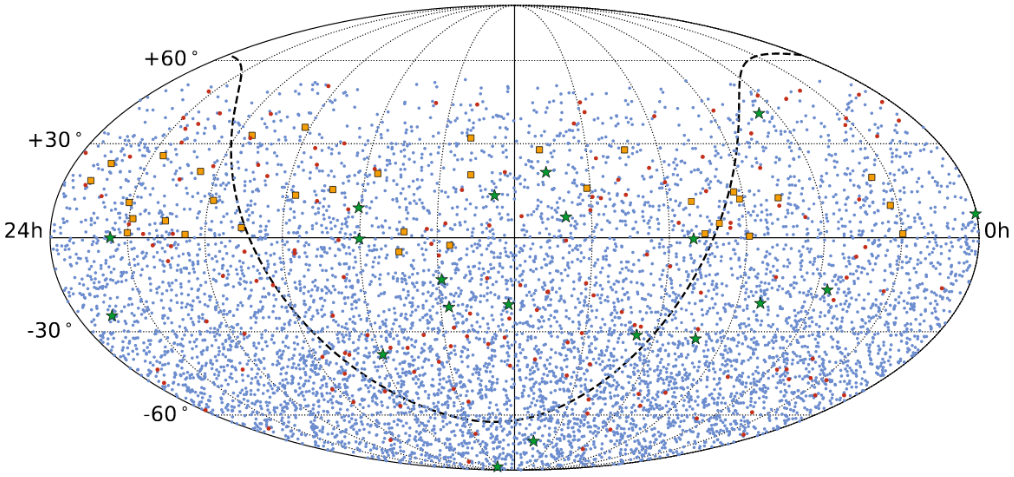


Figure 3.3: *ANTARES sky map of neutrino events, in equatorial coordinates (the Galactic equator is the dashed line). Blue and magenta dots are track and shower events, plus events from IceCube (yellow) and HESE (green). Source [45].*

3.3 ORCA

ORCA stands for Oscillation Research with Cosmics in the Abyss. Its main scientific goal is to determine the neutrino mass ordering via the detection of the neutrino oscillations, using this mechanism to calculate several of the fundamental parameters of the neutrinos. Placed 40 km off the coast of Toulon, France; at 2.5 km depth, ORCA will probe the peak of atmospheric neutrino distribution (see in Fig 1.4), which corresponds to the lower end of the energy range between 1 and 100 GeV .

As we have seen in sections 2.2 and mainly in 2.2.3, the neutrino oscillation probability depends on L/E. Atmospheric neutrinos of low energy (< 10 GeV), apart from a nice flux (Fig 1.4 again), happen to have the perfect energy to produce neutrino oscillations at the relevant baseline (distance travelled), see 3.4 below.

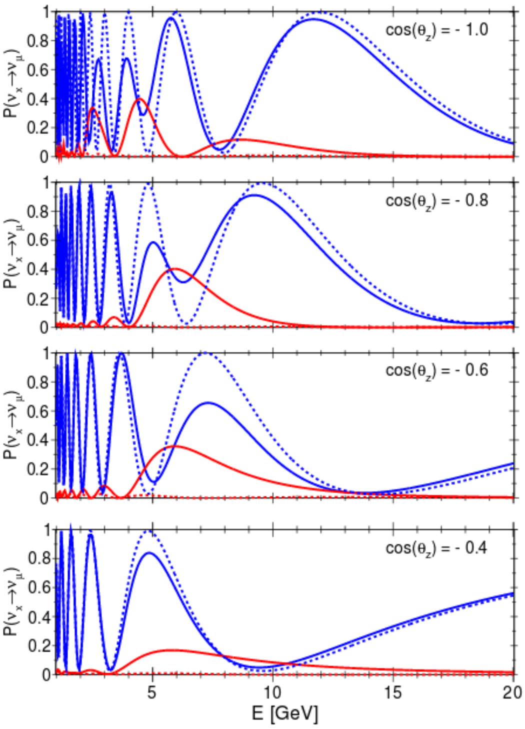


__Figure 3.4:__ Oscillation probabilities depending on the energy and the zenith angle for neutrinos (left) and antineutrinos (right). In ORCA, the Earth-crossing length is defined by the zenith angle θ_z, the angle between the neutrino event and the vertical z axis. This is thanks to the known position of the detector and a assumed cylindrical symmetry around it. The probability depends on the initial state: ν_μ (survival, blue) vs ν_e (oscillation,red) and the hierarchy: normal (solid line) vs inverted (dashed line) hierarchy. From [32], pg 60.

What determines how much its oscillation probability has changed is essentially the exact distance under the Earth that a neutrino crosses. This is because the oscillation calculation needs to take into account matter effects, described in 2.2.3. The biggest difference within an oscillation maximum shown in 3.4, with the Earth radius distance scale, is in the 4-8 GeV energy range, because the matter effects of the core of the Earth (the densest) cause resonant production [32], pg 61.

From the real expected probabilities in Fig 3.4, above, the differences are so small that, instead of the net event difference between inverted and normal hierarchy $(N_{IH} - N_{NH})$, an new parameter is used to calculate the ratio of events appearance, the asymmetry A':

$$A' = \frac{N_{IH} - N_{NH}}{N_{NH}}. \tag{3.1}$$

This parameter is used to define the amplitude of the difference in the signal that we expect to record, see Fig 3.5 below. While the magnitude of the this asymmetry is small, just measuring the sign of the asymmetry in both channels is enough to have a guess on the mass hierarchy.

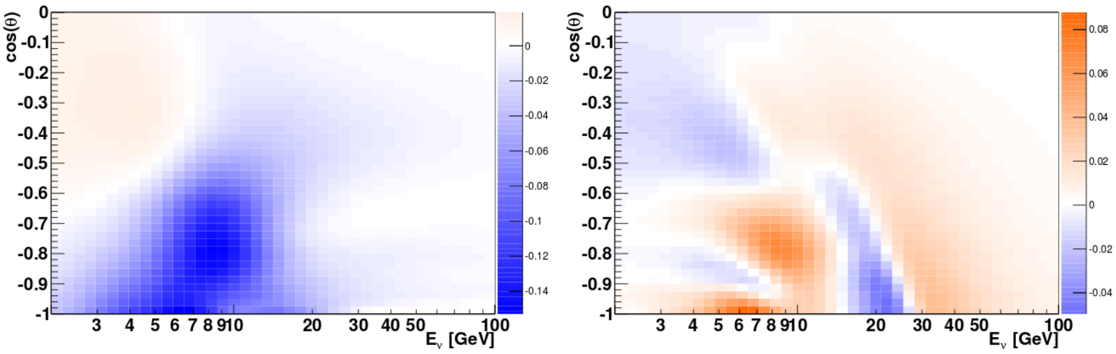


Figure 3.5: *Asymmetry between CC interactions for electrons (left) and muons (right), for different energies and zenith angles. This result is somewhat realistic, with energy and angle uncertainties imposed ad hoc. From [32], pg 56.*

Computing this probability distributions for muons and electrons is conceptually simple, but the recorded track-shower signal will not directly like this until the NC events, and the tau events have been untangled. Since we assume that it is not possible, this asymmetry distributions are combined for the different showers, adding more errors and smearing (even) more the small signal.

A pure muon and electron identification could be of the possible studies that currently are not available and could be possible if a 4 categories PID is built. This is one of the few use cases that motivates this work. Even if an improved PID places a more restrictive cut on our data, the better defined picture requires less statistics are needed to compute a good asymmetry estimate. The challenge here is to correctly identify muon and electron events between 4 and 20 GeV, since is where the asymmetry is the highest and the biggest difference between the two expected. We will see that below 20 GeV, the classification algorithms have a performance linearly dependant with energy, because we are in a range where so little light is emitted and/or recorded that a good classification is very hard to obtain.

3.4 Software Infrastructure

3.4.1 Triggering

The most basic electronic signals are the analogue current pulses from the PMTs in the electronics in the DOM. The signals, coming from the recorded photons, is almost constant. Only when the current excess a preset threshold, the particular hit is triggered. This single-PMT, low-level trigger is called "L0". KM3NeT operates via an "all data to shore" approach, which means that the L0 signals are sent to the shore station, including relevant information such as the timestamp. The expected data rate is 25 Gb per second per string, which raises two issues: One, that amount of data is too large, so it cannot possibly be stored continuously for several years, and two, we know that most of the data is useless. With the low neutrino event rate expected from the flux, all the other data is coming from backgrounds. Despite the shielding provided by the seawater, there are three main backgrounds: the decay of radioactive potassium 40 (K^{40}) dissolved in seawater, the high energy atmospheric muons that are able to reach our detector over 3 km of water and bioluminescence from marine creatures. To reduce the data, we look for coincidences between PMTs. The L1 or level one filter looks for two or more (up to eight) PMT coincidences in the same DOM in a small time window (typically, 10 ns). This still gives a L1 rate of 1000 Hz, out of which 600 Hz is K^{40} decays. Higher level filters, might use the orientation in the DOMs to further reduce coincidences. In addition, finer trigger trigger algorithms can be developed (biolumiscence is periodical, atmospheric muons are always downward-going). For this work, were we are using simulated data, we can work in the assumption that the events can be perfectly separated from the background. Since background events hardly contain any structure in the hit distribution, we can reject all backgrounds by imposing event selection. Event selection makes sure that only relevant signals are used in the analysis. A good solution, even if it is not perfect and misses to find actual events. This is especially an issue for sub-20 GeV energy events (see Fig 3.6).

3.4.2 Reconstruction and Particle Identification

When a event is flagged as L1, a whole snapshot of the detector is taken, with some margin to make sure no information is lost. Then, the event goes through both track and shower reconstruction algorithms: JGandalf for track, Dusj for shower[2]. The reconstruction aims at finding the relevant information about the event, like the energy and the direction (3-momentum). They both provide their reconstruction estimates, which is used for the posterior particle identification. Event reconstruction is a hard problem given the little information we have about an event, which makes it very sensitive to external conditions, meaning that the reconstructed output has a highly non-linear response with the signal recorded. For example, current direction resolution for the reconstructed track events is about 0.1 ° for ARCA (@ 100 TeV) but ~ 20 ° for ORCA (@ 20 GeV) ([32], pgs. 24 and 55). Reconstruction improvement is the current focus of much of the offshore work before the detector is fully operational. All the details about the reconstruction can be found at [32], sections 4.3 and 4.4.

[2] Dusj means shower in Dutch, and JGandalf means that somebody is a Tolkien fan.

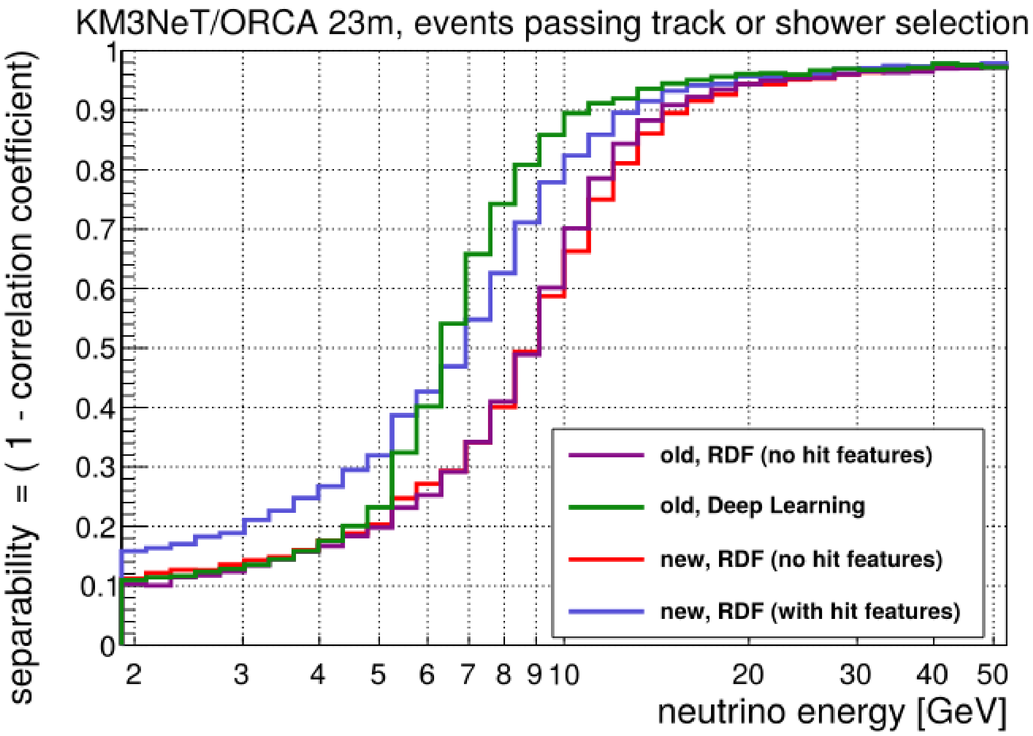


Figure 3.6: *ORCA's current performance of selection capabilities using two different Machine Learning algorithms. Separability is explained in section 4.3.2, while the RDF means Random Decision Forest. Credit: Stephen Hallmann, ECAP*

The current particle identification standard is a Random Decision Forest (RDF), a Machine Learning algorithm very efficient in the binomial classification (track-shower) problem. The RDF is an example of what is called "Shallow Learning", because it makes use of explicit variables (features) that we are giving to it. It is called a forest because is a combination of a hundred different decision "trees". Every tree takes a different subset of the reconstructed values in a random order and tries to find the best threshold in those variables that separates between track and shower events. A tree can learn a really specific combination of good features that hold information about the distinction, but generalizes badly to every event. This is why a "forest" is used. Let all the trees "vote" which category they predict and determine a final event classification probability simple as the fraction of trees that voted for the category.

3.5 Mass ordering fit

The mass ordering fit is the main analysis of the ORCA experiment, which aims to solve the mass hierarchy problem and constrain the fundamental neutrino parameters that we are are sensitive to: Δm_{23}, θ_{23} and δ_{CP} [40]. The mass ordering analysis essentially needs the information about the flavour of event, the energy and the angle, so a successful reconstruction and classification are required. This important and complex analysis is detailed in section 4.6 of [32], and here only the most significant details are covered.

Assuming enough ORCA events (MC or otherwise), the whole goal is to build a probability distribution function (PDF) or oscillogram (see Fig 3.7) in three dimensions corresponding to the number of events of one flavour vs the two free parameters: E, the energy and the distance $L \leftrightarrow \theta_z$.

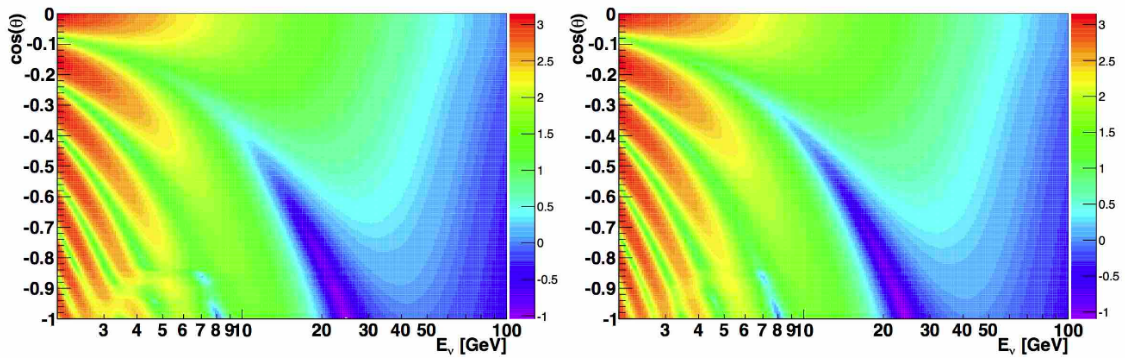


Figure 3.7: *ORCA Oscillogram for muons ($\mu + \bar{\mu}$), with the event rate in log scale and units of $GeV^{-}1$, $sr^{-}1$, $yr^{-}1$ for both normal hierachy (left) and inverted hierarchy (right).*

With the oscillogram characterized, the following step is to find the best fit to it for the neutrino parameters (the mass hierarchy, the mass differences and the oscillation angles) within the boundaries that we already know. This best fit is determined by a standard log-likelihood ratio (LLR) computation. As always, the goal is to reject the wrong hypothesis, so for the calculations using MC data, this data is generated assuming the opposite hierarchy than the one that is going to be fitted. After the 'best fit', we can reject it if the separation between the two distributions is 3σ or more. For real data, without prior information, we simply fit the data to both hypotheses and see which one is rejected. The sensitivity, the summary of this process, is defined in terms of the separation of the Gaussian likelihoods for the two hypothesis as

$$S_{true} := \frac{\mu_{true} - \mu_{alternative}\sigma_{alternative}}{,} \tag{3.2}$$

with μ and the mean and width of the LLR of the (assumed) true and alternative hierarchies, where normal hierarchy is usually considered as true.

However, this calculation, summarized in 3.8, is far from a trivial computation. The differences in Fig 3.7 are very subtle, resulting in a very complex analysis required to achieve the needed resolution. It involves an end-to-end recreation of the whole experiment, from neutrino flux to reconstruction, and needs to factor all the possible variables that will affect the result, including the associated sources of error and uncertainty (called systematics, as opposed of the statistical uncertainties).

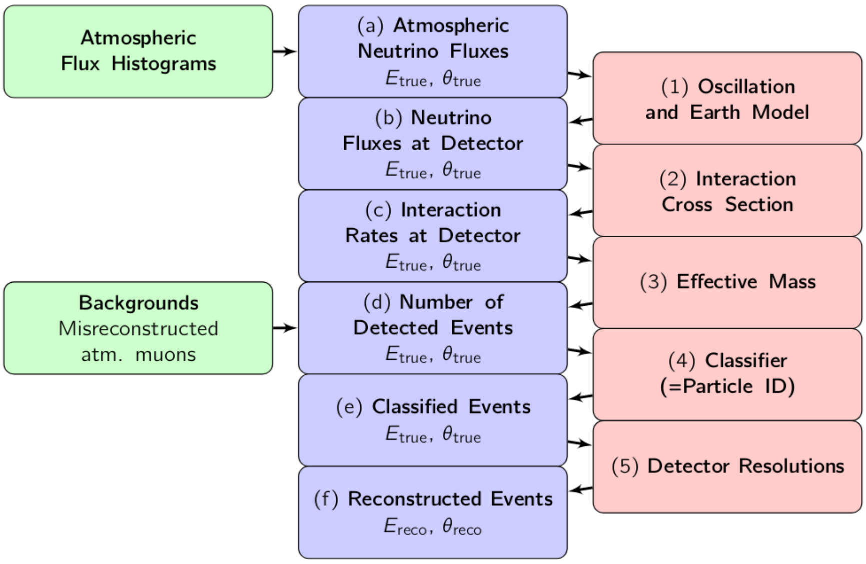

Atmospheric Flux Histograms
(a) Atmospheric Neutrino Fluxes
E_{true}, θ_{true}
(b) Neutrino Fluxes at Detector
E_{true}, θ_{true}
(c) Interaction Rates at Detector
E_{true}, θ_{true}
(d) Number of Detected Events
E_{true}, θ_{true}
(e) Classified Events
E_{true}, θ_{true}
(f) Reconstructed Events
E_{reco}, θ_{reco}
Backgrounds
Misreconstructed
atm. muons
(1) Oscillation and Earth Model
(2) Interaction Cross Section
(3) Effective Mass
(4) Classifier (=Particle ID)
(5) Detector Resolutions

Figure 3.8: *Computation chain required for mass sensitivity studies. The top half, until (d), is the event generation part, and involves (a) The neutrino atmospheric flux, taken from the Honda tables; (1) the raw oscillation probability, calculated from OscProc software and including matter effects; (2) the cross sections, from GENIE software and (3) the effective mass of the detector, known from MC simulations. From [32], pg 95.*

After the very expensive and difficult process of generating events, or the even harder process of building the detector and taking data, the fact is that most events lay in the low-energy, high-flux range. This means that they will be reconstructed poorly, and then rejected for the following mass hierarchy fit. This is why, even with hundreds of expected events per day in ORCA, the sensitivity studies are still measured in years of data taking.

A better classification, however small, is key to improving the mass hierarchy fit. Any improved classification, especially in the lower energy range, will provide a lot more events for a posterior mass hierarchy fit study. This is why, for this work to have a positive contribution to ORCA, it does not need to bring revolutionary changes. Even a small improvement in the current track-shower scheme in a key area like the low energy range can mean a significant boost to our analysis capabilities. This can be translated in a estimated improvement in the sensitivity of months to a year. A previous attempt to bring the flavour information via parametric reconstruction [36] showed a minimum sensitivity estimate from 3.22 to 3.28σ, a 7% improvement.

4 Methods

This section covers the description and discussion of the specific tools used for this work, both for coding and for statistical analysis. For a general introduction to what Machine Learning (ML) and Deep Learning (DL) is, refer to appendix A.2 and for an idea of the steps required to replicate this work, please refer to appendix C.2.

The main deep learning based framework for the ORCA detector in the KM3NeT collaboration has been developed in ECAP, Germany by Michael Moser, Stefan Reck *et al.* It contains a preprocessing software called OrcaSong and a DL toolset called OrcaNet. OrcaNet is sometimes used to refer to the full framework, but in the following the difference mostly conserved for clarity. OrcaNet is the first effort that aims to establish a shared, coherent methodology and toolset for this type of projects with the goal of sharing efforts and making projects and ease into continuing current projects. This is why OrcaNet is used here, along with the good state of the development, documentation and support.

4.1 OrcaSong

OrcaSong is the software library used to produce "images", that can be later processed by OrcaNet. The "images" simply refer to 2,3 or 4 dimensional binned histograms filled with ORCA detector information. A general, almost trivial, concept in ML is that the algorithm (OrcaNet) will perform better if it has more information available to it. This implies that OrcaSong's goal is to format and pass all the information from the hits that is available to us with the minimal amount of losses and, preferably, in the most efficient way for the posterior computation.

For every hit recorded, there is 5 dimensional array of data (X,Y,Z,T,P): the 3D position of the DOM, the time of the hit (inside the recorded event) and the PMT channel within the DOM. Due to design constraints on the underlying software, OrcaNet cannot currently process more than 4D data, which forces us to reduce a channel either by a) lose some information, b) taking lower dimensional projections (and still losing some information) or c) combine two channels. The naive solution used in this work, due to computational constraints, was a). The P channel was discarded, making this work's reconstruction limit at DOM level. This decision was taken to reduce the required computational preprocessing of the data, since DOM level resolution is enough for the goal of this project, to prove that we can reconstruct different shower signals, in spite of obviously limiting our maximum possible performance.

The option used instead of losing the P channel in other cases is b). Keeping in mind that all lower dimensional projections do not equate a higher dimensional figure, we can still strive for better performance than losing the channel altogether. OrcaSong is able to produce 2D, 3D and 4D images. Decomposing the 5D image of an event on the detector can be done in 10 different 2D images (x-y, x-z, etc.), which, if combined carefully, should minimize the information loss. The same 5D image is also possible to be decomposed in 2 4D images (XYZT, XYZP), a much favorable decomposition with an immensely higher amount of information and computationally faster. This is why 4D is the standard for most of the Deep Learning work in OrcaNet.

Dimension reduction, the c) option, does not have a clear implementation. For example, integrating the PMT distribution within the DOM for more precise position coordinates. However, this increases our number of spatial data points by a factor of 30 (since there are 31 PMTs per DOM) in exchange for corrections in the order of a few cm. This idea is disfavored since the simple model that we are using is already limited in computation power and time. Also, the position uncertainty is as big as $\simeq 30$ cm, from the associated time resolution of ~ 1 ns. A very new solution keeps P channel as an independent variable and simply stacks the two images along the 4th channel, making XYZ(T+P) images, twice as big with minimal information loss. The neural network has no physical meaning behind the channel dimensions, so OrcaNet (if it has enough free parameters) should be able to find the respective time and direction correlations with the two halves of the provided images.

To summarize, here OrcaSong makes a image of the detector at DOM level, binning the data in a 1 bin = 1 "pixel" = 1 DOM principle, and each bin contains the hits per DOM (summed over all PMTs) during the time of an event, making it a XYZT image. For a full, 115 string detector, the minimal spatial binning (X,Y,Z) that does not change the spatial orientation of the DOM is 11x13x18 bins, see 4.1.

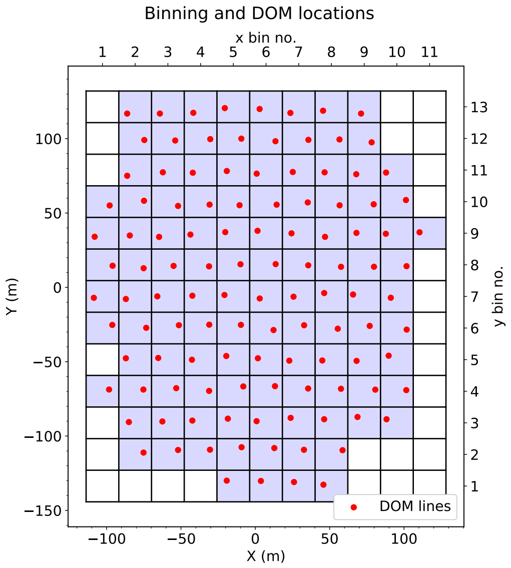


Figure 4.1: OrcaSong spatial binning scheme. Credit: Stefan Reck, ECAP.

The time dimension will receive the same binning treatment as the spatial dimensions, but since it is not fixed by geometry (DOM positions), its binning is, in principle, a free parameter. How do we find an optimal time binning?

First, let's look at the time duration of an event. Different events obviously have different times, mostly depending on the energy, so an event snapshot covers 3 μs to make sure that no information is lost. This is much more than the longest time an event would take to pass through the full detector. If we enclose the ORCA volume in a orthogonal box, the dimensions will be, roughly, 200 x 200 x 160 m, which has a diagonal of 325 m. If a relativistic particle crosses the diagonal, its Cherenkov cone, moving at the speed of light on the water ($2.25 * 10^8$ m/s) would take 1.444 μs. This means that, in the best case scenario, at least half of the event snapshot is empty of signal hits, just recording the background noise. A more realistic average time scale for the (simulated) events can be found via the attenuation scale. The attenuation of the light propagation in seawater is dominated by absorption. The absorption length of blue light in water measured for ANTARES [27] is 60 m, which translates to a time scale of $\simeq$ 260 ns. This means that, after the photons are emitted, 98.2% of them will disappear in 4 interaction lengths, or roughly 1 μs ($0.982 \simeq 1 - e^{-4}$). Coincidentally, this is close to the effective scattering length of blue light in water, as measured by ANTARES, which is 265 m. The conclusion is that an average ORCA event will have a duration of 3 absorption times ($\sim 750ns$) after the light is emitted (at the interaction point for showers, while for tracks it depends on the muon path length).

Second, we will look at time resolution. The minimum binning is found in the intrinsic time resolution expected for ORCA, around 1 ns. But using 1 ns resolution implies having images with 750 to 3000 bins *per image*, clearly far from the optimal value considering the time it would take to train that many bins. 1 ns resolution also guarantees that most of the event bins have very few signal hits within them. A good time resolution has to be so small that does not lose spatial information by overlapping hits. The lowest spatial distance between 2 neighbouring DOMs in a string is 9 m, which means a light resolution of $\sim$ 40 ns.

Third, you need to establish good time cuts, in order to provide time regularization (same time width per bin, so all bins are equivalent), but mainly to ensure that the event has as much as signal hits (from the interaction) and as little as background hits as possible. OrcaSong's solution is to select the peak of the amount of hits (photons) of the event and cut according to the mean distribution of hits calculated from the data.

This makes the cuts for this work in the [-250, 500] ns range over the peak. This means 750 ns total ($\sim$ 3 attenuation lengths) and, using 60 time bins ($\lesssim 100$ is a computationally sensible choice), it provides the nice resolution of 12.5 ns/bin. This setting is called tight-1 in OrcaSong, see C.2.

The PMT channel is a discrete number and needs no special binning when used.

The binning choices used in this work are not the only possible ones. Similar efforts for ARCA had different principles: M. Post [50] used a TXYZ = 50x13x13x18 binning (with a time bin width of 400 ns) to map the hexagonal layout into a square one, trying to retain the spatial correlations between different strings, while C. de Sio *et al* used a TXYZ = 75x16x15x18, (with time bin width 12 ns) system, in an effort to truthfully recreate the hexagonal distribution. These efforts increase resolution, but at the cost of producing highly sparse images XY plane: Post leaves $13*13-115 = 54$, or 32% of unused pixels while de Sio leaves 125 unused pixels, or the 43% of the total. These always-empty pixels are a hindrance in computation, as they require computation memory and time dedicated to them, that is why binning used here is a more coarse, but efficient one, with only 28 empty pixels or the 19.6% of the total.

4.2 OrcaNet

OrcaNet is a software package that helps building a Convolutional Neural Network (CNN) architecture using the images from OrcaSong. It is a complete framework that controls all the steps of the model, keeping track of the state of the model from the initial configuration, during training, until prediction. OrcaNet is built upon Keras, one of the most popular high level Machine Learning libraries due to the user friendliness, ease of use and freedom to choose any ML framework. The back-end framework used here is the most popular today, Tensorflow (TF), but thanks to multi-backend solutions like Keras this choice is less and less relevant.

A technical introduction to the CNN model used here can be found in appendix A.2, so this section is meant to discuss the specifics details relevant for this work. Let's start by stating that CNN are essentially the ML models behind computer vision. The qualitative understanding is that a computer can "scan" an image to obtain complex information, this "scan" operation is mostly the convolution operation, hence the CNN name.

Since it was developed for image recognition, Tensorflow convolution operation only works up to 3 dimensions, namely the dimensions for a color (RGB) picture or greyscale video. A 3D convolution in TF allows for 4 dimensional input, with the 4th dimension used to determine the kernels, which are the different filters used for the fist convolution (scan) of the data. This last dimension, called channel dimension, is not convoluted and is influence on the model is reduced to the first layer[1]. The fact is that non-convoluted information from the channel dimension will not be incorporated into the data equally compared to the other dimensions, requiring at least as many filters as input channels to attempt to capture the information without loss. Hence, the lack of 4 and 5 dimensional convolutions is the most limiting factor[2] in our current models and what forces us to do the OrcaSong preprocessing choices discussed above. A change of backend to another that allows for 4D or 5D, thanks

[1] The system has been nicknamed "3.5D convolution", where 4 dimensions are taken but 3 are convoluted.

[2] On a technical note, this issue is just a lack of software implementation, since the underlying technology called CUDA allows convolutions until 8 dimensions.

Category	Train	Validate	Test
elec - CC, 1-5 GeV	1 - 250	301 - 400	401 - 500
elec - CC, 3-100 GeV	1 - 750	751 - 1000	1001 - 1250
muon - CC, 1-5 GeV	1 - 250	301 - 400	401 - 500
muon - CC, 3-100 GeV	401 - 500	751 - 1000	1001 - 1250
tau - CC, 1-5 GeV	——	——	——
tau - CC, 3-100 GeV	1 - 750	751 - 1000	1001 - 1250
elec - NC, 1-5 GeV	1 - 250	301 - 400	401 - 500
elec - NC, 3-100 GeV	1 - 750	751 - 1000	1001 - 1250
Total event no.	13638226	3507574	2539162

Table 4.1: *Events file numbers from the ORCA, 115lines, 23x9 spacing, 2016 production used for this work. There are no low energy tau events simulated since the tau rest mass is 1.776 GeV. The NC event files are simulated with electron neutrino but are obviously universal to the 3 flavours, since flavour does not play a role.*

to Keras, is among the next steps to increase OrcaNet performance.

After the fundamental considerations, let's discuss the model itself. A CNN model is defined by its architecture, which is the name for the nature and the order of the layers used. The architecture for the model here and its layers are described in the appendix A.2. About the architecture, two main details are relevant, is one of the best for image recognition, and is very slow to train. The model used here has 45 layers (only 10 actually perform convolutions), and the number of events used here is less than 20 million total, see table 4.1 below. This makes our model on the small side for the current standards, and still the average training for the model took around a week, with $\sim$ 12 hours per epoch of training and $\sim$ 10 epochs to converge (an epoch is a fully cycle over the data). This is still acceptable for the time constraints, but on the longer side of an effective training.

The computational power is not the issue here. The training was realized on an Nvidia Geforce GTX 1080 GPU of 2018, still a powerful option to run Tensorflow. Increasing the power to a multi-gpu setup is not an option in this project. However, the computational needs of the model are the main factor behind the constraints on the performance of this work. This constraints have been mentioned already due to its influence on the shape and resolution of our data, but, more importantly, they limit the total amount of events used here (since this work was based on a fixed amount of training time, not necessarily the best choice for flexibility and speed but great for cross-checking the different iterations of the models).

The data is the fundamental factor behind the final performance of the model. An objectively worse model, or less optimized one might have a better final performance if it has in the order of ten times more data. Having as much data available as possible for the training is best practice in ML and crucial for DL. However, our case is one of the few where the data is not the limiting factor since there are enough simulated events in ORCA (in the order of tens of millions of events per flavour), and more simulated events can be produced upon request. This abundance is also why, if the simulated data is a realistic recreation of the physical phenomena, the use of DL models in this experiments might be the best option, performance wise. However, since the before mentioned computational constraints from the model are a huge limiting factor, only a fraction of the total events was used. The inclusion in the OrcaNet code of newer model architectures (like the ResNets), should they be more resource efficient, would translate to an improvement in performance, especially if they can train with more data in the same time.

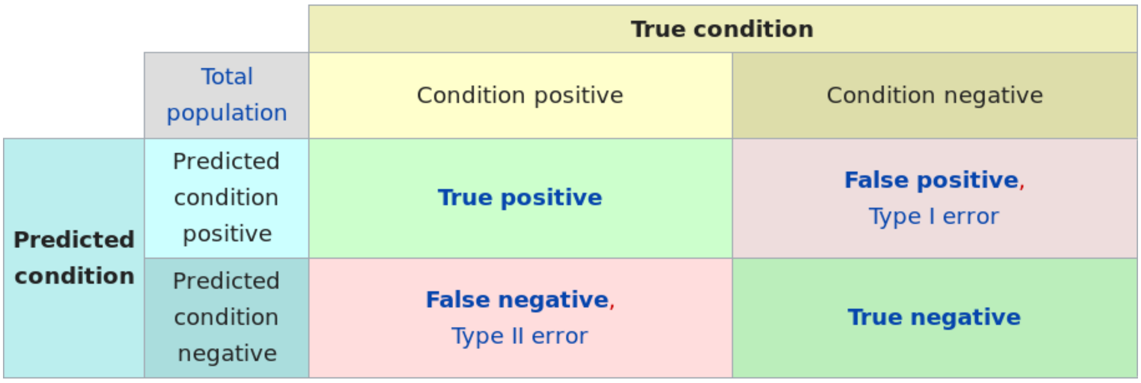
The events used in this work (see in table 4.1) are already defined in the train-validate-test categories when passing them to OrcaSong to make the images. The amount of files per category was a rough estimation of 75% for training, 25% for validation, and then a secondary test dataset roughly the same size as the validation. The low-high energy file ratio was chosen in accordance of the total file ratio ($\sim 1/3$) and in order to have as much as low energy events as one of the other categories. Note that the numbers refer to event files, where each event file contains a similar number of events. The total number of events are 19684962, a bit shy of 20 million, split into approx 13.5 M - 3.5 M - 2.5 M for train, validate and test samples, respectively. Smaller datasets were used for testing and debugging, but their results are of no interest here.

The events are generated in two different energy ranges (1-5, 3-100 GeV) because they obey different power law spectrums (number of events vs energy). Combining the two productions means that there are some more events in the 3-5 GeV range than a quasi-realistic distribution. This can be fixed with posterior modifications to the dataset, but it is not done since it should have no effect on the training, given that it does not see the energy as a training parameter, so it should not pick the associated distribution (If it did, it would mean that it is reconstructing the event's energy far more efficiently than JGandalf and Dusj). Also, due to the needs of the oscillation analysis, the most populated bins are in the 4-7 GeV range. A flattening of the energy spectrum can eliminate a significant fraction of useful events, so it is not recommended to do so, especially if it is not warranted by data.

This thesis contains the results of three different trained models, where OrcaNet has 2,3 and 4 output options for classification: 2 category is track-shower, 3 categories is track-EM shower and the rest (NC-showers and tau showers) and 4 categories is for every different signature. The goal here is to establish if attempting to reconstruct more signals benefits (or harms) the reconstruction power. Thus, the 3 case models are compared under the same conditions and data, in order to establish a clear connection between the number of categories and the difference in performance.

4.3 Statistical Classification

The goal of our analysis is that our results, coming from a simple category classification can produce conclusions with the statistical power of a maximum likelihood analysis. Starting with binary classification (2 categories), we can express our millions of tested events as a simple confusion matrix (Fig. 4.2), the first step towards the construction of any statistical metric.



		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Figure 4.2: *The confusion matrix, also known as error matrix or 2x2 contingency table. Any 2 category classification system can be expressed as a condition test. In our case, our true condition will be to be a track, so condition positive = track and condition negative = shower. Detail from [52].*

While we can in principle build a contingency table for any number of categories n as a nxn matrix, that is useful to see what are the biggest confusions during identification. What we want in this case is to treat each category independently during analysis, to study the appearance of a single category vs "the backgrounds". This is possible by simply building n different confusion matrices, one for each category. When we are with 4 categories, 4 2-category tables also simplifies some posterior computations.

4.3.1 Threshold

A confusion matrix requires binary predicted output per category per event (the event has either a true or false value for every category). Meanwhile, our prediction output for an event is an array $\bar{p} = [p_1, ..., p_n]$ with the probabilities p_i for every category $i \subset [1, n]$, where $\Sigma(p_i) = 1$ (also known as *one hot encoding* system). A trivial idea would be to assign to every event the category i with highest probability, $p_i = max(p)$. The issue is that if $n > 2$, then is not possible to know the maximum from a single p_i, which makes you evaluate the whole event, and that complicates the construction of the n confusion matrices (we need to evaluate millions of events).

The solution is to arbitrarily set a threshold probability value p_{th}, and reduce the probabilities in $\bar{p}$ to a binary output by applying the step function: $\Theta(p_i - p_{th})$, with category i true if $p_i > p_{th}$. The threshold is a free parameter, but if it is too low, more than one value of $\bar{p}$ might be above it, that is, the same event might end up being classified as predicted for more than 1 category, and will thus account as a false positive instead of a true negative. The extreme case would be when the cutoff threshold is lower than the random guess ($p_{th} < min(p_i = max(p)) = 1/n$). This would give the trivial, wrong solution that all categories are true.

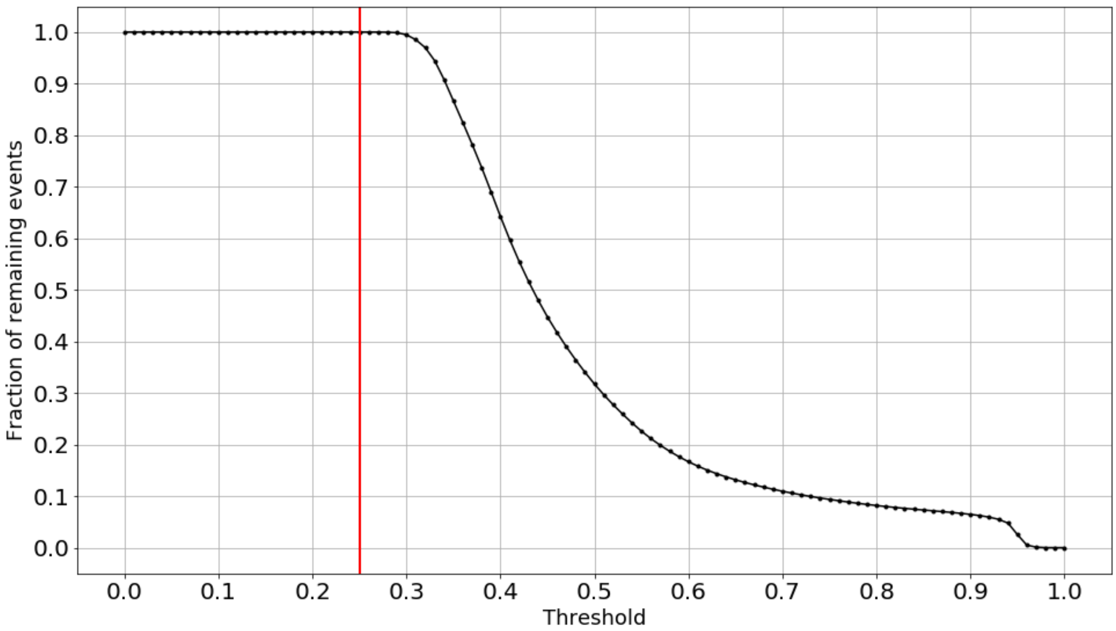


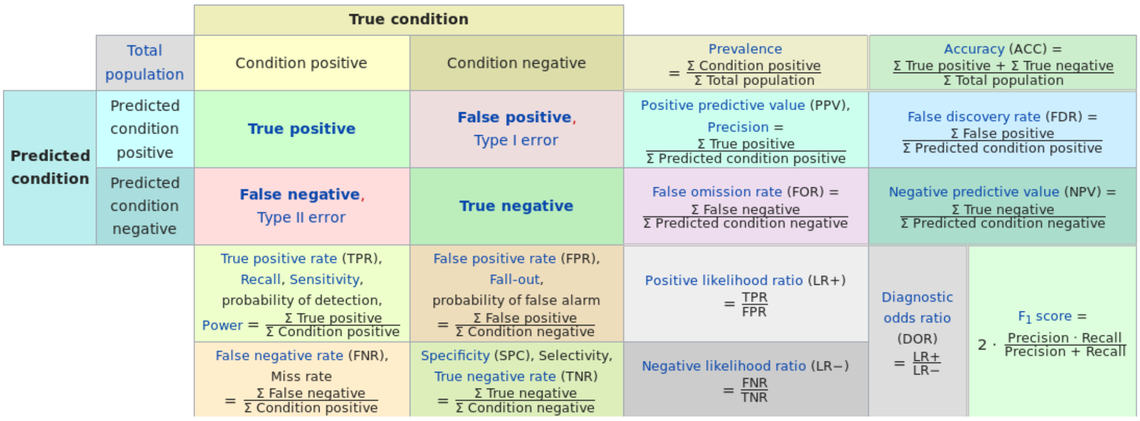
Figure 4.3: *The event loss due to the threshold for our predicted values. The red line is the random guess $p = 0.25$ for 4 categories, where below no event loss can happen. We can see a significant sudden decay, so every posterior analysis must keep in mind the effective number of events where the statistics are happening.*

To avoid this problem, we can set $p_{th} \geq 0.5$. This way we know that if $p_i > p_{th}$, then $p_i = max(p)$. It is no coincidence that 0.5 is the the random guess value for the 2 category system. However, if $p_i = max(p) < p_{th}$, the event (with all its probabilities low, even close to the random guess) might not account in any of the confusion matrices. This means that, for $n > 2$, our confusion matrices might not include all the events from our test dataset if the threshold is higher than the random guess. The events that are discarded (ignored) are those whose maximum probability is close to the random guess. Thus, the error introduced by a higher threshold is just a decrease in our statistics, however it eliminates the worst examples in the data. This "worst example" events are literally the ones with the worst prediction from OrcaNet. This decrease is, then, acceptable, as long as it does not interfere with the confusion matrices. This is also why our evaluation datasets range in the order of million of events, so that the lack of statistics is not an issue even if we use only a small percentage of the events.

What is more, the threshold is not connected to the ML model, it is just an extra choice that we impose to analyze the data, so we will need to explore its effects on the estimators and its values, since it might significantly alter our statistics and certain threshold values might not make any sense, even if they alter favourably our statistics (see ROC section).

4.3.2 Estimators

From the confusion matrix stems a plethora of different ways to estimate the performance of the classification. An estimators is essentially a linear relation between two or more of the 8 categories in the confusion matrix shown in Fig 4.2. True condition, Predicted condition, True label and False label, where each of them can be positive or negative.



		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	True positive	False positive, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	False negative, Type II error	True negative	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Figure 4.4: *The confusion matrix, extended to include the most simple and straightforward estimators, usually relating 2 of the categories in the confusion matrix or other estimators. Credit [52].*

As we can see in Fig 4.4 above, the most simple 8 estimators come from comparing the true and false positives with the predicted categories (PPV, FDR, FOR and NPV) and with the true condition or MC truth (TPR, FPR,FNR,TNR). Estimators are commonly referred as "accuracy measures", and their names usually interchanged with accuracy. Accuracy is the most popular and universal measurement of performance, but it has its own definition. In cases like this one, a careful choice of estimators and their proper naming is fundamental to avoid confusion. Besides, in complex classification cases like the one in this work, the accuracy is not even the best estimator possible.

For example, consider a dataset containing real shower events. Assume that it is composed of 90% electronic showers and 10% of tau showers. If we try to benchmark different electron vs tau showers classifiers, a very bad classifier that might ignore the tau signal and return 100% for the cases as electronic shower, with an associated accuracy of 90% (since all the electrons are correctly identified).

This is why some complex estimators have been defined in terms of other, more simple estimators. If a "simple" estimator uses the information of 2 categories, and thus reflects a single aspect of the data, combining these estimators brings more information together, making it less intuitive but more representative of the underlying data. A great example is the F_1 Score, the harmonic mean of the precision (PPV, rate of true positives among the predicted positives) and the recall (TPR, rate of true positives among the MC positives). This is a balanced estimator, more accurate and less prone to be biased in case the underlying distributions are uneven. The generalized version of the F_1 score is the F_β score, where the weighted average would give more relevance to one of the components. In our example before, the F_1 would somewhat penalize the missed tau events; and a F_β can be defined to penalize stronger the false positives, so that the classifiers that did not miss the tau events would score higher. The F_1 score has been used plenty during the development and training of the network, and it might as well be main training metric for OrcaNet in the future.

The chosen performance estimator for this thesis is the one considered to be the best single metric for a confusion matrix: The *Matthews Correlation Coefficient* (MCC) [28], which takes this non-linear form:

$$MCC = \sqrt{PPV * TPR * TNR * NPV} - \sqrt{FDR * FNR * FPR * FOR}$$
$$= \left\{PPV{*}TPR{*}TNR{*}NPV\right\}^{1/2} - \left\{(1-PPV)(1-TPR)(1-TNR).(1-NPV)\right\}^{1/2} \tag{4.1}$$

The MCC combines the 8 basic estimators mentioned above, while the F_1 score combines 2 only, so it effectively includes the full information of the confusion matrix, weighting equally the appearance of positive and negative cases. The consequence of this is that meanwhile the other estimators range from 0 (random guess) to 1 (perfect classification), the MCC ranges from -1 (Perfect misclassification) to 1 (Perfect classification), with 0 the random guess value; so it is able to measure worse-than-random performances. The results in this work use the separation variable, which is simply 1 - MCC. This is because we are not interested in the correlation or overlap between underlying distributions but the separation between them, where the distributions are the different event signatures. This simply means that we ignore the below zero correlation values (not present but not relevant anyways) and that we invert the correlation plot, to keep it in accordance to the general estimator definition of 0 = worst performance = random guess, 1 = best performance = perfect classification.

The MCC can be understood as the application of Pearson correlation coefficient for the contingency matrix. This means that the MCC is a familiar concept, that can be compared to other Pearson correlations from a completely different analysis or be related to the chi-square (χ^2) using the total number of observations n:

$$|MCC| = \sqrt{\frac{\chi^2}{n}}. \tag{4.2}$$

Technically, the MCC comes from the geometric mean two estimators $MCC = \sqrt{J * \delta p}$. Youden's index (J) or *Informedness* the "correlation", the fit to the linear regression of the predicted and the true distributions and its dual. δp, the *Markedness* is its dual or complementary, and signs the causality strength of the correlation. This definition makes MCC a resilient estimator for when the data in the confusion matrix is unbalanced, since it is making an attempt to describe the shape of the underlying data distributions, not the performance of a particular measurement. What is more, as long as the confusion matrix is balanced (as is this case), the value of the MCC is equivalent to the actual accuracy estimator, and can be loosely referred to as an "accuracy measure".

4.3.3 The receiver operating characteristic

The receiver or relative operating characteristic curve (ROC), and is the sensitivity or recall (y axis) vs fallout (x axis) plot. [28] The sensitivity represents the "accuracy" for the positive class, how much is the classifier right and the fallout the false positive rate or how much the classified is fooled into a positive classification, the contamination.

If we can combine the values, as the MCC does, why using a 2D plot? Because the ROC helps us visualize the impact of the threshold. For any probability dataset in the (0,1) range, a threshold of $p_{th} = 1$ will turn zero correct guesses (TPR = 0) and zero incorrect guesses (FPR = 0), while a $p_{th} = 0$ turns all correct guesses but all possible mislabelling (TPR = FPR = 1). This means that all ROC curves have their extremes fixed at (0,0) and (1,1).4.5, and that curve can be drawn by moving along the different threshold values.

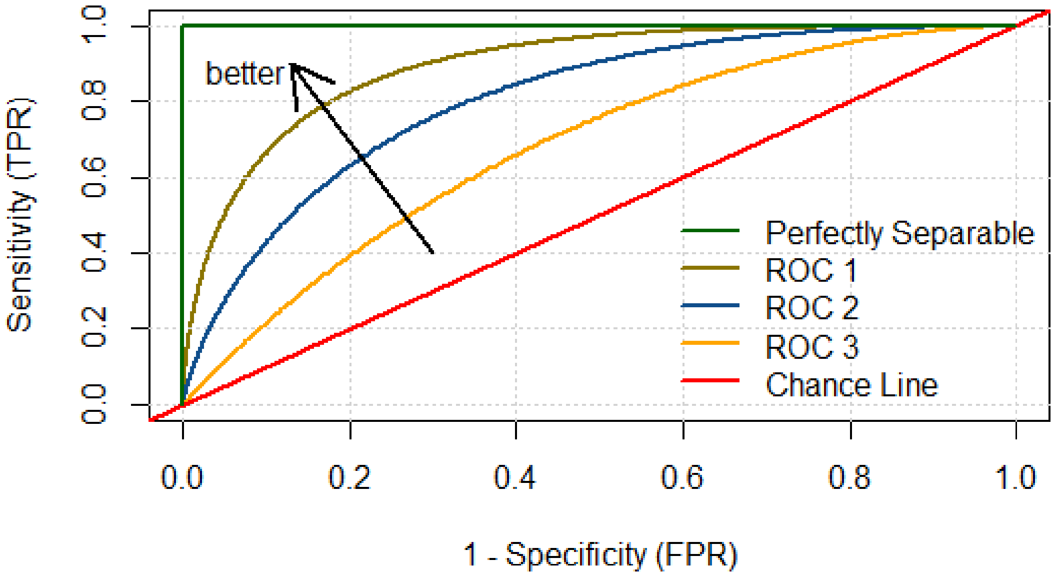


Figure 4.5: *The ROC curve space. The red line is the chance or random guess line and represents TPR = FPR. The other colors represent different performances, different metrics or different datasets. The green line is the perfect case, where the curve reaches the point (0,1). The threshold decreases while we move along the curve from left to right. From [49].*

It is remarkable that curves only move along the $y = x$ random line, since any value below would mean worse than random. This has motivated an estimator associated with the ROC, the AUC (area under the curve), that can be found by integrating the ROC between 0 and 1. This metric is closely related to the MCC, as they both try to measure the "overlap" between the true and the predicted distribution. In fact, the distance between the ROC and the chance line (that depends on the threshold) is the informedness (J), one of the components of the MCC. The MCC also focuses on the inverse relation (the "significance" of the measured overlap), so it is still more adequate to use, making the use of the AUC pretty redundant in this case.

5 Results

After the model has been trained until convergence, we are ready to see how well it is able to separate the event signatures. The 3 categories (track/shower/tau+NC) and the 4 categories model (track/shower/tau/NC) give very similar results. Whenever it was redundant, the 3 categories figures have been omitted.

5.1 Prediction probability distributions

5.1.1 Track probability distribution

The first step is to visualize the raw probabilities of the predictions for the data. Orcanet's prediction probability output is an array with the associated category probabilities per event (one hot encoding). The track category (one of the columns in this one hot encoding array) is, then, the same as the track quality score in the standard particle identification, seen in 5.1 below.

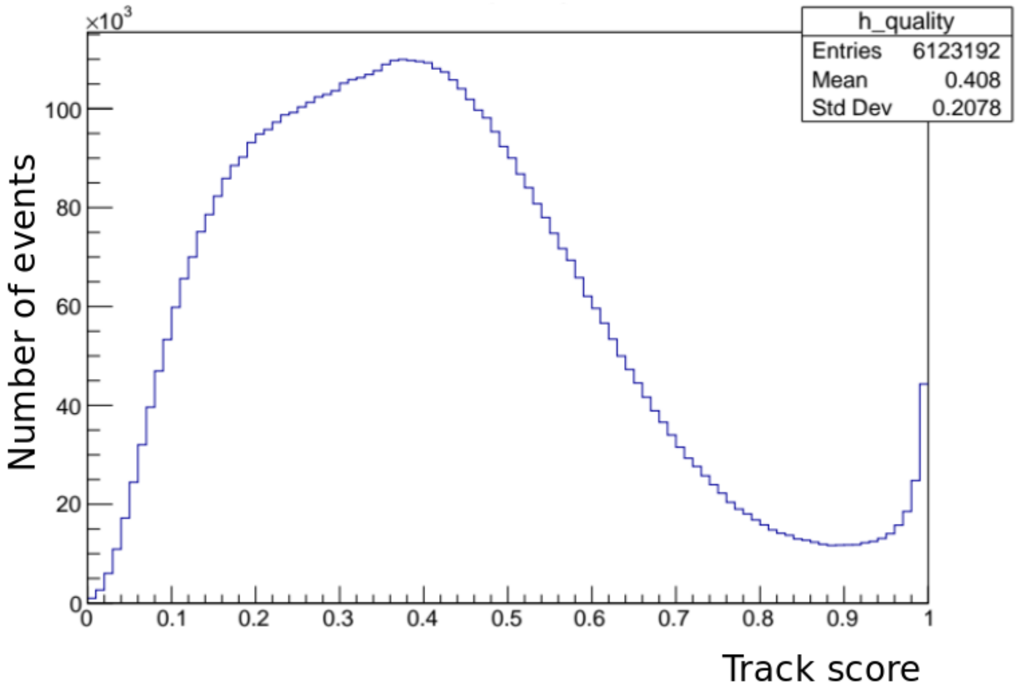


Figure 5.1: *Results from the random decision forest used in the current KM3NeT's PID, see section 3.4.2. Recall that, in this system, an event is classified as a track if its track score 0.6. You can see the problem mentioned in the motivation: The same amount of events per signature produces an uneven event distribution in track-shower classification.*

Let's compare this results from Fig 5.1, with the best results obtained with OrcaNet for track-shower, in Fig 5.2, below.

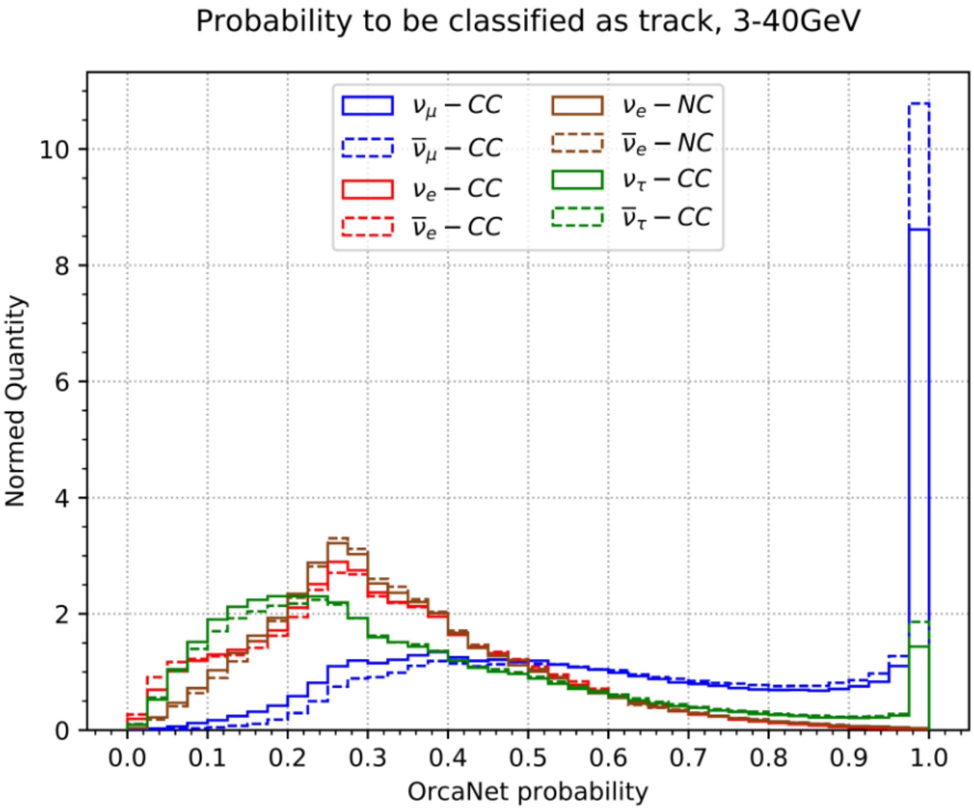
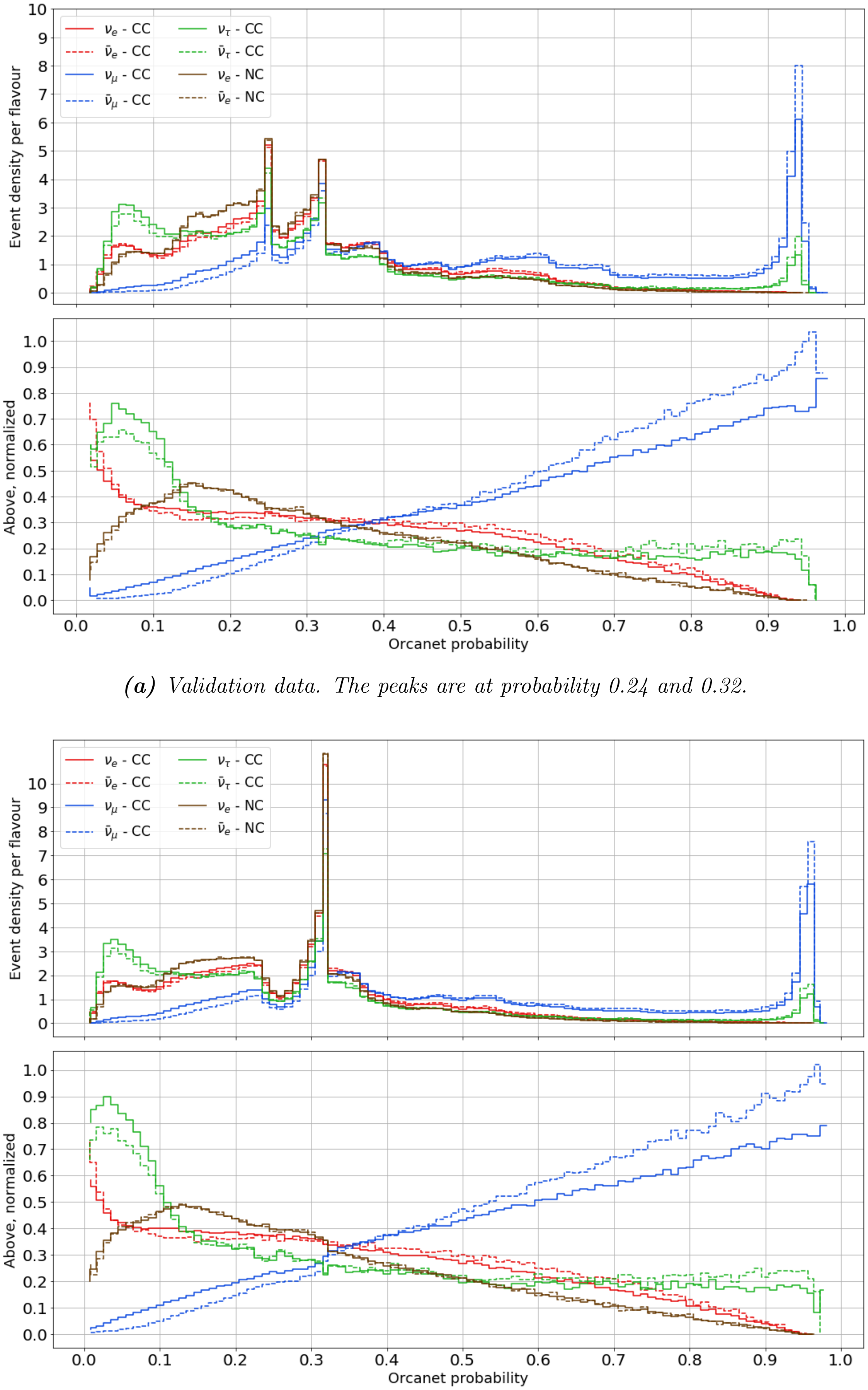


Figure 5.2: *State of the art PID in ML, done by Moser et al.*

This improved result, despite working with track-shower classification, looks at the flavour decomposition (actually, in flavour + charge decomposition) of the "showers" for details, instead of treating them as a monolithic category. This allows for a better prediction distribution between the "shower" (low track probability) and the "track" areas (high track probability/score), showing the muon distribution being highly predominant in the track score.

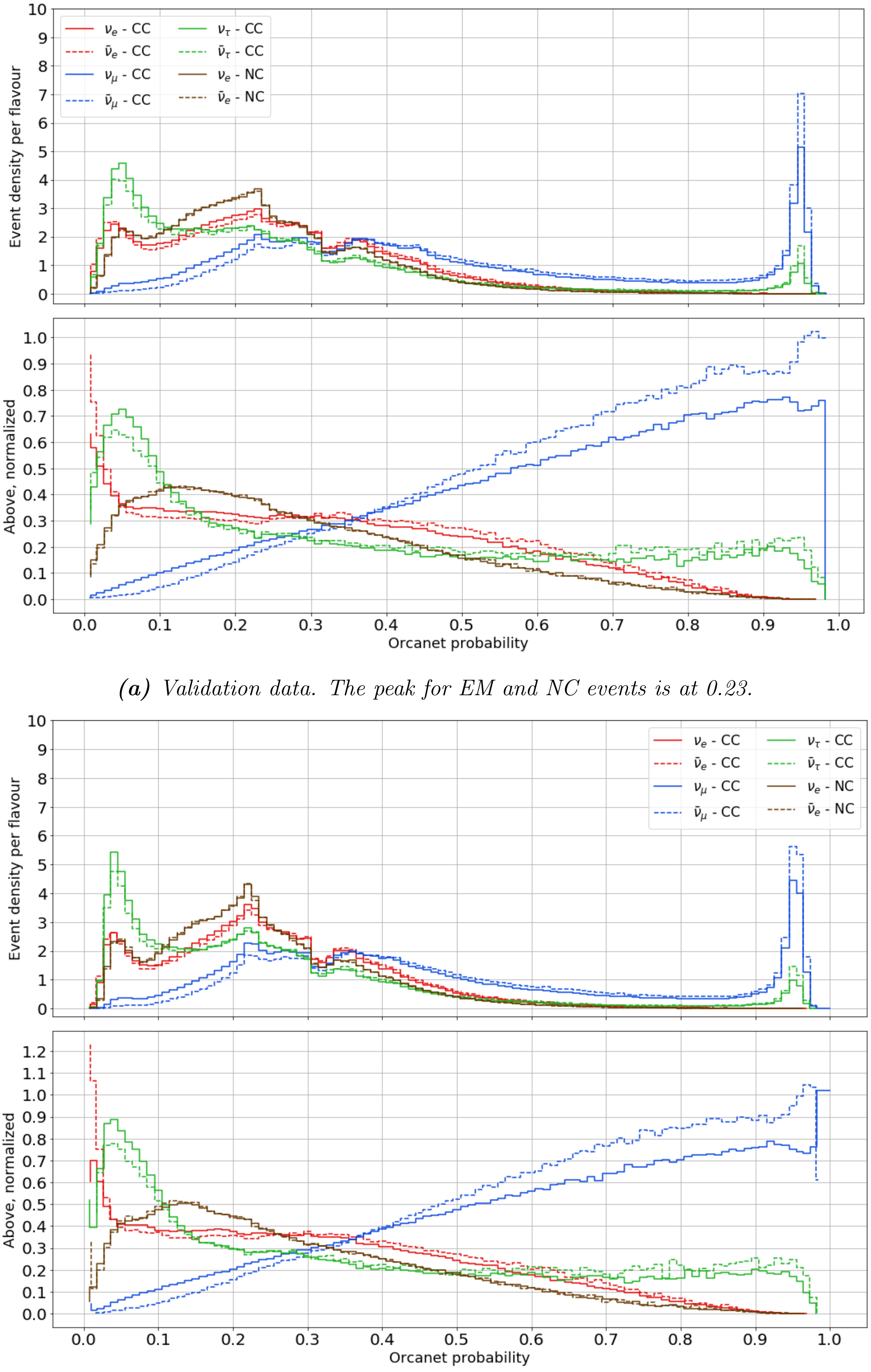
The comparisons between this previous results (5.1, but mainly 5.2) and the following ones (this work) in terms of performance will not be fair since these results have a totally different data selection an preprocessing. For example, Fig 5.2 has used with around twice the MC data than this work for training. To train with that much data would require double of the computer power that could be allocated within the scope of this work. It also includes heavy processing of the data outside OrcaNet: before training, flattening the energy range and discarding the tau events (which usually have the worst OrcaNet prediction performance). After the training, and with the goal of avoiding OrcaSong binning biases, the final result comes from combining several outputs from different OrcaSong images, made by adding the PMT information, cycling data over the 4th dimension and playing with the time binning.



(a) *Validation data. The peaks are at probability 0.24 and 0.32.*

(b) *Test data. The peak is at probability 0.32.*

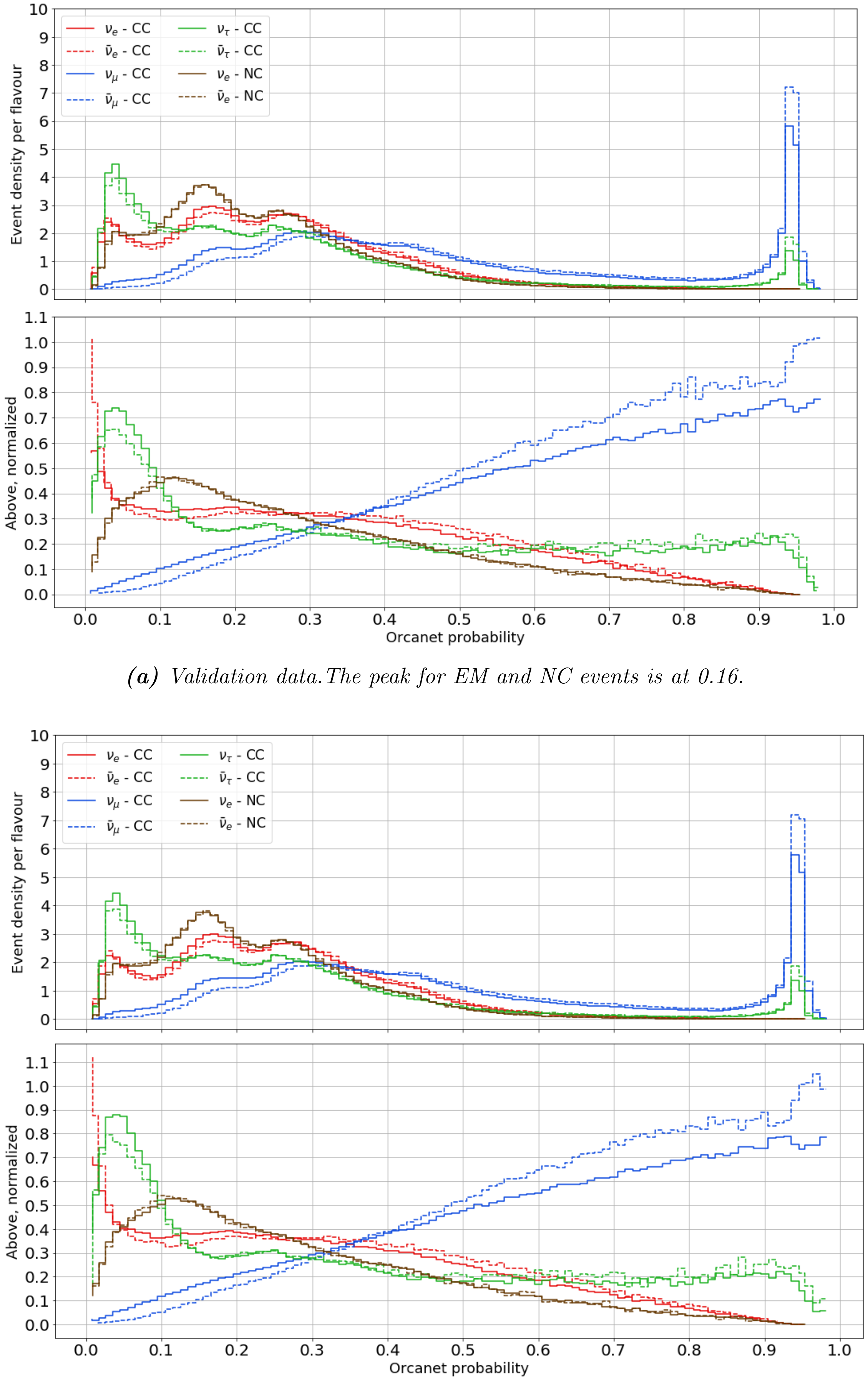
Figure 5.3: *OrcaNet prediction probability distribution for track/shower classification.*



(a) *Validation data. The peak for EM and NC events is at 0.23.*

(b) *Test data. The peak for EM and NC events is at 0.23.*

Figure 5.4: *OrcaNet prediction probability distribution for track/EM/tau+NC classification.*



(a) *Validation data.The peak for EM and NC events is at 0.16.*

(b) *Test data. The peak for EM and NC events is at 0.16.*

Figure 5.5: *OrcaNet prediction probability distribution for track/EM/tau/NC classification.*

First, let's describe our results. This previous 3 figures (5.3, 5.4, 5.5) refer to the 3 classification schemes in study, for two datasets (validation and test), and for each dataset, they have the results shown in two different views. In the top graph, **every line (representing a flavor and charge component) is corrected by the total amount of events of that component**, so the irregular composition of our dataset was compensated for. Although the event selection from MC event files tried to be balanced between the different categories (shown in Table 4.1), there are no low energy (1-5 GeV) tau MC files. Our final event distribution has only $\sim 9\%$ tau neutrinos, while the other 3 classes make up around 30% of the distribution each. Exactly the same happens for antineutrino events. The data used here was simulated following the expected atmospheric distribution of events, so the ratio between particle and antiparticle has been set 2:1. Even if we compensate for these deficits in these results, it is vital that future work tries to fix this imbalance to have an improved neutrino training and classification.

Second, from this event densities per flavour, we can make some useful visual comparisons with the previous result (Fig 5.2), as a safety check that ensures that our results have no intrinsic errors. Our results, Figs 5.3, 5.4, 5.5 share a very similar overall structure, with the tau events peaking at the lowest values (before probability 0.1), the EM and hadronic showers peaking at the same value within the same model, and a very high (> 0.9) probability muon peak. Also worth noting the small tau peak that always shows at a high track predicted value. This is something to be expected, since 17% of all the tau events decay into a muon, and said muons may leave a strong track that is prone to be misclassified. Even finer details, such as the antiparticle behaviour with respect to the particle, are still present. NC events show the same distribution for particle and antiparticle (to be expected, since the interaction is the same). for the rest, antiparticles do underperform their respective particles slightly at the lower probability end and perform better at the higher probability end, this is clearly visible in the muon peak.

Next, let's turn to the differences between validation and test datasets. Validation data is not seen by OrcaNet's training (It could eventually be "seen" through extensive fine-tunning, something that did not happen in this work), so its results should be identical to those of the test dataset, within fluctuations. We can see that this very much true in the 3 and 4 category distributions, Figs 5.4 and 5.5. The minor differences are compatible within statistical Poisson fluctuations, not included here as error bars for clarity's sake (a clear example is shown later at Fig 5.19). This similarity is why, in the following cases is validation data is shown only, since it has 1 million more events.

There is a single feature that makes validation and test data behave differently, in Fig 5.3. This is the single-bin dominating peaks, one in the test data (5.3a) and two in the validation data (5.3b). Despite their size and seemingly relevance, these peaks are just an artifact of Deep learning training. This peaks were found during the testing phase, in the hidden feature check described in B. Although first thought of a consequence of the *black-box* behaviour of OrcaNet/DL[1], this peaks have

[1] After all, there is no reason to argue that the number of events should be smoothly distributed in the probability space, since we cannot know the factors that influence a single event prediction.

been found to vanish with the flattening the hits distribution previous to training. The current hypothesis is that the single-bin peaks form as a consequence of an irregular distribution of hits among the categories in the training data. Not to be confused with the extra 3 to 5 GeV excess in energy mentioned before, this is a flavour imbalance in the phase space, not an irregular variable distribution. This artifacts have not been eliminated in this result for two reasons: preprocessing cost and the fact that this artifact does not seem to affect our posterior results.

Not only these peaks are not present in the other classification schemes (which have been trained and tested with the same data!), but we can get rid of the peaks appearance by normalizing our distribution along the x axis. This is why this results have included the bottom graph, which is just the same as the above result where each value has been normalized be the total amount of events per bin. Since we have the event densities and not the raw number of events, the "normalized" version do not sum up to 1 per bin, which would have been the case if we had normalized the number of events. This is another reason to have a good event distribution to train and predict, instead of having to correct per category differences. Looking closely at the normalized versions of the previous figures, we can see that now not only the resulting distributions are oddly similar between the different classification schemes (considering the very different distributions that they present in the normalized version), but now the artificial single-bin peaks have disappeared. What is more, the results between the validation and test dataset for any of the schemes seem to be almost identical. The normalized representation is a more abstract visualization of the event distribution, but much less misleading than the event number or density.

Last, but not at least, let's try to have a comment on the performance of the classification schemes, and make an educated guess of which classification is performing better. OrcaNet's training goal is the separation of the signals, not ensuring the best muon fraction of events. This means that the algorithm deems it acceptable to have a lower amount of high-quality muons in exchange for lower backgrounds. Thus, in order to compare the prediction distributions between Figs 5.3, 5.4 and 5.5; or even with 5.2 and 5.1 (knowing that the comparison with this last two is just an observation from which a conclusion cannot be drawn) we will not look at the height of the peaks but to their x-axis position. We can see that in the standard PID case (Fig 5.1) the peak of the distribution is just below 0.4. In our results (Fig 5.3a) the artificial peak is at 0.32 and the peak outside the artificial single bin effect is closer to 0.3. The best-case OrcaNet PID shows the shower peaking below 0.3, look at Fig 5.2. The 4 category system (Fig **??**, shown in next section) peaks at 0.15, with a secondary peak below 0.3. This means that even if the muon events are not that clearly identified, the backgrounds (false positives) are nicely suppressed in this last case. Tentatively, we can say now that *the average prediction value in every category for the wrong type of events is lower with a 4 category classification.* This is not a final conclusion (yet) but rather an observation that has to be corroborated by strong statistical metrics (see the next result sections).

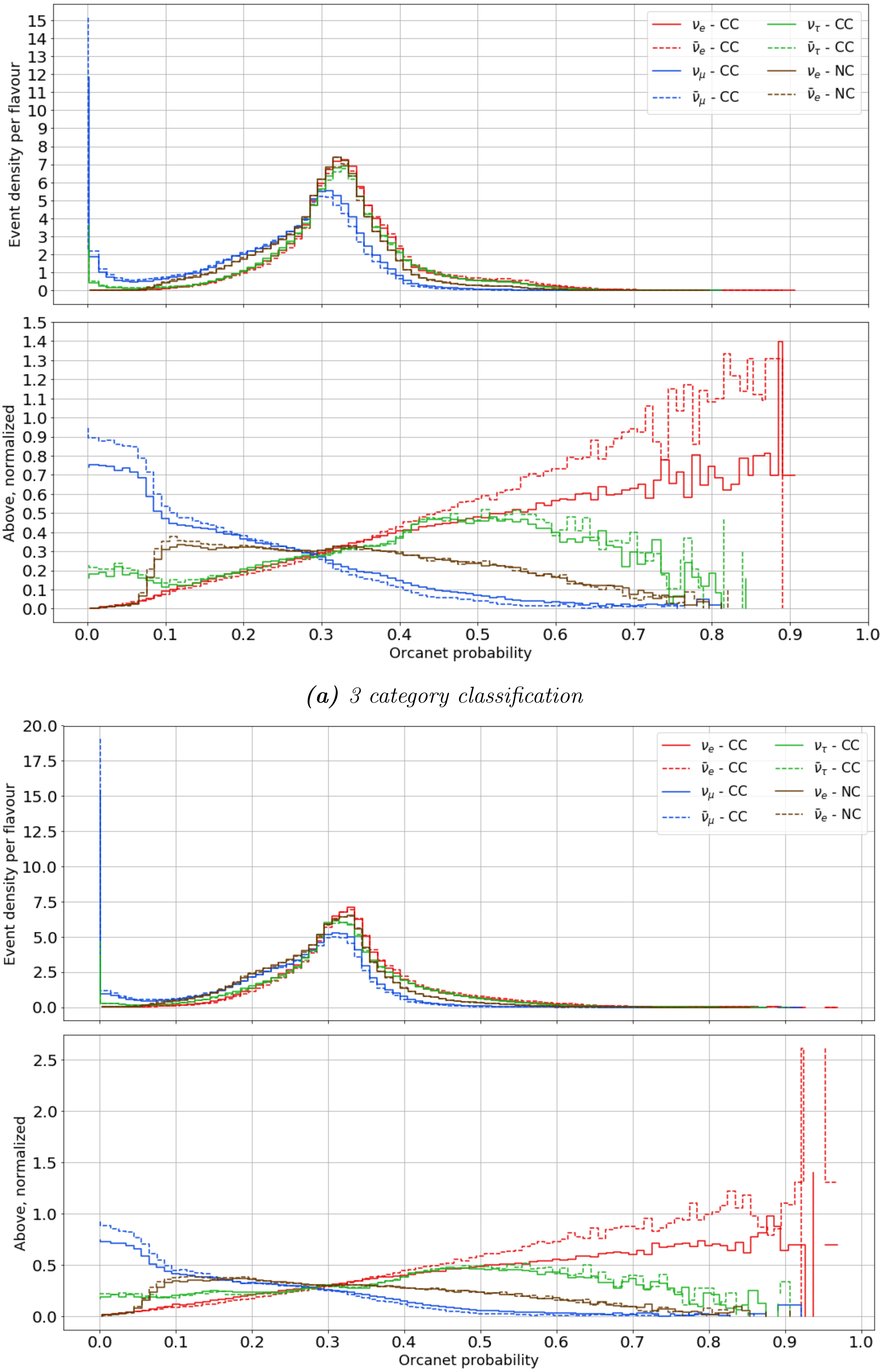
5.1.2 New probability distributions

In any classification scheme, the sum of all the categories have to add 1. For the track-shower (2 label) system, this means that there is no shower quality score in the standard reconstruction, since it is simply the inverse of the track score. However, in the newer proposed models, 3 and 4 flavour categories, it makes sense to have 3 or 4 different plots, because, even if the sum of the probabilities is still 1, you cannot easily visualize a category from the distributions of the other 2 or 3.

The following results are only shown with the validation dataset, since the close similarities have been shown above. Fig 5.6 includes the probability distribution for the EM shower for the two new cases, the 3 and 4 category models. The remaining results from tau and NC events are 1 single result in the 3 category case, Fig 5.8 and as 2 separate cases in the the 4 category model, Fig 5.7. This are the new results opened by this alternative classification schemes, invisible to a track-shower classification. The most groundbreaking result overall that we can get from all of them is that the performance is not unlike the previous figures, which would not be the case if the classification for this schemes was impossible.

Actually, the identification for the different shower categories, is clearly worse than the muon identification. This can be seen in something as simple as the maximum predicted value for a distribution. Only muon events, with the corresponding missclassified tau events, surpass the ~ 0.95 probability mark (seen clearly in the normalized plots in Fig 5.5). This means that this system is able to predict, at that confidence level, a few thousand muon events with the sole background of a few hundred tau events (the actual numbers had to be looked from the dataset) that the model itself is rejecting as showers. We cannot know for sure that all those events correlate with actual muon or tau events right now, but in cases where we are lacking the prior information, the threshold above which the network is not mistaken anymore is the next best value for a high quality event. In the EM shower case, this is seen for the 3 category model (Fig 5.6a), but at the lower threshold of 0.85 and non existent for the 4 category model at all (Fig 5.6b). It is also not seen in any of the tau and/or NC classifications for both schemes.

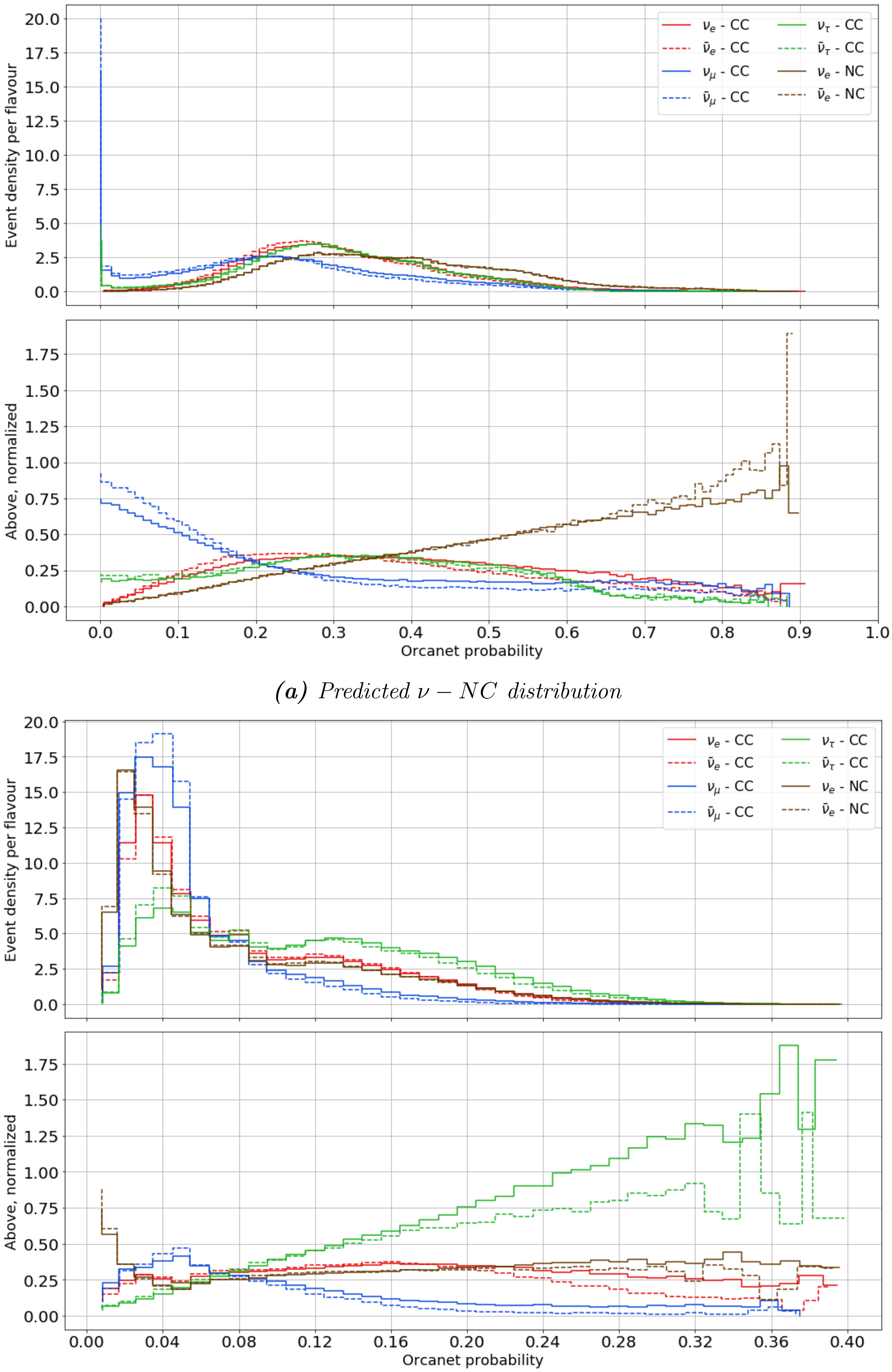
Moving on to inspect the showers, both the electron category in both schemes (in Fig 5.6) and the neutral current category in 4 categories (in Fig 5.7a) show a similar feature: the muon neutrino events have a very large peak at 0 probability, which actually is around 10^{-3} upon inspection. This seems to be the most clear benefit of the extra categories: they are helping OrcaNet to achieve better internal definition of what a track is. Even if the extra categories themselves perform badly, with low probability values and a lot of misclassification, they are allowing OrcaNet to have less confusion between tracks and showers, rejecting more clearly the track hypothesis for shower events. This is a signal of the improvement within the track-shower scheme mentioned before.



***(a)** 3 category classification*

***(b)** 4 category classification*

Figure 5.6: *OrcaNet prediction distribution for as EM shower ($\nu_e - CC$).*



(a) Predicted $\nu - NC$ distribution

(b) Predicted $\nu_\tau - CC$ distribution in the 4 category model

Figure 5.7: *OrcaNet prediction distributions for the 4 category model.*

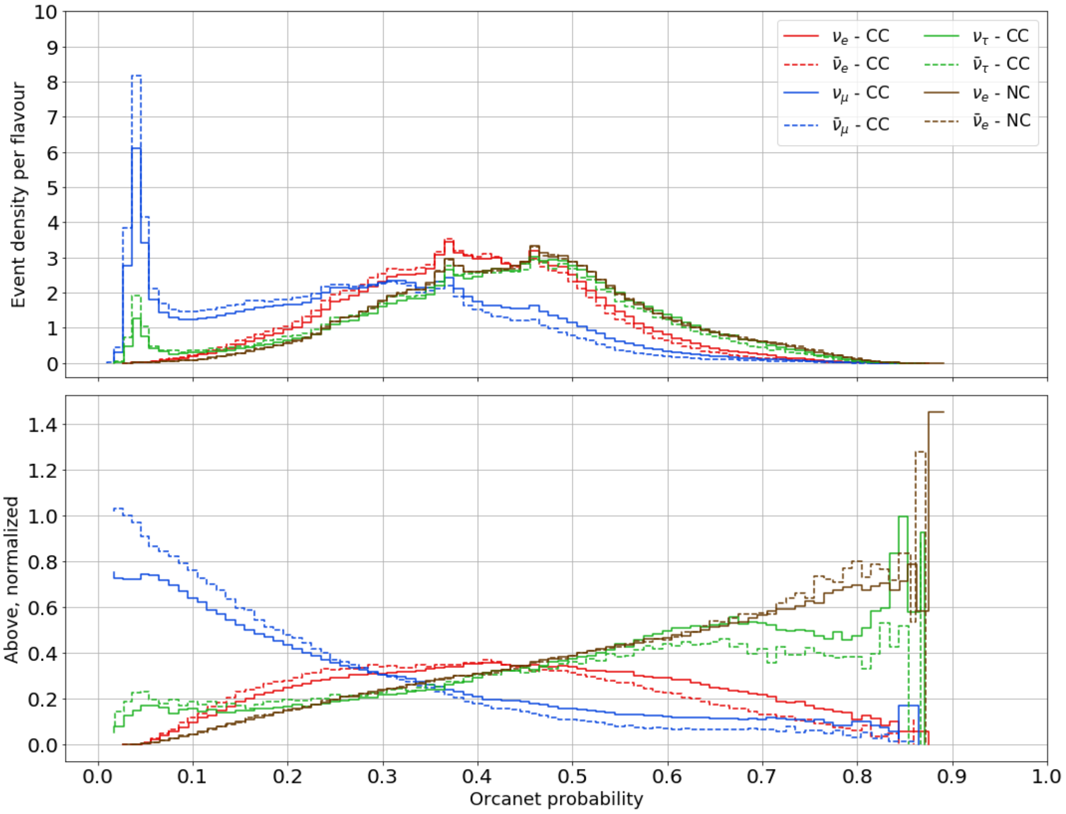


Figure 5.8: *Predicted $\nu_\tau - CC$ or $\nu - NC$ distribution in the 3 category model.*

As for the structure of the electronic showers, we can see two very similar pictures. Similar shape with a peak at 0.32. We could claim, even if we need further corroboration, that the EM shower classification for both 3 and 4 categories seems to be identical. In both cases the backgrounds are as present as the correct category around the distribution peak. The backgrounds are not suppressed (reduced compared to the correct identification signal) until, at least, 0.6 probability, this can be clearly seen in the normalized view of the data. This means that before a quite high high probability prediction, a EM shower event still has a very high chance of being misclassified. Indeed, shower identification does not perform as well as track identification.

Finally, a very important detail to look for is the high fluctuations that the normalized views present, seen clearly in the high probability bins of 5.6b and 5.7a. This fluctuations in the data stem from almost empty bins, where the fraction of events jumps between 0 (no events/bin) to 1 (< 5 correctly guessed events/bin). Error bars could be added to show this effect, but they confuse the figure where multiple lines are shown. The fluctuations are a consequence of the power-law energy spectrum of the data, and it can be fixed by using more events in the high-energy end of the test data. We have only seen it in the high probability end of the spectrum, not in the energy spectrum, but it makes intuitive sense that at higher energies the predicted score would be higher, since there is more light and more hits for a better identification. Anyway, in the next section (visible energy dependency), this correlation is clearly shown.

The tau case in the 4 category classification (Fig 5.7b) is quite remarkable on itself. If you look at its distribution, especially at the normalized version, it seems to perform better than the other categories, with the backgrounds clearly suppressed versus the tau signal. However, notice that no event has a probability of being a tau event higher than 0.4. The normalized figure is compensating for the lack of tau events compared to other categories, so the tau performance is heightened. Actually, if we look at the raw event number all the background classes have more events mislabeled as tau events than tau events correctly identified until 0.36 (out of 0.4!). The most probable cause for this is the lack of tau events in training, which gives a weak signal to OrcaNet to train on. The corrected plot hints at a much improved tau classification for 4 categories should the training be done on an more homogeneous dataset.

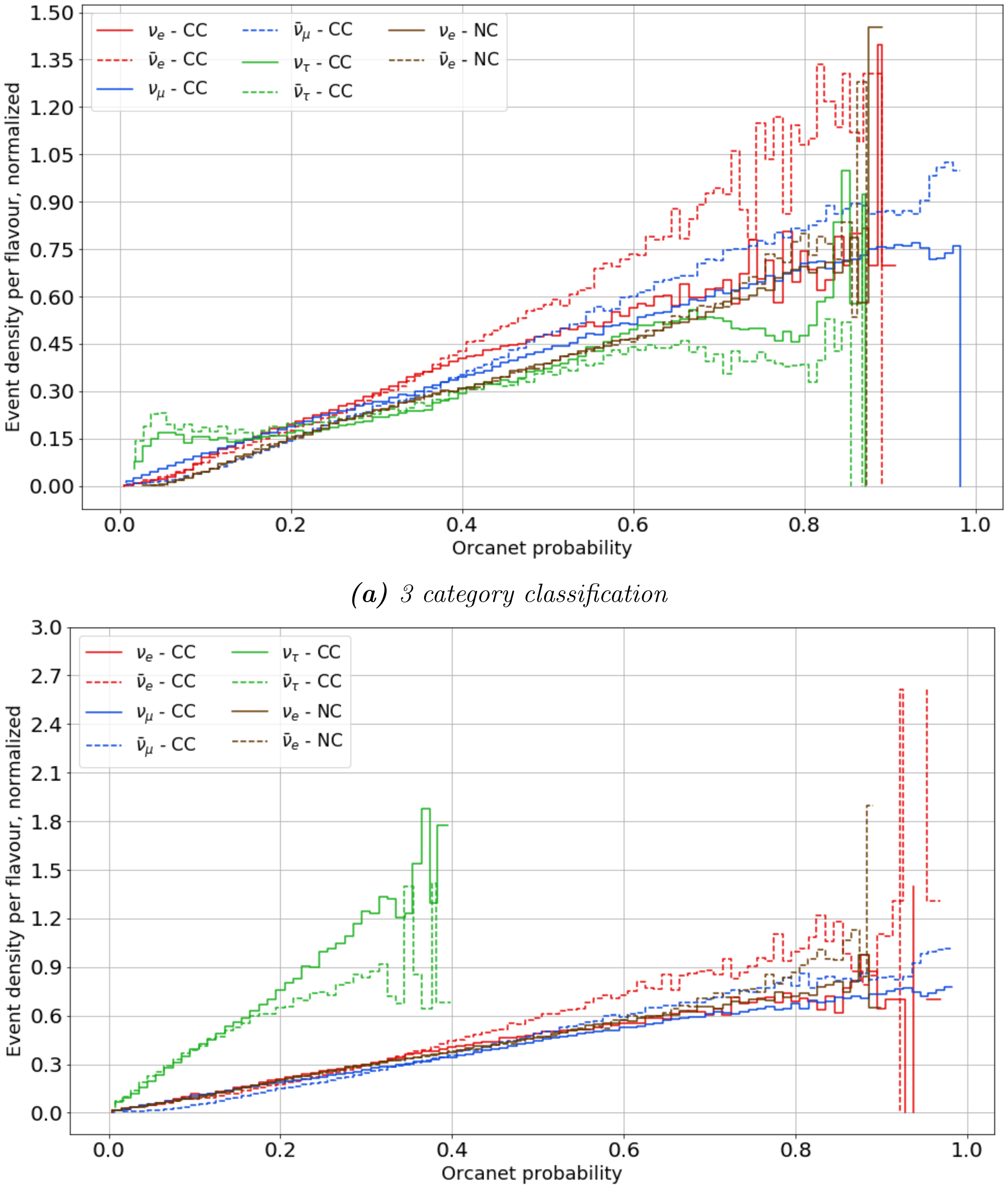
The previous figures in this section, especially the normalized figures, include one overlooked aspect of the data: the charge difference (particle vs antiparticle). The charge separation alongside the flavour in previous results was introduced following the format of 5.2. Antiparticles do not show a very different signal than their respective particles in these plots, but if we looked at the raw event number performance, the antiparticle events (again, only a third of the total) present a much lower, almost secondary performance. The density plots are required to understand that the differences are not intrinsic to OrcaNet but due to the training and validation datasets, as they vanish when we correct for them.

We know from theory that the two signals must be different, since particles and antiparticles differ in their inelasticity distributions with energy. The bjorken-y or inelasticity is the fraction of the energy going into the hadronic shower,

$$y_{bj} = \frac{E_{shower}}{E_{\nu,in}} = \frac{E_{shower}}{E_{shower} + E_{\nu,out}}, \tag{5.1}$$

since the outgoing neutrino does not produce light. The problem is that for the same energy, it is a fact that antineutrinos have higher inelasticity, so they are going to deposit more energy into the light-producing primary shower, so for all cases more signal hits should be available for identification, and thus perform better. This is not seen here due to the normalization, but it is going to be the case in the following results. A next version of this models should be trained in the same amount of particle and antiparticles to see if a difference in identification appears or not without posterior corrections. If it appears, it might be an indication that the signals from the particles and antiparticles are being reconstructed different. If this were to be the case, it would mean that the 4-category model used here could be extended up to an 8 category model that separates in flavour and charge. This hypothesis is unfavored since it confronts the knowledge of the detector's resolution, but it should be pursued if the difference actually exists. Good separation between flavour and charge means an excellent inelasticity reconstruction, so maybe a bjorken-y regression model with OrcaNet might be also a good idea.

All the correct guesses, the flavours that correspond to the predicted category for every single previous figure has a positive linear dependency, both for particles and antiparticles. This dependencies can be compared to each other, as done in the following Fig 5.9 for both schemes. It seems that OrcaNet gives the same fraction of events (normalized density) at the same certainty, independently of the category or the scheme[2]. This could be a hint of the inner rules of OrcaNet for its event determination, both for training or prediction.



(a) 3 category classification

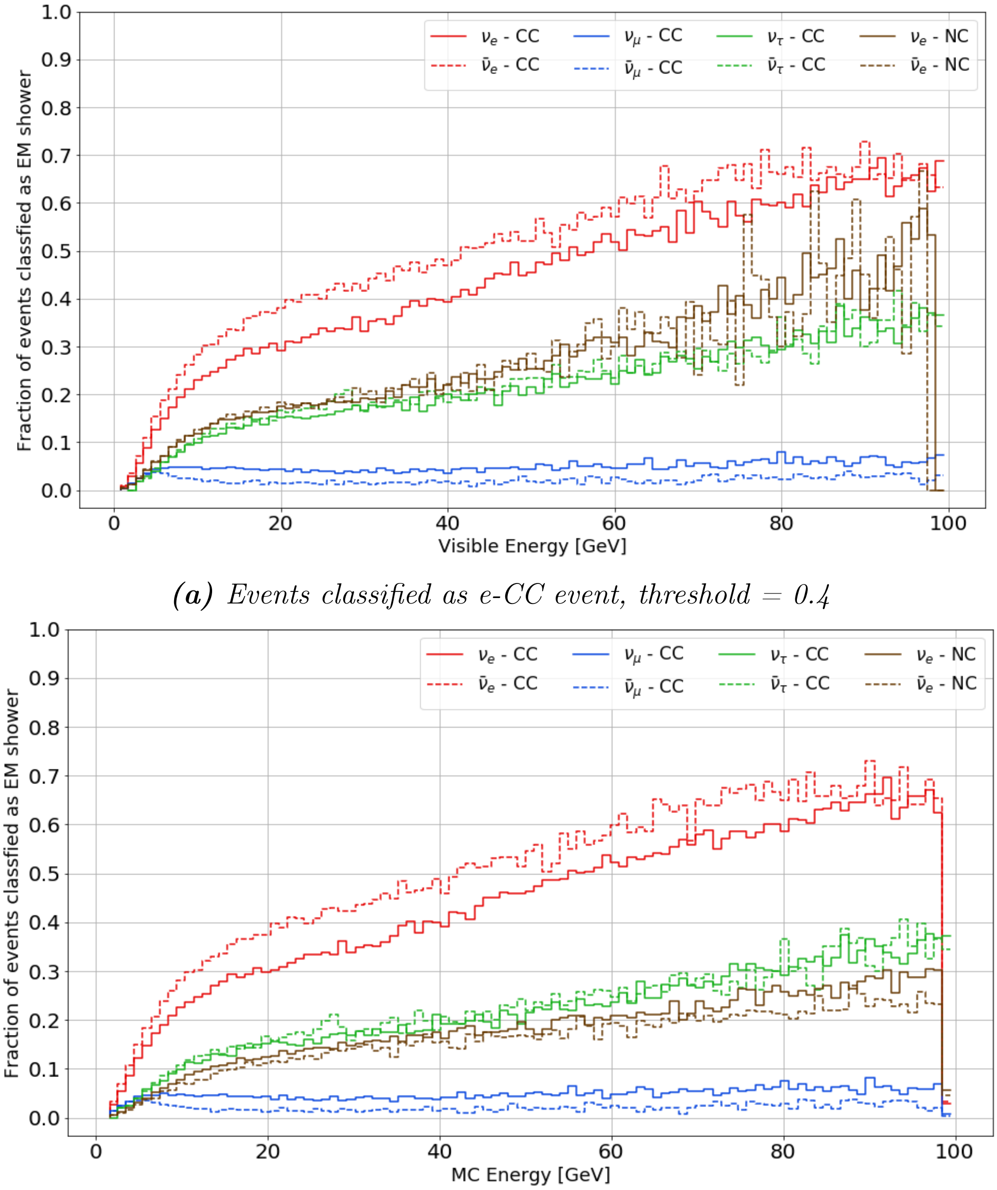
(b) 4 category classification

Figure 5.9: *Normalized distributions for correctly guessed categories.*

[2] With the exception of tau signal, which seems to be comparatively better. This advantage can be because of the actual weak signal and should not be trusted until a better tau identification is achieved.

5.2 Visible Energy dependency

The visible energy is defined as the MC energy for the CC events and the MC energy times the bjorken-y for the NC events. This definition is chosen here because it is realistic: OrcaNet only has the "detected" energy, the the energy in the event that produces light (hits), while standard PID relies on energy reconstruction, closer to the MC Energy. The visible energy is an effective approximation to introduce this difference in our results by simply combining the 2 most relevant MC variables (energy and inelasticity), even if this change only affects the NC events in the following results (see 5.18). For other variables, check the hidden features section, in appendix B.



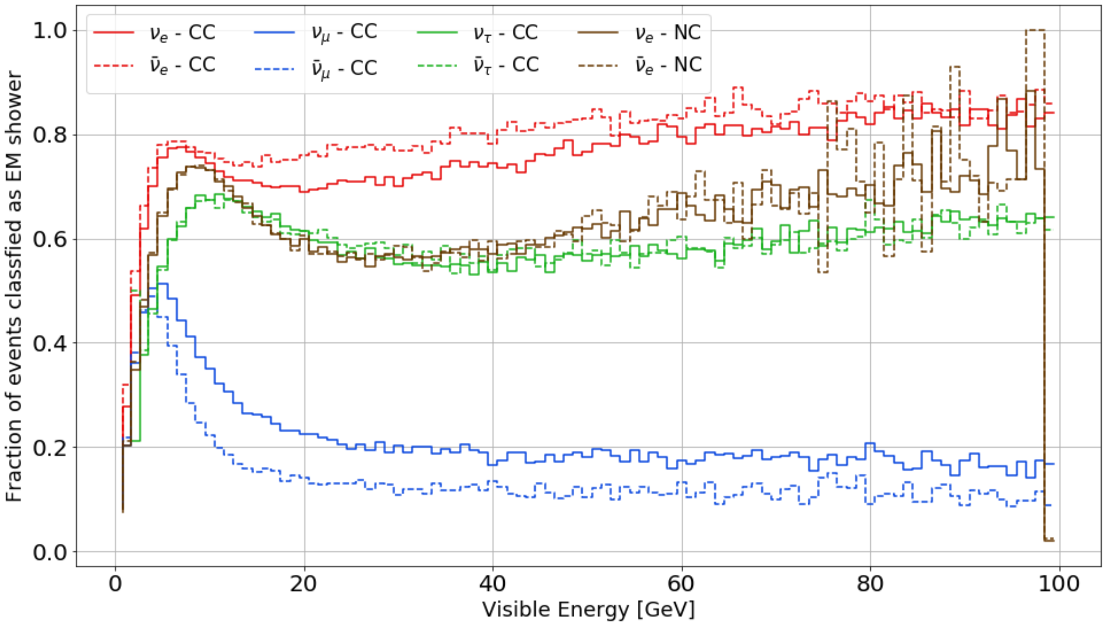
(a) *Events classified as e-CC event, threshold = 0.4*

(b) *Events classified as e-CC event, threshold = 0.4*

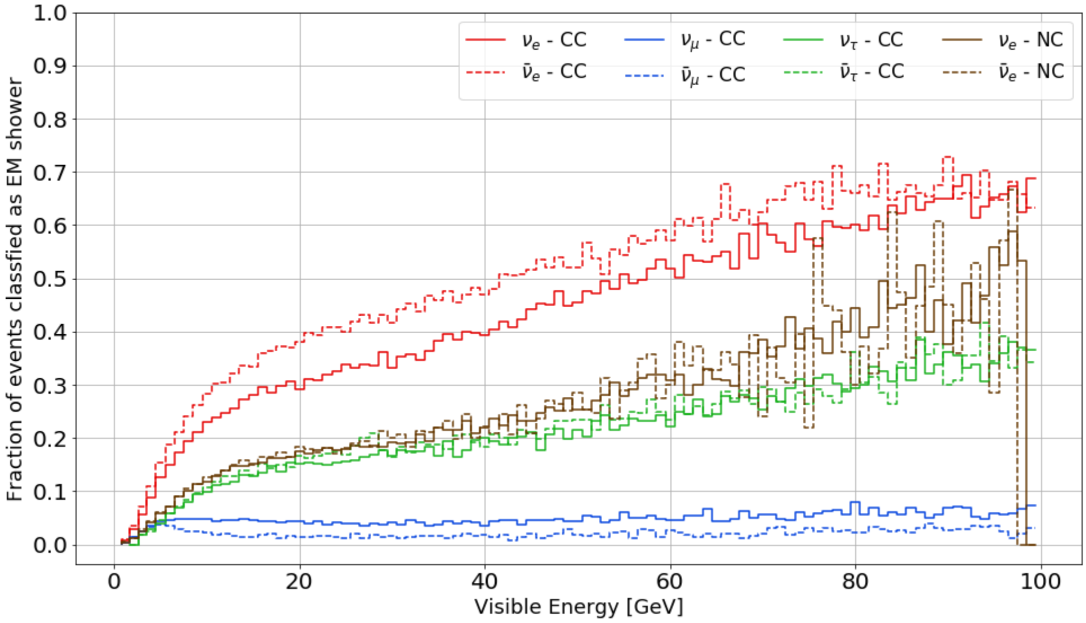
Figure 5.10: *Influence of the plotting variable choice for 4 category classification with same data. Using the MC energy should give better results, since you are assuming a good energy reconstruction. This affects equally to all classification schemes.*

5.2.1 Threshold effect

The threshold, the minimum predicted value required to assign an event to a category in the confusion matrix (defined in 4.3.1) is necessary for a correct understanding and interpretation of the data, as the following results are non-linearly dependant on the threshold (see Fig (5.11). A better explanation about its nature is given on section 5.5.



(a) *Events classified as e-CC event, threshold = 0.3*



(b) *Events classified as e-CC event, threshold = 0.4*

Figure 5.11: *Influence of the threshold in the final event distribution. This is an illustrative, extreme case that tells that a huge fraction of EM showers are given a prediction score between 0.3 and 0.4.*

5.3 Categories comparison

The energy dependency with the fraction of events make for a more accurate comparison of the actual classification performance. The OrcaNet results of this work (in Fig 5.13) can be compared against two different references, the standard PID via RDF (Fig 5.12, right) and "state of the art" OrcaNet (Fig 5.12, left) as implemented by the ECAP team and optimized as described in the previous section (this figure follows from Fig 5.2).

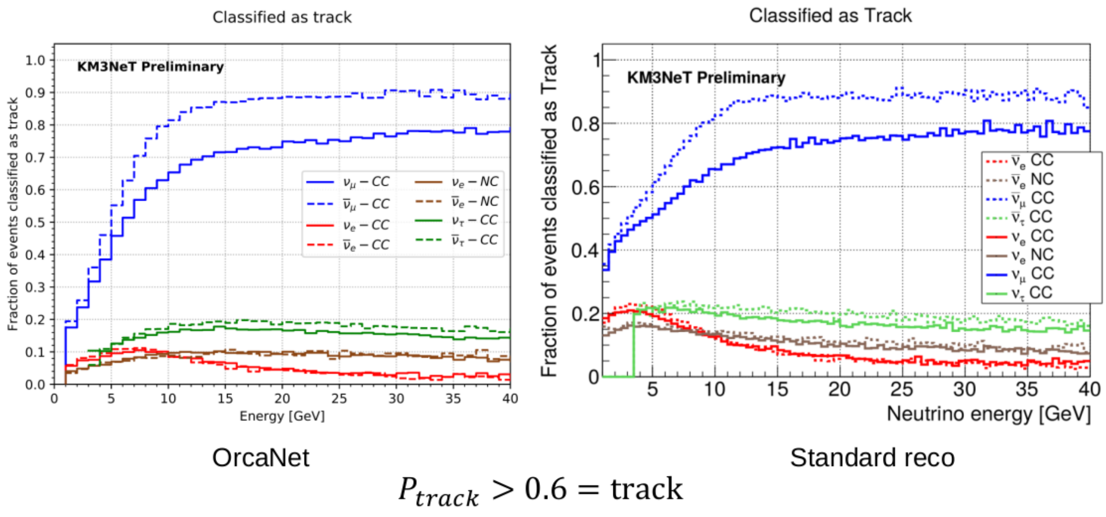


Figure 5.12: *Standard reco and state-of-the-art OrcaNet MC energy dependency. Note the use of MC energy instead of visible energy, so they assume some degree of reconstruction in the data. This illustration shows that the two results are basically equivalent.*

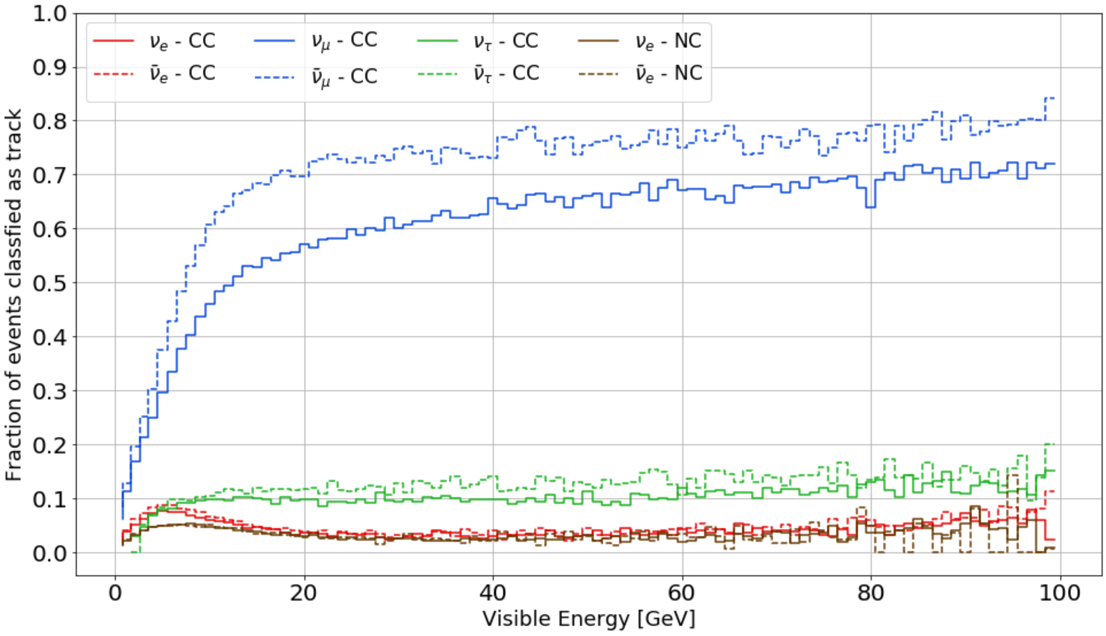


Figure 5.13: *Classified as μ-CC event for 2 classification categories, threshold = 0.6*

As already mentioned, before comparing the two results we have to note that these results have different event selections, preprocessing, etc. The comparison here is done with an illustrative purpose, of how our absolute score is worse (around a 10% lower for all lines). Lower signal and lower backgrounds might mean that our OrcaNet training has found an equivalently good solution in terms of separation. However, the extra increase in in the total muon fraction requires some extensive parameter tuning and preprocessing, given that during the testing phase of this work (where different cuts and flattening options were applied) the increase in the muon fraction was, at best, 2%.

Now, we turn out our attention to the 4 category model, seen in Fig 5.14. The figure, and its threshold value, have been chosen to reproduce the muon curves for the previous 2 category result, in Fig 5.13.

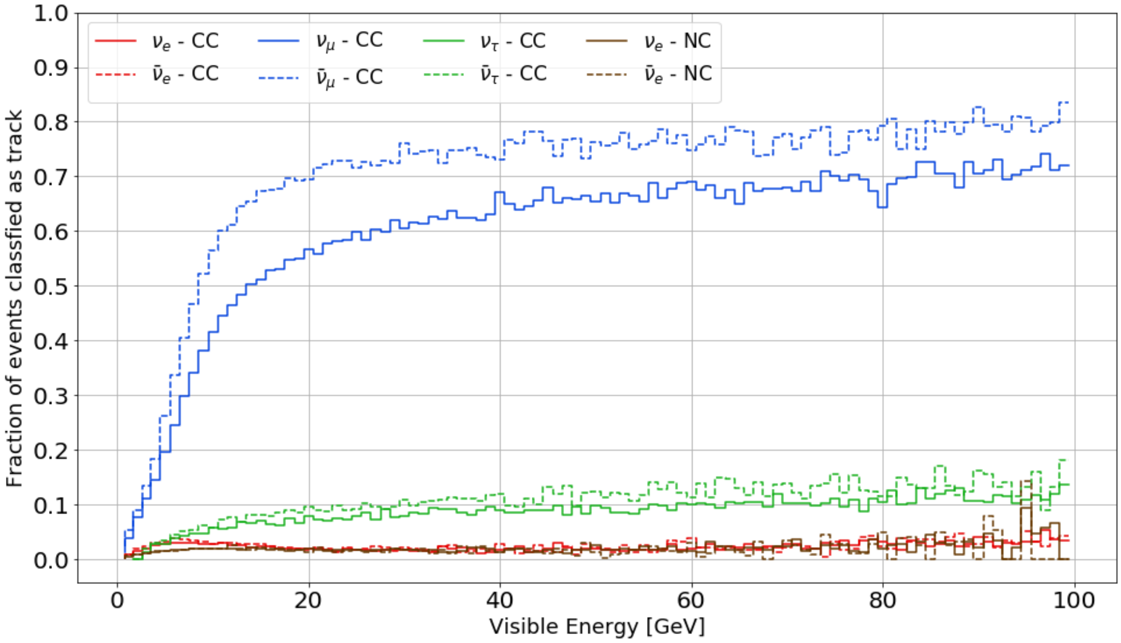


Figure 5.14: *Classified as μ-CC event for 4 classification categories, threshold = 0.4*

The 2 results come from different models that have been trained and evaluated on the same data, so we are expecting that no great differences should exist for the track category, since it has been defined identically. However, the curves only take the same shape (defined by visual markers like crossing fraction = 0.7 at 20 GeV) when the 4 categories model has a threshold of between 0.5 and 0.4. This means that the 4 category model is providing the same results as the 2 category models at a 10 to 20 % lower threshold. This would not be relevant but for the fact that, as mentioned in section 4.3.1, lower threshold implies higher statistics, especially in this 0.4 - 0.6 range, see Fig 4.3.

There are other minor changes in the shower distributions. All showers but the tau are further suppressed in 5.14 than in 5.13. This is most significant in the lower energy peak totally vanishing. Most of the neutrino flux happens at low energies, so the analysis tend to focus on the lower energy side. Any improvement there is intrinsically harder (less raw information) and more useful. The tau events are the main background, and this one has not been suppressed. This can be explained by the aforementioned fraction of tau events decaying into a muon. To fully characterize this background, it would be nice to have access to the signal hits or the MC info about all particles involved in the event, to see if a correlation between the misclassification background and the decay channel of the tau can be established. There was an unsuccessful attempt to recover this information in this work, but the current configuration (the new OrcaSong2) might make it feasible, see appendix C.2.

5.3.1 Hadronic and Tau shower

The results from the last two categories, NC or hadronic events and tau events, are going to be covered now to understand the full picture of the 4 category distribution.

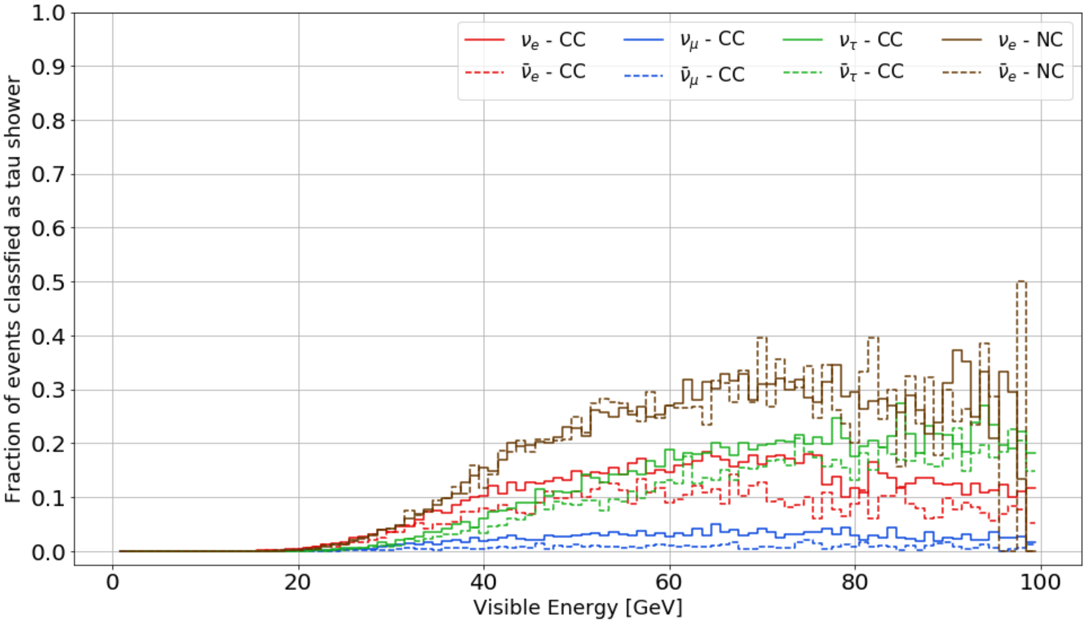


Figure 5.15: *Events classified as $\tau - CC$, threshold = 0.25*

Tau event, seen here in Fig 5.15, are the least capable of being identified. This was known from the previous section, where no tau event has a chance of belonging in their own category bigger than 0.5. To show as much signal as possible, the threshold for this case was lowered to the minimum threshold that makes sense, the random guess value. In spite of being in situation with the least stringent threshold, the tau events are dominated by hadronic shower misclassifications at every point. Furthermore, no event below 20 GeV is flagged as tau whatsoever. This might be influenced by the lack of a low energy tau events file, but there should be tau events beyond 5 GeV. Another reason might be the dominance of the neutral current events in the same energy range, seen in 5.16. In fact, neutral current is the classification category that dominates in the first 20 to 50 GeV of the energy range.

It seems confusing the fact that there are such high values in 5.16 at lower energies. One might wrongly assume that for any of the previous figures, (Figs 5.14 to 5.16) all the lines within each figure had to add 1. That is not what should happen. Instead, all the lines (the distributions) have to add 1 *added over the 4 different figures for* *the same threshold.*

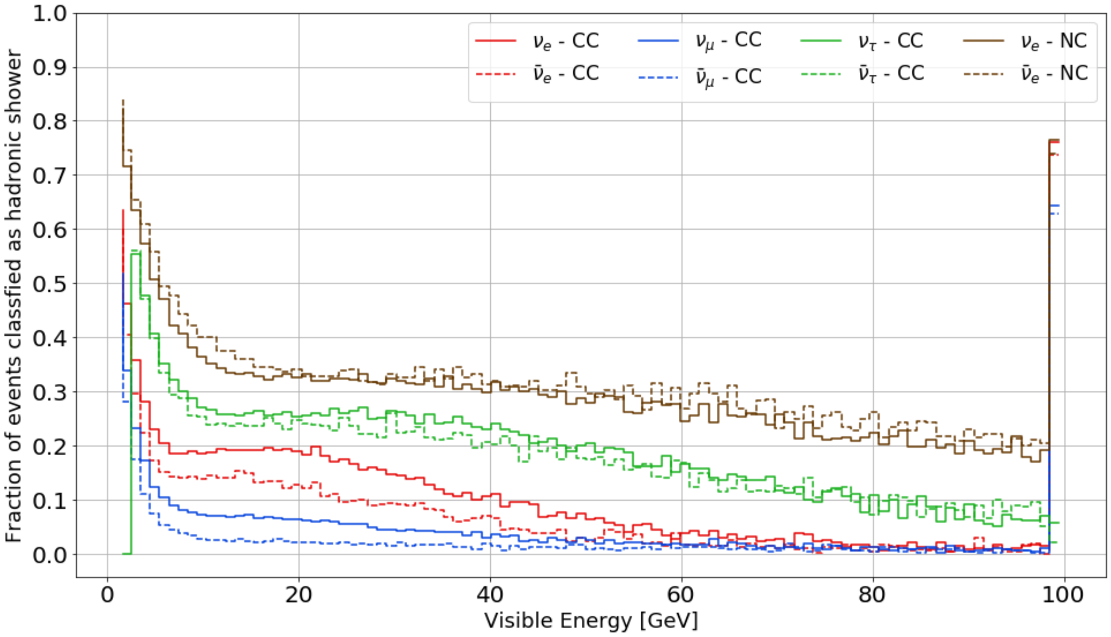


Figure 5.16: *Events classified as a $\nu - NC$ event, threshold = 0.4*

The neutral current classification for every type of event in the low energy range has a perfect reason. There are simply less hits in the detector at low energies, and the class that emits the least amount of light is learned to be the neutral current by OrcaNet. This is mainly the reason we are using here the visible energy. It is very relevant to the results that the model sees the hit number and distribution, not the energy. If this aspect is not corrected (by this variable definition or otherwise), our results would be biased.

The neutral current shower distribution seems to have an complementary behaviour with respect to the tau events distribution. The neutral current fraction of events starts to linearly decline at 60 GeV (see **??** again), the same point where tau events start to get nonzero values at 0.4 threshold. (Image not included due to the little content, most results are suppressed at 0.4). This correlation could be attempted to be established with respect to the EM shower distribution 5.11b, but the EM shower shows a very clear linear increase all throughout the energy range, something that is not reflected in the hadronic counterpart. This is what lead to a 3 category model, as the hypothesis is that hadronic and tau events might be complementary under the current OrcaNet identification. That could explain the fact that tau events are the main background for hadronic distribution and viceversa. The 3 category model tries to unify the two worst performing categories. If they are a single category for OrcaNet during identification, the unification should come with a big improvement in performance for classification.

5.4 Separation

The separation as used here comes from the Matthews Correlation Coefficient for the multinomial classification, explained in section 4.3.2.

5.4.1 2 categories: e-CC vs e-NC

The separation between e-CC and e-NC, shown in 5.17, was the conclusion of the first tests with OrcaNet, made to prove that OrcaNet is capable of a higher resolution than track-shower. The separation metric as adopted from the other study within KM3NeT/ORCA that has attempted to include the flavour of the interaction, seen in Jannik Hofestaet's PhD thesis [36] and reproduced here in Fig 5.18a. We can compare better both results on Fig 5.18.

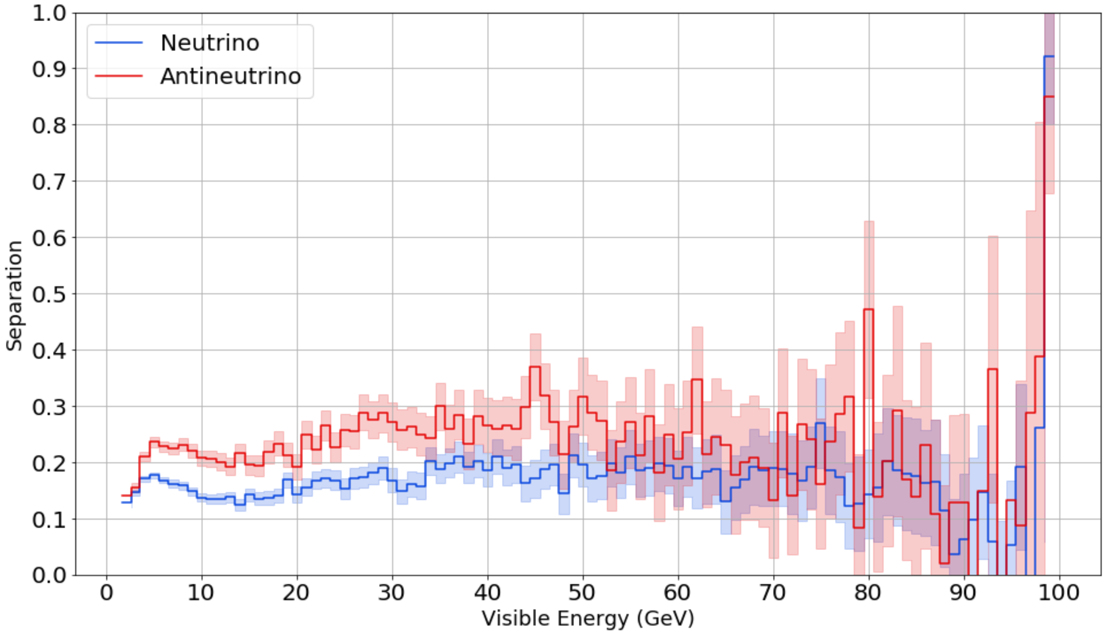
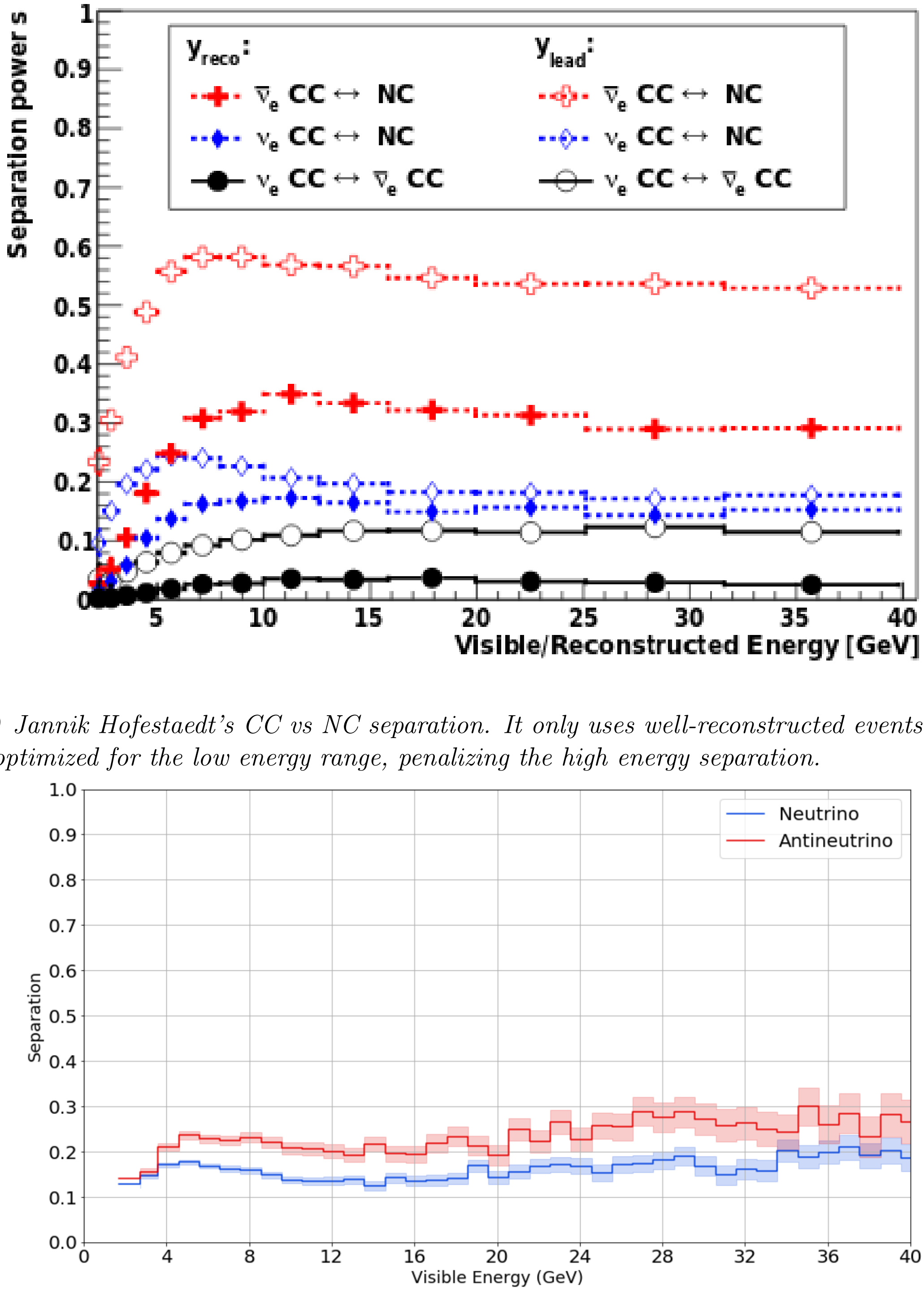


Figure 5.17: *e-CC vs e-NC events separation. Separation of 0.2 means that around 20% of the events, could be to be correctly identified. This translates to an average accuracy of 60% instead of a random guess (50%). This result is also proof that the antiparticles are more separable than the particles.*

Jannik's result was computed in a non-ML way, using a maximum likelihood analysis on the reconstructed inelasticity (bjorken-y), using 5 bins and fitting to the different inelasticity profile of the different flavours. We have already seen the influence of the inelasticity in our data in section 5.1.2, and we can notice that in the results in the previous sections 5.2 to 5.3, the antiparticles always perform equal or better to their counterparts, as expected from theory.

This comparison figure (5.18) is also included here since Hofestaedt's result is the one mentioned before in 3.5, the result that has an estimated 7% of improvement in sensibility, should the information be included in the mass hierarchy fit. This gives an idea of how the small separation, like the one achieved here, could translate in a posterior improvement.



***(a)** Jannik Hofestaedt's CC vs NC separation. It only uses well-reconstructed events, and is optimized for the low energy range, penalizing the high energy separation.*

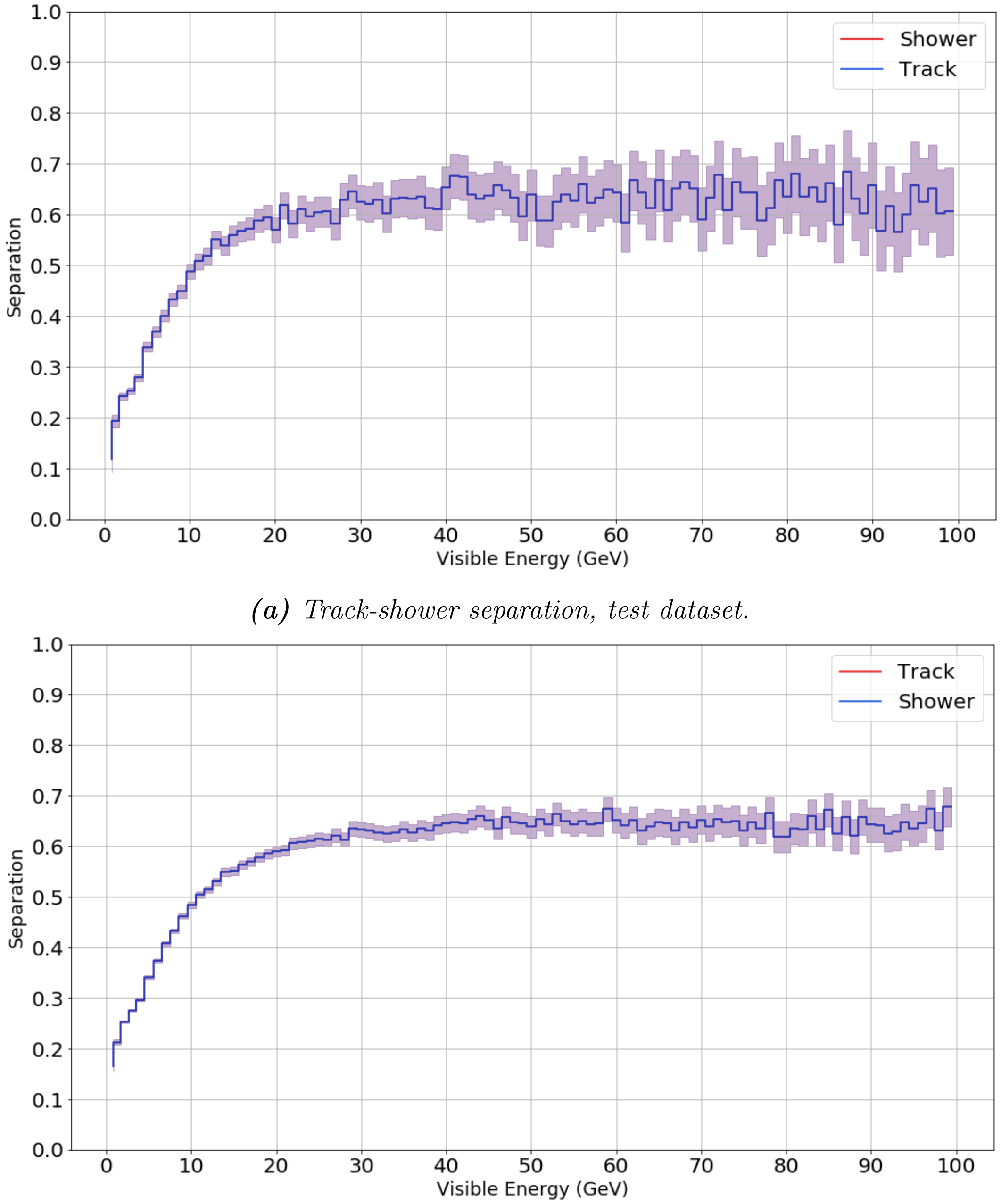
***(b)** Detail from 5.17. This was a small (> 8 million events) testing model with no data preselection or low energy (1-5 GeV) files.*

***Figure 5.18:** Comparison of CC vs NC separation results. The comparison should be done with the full markers in the right (the reconstruction) and not with the hollow markers, which indicate the theoretical maximum separation for a perfect reconstruction.*

This work's result (5.18b) is clearly inferior to Jannik's (5.18a). Instead of improving this result, this work was not continued in favour of the 4 category analysis, since this case was a test rather than the focus of this thesis.

5.4.2 2 categories: Track-shower

This work's separation for the 2 categories, track vs shower(s), can be seen below in 5.19.



(a) Track-shower separation, test dataset.

(b) Track-shower separation, validation dataset.

Figure 5.19: *The same result from two different datasets. This example shows that validation and test dataset give equivalent results with different fluctuations. The validation example (below) has approximately 1 million events more, so the statistical fluctuations are reduced and the value is more clear.*

Separating tracks from showers is the same as separating showers from tracks, so the two measures are completely equivalent in this case and they overlap (notice the legends).

5.4.3 3 categories: track, EM shower, tau + NC

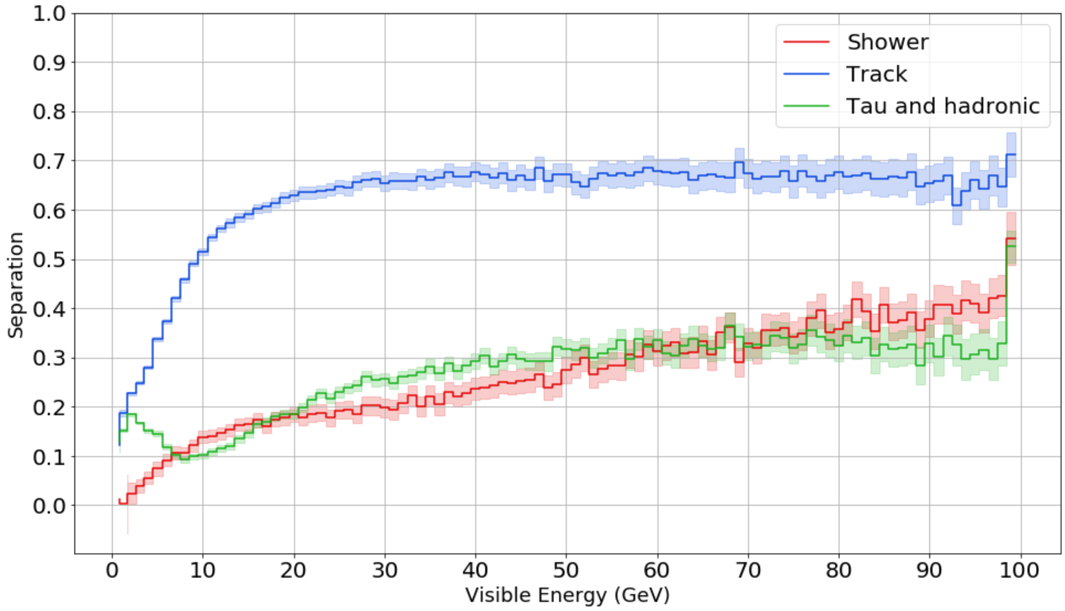


Figure 5.20: *track vs shower vs tau+hadronic events separation, threshold = 0.5.*

The 3 category model, in Fig 5.20 above, presents track separation close to previous model (5.19), and also linearly increasing separation for showers, both electronic and hadronic+Tau. Every "flavour" (line) here is separated against the rest of them (background), this is why the separation differs now.

5.4.4 4 categories

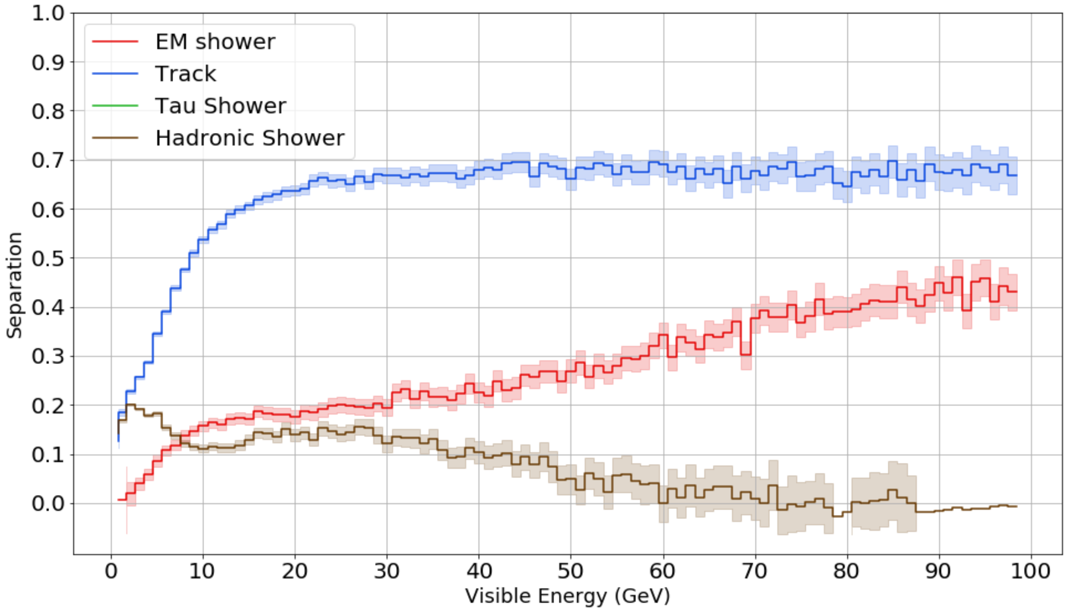


Figure 5.21: *Track vs shower vs tau vs hadronic events separation, threshold = 0.5. The tau line is missing since it did none of its events passes the threshold.*

The 4 category model (Fig 5.21) shows similar track and EM shower results to the 3 category model (5.20), but the tau signal is fully missing (as expected for this threshold) and the neutral current showers surprisingly shows a linear decrease after $\sim$ 50 GeV.

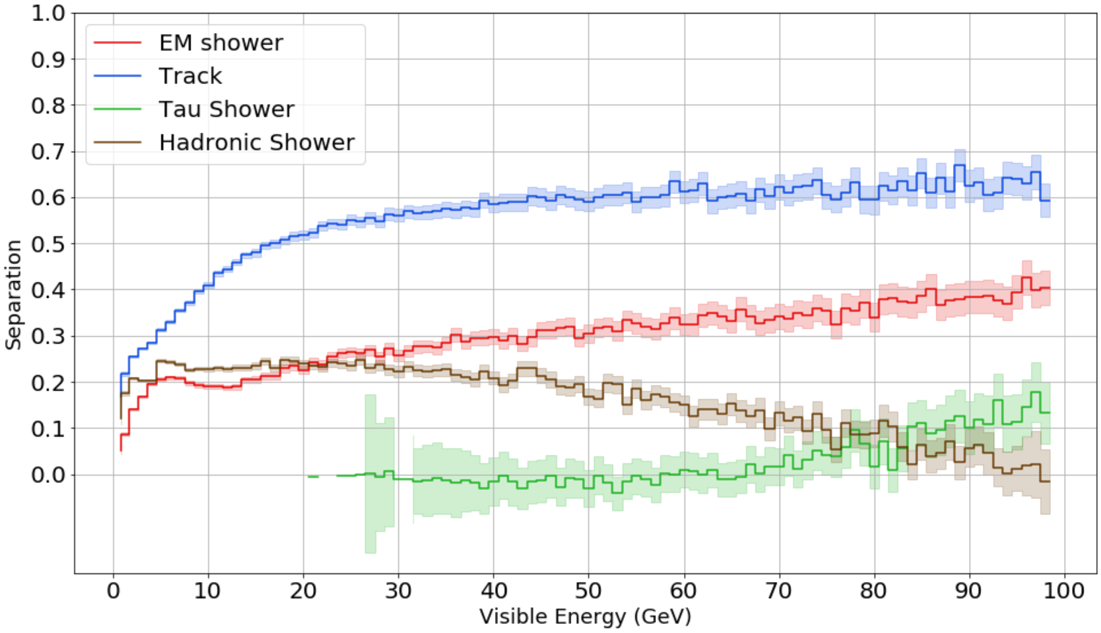


Figure 5.22: *Same result as Fig 5.21, above, with threshold = 0.3. With lowered threshold the tau signal (green) appears again.*

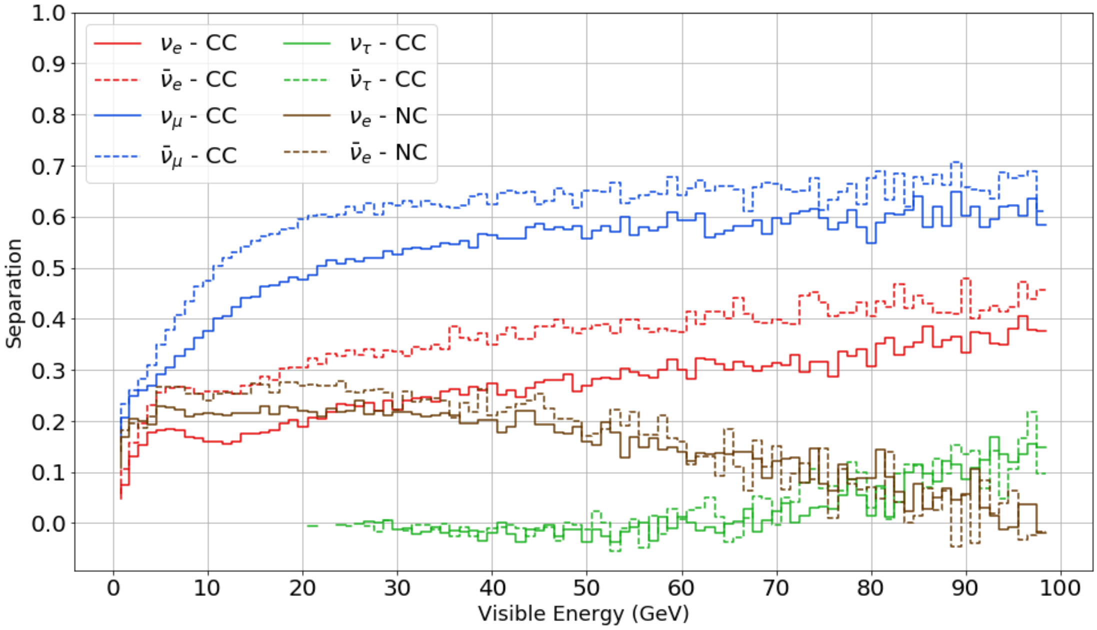
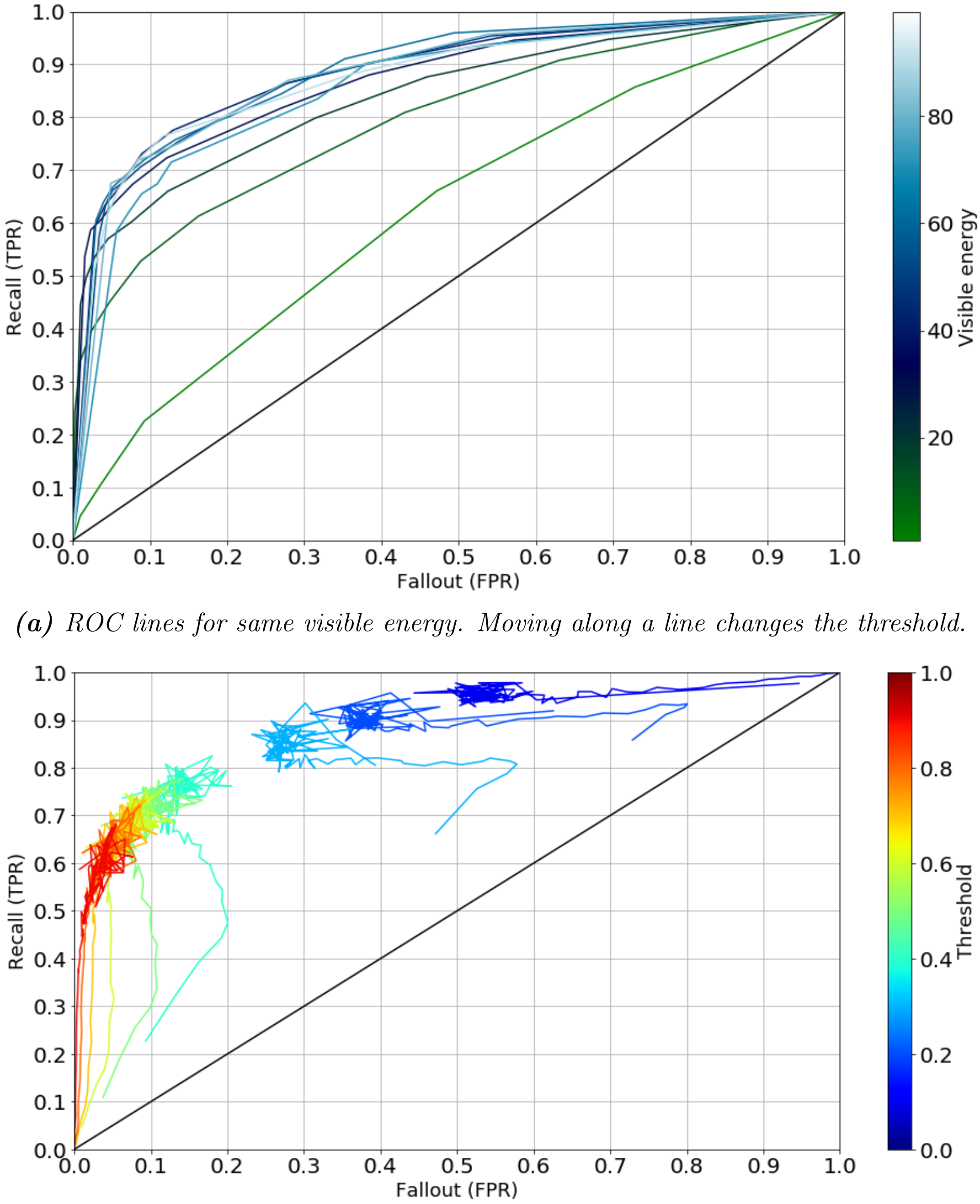


Figure 5.23: *Events separation by interaction flavour and charge, threshold = 0.3. When the flavours are split by charge, we can see the antiparticles performing better again.*

If tau events can be classified as such as with as little as 20 GeV, the classification holds no power (compatible with a random guess) until 70 GeV.

5.5 ROC curves

Previous results are clearly dependant on the threshold, a "free" parameter introduced in the analysis without physical meaning. If we want to inspect the threshold to find the optimum value, we need to look at the ROCs, described in section 4.3.3. For the standard track-shower PID, the best threshold was determined to be 0.6, and used as convention.

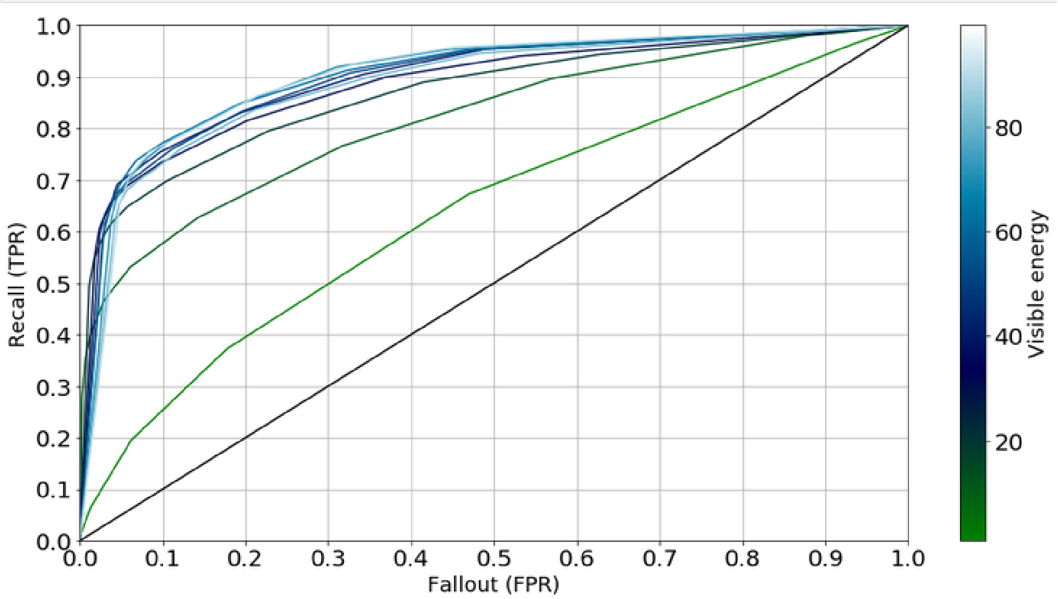


(a) ROC lines for same visible energy. Moving along a line changes the threshold.

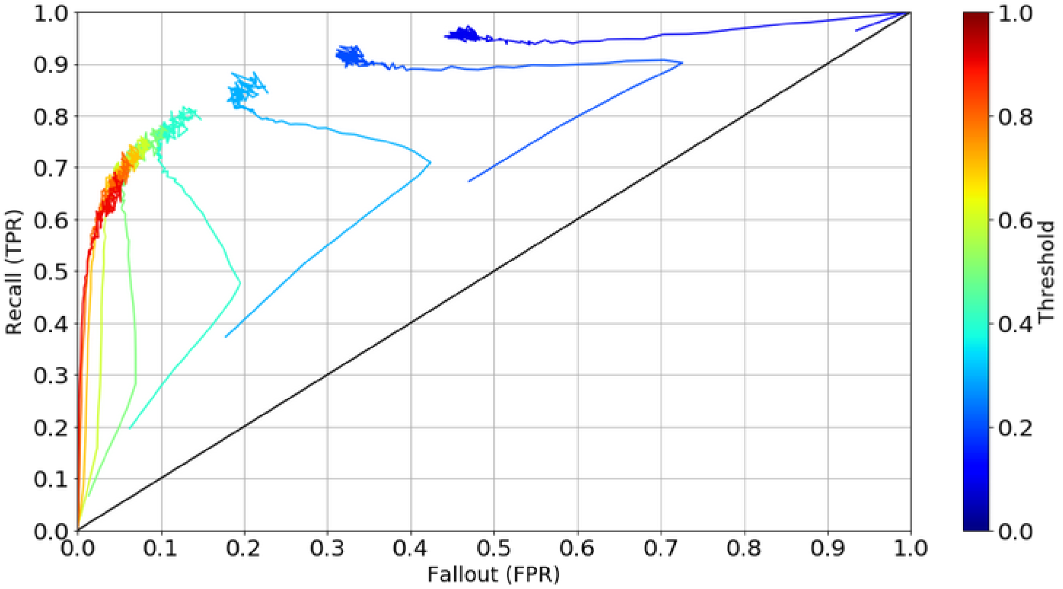
(b) ROC decomposed in isothreshold lines (same threshold). Moving along a line changes the energy. Every line here holds a different number of statistics.

Figure 5.24: *Track-shower ROC space, two different views of the same data. Values are connected by same energy (above) and by same threshold (below).*

The ROCs of the 2 category (5.24) and the 4 category (5.25) models' results look very similar. Looking at the energy decomposition (5.24a, 5.25a), increasing the energy increases performance until ~ 30 GeV, where the values converge. Looking at threshold (5.24b, 5.25b), we can try to estimate a best threshold (for visible energy higher than 30 GeV), of around 0.4 or 0.5, since that is the point with highest recall (less events thrown out). For the 2 (4) category model, any threshold below random guess, 0.5 (0.25) makes no sense. Not all results in this space are necessarily acceptable.



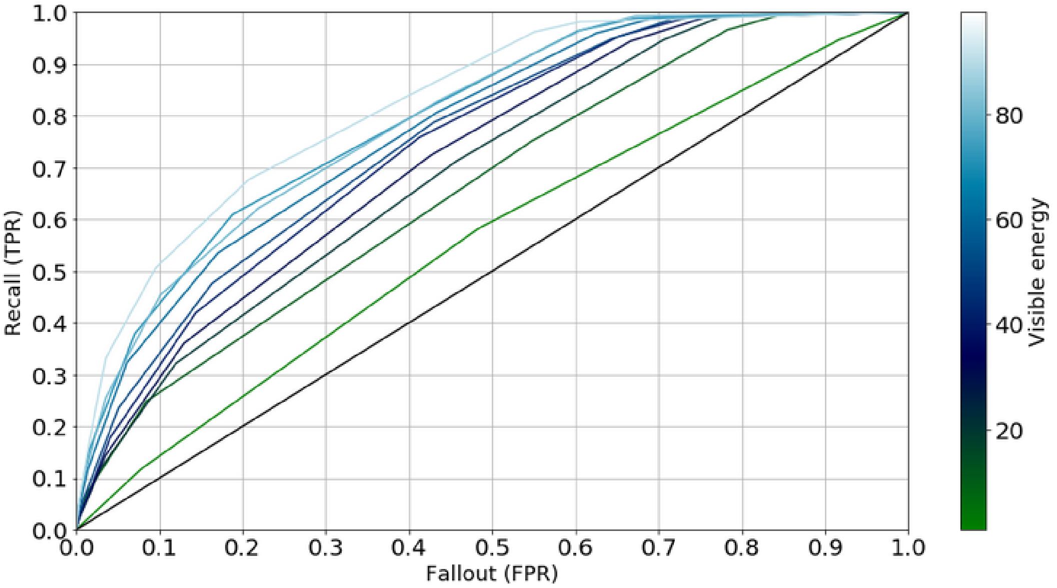
(a) *ROC, decomposed in lines of same visible energy.*



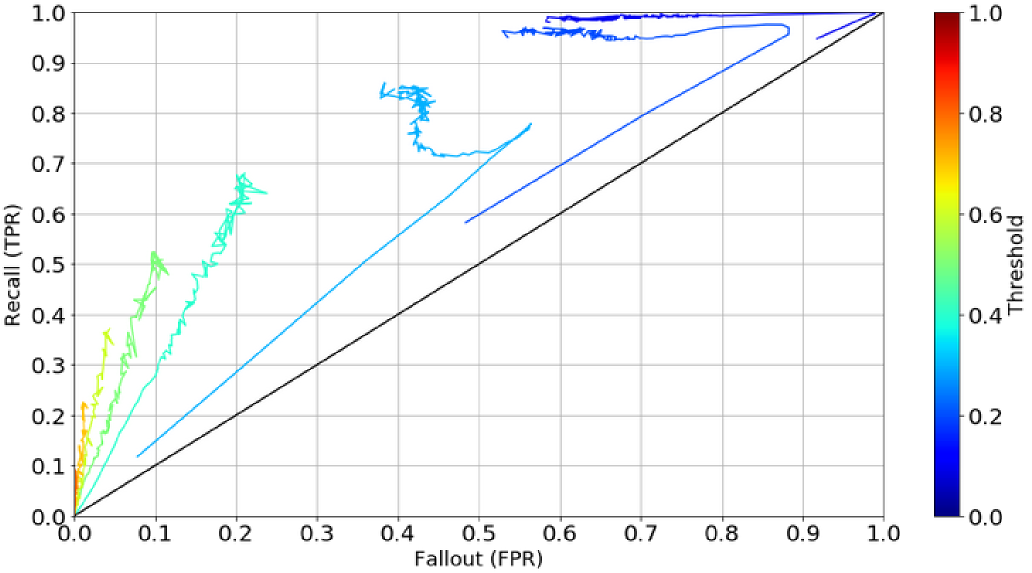
(b) *ROC, decomposed in lines of same visible energy. Isothreshold view.*

Figure 5.25: *4 category, $\mu - CC$ ROC space.*

The EM shower associated ROC, 5.26, does not converge with energy but rather reproduces the linear behaviour seen in the separation plot 5.21. This is clearly seen in the isothreshold lines (5.26b), which underline the fact that the threshold does not give a constant performance, but one linear with energy.



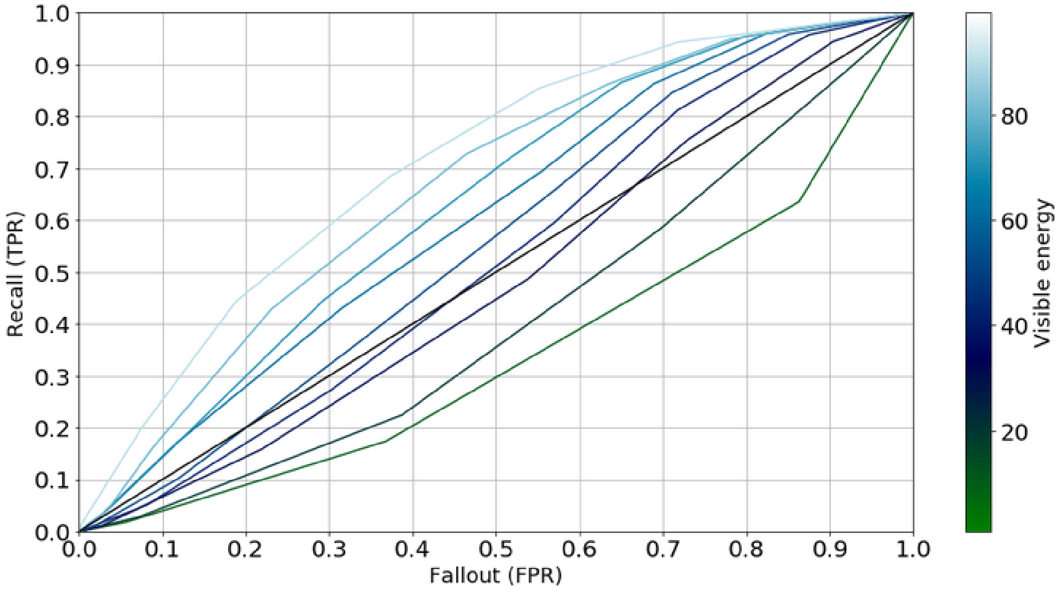
(a) *ROC, decomposed in lines of same visible energy.*



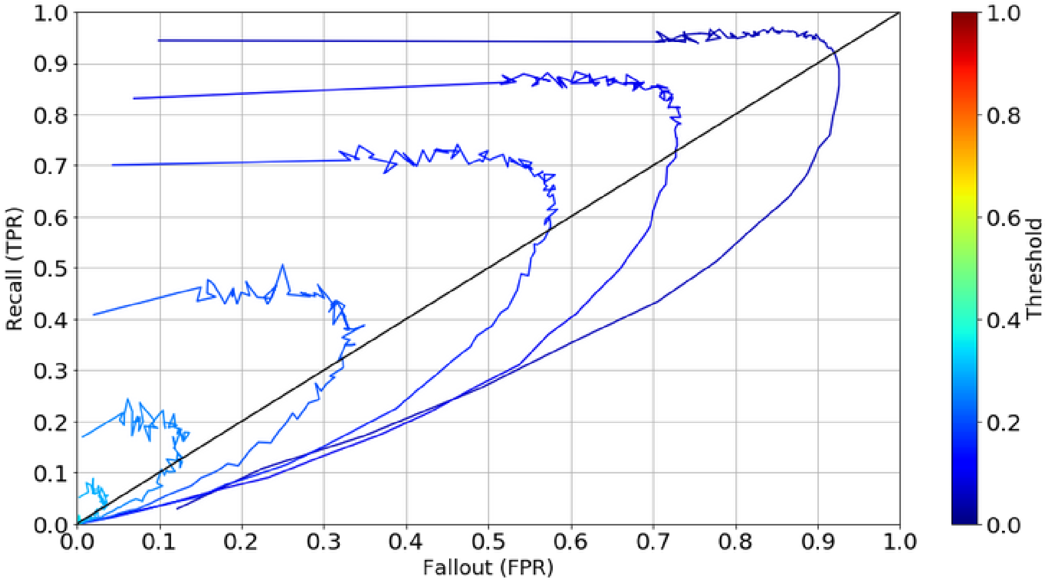
(b) *ROC, decomposed in lines of same visible energy. Isothreshold view. The highest energy bin (top end of the lines) has a very nice performance, but this is due to binning effects that left few statistics in that bin.*

Figure 5.26: *4 category, $e - CC$ ROC space.*

The tau results (Fig 5.27) show an interesting behaviour: more than half of the energy lines (5.27a) lie at some point below the chance line, which means worse than random performance. We know from the separation (5.22) that above 70 GeV the results are just about better than random performance. The threshold curves (5.27b) show again that no result is given above 0.5 probability, and that for lower thresholds, the only part that is better than random is the higher energy end of the spectrum. Please note that the first two isothreshold lines, which perform the best, should not be used, since the threshold would be below 0.25.



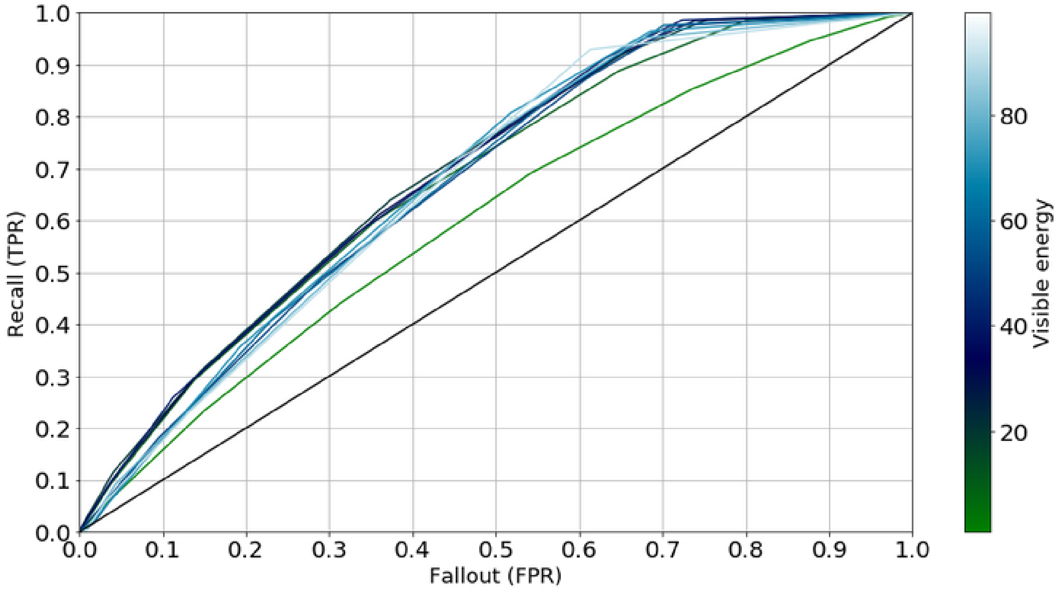
(a) ROC, decomposed in lines of same visible energy.



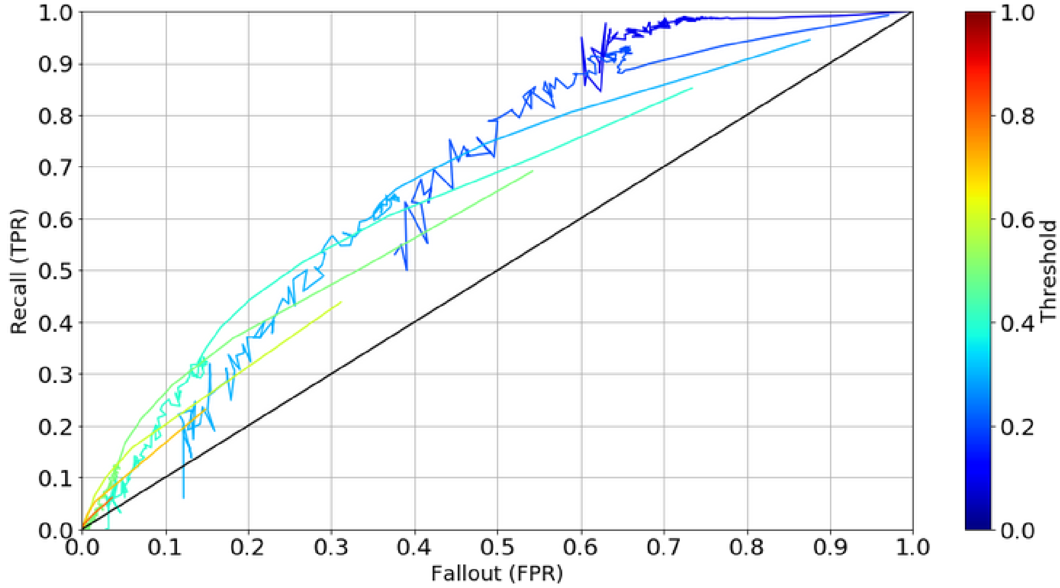
(b) ROC, decomposed in lines of same visible energy. Isothreshold view.

Figure 5.27: *4 category, $\tau - CC$ ROC space.*

The neutral current results (fig 5.28) are also quite notable. Almost all the ROCs in the energy decomposition (5.28a) are overlapping, which means that the performance varies very litte with the energy. A closer inspection shows that the results of the separation (from 5.22) are also here, with the higher energies performing worse. It is in the isothreshold view (5.28b) where this unique behaviour starts to make some sense. All the isothreshold lines here do not converge but seem to move around the same isochance range[3]. This implies that, within some small margin, all values of threshold give the same performance.



(a) ROC, decomposed in lines of same visible energy.



(b) ROC, decomposed in lines of same visible energy. Isothreshold view.

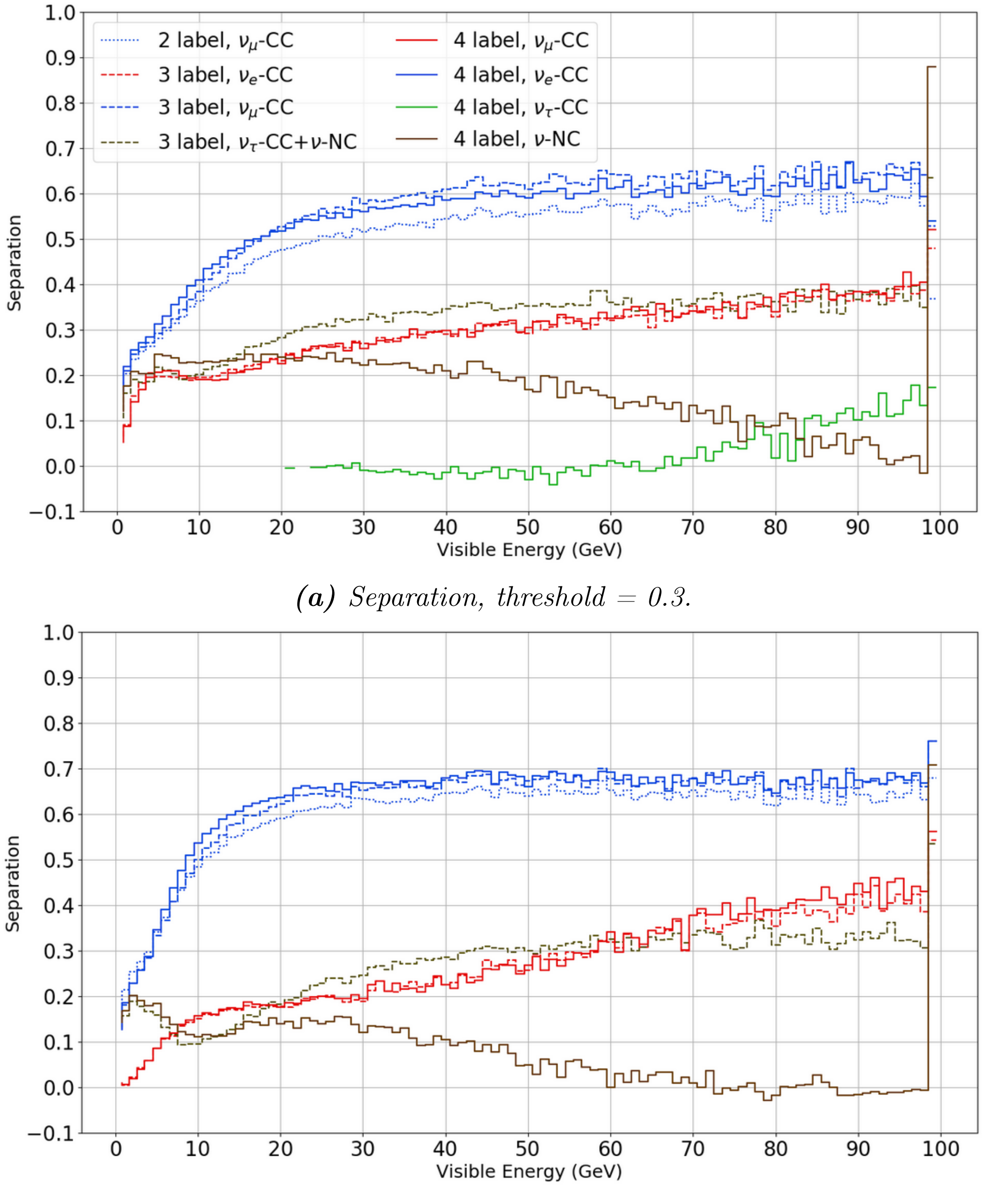
Figure 5.28: *4 category, $\nu - NC$ ROC space.*

[3] A isochance line runs parallel to the random guess line and contains all the values with the same chance accuracy vs contamination ratio. The range is just a isochance line with a certain margin around it.

6 Conclusions

6.1 Research Objectives

Which classification scheme is the best?



(a) Separation, threshold = 0.3.

(b) Separation, threshold = 0.5.

Figure 6.1: *Superposition of previous separation results.*

As mentioned in the motivation, this work, the study of the performance of OrcaNet in the event identification problem, had two different questions in mind: Can we do better *in* the track/shower classification? Can we do better *than* the track/shower classification?

The answer to the first question is yes. Not only OrcaNet performs equal or better than other particle identification schemes already (as shown by the ECAP team, see Fig 5.12), but Fig 6.1 (above) shows that **both 3 or 4 categories produce a small but inherently better track separation than track-shower under the same conditions**. This could be because at least 1 more category allows for an improved classification due to a better definition of a track during OrcaNet's training, as seen in the improved track rejection of Figs **??** and **??**. Furthermore, it remains unclear if 3 or 4 categories performs better, as it seems to be depending on the threshold value. In the case of track separation, 4 categories seems to have an small advantage at low energies, but at higher energies both results are fully compatible within fluctuations.

The answer to the second question is also yes. **The sole presence of a bigger than zero separation for the other categories different than track is a piece of information about the event that is lost with the track/shower model, and that can be recovered not at the cost of track classification but at its benefit.** Let's examine the strength of those signals:

The EM showers ($e-CC$ events) range between 20% separation for low energies to an expected 40% in high energies, increasing linearly with energy[1]. This result is the same for 3 or 4 categories, which makes this category well defined and identified, as it is not affected by the changes to other types of events.

For this identification model, tau identification is the worst performing. Only appearing low thresholds, the separation is statistically significant beyond 70 GeV and just up to 10% at 100 GeV. With our current resolution, only very high energy tau events are expected to give a signal clear enough to be separable. However, the tau signature (double bang) identification relies heavily on the direction of the light (it is the only signal that can produce light in the opposite direction of the primary shower), so it is expected to suffer a huge increase in identification when the PMT-level reconstruction is implemented.

NC events have a 20% separation at lower energies, independent of the threshold, and this decreases linearly with energy. However, if it is combined with the tau class (as seen in the 3 categories model), their separation goes up, performing even better than pure EM showers, see Fig 6.1. Thus, it is a reasonable idea that these events might be entangled during identification, given that the two flavours perform well as a single category but badly as two independent categories. This might be possibly motivated by the absence of events during training. We have mentioned before the lack of simulated low energy tau events, but there is also a lack of high visible energy NC events. NC events are simulated at $>$ 100 GeV in order to have a visible energy

[1] This is an expected behaviour from the theory: more energy means more light and more hits, so more information to correctly identify the event.

of a 100 GeV, so its signals in the detector can be comparable to those of a $e - CC$ event. This is why in the 4 category approach the separation decreases with energy, since there are fewer NC events at high energies that allow for a good training. Also, a very high energy NC event that saturates the detector could be mistaken as a CC event, but there are no such signals that that might be happening in Fig 5.16.

The track-shower scheme of event classification is a strong standard, shared with ANTARES and IceCube and, thus, is the basis of all reconstruction. However, these separations are significant enough to defend the case that it exists a better identification outside just "track" classification. Perhaps the most enlightening conclusion is the fact that all the new, different classification categories are fully compatible with the track-shower framework. The track-shower framework basically establishes a "track score" for the events, as is the case with the current PID. This new system would simply attach "$e - CC$", "NC" and "$\tau - CC$" scores. Should you want to work with 2 categories, or the event reconstruction happens to be bad, you can always just look at the track score and bundle the other categories' results together for a "shower score". This would leave the track-shower separation as the "minimal" or "first order" identification. If in a near future, KM3NeT's particle identification is upgraded from the actual random decision forest to a Deep Learning solution based on OrcaNet, this work shows that implementing more than 2 categories, is a straightforward, no-cost move that gives you a small but noticeable gain in track classification, and a more accurate classification for non-negligible amount of events.

The number of categories - 3 or 4 - which should be implemented is not unequivocally determined from this work. The 3 categories scheme seems to have the best separation for the current resolution, as all of its classes improve their behaviour with energy, and usually have the highest separation score. However, the 4 categories system reproduces the results of the 3 category system with EM showers and track (if not improves this last one), and still makes a good attempt at a finer level of classification. A poor tau performance is not desirable but even having just a few well reconstructed events can be enough to actually explore secondary scientific analyses that requires the 3 flavour identification, something that it is not possible otherwise.

A more precise identification very surely will result in improvements in posterior reconstruction phases. In the current PID chain, the energy of the event is reconstructed first and used as a prior to aid the identification, and no further reconstruction happens after the identification. In this work, identification is happening without prior reconstruction, so the event type can be used as a prior on a future, improved, event reconstruction.

Even if we do not factor such hypothetical changes, there ir a raw improvement in future studies due to the increase in available statistics. A 4 category model which gives the same level of separation as a track-shower model at a 10 to 20% lower threshold will keep a significant amount of events that would be ignored otherwise, see fig 4.3. This increase on the available statistics for posterior studies might prove crucial in the success of certain studies that are constrained by the expected amount of events recorded in our detectors.

This is not an automatic advantage, since we need to factor the large influence that the threshold has on the final event distributions. We can see in Fig 6.1 that lowering the threshold lowers the track separation and improves the separation of the rest of the categories. There is a trade-off between the quality and the number of events when the threshold is changed, and there might not always be a single good value.

In fact, looking at the ROC curves we can see that the correct answer is that there is not an universal best value of a threshold. Even if we assume that the threshold is independent of the data, and its effect is applicable to new, unseen data, it is still a free parameter for the analysis. What this work has shown is that now we can understand the meaning behind a threshold choice. If you take an OrcaNet model (or any other classification model, in fact) and a MC dataset, you can compute its ROC. Now, you are free to chose any threshold while being fully conscious of the model's performance on unseen data. If we look at the ROC 5.25, we can now know that a threshold of 0.3 will give you 0.9 accuracy and 0.3 contamination for energies above 20 GeV, this is, it will find 90% of the muons in your test distribution but it will also return 30% of showers by mistake. However, at threshold 0.3, for low energies the contamination goes up to 0.7. You could also pick a threshold of 0.5, that will give you only 75% of the tracks but the contamination is below 10% for all energies. This way, the best threshold is now a conscious choice that depends on the requirements of the posterior analysis, and its needs with respect to amount of statistics and contamination tolerance.

Ideally, since the ROC characterizes the performance for any given threshold and visible energy, we could in principle make use of Bayesian statistics to find a more accurate probability, applying the ROC score as a prior on the raw predicted output.

As a final note, I want to stress that this work has had a more exploratory aim than a precise characterization goal. The OrcaNet implementation used here is a simple one, without a lot of optimization and fine tuning. For example, not all data was used, but rather as much as allowed for a relatively fast training. The performance plots used here are meant to draw conclusions on the relative performance of OrcaNet and ORCA, not to increase the current maximum of any of its metrics. This is not to say that the performance is bad in this model, but it could be quite easily improved with a fully optimized model.

6.2 Practical Implications

The results presented have shown a clear improvement in particle identification, albeit a small one in the track category. Whether this improvement in particle identification can be translated to a quantifiable improvement in sensitivity is a question without an obvious answer. The simple idea is that the lowering of threshold allows for less discarded events for the mass hierarchy fit, and thus less data-taking time. However, changing our reconstruction algorithm scheme, and how the events are used, means that the relation between event number and performance is not linear and cannot be predicted beforehand.

What is left as the very next step is to run a mass hierarchy fit and a sensitivity study with these models and these results, to compare and show the net gain in significance from this new particle identification technique. This final test was considered but goes beyond the scope of this work. The mass hierarchy software used, MONA (Mass Ordering Nikhef Analysis), was just recently adapted to the Python programming language, so it would be compatible with OrcaNet output with little effort. The results from MONA tests would show the expected improvement and the benefit of the OrcaNet/DL implementation; this is definitely the first step in the direction of any possible changes in KM3NeT's reconstruction scheme.

The best method on how use the event information is still an open question, with a lot of progress in all the reconstruction steps. Track-shower identification is done with a single variable ("track-like" score). The ideal muon-CC vs electron-CC or tau-CC that stems from the oscillation probabilities provides less theoretical uncertainties when building the oscillograms, and thus a more clear analysis. A muon-electron oscillogram also requires another extra variable to separate electrons. While this work shows that it is possible, this is only true for a small fraction of all the showers recorded. Discarding most of the shower events in pursue of a better quality might not be the most useful approach.

The true power of this new reconstruction method is not shown in a single metric, but in the realization that we can extract more information from our events than previously thought. Even if the performance is limited, this opens up not only the possibility of an electron-muon-tau reconstruction, but the new dimensions in the information opens up the door to a new unexplored methods of reconstruction that make the most about all the information contained in a single event. For example, the probability for different flavours comes now from raw hits, so it might be used before the reconstruction (as a prior hypothesis), instead of after the reconstruction, as is done now.

The fundamental conclusion is that, given that the detector is under construction, we now need to understand its the resolution and limits, so we can develop tools to fully exploit all the information that they will provide. This work looked at the number of categories in PID, and even if there is promising improvements for a higher number of categories, the great influence of the threshold might be the single factor that determines the final performance. Thus, a precise optimization of threshold will be needed for future analyses.

6.3 Future Research

The code has been developed, the technology is up an running, and the methods have proved their potential for usefulness beyond reasonable doubt, so I am of the opinion that the work in the OrcaNet/Deep Learning should be continued. This is just an outline of possible future work ideas that could be attempted:

- **Tau - NC entanglement study**: Train a model in a fully homogeneous dataset. This means the same amount of events for flavour and charge, and hopefully also for energy or visible energy. This might require to prepare or even simulate new events to have enough statistics. This is the starting point for a better benchmark of this tools or even a OrcaNet implementation specialized in tau events.

- **MONA Integration**: As mentioned before, adding MONA to OrcaNet's pipeline will provide a clear, quantifiable measure of the improvements in reconstruction done by OrcaNet. This is also true for the OrcaNet reconstruction implementations, not only in flavour but for energy and direction thanks to variable regression beyond simple classification.

- **ARCANET**: OrcaNet itself does not incorporate any ORCA specific coding, as the images that is sees do not distinguish between the two instruments, thanks to the modular design and the binning in OrcaSong. Adding the ARCA detector geometry and its corresponding binning (Perhaps implementing the work done in M. Post thesis [50]) should not be difficult, and would open the opportunity to replicate the tests in this work, along with bringing the chance to compare both detector's performance under the same reconstruction techniques.

- **Architectural redesign**: OrcaNet is not limited to a single architecture. The simple implementation here is effective but far from the state of the art in Deep Learning, a field still undergoing a revolution, with new improvements in theory and practice happening every year. Even if it will require some extensions to OrcaNet code, different technologies can be tried to improve the sheer performance of the algorithm. ResNets (Residual networks) are the current champions in computer vision, while recurrent neural networks (RNN) based in LSTM (Long short-term memory) architecture are the current way to go in sequential information (like the time-evolution of an event). Other than drastic changes, very little has been done for optimization of the algorithms in terms of efficiency. Bringing the computation time and/or the amount of data processed up is the way to achieve more performance in Deep Learning.

- **Specialized event filtering**: There is a lot of additional research that could be done with the KM3NeT instruments, such as tau physics, cosmic ray physics, Dark Matter decay searches from single sources, supernova alarm system, etc. For all this ideas and more that might be difficult to isolate and do not warrant a lot of time or investment, OrcaNet can mean a quick way to set up an independent trigger and reconstruction method specific for what is required in each case. As long as there is data, real or simulated, a model can be trained and prepared.

Acknowledgements

This thesis is the result of one year of work at Nikhef and two years of a Master's degree in the Netherlands, and I am thankful for this experience and all the amazing people I have had the chance to met. With no special relevance or order:

Thanks to Paul, my supervisor. Thank you for letting me be part of KM3NeT for this last year, thank you for the freedom and encouragement to take an overly ambitious and optimistic project, but, mostly, thank you for your guidance.

Thanks to the amazing friends that I made this year. First, to my office colleagues Brían Ó Fearraigh, Lodewijk Nauta and Karel Melis. Thank you for your patience and your dedication to answer my incessant, interrupting questions. Thank you for the jokes. Thanks for your help proofreading this thesis. Big shout-out to Brían for proofreading this whole thesis while dealing with the longest running software error of the team to date[2]. Then, thanks to the rest of you: Jordan, Thijs, Max, Rasa, Bruno, Alfonso and Milo. Thanks for making me part of KM3NeT and thanks for the fun times, especially for those around the coffee machine and the table football. Please, make sure you keep being Nikhef's best for when I come visit.

Thanks to the one person without whom this thesis probably wouldn't happened: Michael Moser. Thank you for your constant technical support with the coding problems all over this year, and for the excellent tools you have made. Additional thanks to the computing team at Nikhef and to Jannik Hofestädt for his help with the separation questions.

Thanks to the rest of the KM3NeT team! Ronald, Dorothea, Maarten, Aart and Daan. You are making an amazing project come true.

Thanks to Gonzalo Contreras Aso for his friendship and his hospitality this last two years.

To all of the above I wish you best of luck and always remember:

"It doesn't stop being magic just because you know how it works."

My final thanks go to dad and mom, the giants on whose shoulders I am have stood in order to reach my dreams. All of this has been possible thanks to you.

[2] Of course, it had to be a missing semicolon. Yes, really.

Bibliography

[1] K. Fajans, “Radioactive transformations and the periodic system of the elements”, Berichte der Deutschen Chemischen Gesellschaft **46**, 423–432 (1913).

[2] F. Soddy, “The radio-elements and the periodic law”, Nature **91**, 57–58 (1913).

[4] J. Chadwick, “Possible existence of a neutron”, Nature **129**, 312–312 (1932).

[5] H. BETHE, et al., “The “neutrino””, Nature **133**, 532–532 (1934).

[6] W. Heisenberg, “Zur theorie der "schauer" in der höhenstrahlung”, Zeitschrift für Physik **101**, 533–540 (1936).

[7] F. REINES, et al., “The neutrino”, Nature **178**, 446–449 (1956).

[8] B. Pontecorvo, “Mesonium and Antimesonium”, Soviet Journal of Experimental and Theoretical Physics **6**, 429 (1958).

[9] K. Greisen, “Cosmic ray showers”, Annual Review of Nuclear Science **10**, 63–108 (1960).

[10] M. Markov, et al., “On high energy neutrino physics in cosmic rays”, Nuclear Physics **27**, 385–394 (1961).

[11] G. Danby, et al., “Observation of high-energy neutrino reactions and the existence of two kinds of neutrinos”, Physical Review Letters **9**, 36–44 (1962).

[12] Z. Maki, et al., “Remarks on the unified model of elementary particles”, Progress of Theoretical Physics **28**, 870–880 (1962).

[13] C. V. Achar, et al., “The Kolar Gold Field neutrino experiment.”, International Cosmic Ray Conference **1**, Provided by the SAO/NASA Astrophysics Data System, 1012 (1965).

[14] F. Reines, et al., “Evidence for high-energy cosmic-ray neutrino interactions”, Physical Review Letters **15**, 429–433 (1965).

[15] K. Greisen, “End to the cosmic-ray spectrum?”, Phys. Rev. Lett. **16**, 748–750 (1966).

[16] G. T. Zatsepin, et al., “Upper limit of the spectrum of cosmic rays”, JETP Lett. **4**, [Pisma Zh. Eksp. Teor. Fiz.4,114(1966)], 78–80 (1966).

[17] S. Weinberg, “A model of leptons”, Phys. Rev. Lett. **19**, 1264–1266 (1967).

[18] V. Gribov, et al., “Neutrino astronomy and lepton charge”, Physics Letters B **28**, 493–496 (1969).

[20] G. Arnison, et al., “Experimental observation of isolated large transverse energy electrons with associated missing energy at”, Physics Letters B **122**, 103–116 (1983).

[21] E. Amaldi, "From the discovery of the neutron to the discovery of nuclear fission", Physics Reports **111**, 1–331 (1984).

[22] S. Mikheyev, et al., "Resonant neutrino oscillations in matter", Progress in Particle and Nuclear Physics **23**, 41–136 (1989).

[23] K. Lande, et al., "Results from the Homestake solar neutrino observatory", Conf. Proc. **C900802**, 867–675 (1990).

[24] N. Cooper, *Los alamos science, number 25 – 1997: celebrating the neutrino*, tech. rep. (Dec. 1997).

[25] Y. Fukuda, et al., "Evidence for oscillation of atmospheric neutrinos", Physical Review Letters **81**, 1562 (1998).

[26] Q. R. Ahmad, et al., "Measurement of the rate of ν e+ d→ p+ p+ e- interactions produced by b 8 solar neutrinos at the sudbury neutrino observatory", Physical Review Letters **87**, 071301 (2001).

[27] T. A. collaboration: J.A. Aguilar *et al*, "Transmission of light in deep sea water at the site of the antares neutrino telescope", Astroparticle Physics **23**, 131–155 (2005).

[28] D. Powers, "Evaluation: from precision, recall and f-factor to roc, informedness, markedness & correlation", Mach. Learn. Technol. **2** (2008).

[29] U. Katz, et al., "High-energy neutrino astrophysics: status and perspectives", Progress in Particle and Nuclear Physics **67**, 651–704 (2012).

[30] K. W. Melis, "Reconstruction of high-energy neutrino-induced particle showers in km3net.", masterthesis (Universiteit van Amsterdam (UVA), 2014), p. 66.

[31] K. Simonyan, et al., "Very deep convolutional networks for large-scale image recognition", arXiv 1409.1556 (2014).

[32] S. Adrián-Martınez, et al., "Letter of intent for KM3net 2.0", Journal of Physics G: Nuclear and Particle Physics **43**, 084001 (2016).

[34] S. Sheehan, et al., "Deep learning for population genetic inference", PLOS Computational Biology **12**, edited by K. Chen, e1004845 (2016).

[35] B. Ó. Ferraigh, "Constraining sterile neutrino parameters using oscillation experiments and cosmology", Master's thesis (The University of Manchester, 2017), p. 86.

[36] J. Hofestädt, "Measuring the neutrino mass hierarchy with the future km3net/orca detector", doctoralthesis (Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 2017), p. 239.

[37] Z. Lu, et al., "The expressive power of neural networks: a view from the width", in *Advances in neural information processing systems 30*, edited by I. Guyon, et al. (Curran Associates, Inc., 2017), pp. 6231–6239.

[38] E. Pinat, "The icecube neutrino observatory: search for extended sources of neutrinos and preliminary study of a communication protocol for its future upgrade", doctoralthesis (Université Libre de Bruxelles (ULB), 2017), p. 161.

[39] M. Ahlers, et al., "Opening a new window onto the universe with icecube", Progress in Particle and Nuclear Physics **102**, 73–88 (2018).

[40] S. Bourret, et al., “Sensitivity of orca to the neutrino mass ordering and oscillation parameters”, (2018) 10.5281/zenodo.1300771.

[41] I. Collaboration, “Neutrino emission from the direction of the blazar txs 0506+056 prior to the icecube-170922a alert”, Science (2018) 10.1126/science.aat2890.

[42] G. Fantini, et al., “Introduction to the formalism of neutrino oscillations”, Adv. Ser. Direct. High Energy Phys **28**, 37–119 (2018).

[43] A. C. O. Santos, et al., “Lower mass bound on the w′ mass via neutrinoless double beta decay in a 3-3-1 model”, Advances in High Energy Physics **2018**, 1–7 (2018).

[44] M. Tanabashi, et al., “Review of particle physics”, Phys. Rev. D **98**, 030001 (2018).

[45] A. Albert, et al., “ANTARES neutrino search for time and space correlations with IceCube high-energy neutrino events”, The Astrophysical Journal **879**, 108 (2019).

[46] A. Avrorin, et al., “Status of the baikal-GVD neutrino telescope”, EPJ Web of Conferences **207**, edited by C. Spiering, 01003 (2019).

[47] I. Esteban, et al., “Global analysis of three-flavour neutrino oscillations: synergies and tensions in the determination of 23, δCP, and the mass ordering”, Journal of High Energy Physics **2019** (2019) 10.1007/jhep01(2019)106.

[50] M. Post, “”km3nnet” a neural network for triggering and classifying raw km3net data”, Master’s thesis (University of Amsterdam (UVA), 2019), p. 48.

Online References

[3] W. Pauli, *Pauli’s letter*, Made available by The MicroBoone Collaboration, (1930) http://microboone-docdb.fnal.gov/cgi-bin/RetrieveFile?docid=953;filename=pauli%20letter1930.pdf.

[19] R. Hudson, *Reversal of the parity conservation law in nuclear physics*, Accessed 2019/07/7, (1982) https://nvlpubs.nist.gov/nistpubs/sp958-lide/111-115.pdf.

[33] NIST, *The reversal of parity law in nuclear physics*, Accessed 2019/07/7, (2016) https://www.nist.gov/pml/reversal-parity-law-nuclear-physics.

[48] T. Gal, *Km3pipe: analysis framework for km3net related data, heavily based on numpy.* Last visited: July 2019, (2019) https://git.km3net.de/km3py/km3pipe.

[49] M. R. A. Khan, *Rocit: an r package for performance assessment of binary classifier with visualization*, Last visited: July 2019, (2019) https://cran.r-project.org/web/packages/ROCit/vignettes/my-vignette.html.

[51] E. H. Santiago, *Orcanet personal fork*, Last visited: July 2019, (2019) https://git.km3net.de/ehuescasantiago/OrcaNet.

[52] Unknown, *Confusion matrix*, Last visited: July 2019, (2019) https://en.wikipedia.org/wiki/Confusion_matrix.

A Introduction to Machine Learning

Machine Learning (ML) or Statistical Learning is the subfield in artificial intelligence that studies how a computer system can use statistical methods and algorithms to produce advanced data analysis without using explicit instructions. While the term "Machine Learning" was coined by Arthur Samuel in 1959, the field has been constantly developing since the 1980s - 1990s, with the main explosion happening in 2010 with the appearance and popularization of the GPU-based computation that has made the Deep Learning (see below) technologically feasible. Despite the many different stages of the field, all ML algorithms share a main goal: To make a computer program that is capable of performing a single task without using specific, previously coded, instructions as efficiently as, or better than, a (trained) human.

ML is deeply rooted in the statistical analysis field, where most of the principles and methods stem. From the statistics point of view, machine learning is just a new, efficient computational implementation of the traditional statistical tools. This means we can formulate our ML goal in a more precise way. For any given task, we can define a random behaviour as our background hypothesis (H0) and a perfect performance on the task as our ideal "signal" hypothesis (H1). Thus, a ML model is, in essence, any algorithm that starting in H0, and with the sole knowledge of H1 and its own past performance, is able to move towards H1. The most simple and popular example, the case of this thesis, is the binomial classification between a signal (an event signature) and a background (the rest of the signatures).

This is done by allowing the algorithm use using concrete examples of the task where the proper value or label is already known[1]. This is called training data, since the whole "learning" idea is motivated by this feedback principle, where the information of the performance is used to increase the future performance, which resembles how humans learn from successes and mistakes. An algorithm, instead, is defined by its loss or cost function. The loss function is a mathematical function that uses the information from the training data as an input and gives a measure of performance as the output. The statistical origin of ML can be seen in the loss function. In the most popular ML application, the multinomial classification (2 or more independent categories), the loss is inversely related to the likelihood. This learning process is then, reduced at adjusting the free parameters in the loss function to the minimal value (optimal configuration), by the use of iterative trial-and-error, so it is equivalent to find the maximum in the likelihood in our H0 vs H1 problem.

[1] Quantifiable data is trivial to add to a function, but abstract concepts such as labels are encoded with a system called 'one hot encoding', where the inputs and outputs of your data are arrays, so that the categories form an orthogonal space with every category as an eigenvector.

The information concept is crucial, because it means that the ML algorithm is fundamentally defined and determined by the data available to it. In fact, the H0 hypothesis is formally defined in terms of the information encoded in the training data: If the data is totally random, then the algorithm will not perform better than a random prediction and if the information in data is only enough to warrant a partial separation, then that is the H1 of the problem. This immediately poses the fist problem that the ML faces: What if my training data is not representative of the whole dataset I want to apply my algorithm on? When the algorithm is so optimized that only is useful for its training data, it is call overfitted, and is in detriment of the usefulness of the ML (you don't want to have a machine that only works on past data and not in future one).

Thus, one of the requisites of your algorithm is the *generalization*, this is, the performance on the training data is not as useful as the performance on the unseen data. This is why two more datasets are used: The validation and the test dataset. Their purpose is the same, but the validation dataset is reused all along the training process (which is okay, as long as the information from the validation is not used in the training). And the test dataset, which is used only once for the final check on performance. The best practice is the train-validate-test data splitting, even if is very common to combine validation and test in the low statistics case (which doing so should not imply loss of generalization). The recommended practice is use the test data once, and then find or obtain new test data, so that you can add the "used" data to the training data.

Validate and test data need to be as small as possible, so that most data can be used for training and improve the performance. However, it needs to contain enough events that it does not induce inhomogeneities and/or statistical fluctuation effects. This is especially relevant in the cases like this one where a huge percentage of the data is discarded (see threshold section). While there is no perfect configuration, the usual relative ratio of events for train-validate-test goes from 60%-20%-20% for low amount of events to 95%-5%-1% when millions of data samples are available. This means that our case is far from employing the most efficient data distribution, but, as mentioned, we are in one of the rare cases that data is not a limiting factor, so this is not a concern.

Machine Learning algorithms are historically separated in supervised and unsupervised learning, referring to the fact that the data used for study is labelled (categorized); however, most of the Machine Learning nowadays is supervised, as the supervised techniques, classification and (linear) regression are more useful and efficient than the unsupervised ones, clustering and dimension reduction. The separation that is getting more and more relevant is the *Shallow Learning* vs *Deep Learning* (DL)[2], which makes reference to the architectural design of these algorithms. Shallow learning covers the models that obtain the final value (prediction, loss, etc) from a single set of computations with the data. This models are older and not currently favored, but still useful when the computation power is limited. Some shallow learning models (or family of models, since the same principle can have several slightly different implementations) are K-Nearest-Neighbour (KNN), Decision

[2] The proper name for Deep Learning is representation or hierarchical learning, but since such technicalities difficult the understanding, the more popular terminology is chosen here. Keeping in mind that in a more formal context this convention is inexact, it still can be used here.

Trees (including Random Forest), etc. Deep Learning, on the other hand, is based on Artificial Neural Networks or simply Neural Networks (NN).

A.1 Neural Network

A neural network is a complex ML algorithm inspired by the workings of a biological brain, but without any resemblance to it. It is called network because is the interconnected combination of "neurons", where a neuron is a mathematical 1-dimensional function, that combines a series of input values into a single output value.

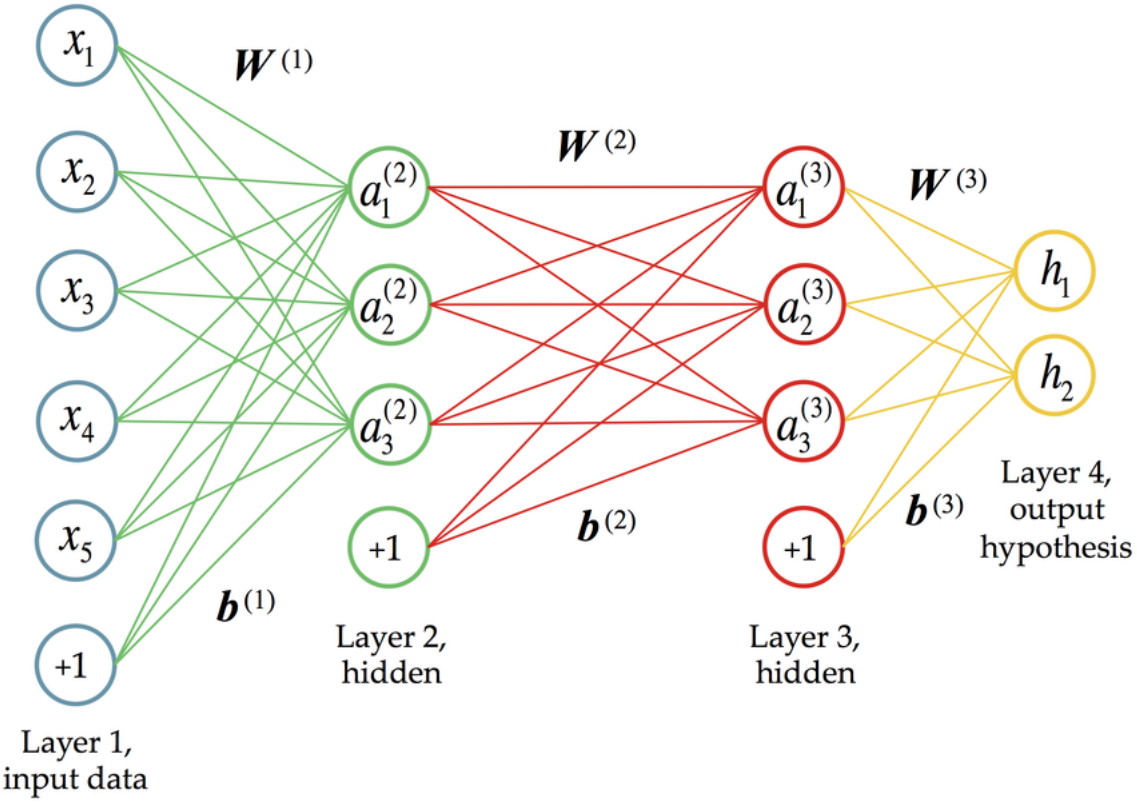


Figure A.1: *Visualization of a simple neural network architecture. x_i are the data points, a_i are the neurons, and h_i are the outputs. The configuration the network is called architecture. The neurons are grouped in "layers", in such a way that every layer includes the neurons with the same, single purpose. A layer takes the output of the previous one and feeds the following one. The neurons among a single layer are never connected, and besides their specific function they can share a bias, used to highlight the importance of the layer in the final computation (the +1 in the bottom). Since the user only needs to act on the first layer (that feeds the information) and the last layer (that contains the output), the middle layers are called hidden layers. This is from where the idea of "deepness" in DL stems from, where more layers = more 'deepness'. Source: [34], with Creative Commons license.*

The fundamental idea that supports the neural networks, and the modern Deep Learning, is the universal approximation theorem. In its most modern formulation [37], the universal approximation theorem states that any continuous and convex function (like the inverse likelihood function) of n dimensional input variables can be approximated by a NN with 1 single 'hidden' layer of ReLU neurons whose size is

$n + 1$. This means that the likelihood landscape can be replicated within the NN encoding enough detail that a minimization algorithm can find the best value within. ReLU stands for Rectified Linear Unit, which is a neuron using the ramp function ($f(x) = max(x, 0)$, where x is the input). The neurons functions are usually basic, and only have the task to determine if the neurons in the next layer receive signal (this is where the artificial neurons imitate the behaviour of the biological ones). This is why the functions in the neurons are called activation functions. Every neuron usually includes a single free tunable parameter, the strength of the activation, that gives the whole network the flexibility to adopt many different configurations.

The most popular type of Neural Network is the Convolutional Neural Network, the technique behind the recent revolution on computer vision, the ability for a computer to recognize and identify objects and images. A Convolutional Neural Network (CNN) is the type of neural network where (at least) some of the different layers are connected by a convolution (technically, a sliding dot product).

A.2 Model's Architecture

The architecture used for all the OrcaNet implementations in this thesis is based on the VGG [31], one of the most conceptually simple architectures (named from its creators, members of the Visual Geomerty Group of the University of Oxford). The VGG is a very popular architecture today, considered still of the best architectures for deep learning image classification. The basic elements of the network are the activation layer, the pooling layer, the flatten layer and the fully connected or dense layer, and the extra layers used for regularization are normalization layer and the dropout layer. The configuration uses 45 layers: 12 activation, 10 convolutional, 10 normalization, 10 dropout, 2 dense and 1 flatten.

- **Activation layer**: The layer with the activation "neuron" mentioned above. In our case, the activation uses rectilinear unit neurons (relu).
- **Pooling layer**: The pooling layer is dedicated to combine the result of several adjacent values into one, reducing the dimensionality of the data. This is done either by taking the average of the values or the maximum value (maxpool). In our case, the pooling layer is maxpool of size 2 in the 3 dimensions.
- **Flatten layer**: The flatten layer simple reduces the dimension of the previous layers (2 or 3 range matrices) to a 1 dimensional vector. This is needed only once to be able to give a single vector with the output probabilities ("one hot encoding"), so it is usually placed towards the end of the configuration.
- **Dense layer**: A dense layer is simple a fully connected layer, where every input is connected to every output. It makes sure that the all the values from the previous layers are able to contribute to the final output layer. Placed at the end of the network, at least 1 is required. In out case, we use 3 dense layers (including the final output layer) to allow fine tunning (since introduces more weights). It is usually useful to add an activation layer afterwards.
- **Normalization layer**: Scales the previous layer so it has zero mean and unit standard deviation.

- **Dropout layer**: It "drops out" or deletes a random subsample of the previous layer. This counterintuitive measure is taken to increase the generalization of the network output.

The regularization layers are extra operation that makes the training more resilient training, so they are useful every few layers. For example, if the network were to receive as information the energy of your dummy particles, applying normalization in the early layers simplifies the data, since the machine does not know nor care what "energy" is, so the exact value is not as useful, since the same information can be encoded in a continuous range, allowing for faster training (simpler operations). Dropout is used to avoid overfitting. Overfitting happens when the prediction power of the network is powerful enough (enough free weights in the model), it starts to learn random small scale patterns in the data as outlier behaviors and mistaken for relevant information, which improves performance for the exact training data configuration but it is harmful in unseen data. Dropout randomly deletes information, so it forces the machine to relearn certain weights periodically, effectively erasing the short scale fluctuations. Dropout is the reason that the loss (accuracy) of validation data converges in our results; without dropout, the loss (accuracy) of the validation instead would decrease (increase) at first, but then it would start to increase (decrease), when the overfitting starts to happen.

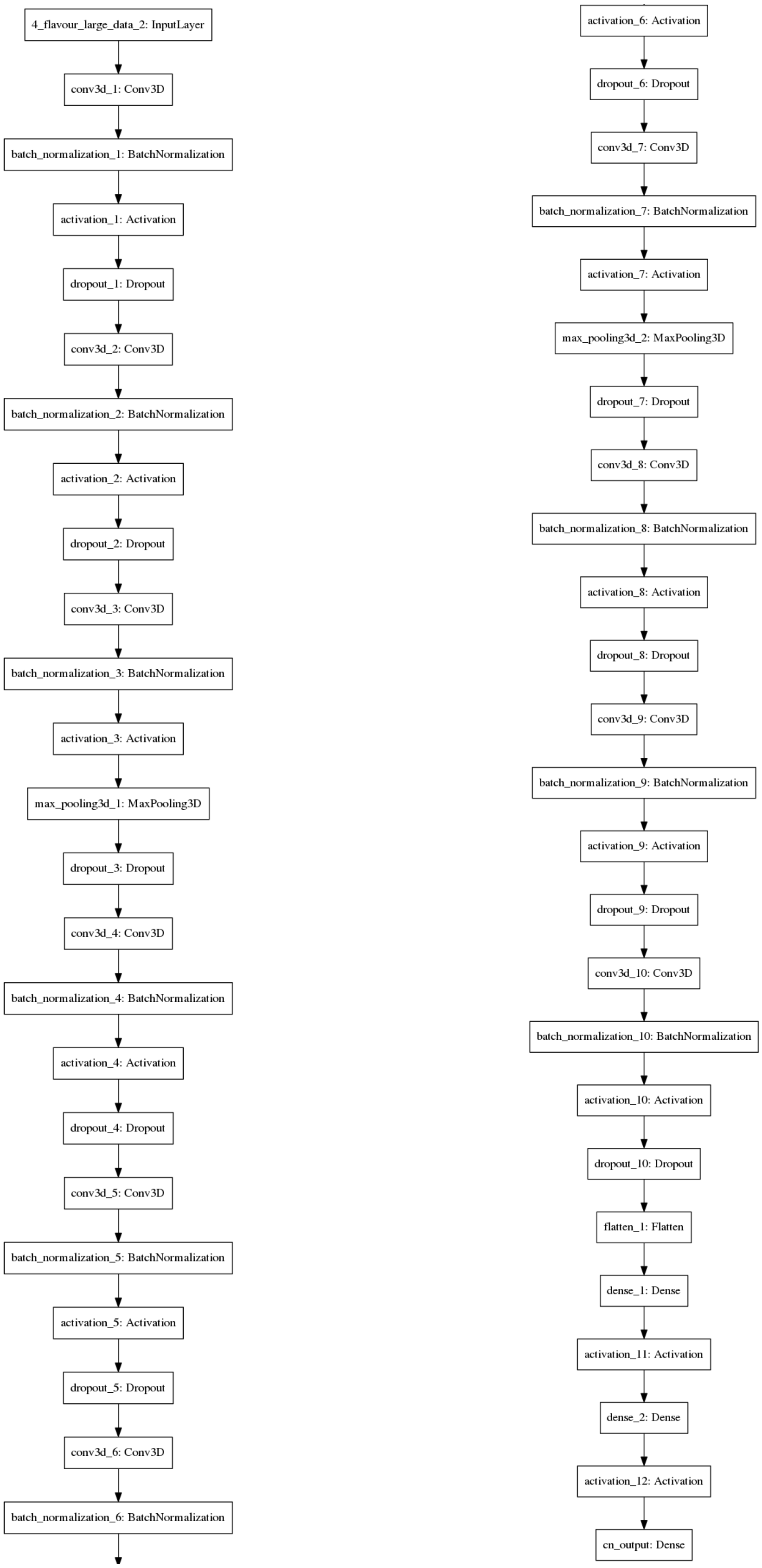


Figure A.2: *Architecture used for the OrcaNet model.*

B Hidden Feature Check

Deep learning training is based in is a continuous evaluation process (backpropagation) and frequent validation, but this process is fully automated and independent. This derives in the so-called black box behaviour, where the understanding of the inner processes is forfeit in exchange for speed and efficiency. This automatic training of the millions of free parameters in your layers means that a particular parameter no longer has any meaning attached to it, so even if we look at the weights of the model, no information can be extracted from there. However, that is not to say that there is no way to try to understand the behaviour or the configuration f the algorithm. The MC information is incredibly useful to probe the response of the algorithm and learn where the training might be failing or how sensible is to the features.

This process is done simultaneously and with the same data visualization tools than the results from chapter 5, but there is an important difference. This results, which can be derived from the ones before, are not meant to reflect anything on the underlying physics or goal but rather on the nature of OrcaNet's model.

A hidden feature is a variable within the data that the model is using for training, but the user is unaware of its existence. These variables are only possible in Deep Learning, where the features are not coded beforehand and the machine acts like a black box. A DL algorithm will use all the information that it finds in the data to improve the performance. A hidden feature might be a new information that emerges as a consequence of the DL study, but in most cases a hidden feature is information about a known variable or parameter that has leaked to the training data due to a mistake. The most obvious example for hidden variable this might be an unknown bias or a bug in the image making process that leaks information information that the model should not have available.

Hidden features might be the signal that causes a nice performance on training, leading to think that the model is working well, while it might fail in a real case scenario. This is one of the most famous errors within the ML field, and especially in optical recognition, where secondary features like the orientation and/or the colour coding are used by the algorithms to drive performance up, instead of the spatial relations.

These features must be suppressed for the DL model to work properly. A hidden features check is good practice whenever a new model or a whole new dataset is being used, to make sure that the model is working properly before starting the optimization. This check can be as simple as, using the dummy data, make sure that the performance only depends on the variables that actually should have an influence, and that is returning a flat spectrum for the rest of variables. The most clear cases can be seen in B.1.

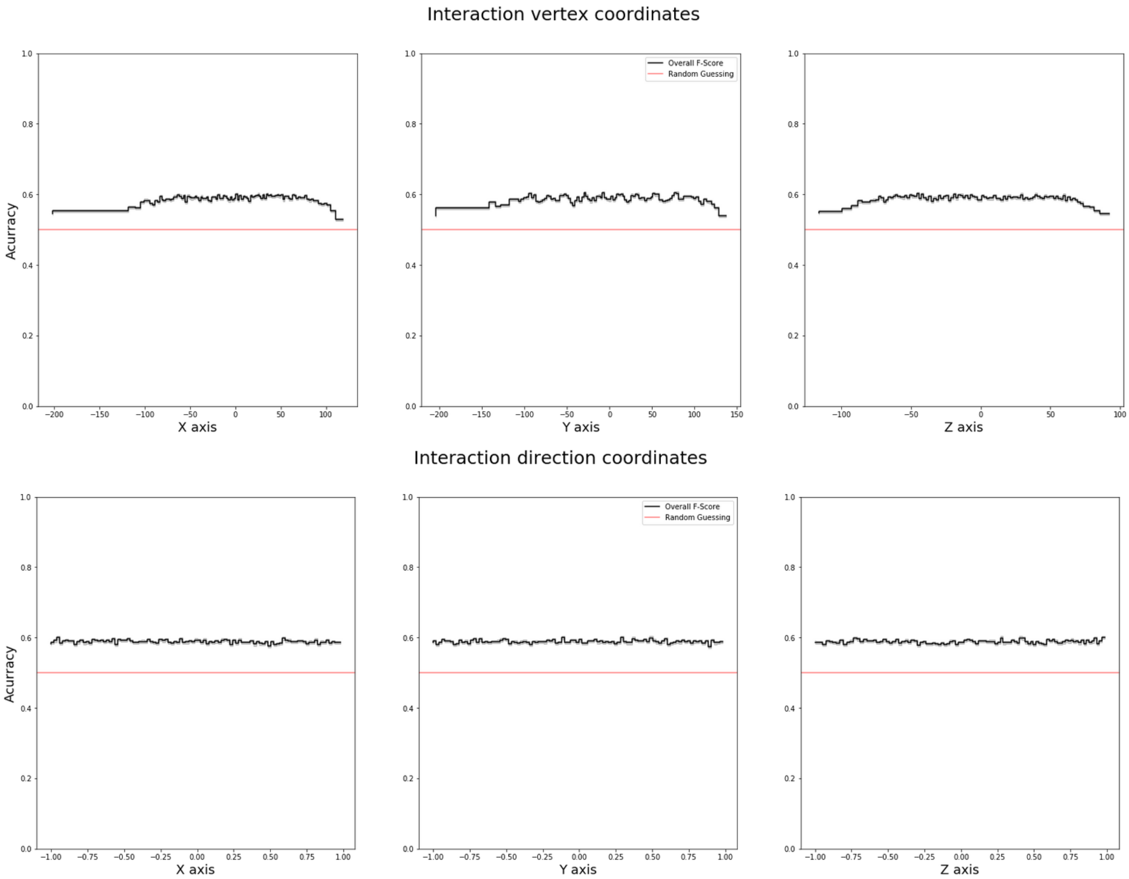


Figure B.1: *Hidden feature check for spatial dimensions. The accuracy metric used was F1 thoughout the testing, the y axis label was included for simplicity when sharing the results.*

We have an homogeneous and isotropic distribution of events in our simulated data, so the corresponding variables (3-dimensional position and direction, respectively) should not contribute to the learning, and thus their performances have to be flat. This is indeed the case, within statistical fluctuations.

The hidden feature checks were required in this work during the first iteration of tests (the electron CC vs NC test shown in 5.18), because of linear correlations that appeared within some of the available variables. This correlations, along with large overfitting (due to the small amount of data used for the tests) led to the hypothesis that the machine was using the total number of hits as a feature, instead of their spatial distribution. The total number of hits might be a hidden feature, and some corrections might have been needed if that were the case.

Of course, there was an event dependency and a overfitting in the hits spectrum, but this was caused by a uneven distribution of the events in the number of hits space. CC and NC events were chosen almost equally in the MC energy distribution, but both times have different average visible energies for the same MC energy B.2.

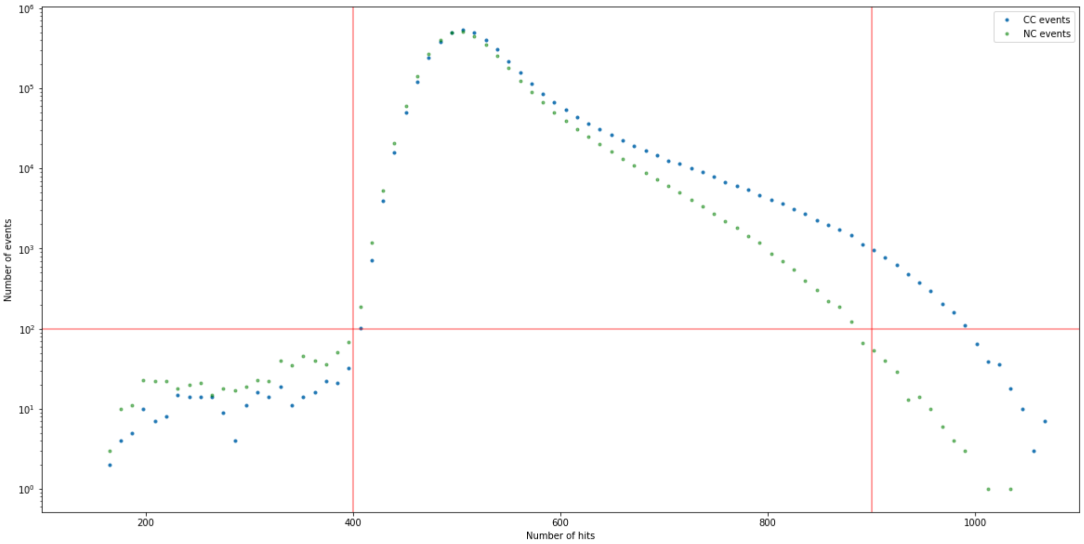


Figure B.2: *Number of events (log scale) vs number of hits for both flavours in the test dataset. The horizontal red line marks the 100 events line, and the vertical red lines, used for cuts, lie on 400 and 900 hits.*

The two distributions are very different on the information available to OrcaNet, and this was the driving factor in the training. This differences were suppressed by eliminating the all the extreme cases with very low statistics, below 100 events (B.2, horizontal red line) and then, suppressing most of the difference in the surviving range (400 to 900 hits) in a single hit bin basis.

After this preprocessing, the model was retrained and checked for hits dependency, seen in B.3.

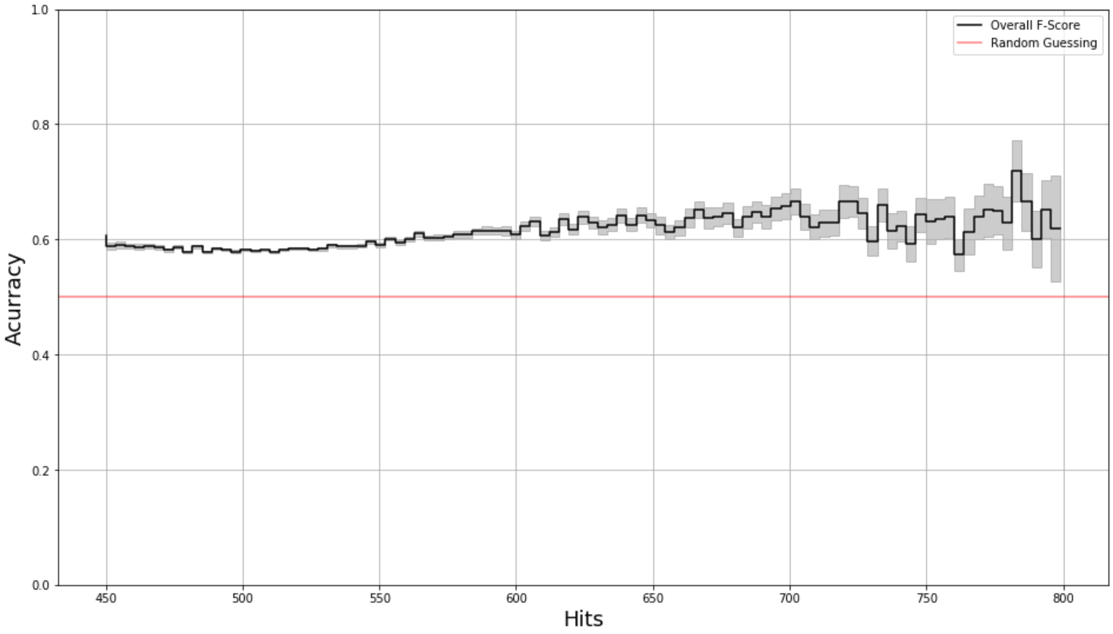


Figure B.3: *F1-Score dependency on the number of hits. The term accuracy (as "accuracy measure") is used here in the y axis for clarity.*

Where the hits dependency has (almost disappeared). Rather than breaking down the model, making the performance disappear, the drop in performance with respect of the not flattened data was was less than 2%. Thus, the model is not looking at the raw number of hits in the images but rather learning from their spatial relations, as expected. All the linear behaviours seen other variables do come from the effect of the physical variable on the original event simulation, reconstructed by OrcaNet.

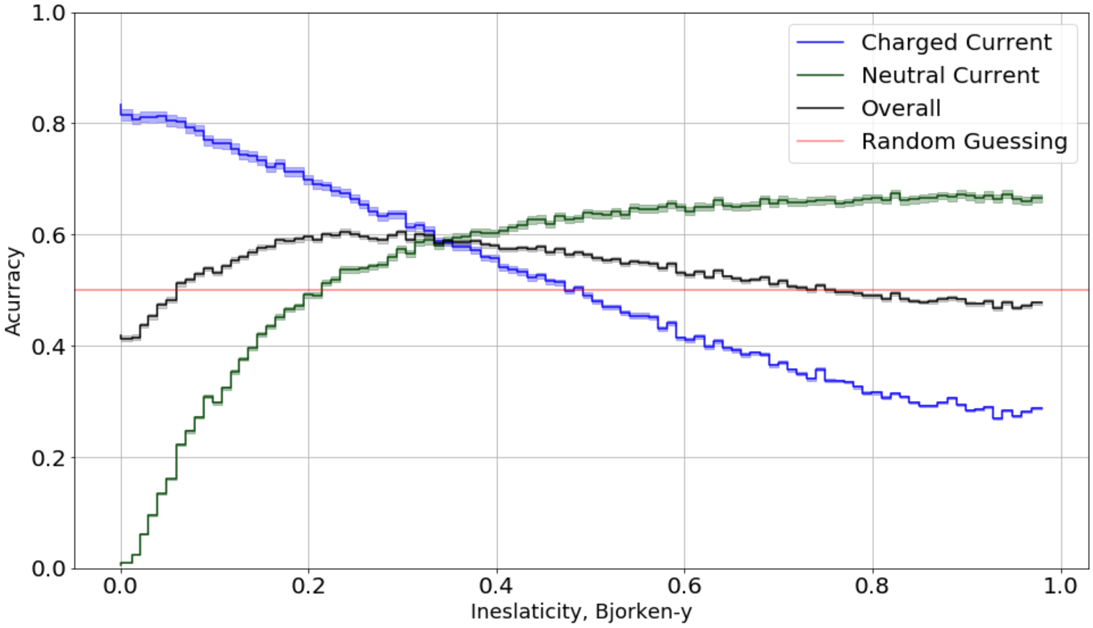


Figure B.4: *F1-Score dependency on the bjorken-y. The term accuracy (as "accuracy measure") is used here in the y axis for clarity.*

The most clear example of this is is the inelasticity, as seen in Fig B.4. The blue line is the reconstructed bjorken-y for the charged-current events, and it follows the expected shape of the real bjorken-y (Bj-y) distribution of the events: when the Bj-y is very low, all the energy goes to the outgoing lepton, so the accuracy of the CC events is the maximum and the NC events is the minimum; while the Bj-y increases, more light goes to the primary shower and more closely related look the two interactions. Beyond Bj-y ~ 0.42, this reconstruction was not able to distinguish between CC events, and for high Bj-y the reconstruction algorithm was almost always wrong categorizing CC events.

Although very useful to understand the data, this extra step is computationally very expensive and it is only needed for when you absolutely need to guarantee the maximum generalization of the model by getting rid of all inhomogenities in the training data. This is only applicable when you are training on all your available data, since more data is the preferred way to improve performance and generalization. Changing the training dataset requires to rerun all the preprocessing steps of flattening your data distributions. In the case of this thesis, the work was not performance oriented (just for the sake of training speed) and the final results do not contain any preprocessing. However, data flattening on your largest dataset and hidden feature check are necessary steps if any Deep Learning model wishes to reach best performance and good generalization.

C Technical Details

C.1 The Pipeline

In order to properly explain all the necessary details, I will start by giving the general overview of how to develop a Deep Learning project using OrcaNet from scratch. For clarity's sake, this won't cover any computational details, I will just explain the necessary steps.

1. **Obtain the raw data**: The KM3NeT collaboration stores its MonteCarlo (MC) simulations of their detectors, and the current and future data recorded by the instruments in the IN2P3 Computing Centre in Lyon, France. While a more accessible version might be available, the long term storage backups of all the data are in the magnetic tape storage system iRODS. A fully simulated event from a production look like this:

```
JTE.KM3Sim.gseagen.elec-CC.3-100GeV-1.1E6-1bin-3.0gspec.
ORCA115_9m_2016.1029.root
```

This naming convention includes the maximum information possible about its contents. In reading order: `JTE` refers to JTriggerEfficiency, the triggering software; `KM3Sim` is the software used for the photon tracking and propagation; `gseagen` is the software used to simulate the underwater neutrino interactions (based on GEANT); `elec-CC.3-100GeV` is the important information for us, given that it tells us the flavour of the interactions in the file and the energy range of the generated neutrinos; `1.1E6` is the starting number of events simulated in the file, the actual number is much less since a lot of neutrinos do not interact; `3.0gspec` refers to the spectral index of the power-law energy distribution chosen for this the events; `ORCA115_9m_2016` refers to the configuration for this production, 115 lines (full detector), 9 meters of vertical spacing (and 23 of horizontal spacing), done in 2016; finally, `1029` is the number of the file withing the production. Of course, .root means that the files are stored in the ROOT dataformat.

Right now, the 2016 production has been superseded by a newer production in 2018 - 2019. This production has improved quality and a better balance between the different flavours (every flavour has 1000 runs), but it has been designed with a different geometry in mind. The horizontal spacing between the strings is reduced to 20 m, allowing for a more dense configuration, but making it not comparable with all the previous results. This is why the 2016 production was used for this thesis.

2. **Convert the data**: The ROOT data format, the standard in high-energy physics, is not compatible with the ML standard, based in Python. The proper data format to store very large size ($\sim$ Gb) data files in a way that is readable by Python is called hdf5 (Hierarchical Data Format), since it has a . The hierarchical component means that, inside of a file, the information is stored in a "folder" or directory system; which means that the typical ROOT substructure system of Tree > Branch(es) > Leaf can be replicated here without the loss of information.

In the KM3NeT collaboration there is a tool ready to convert between the different formats, `tohdf5`, part of the `km3pipe` toolset, based on the JPP and aanet frameworks [48]. This way, we will have the contents of the .root file in the .h5 format, where we can easily visualize it (for example, using km3pipe's `ptdump` command), access and manipulate it from Python scripts with the `h5py` and `pytables` libraries.

3. **Calibrate the data**: The calibration step consists simply in the addition of the x,y,z positions, something that does not come in the production files and is obviusly needed to generate the images. km3pipe's `calibrate` tool is ready for this, and it just needs the corresponding detector file (.detx). OrcaSong, does perform a calibration check and will perform the calibration if it has not been done before, but performing it one beforehand will save time in the subsequent calls from OrcaSong.

Until now, all the steps are necessary, but they only have to be done once. If there is enough local storage available, in the order of 1 Tb, it is strongly recommended that all this preprocessing is done beforehand for all the data in a production since it will guarantee a great reservoir of MC data for any future ML projects. If the work is going to be done in the In2p3 computing centre, where there are storage constraints, not all data might be ready at once.

4. **Produce the images with OrcaSong**: OrcaSong is the library that will turn the descriptive data from the .root (now .h5) files and make the images for the network. Check its own section for details.

5. **Train and test with OrcaNet**: OrcaNet contains all the Machine Learning processes related to training an model and using it to evaluate unknown data (see the OrcaNet section).

6. **Visualize the results**: When the model is done evaluating, we have a final .h5 file which stores in three different datasets the prediction results, the MC truth per event, and all the MC info present in the previous files. All this information is needed in order need to check its performance using statistical estimators, explained in section 5, I have written from scratch a few Jupyter notebooks, available here [51]. This notebooks, also based in Python, take the results from a model at once and compute and plot all relevant results at once. The Jupyter notebook system was chosen for visualization thanks to its interactivity and enormous flexibility to adapt to different use cases and needs.

C.2 OrcaSong 2.0

OrcaSong has another task apart from making the binned images, and it is to extract the relevant MC information about every event from the simulated data file (in ROOT format) and attach it to the image file to be transferred along the to future steps of the analysis. This information should contain all the relevant variables about the event that will be useful for the evaluation process, see sections 5.2 and B. Of course, this information should not be seen at any point be the DL OrcaNet model, who has to learn from the spatial and temporal images.

The first version of OrcaSong, used for this thesis, requires to run a script (`make_nn_ images.py`) and provide a configuration file in the .toml format. This restrained the configuration options (where most of them were just legacy options, not usable anymore) within to the hard-coded values, like selecting only one of the following time cuts:

- **all**: [-350, +850] ns, 1200 ns total ∼ 5 attenuation lengths. 20ns/bin.
- **tight-0**: [-450, +500] ns, 950 ns total ∼ 4 attenuation lengths. 12.8 ns/bin.
- **tight-1**: [-250, 500] ns, 750 ns total ∼ 3 attenuation lengths. 12.5 ns/bin.
- **tight-2**: [-150, 200] ns, 350 ns total ∼ 1.5 attenuation lengths. 5.8 ns/bin.

For the MC information, OrcaSong had a somewhat fixed setting of what variables are saved for every event: The particle type, the energy, the nature of the interaction (charged or neutral), the ineslaticity or bjorken's y, the direction components (x, y, z), the position of the event (interaction vertex x, y, z), the JTE interaction time and time residual, plus some other identification information (event id [aanet's frame index], production id, group id and run id). This seems like a lot of variables, but only the first 4 have been used in the posterior analysis 5.2. At the same time some potentially useful variables, like the amount of signal and total hits, were missing. Some of this is a requirement to ensure the proper working of the model, see appendix B.

To resume, this system is a bit inflexible, since any change to the configuration option need to be hard-coded into your OrcaSong implementation. Changing the label extraction code from OrcaSong (since the simulated data .root file included a lot more of MC variables than the ones extracted bt OrcaSong) was attempted unsuccessfully at the beginning of the project. Adding information from the root file (outside of OrcaSong) is a far from trivial process due to the intense data manipulation, a rigid, structured process with heavy compression and resource use. Only the total (signal + background) number of hits per event was added to the image data file from the respective OrcaSong image. The issue has been fixed since during this last year, and now OrcaSong 2, a more open and flexible version exists to overcome these limitations.

OrcaSong 2.0, the version of OrcaSong currently available is based along the new class FileBinner. It is based on the km3pipe's pipe and blob system, and can be imported into your processing script. In order to work, you only need to define your own spatial and temporal bin edges and the custom function that retrieves the MC information to pass it forward the pipeline. In this case, you have to determine beforehand the bin edges, but there is a script in OrcaSong that helps you computes the bin edges given the number of bins, `geo_binning.py`. For the temporal binning, the binning used in OrcaSong 1.0 can be used as reference. The fantastic benefits of using this new system is that the flexibility opens up the possibility to do optimization searches and the customization of the MC information included with the image in you files. New variables, not present in OrcaSong 1.0 configuration, might be useful for later analysis, like the said number of MC signal hits or the PMT channel (as data not seen by the MC).

The number of signal hits is an interesting example of a addition since it would allow to visualize the correlation between the prediction and the amount of signal hits, quantifying this way the performance of OrcaNet as an offline background discriminator, how good OrcaNet is at ignoring random noise from backgrounds and looking at the relevant data.