New Machine Learning Approach for Pulse Shape Analysis in LEGEND with Feature Importance Supervision

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Katharina Sophie Kilgus

aus Freudenstadt

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	09.07.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Josef Jochum
2. Berichterstatter/-in:	Prof. Dr. Tobias Lachenmaier

Für meine Familie

Mit euch ist
kein Weg zu weit,
kein Sturm zu wild,
und kein Huhn jemals allein.

Abstract

The Large Enriched Germanium Experiment for Neutrinoless $\beta\beta$ Decay (LEGEND) searches for neutrinoless double beta decay ($0\nu\beta\beta$ decay), a process that would provide essential insights into the formation of the universe and the nature of neutrinos. Since the $0\nu\beta\beta$ decay would be an extremely rare decay, high sensitivity is crucial for LEGEND.
To improve this sensitivity and open up new possibilities in terms of data analysis, this work investigates the use of a new Machine Learning approach called Feature Importance Supervision (FIS) for Pulse Shape Analysis (PSA).

FIS builds on the idea of incorporating physical knowledge directly into the Machine Learning model. Using this method can offer several new possibilities, including a higher degree of freedom in the choice of training data and the incorporation of different corrections and cut parameters from the classical PSA in LEGEND.
It is successfully demonstrated that an energy-independent PSA is possible using FIS, even when the training dataset contains events at different energies. This provides more flexibility in the selection of the training dataset. Additionally, three different versions of FIS are investigated, each incorporating a variable quantity of information. Between these versions, physically meaningful differences in the model outputs are observed, all correlated with the given knowledge.

At the current stage, the output of the FIS model carries large uncertainties, which are presumed to arise from impurities in the training dataset. The exact composition of the training data is investigated and can be categorised into several event classes by using and combining different parameters from classical data analysis. This investigation can be used in a continuation of the work to minimise these uncertainties.

Zusammenfassung

Das Large Enriched Germanium Experiment for Neutrinoless $\beta\beta$ Decay (LEGEND) sucht nach dem neutrinolosen doppelten Betazerfall ($0\nu\beta\beta$ Zerfall), einem Prozess, der wesentliche Erkenntnisse über die Entstehung des Universums und die Natur der Neutrinos liefern würde. Da es sich bei $0\nu\beta\beta$ Zerfall um einen extrem seltenen Zerfall handelt, ist eine hohe Sensitivität für LEGEND entscheidend. Um die Sensitivität zu verbessern und neue Möglichkeiten der Datenanalyse zu eröffnen, wird in dieser Arbeit die Verwendung eines neuen Ansatzes des maschinellen Lernens namens Feature Importance Supervision (FIS) für die Pulsformanalyse (PSA) untersucht.

FIS basiert auf der Idee, physikalisches Wissen direkt in das Modell des maschinellen Lernens einzubeziehen. Die Verwendung dieser Methode bietet mehrere neue Möglichkeiten, darunter eine größere Freiheit bei der Auswahl der Trainingsdaten und die Einbeziehung verschiedener Korrekturen und Schnittparameter aus der klassischen PSA in LEGEND. Es wird erfolgreich demonstriert, dass eine energieunabhängige PSA mit FIS möglich ist, selbst wenn der Trainingsdatensatz Ereignisse mit unterschiedlichen Energien enthält. Dies bietet mehr Flexibilität bei der Auswahl des Trainingsdatensatzes. Zusätzlich werden drei verschiedene Versionen von FIS untersucht, die jeweils eine unterschiedliche Menge an Informationen enthalten. Zwischen diesen Versionen werden physikalisch bedeutsame Unterschiede in den Modellergebnissen beobachtet, die alle mit dem gegebenen Wissen korrelieren.

Zum gegenwärtigen Zeitpunkt ist die Ausgabe des FIS-Modells mit großen Unsicherheiten behaftet, die vermutlich auf Unreinheiten im Trainingsdatensatz zurückzuführen sind. Die genaue Zusammensetzung der Trainingsdaten wird untersucht und kann durch Verwendung und Kombination verschiedener Parameter aus der klassischen Datenanalyse in mehrere Ereignisklassen eingeteilt werden. Diese Untersuchung kann in einer Fortsetzung der Arbeit genutzt werden, um diese Unsicherheiten zu minimieren.

Contents

1. Introduction

The search for Neutrinoless double beta decay ($0\nu\beta\beta$ decay) plays a central role in experimental particle physics. Such a decay would not only allow fundamental conclusions to be drawn about the nature of neutrinos, but would also help to explain the observed matter-antimatter asymmetry in the universe. Large Enriched Germanium Experiment for Neutrinoless $\beta\beta$ Decay (LEGEND) aims to significantly increase the sensitivity to this rare decay.
One of the central challenges of LEGEND is the effective suppression of the background. An essential part of this is Pulse Shape Analysis (PSA), with which signal-like events can be distinguished from background events.
This thesis investigates the application of modern Machine Learning (ML) methods to pulse shape discrimination. In general, the aim of this ML methods is to achieve higher automation and improved adaptability for future phases such as LEGEND 1000 (L1000). Since every new experiment will come with an increase of the number of germanium detectors, it is helpful if as little manual tuning as possible is required. Furthermore, an additional, stable and well performing cut can serve as an additional cross-check or offer new possibilities in terms of further detector development.

The Feature Importance Supervision (FIS) method presented in this work makes it possible to incorporate prior physical knowledge into ML models. On the one hand, this allows a freer choice of training data, without introducing an energy dependency into the model. On the other hand, it can open up the possibility to combine multiple analysis cuts and corrections by giving the model the underlying information of the non-ML analysis parameter.

After the adaption of FIS to Germanium pulses, the resulting model allows more flexibility in the choice of training data, but still returns to be energy independent.
In addition, a proof-of-concept for the incorporation of detailed PSA knowledge is presented. Depending on the physical knowledge given to the model, clear differences in

the results were observed, which in turn can be explained by the information provided. Initial tests with a method based on the discrimination of γ radiation background provide promising indications that this approach can, in principle, also be transferred to other pulse forms.
However, the results to date suggest that the training data currently used has too high a contamination with non-pure events. The consequences of this observation and possible strategies for improving data quality are discussed in conclusion.

Following this outline, Chapter 2.2 provides a short overview of neutrinos in general and the open research questions in this field, before focusing on the $0\nu\beta\beta$ decay decay and its potential impact on physics if observed. Chapter 3 then explains LEGEND in more detail, especially the operating principles of germanium detectors and the already existing data analysis. As another component, Chapter 4 briefly introduces the necessary ML basics used throughout the thesis.
Starting with an already established model, the first part of this work begins with Chapter 5, which addresses the adaptation of the Artificial Neural Network (ANN) from Germanium Detector Array (GERDA) to LEGEND 200 (L200). In Chapter 6, a new method for PSA, FIS, is proposed and investigated. Chapter 7 then examines the current challenges of FIS, their underlying reasons, and possible solutions.

2. Majorana Particles and Neutrinoless double beta decay

The search for the $0\nu\beta\beta$ decay is a highly interesting part of the field of physics beyond the Standard Model. According to the Standard Model, neutrinos are leptons with three different flavours divided into three generations, ν_e, ν_μ and ν_τ. According to the Standard Model, they are massless particles without any electric charge and are only subject to weak interaction. Through weak interaction, each neutrino can interact with leptons of its own flavour: e^-, μ^- or τ^-.
Since neutrinos are massless in the Standard Model, their helicity is left-handed for neutrinos and right-handed for antineutrinos [1] [2] [3].
However, current research shows observations with significant discrepancies compared to the Standard Model, like a neutrino mass larger than zero. This discrepancies are described in Section 2.1, especially with regard to the reasons why the $0\nu\beta\beta$ decay is being researched. The second part of this chapter (Section 2.2) provides a brief overview of the $0\nu\beta\beta$ decay and the current experimental approaches to search for it.

2.1. Neutrinos Beyond the Standard Model

Research in neutrino physics is generally part of Physics Beyond the Standard Model, since observations like neutrino oszillation or neutrino mass are not in agreement with the Standard Model [4] [5] [6]. This section will focus on the two main points that explain why $0\nu\beta\beta$ decay is such an intriguing area of research.

2.1.1. Lepton Number Violation

In the standard model, all particles are included together with an antiparticle, including the neutrino. The special feature of the neutrino, however, is its electric charge, which is zero. Since an antiparticle is defined as the same particle, but with opposite charge, it is not possible to find the difference for neutrinos with this definition. Neutrinos can

be distinguished from antineutrinos by the rules of the underlying particle interaction or through their helicity [3]. But there is no difference in electric charge, as it is usually the case for antimatter. As a result, it is possible that both particle and antiparticle are actually the same. The neutrino is the only candidate in the Standard Model for such a particle, which is called a **Majorana particle**.
Such a Majorana particle would violate the concept of lepton number conservation, as explained lateron for the $0\nu\beta\beta$ decay (see Section 2.2). This lepton number violation would offer a possible explanation for the matter–antimatter asymmetry observed in the universe [7]. By proving the existence of Majorana particles, the concept of lepton number violation would be shown.

2.1.2. Neutrino Mass

Neutrino Oscillation

One of the first hints the Standard Model does not treat neutrinos in a realistic manner was the observation of neutrino oscillation. Neutrino oscillation is only possible if neutrinos have mass, whereas the Standard Model predicts them to be massless particles. The first evidence for neutrino oscillation was a deficit of solar e^- neutrinos measured in the Homestake Experiment in the 1960s.

Neutrinos are produced and detected in their flavour eigenstates $(\nu_e, \nu_\mu, \nu_\tau)$. These flavour eigenstates are superpositions of their mass eigenstates (ν_1, ν_2, ν_3), each of which has a distinct mass. The mass eigenstates propagate through space with different velocities due to their different masses [8, 9, 10]. These differences in velocity cause phase shifts in interference patterns. As a result, when neutrinos are detected, their flavour eigenstate may differ from the one in which they were produced.

This observation implies that neutrinos must have mass. Since the mass of neutrinos is very small, there currently only exists an upper limit of $m_\nu < 0.45\text{eV}$ for the direct neutrino-mass search [11]. In addition to efforts to measure the neutrino mass directly, various experiments attempt to determine sub-elements of this question. For example, it is not known which neutrino mass eigenstate is the lightest, since it is just known that $\nu_1 < \nu_2$, but nothing about the order of ν_1 and ν_3 [12]. It is also possible to determine limits for the sum of neutrinos using cosmological observations [13]. These starting points are approaches to the problem from different sides with different methods and help to improve neutrino theories and their predictions.

Masses of other particles in the standard model can be determined by the Higgs mechanism [14] and are in the range of MeV. Since the neutrino mass is at least for one mass eigenstate smaller than 1 eV, it cannot be explained by the Higgs mechanism and therefore remains a puzzle. If neutrinos are Majorana particles, they can follow the see-saw mechanism, which can account for their small mass.

See-Saw Mechanism

The see-saw mechanism is a theoretical framework aimed at explaining the extremely small masses of neutrinos compared to other fundamental particles. It is based on the idea that a significant mass difference can arise between light and heavy neutrino states due to interactions involving very large mass scales, in the context of the Majorana nature of neutrinos.

The see-saw mechanism proposes that neutrino masses result from the coupling between the light left-handed neutrinos from the Standard Model and new, heavy right-handed neutrinos. This coupling is described by the neutrino mass matrix, which mixes both types of neutrinos. In the simplest version, known as the Type-I see-saw, the neutrino mass matrix takes the form:

$$M = \begin{pmatrix} 0 & m_D \\ m_D & M_R \end{pmatrix} \tag{2.1}$$

Here, m_D represents the Dirac mass term (similar to the mass terms for charged leptons and quarks) that couples the left-handed and right-handed neutrinos, while M_R is a large mass term representing the heavy right-handed Majorana neutrino masses. When diagonalised, this mass matrix yields two eigenstates: a very light neutrino mass m_ν approximately given by:

$$m_{\nu,L} \approx \frac{m_D^2}{M_R} \ll m_D \quad m_{\nu,R} \approx M_R \gg m_D \tag{2.2}$$

and a very heavy mass state corresponding to the right-handed neutrino. Because M_R is assumed to be extremely large (typically on the order of 10^9–10^{15} GeV), the resulting light neutrino masses m_ν are naturally small, even if the Dirac masses m_D are comparable to the masses of other fermions such as the charged leptons.

With this mechanism, the existence of Majorana particles can explain why neutrinos have such small masses relative to other particles by introducing new physics, balancing the mass of the light neutrinos we observe with the hypothesised heavy right-handed neutrinos. The hypothetical right-handed neutrinos would carry the mass but are rarely

detected. This is because neutrinos are still produced as left-handed and then need to flip their helicity. This helicity flip is directly correlated with the mass and is highly suppressed for low masses. This can easily explain why no right-handed neutrinos have been observed yet.

2.2. Neutrinoless Double Beta Decay ($0\nu\beta\beta$)

Currently, one of the most promising ways to identify Majorana particles is the $0\nu\beta\beta$ decay. It is a hypothetical variant of the two neutrino double beta decay ($2\nu\beta\beta$ decay):

$$(A, Z) \rightarrow (A, Z+2) + e^- + e^- + \overline{\nu}_e + \overline{\nu}_e \tag{2.3}$$

The $0\nu\beta\beta$ decay is a rare decay occurring in certain isotopes, where a single beta decay is energetically forbidden, but a simultaneous double beta decay is allowed. This is illustrated in the left panel of Figure 2.1, which shows the mass excess for a constant mass number $A = 76$ as a function of the atomic number Z. Two parabolas are formed: the higher one corresponds to odd-odd nuclei (odd numbers of both protons and neutrons), and the lower one corresponds to even-even nuclei. Since a single β decay can only proceed toward a lower mass excess, the decay from ^{76}Ge to ^{76}As is forbidden, while the double beta decay to ^{76}Se is allowed.
In the case of a $2\nu\beta\beta$ decay, two antineutrinos and two electrons are typically emitted. If neutrinos were Majorana particles – i.e., if neutrino and antineutrino were the same particle – then a light neutrino exchange could occur.
This would lead to a $0\nu\beta\beta$ decay with the decay scheme:

$$(A, Z) \rightarrow (A, Z+2) + e^- + e^-. \tag{2.4}$$

Instead of a lepton number $\Delta L = 0$, as in the case of $2\nu\beta\beta$ decay, a $0\nu\beta\beta$ decay leads to $\Delta L = 2$.

In total, 35 isotopes are known to decay via $2\nu\beta\beta$ decay, all of them with a half-life-time $T_{1/2}^{2\nu\beta\beta} \geq 10^{18}$. (e.g. in ^{76}Ge, is $T_{1/2}^{2\nu\beta\beta} = (1.926 \pm 0.094) \cdot 10^{21}$ years [16]). In all cases, an observed $0\nu\beta\beta$ decay would result in a spectrum as shown schematically in the right panel of Figure 2.1. It consists of a $2\nu\beta\beta$ decay spectrum and a single peak at the Q-value of the $2\nu\beta\beta$ decay spectrum, originated by the $0\nu\beta\beta$ decay. This is because the Q-value is the endpoint of a β-decay spectrum, where no energy is carried away by the

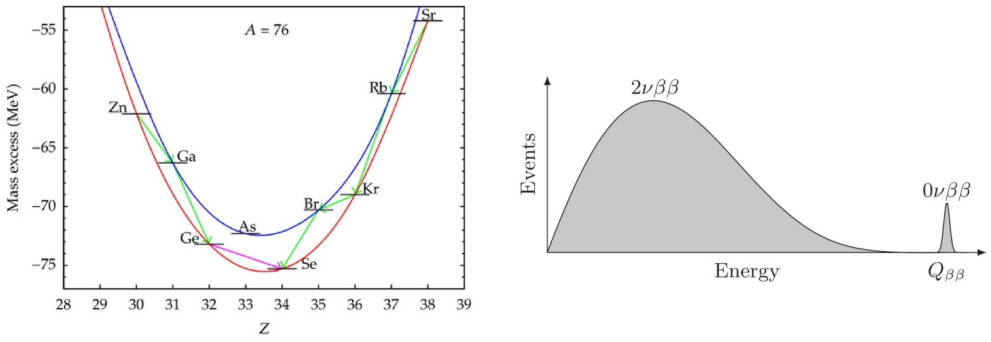


Figure 2.1.: On the left, the condition for a $2\nu\beta\beta$ decay instead of a single β-decay is graphically shown for the example of ^{76}Ge. Since the mass excess of ^{76}As is higher than for ^{76}Ge, a single β decay is forbidden, leading directly to a $2\nu\beta\beta$ decay. The figure on the right schematically illustrates the expected spectrum containing both $2\nu\beta\beta$ decay and $0\nu\beta\beta$ decay. The $0\nu\beta\beta$ decay peak is exaggerated to demonstrate the principle of the measurement. At the current state of research, observing even a few events would be a major breakthrough. (Graphics from [15])

neutrinos. In $0\nu\beta\beta$ decay, all the energy is carried by the produced electrons. Due to this reason, it will produce a peak directly at $Q_{\beta\beta}$.

2.2.1. Experimental Search for $0\nu\beta\beta$

Alongside LEGEND, there are several experiments searching for $0\nu\beta\beta$ decay, employing different techniques and isotopes. Each of them has its own advantages and disadvantages, which will be described in the following section.
Table 2.1 gives an overview about current possible detection techniques and some exemplary experiments for each isotope. When choosing an isotope, a higher Q-value results in lower background, while a higher natural abundance leads to lower costs. The choice of isotope also determines the possibilities of the detector technology.

Time Projection Chambers

One option are **Time Projection Chambers** (TPCs), which use an electric field to separate electrons and other charged particles. As the particles travel through the detector, they produce primary ionisation along their tracks. This allows for 3D event reconstruction. Additionally, it is possible to build large detectors. A common challenge is maintaining the purity of the gas used, typically xenon.

One example of this measurement technique is the Enriched Xenon Observatory (EXO) using ^{136}Xe, which reported a limit of $T_{1/2}^{0\nu\beta\beta,\mathrm{Xe}} > 3.5 \cdot 10^{25}$ yr at 90% confident level from EXO-200 in 2019 [17].

Scintillators

The use of a **scintillator** offers the possibility of event reconstruction. It is necessary to distinguish between scintillator crystals and liquid scintillators doped with a specific isotope. In the latter case, doping is challenging, but it allows for the construction of larger detectors than with pure scintillator crystals. In both cases, the amount of enriched isotopes can remain the same, but the larger volume of the doped scintillator allows for more techniques such as fiducial cuts and provides some self-shielding effects. Since most scintillator experiments use ^{136}Xe, leading to some overlap with the results from TPC experiments, the highest current limit for $0\nu\beta\beta$ decay using a liquid scintillator doped with enriched xenon was published by KamLAND-Zen in 2022. It stands at $T_{1/2}^{0\nu\beta\beta,\mathrm{Xe}} > 2.3 \times 10^{26}$ yr at 90% confident level for ^{136}Xe [18].

Cryogenic Bolometers

Since **cryogenic bolometers** measure the temperature increase resulting from particle interactions, they require very low operating temperatures in the range of a few millikelvin. Background reduction is also challenging, as the detectors are sensitive to any kind of radiation. On the other hand, they offer high energy resolution and good sensitivity.
One experiment using this technique with ^{130}Te is CUORE, which published its current limit in 2022: $T_{1/2}^{0\nu\beta\beta,\mathrm{Te}} > 2.2 \cdot 10^{25}$ yr [19]. Using the same method but equiped with ^{100}Mo is AMoRE, which reported a limit of $T_{1/2}^{0\nu\beta\beta,\mathrm{Mo}} > 2.9 \cdot 10^{24}$ yr at 90% confident level in 2025 [20].

Tracking Calorimeter

A major advantage of a **tracking calorimeter** is that it can be used with any isotope. It utilises a source foil, with a wire-chamber tracker positioned on both sides of the foil. These tracking layers allow for precise track reconstruction of radiation through the detector. Behind the tracking layers, calorimeters are placed to accurately measure the energy of the particles. Due to its complex structure, however, scaling up to larger detector volumes is challenging.
This method is currently used only in NEMO [21] with ^{82}Se on a relatively small scale and serves primarily as a proof of concept for future experiments.

Semiconductor Detectors

In **semiconductor detectors**, ionizing radiation produce a current by generating

Isotope	**Daughter**	$Q_{\beta\beta}$ **[keV]**	f_{nat} **[%]**	**Measurement Technique (Examples)**
^{48}Ca	^{48}Ti	4267.98(32)	0.187(21)	Scintillator Crystals (CANDLES-III)
^{76}Ge	^{76}Se	2039.061(7)	7.75(12)	Semiconductor (LEGEND)
^{82}Se	^{82}Kr	2997.9(3)	8.82(15)	Tracking Calorimeter (SuperNEMO), Bolometers (CUPID0), TPC (IFC)
^{96}Zr	^{96}Mo	3356.097(86)	2.80(2)	Tracking Calorimeter (NEMO-3), Liquid Scintillator (ZICOS)
^{100}Mo	^{100}Ru	3034.40(17)	9.744(65)	Bolometers (AMoRE, CUPID)
^{116}Cd	^{116}Sn	2813.50(13)	7.512(54)	Scintillator Crystals, CdZnTe Semiconductor (COBRA)
^{130}Te	^{130}Xe	2527.518(13)	34.08(62)	Bolometers (CUORE), Loaded Scintillator (SNO+)
^{136}Xe	^{136}Ba	2457.83(37)	8.857(72)	Loaded Scintillator (KamLAND-Zen), TPC (NEXT, EXO)
^{150}Nd	^{150}Sm	3371.38(20)	5.638(28)	Tracking Calorimeter (NEMO-3)

Table 2.1.: Overview of isotopes undergoing $2\nu\beta\beta$ decay decay, including natural abundances, Q-values, and possible measurement techniques with exemplary experiments. [15, 22, 23]

electron–hole pairs in a depletion region. One of their main advantages is their excellent energy resolution, along with the ability to perform pulse shape analysis. A limiting factor, however, is the relatively small detector mass due to the limits of crystal growth and the fact that they need to be fully depleted.

These detectors are currently used in LEGEND and will be described in more detail in the next chapter, as this work was carried out as part of the LEGEND experiment.

From the GERDA experiment, the current limit with semiconductors for $0\nu\beta\beta$ decay in ^{76}Ge is $T_{1/2}^{0\nu\beta\beta,\text{Ge}} > 1.8 \cdot 10^{26}$ yr at 90% confident level ([16]).

3. Experimental approach from LEGEND

One of the largest experiments searching for $0\nu\beta\beta$ decay is the LEGEND experiment. Precursor of LEGEND are the MAJORANA Demonstrator and the GERDA, currently running is L200, while L1000 is already planed for the future. Both experiments collected experience in searching for $0\nu\beta\beta$ decay by using high purity High Purity Germanium Detector (HPGe)s enriched with ^{76}Ge.
The following section will describe the general idea behind the use of Germanium for the search of $0\nu\beta\beta$ decay, the challenges of such a setup and the possibilities of L200 (Section 3.1). Section 3.2 contain information about the full experimental apparatus of L200, the operation mode and the previous characterization campaigns for the HPGes. The last section explains the HPGes in greater detail and the therefore resulting possibilities for data analysis (Section 3.4.2).

3.1. General concept of LEGEND

Fundament of the LEGEND experiment is the choice of ^{76}Ge as the isotope to use for the search for $0\nu\beta\beta$ decay decay. The following section will focus on the general strategy, challenges and possibilities for the $0\nu\beta\beta$ search in general and by using ^{76}Ge in particular.

The long-term plan of the collaboration is to build L1000 with 1t of detector mass. But in the current phase, LEGEND is build as L200 with a detector mass of 200 kg. This will probe the $0\nu\beta\beta$ decay decay of ^{76}Ge with a sensitivity to a half-life time of $T^{0\nu}_{1/2} > 10^{27}$ yr at 90% confidence level (C.L.) within 5 years of measurement time.
This sensitivity can be calculated by

$$T^{0\nu}_{1/2} \propto \sqrt{\frac{m \cdot t}{BI \cdot \Delta E}} \cdot f_{76} \cdot f_{av} \cdot \epsilon \tag{3.1}$$

with the detector mass m, measurement time t, background index BI, energy resolution ΔE, the detection efficiency ϵ and the isotopic abundance of ^{76}Ge f_{76}. [23] In a background-free regime (less than one event expected in the region of interest (1 FWHM)), the sensitivity changes to

$$T_{1/2}^{0\nu} \propto m \cdot t \cdot f_{76} \cdot f_{av} \cdot \epsilon \tag{3.2}$$

With this equations, the sensitivity can be increased by a long measurement time, a high detector mass and a low background index.
The benefit of this choice is a high energy resolution (0.12% FWHM (0.05% σ) at Q value of $2\nu\beta\beta$ ($Q_{\beta\beta}$)[24]) and a good timing determination, as it is typical for HPGes. On the negative side, the current limit in detector mass for a single HPGe is 4 kg, since they need to be depleted. With the current techniques, this is not possible with heavier detectors. Due to this behaviour, the main strategy of LEGEND is a quasi-free background search. To reach this low background, it is necessary to deal with cosmic background as well as with radiation emanating from sources close to the detector like cables, cosmogenic isotopes or contaminations on the detector surface.
Therefore, it is critical to reduce the background index, which is done by the application of several kinds of shielding. To shield the experiment against cosmic rays, it is build in the Laboratorio Nationale de Gran Sasso (LNGS) at a depth of 3600m water equivalent. Also LEGEND features two active veto systems, a water cherenkov detector and a Liquid Argon (LAr) veto.

3.2. Overview over the L200 Setup

Beside the HPGes, LEGEND consists of multiple components and subsystems to suppress background and guarantee a high sensitivity of the full setup. The full setup is described in the following, complemented by information about the calibration runs during the operation time (3.2.1) and the characterization campaigns before the commissioning and operation of L200 (3.2.2).

The first shielding is the rock overburden, which passively reduce the flux of muons to 1.25/(m^{2}h) and eliminate the hadronic components of cosmic rays. [25]

Muon Veto

To get rid of the remaining muon flux, the LEGEND experiment features an outer veto known as the Water-Cherenkov-Muon-Veto, housed within a 10m diameter water

tank. The interior of this tank is lined with VM2000, a highly reflective foil with 99.9% reflectivity, and is equipped with 55 Photo Multiplier Tubes (PMTs) placed on the walls, floor, and in the so-called pillbox directly under the cryostat. These PMTs are designed to detect Cherenkov light generated by cosmic muons in the water. A comprehensive description of the muon veto system of GERDA, which is used further in LEGEND-200, is detailed in [26].

LAr Veto

Nested within the muon veto's water tank is a 4m diameter cryostat filled with Liquid Argon (LAr). The primary purpose of the LAr is to maintain the HPGe detectors at their operational temperature of approximately 200K. Additionally, the scintillating characteristics of LAr makes it an effective material for a secondary veto detector, capable of identifying other surrounding particles than muons, particularly gamma background radiation. Therefore, in the center of the cryostat is a 1374 mm-diameter and 3 m-height cylinder of a WaveLength Shifting Reflector (WLSR), which reflects the scintillation light, produced by particle interactions in the LAr. This scintillation light is further collected by two concentric wavelength shifting fiber shrouds, which surround the Germanium detector strings, and detected by Silicon Photomultiplier (SiPM) on both ends of the fibers. Further information about the LAr veto system can be found in [24], [27].

Germanium Array

The core of the LEGEND 200 experiment are the Germanium arrays, which are the principal detectors for observing neutrinoless double beta decay. These arrays are arranged in 11 strings, as illustrated in figure 3.2. The HPGe are enriched with ^{76}Ge to a level of at least 86% for the detectors reused from GERDA and MAJORANA Demonstrator (MAJORANA), and to a level of $> 92\%$ for all new detectors [24]. With this enrichement, they are the key to identifying potential $0\nu\beta\beta$ decay events in the LEGEND experiment. HPGes are in general able to detect different types of radiation, if it deposit energy inside the active volume of the detector by using the characteristics of Germanium as a semi-conductor. The operational nuances of the HPGe detectors are discussed in detail in section 3.3. Due to the enrichment, the potential source of $0\nu\beta\beta$ decay is part of the detector and can be directly measured.
A critical aspect of the HPGe setup is their free mounting within the LAr veto without any encapsulation around the detectors, minimizing the presence of impure materials and thus reducing potential sources of background radiation. Additionally, the Germanium array facilitates anti-coincidence analysis across different detectors.
To guarantee the best possible data analysis of the HPGes, characterization campaigns

were run before the operation of L200. Also calibration runs are done every week during the operation of L200.

3.2.1. Operation and Calibration procedure in L200

The LEGEND experiment's measurements are divided into distinct periods, while modifications are possibly between two periods. Each period consists of several runs, with each run representing a week of measurement time.
After two periods of commissioning, data taking started in March 2023 with a detector mass of 142 kg total. This lead to 48 weeks of accepted measurement time as released at Neutrino2024.[29]

To get a reference point for the data analysis, calibration runs of each subsystem are performed every week, since a well known detector and steady adjustments are crucial for a precise analysis.
In case of the HPGes, a good statistical base is needed to monitor energy scale and resolution or set cut values for PSA in the low background runs. Therefore, the calibration runs are performed with ^{228}Th sources and the analysis for the physical runs can be prepared by using the resulting spectrum as shown in 3.6 and described in detail in 3.4.2.
To perform the calibration runs, 16 ^{228}Th sources can be lowered along 4 even distributed source insertion systems (SIS) inside the detector parallel to the germanium strings, with an average Activity of 4.3kBq (Feb. 2021). [30]
Between period 3 to 11, SIS 4 was disabled due to hardware issues, resulting in missing calibration sources between string 1 and 2. This impacts the cut performance of the surrounded detectors, including most of the Semi-coaxial detector (Coax) (see Figure 3.2).

3.2.2. Characterization at HADES

Prior to the commissioning of L200, the HPGes underwent a thorough characterization process at the High Activity Disposal Experimental Site (HADES) underground laboratory in Mol, Belgium or at Oak Ridge National Laboratory (ORNL), USA. Goal of this characterization campaigns are to test the performance of all detectors and find the best operation conditions.
Since this work used data from the measurements at HADES, just the details about this

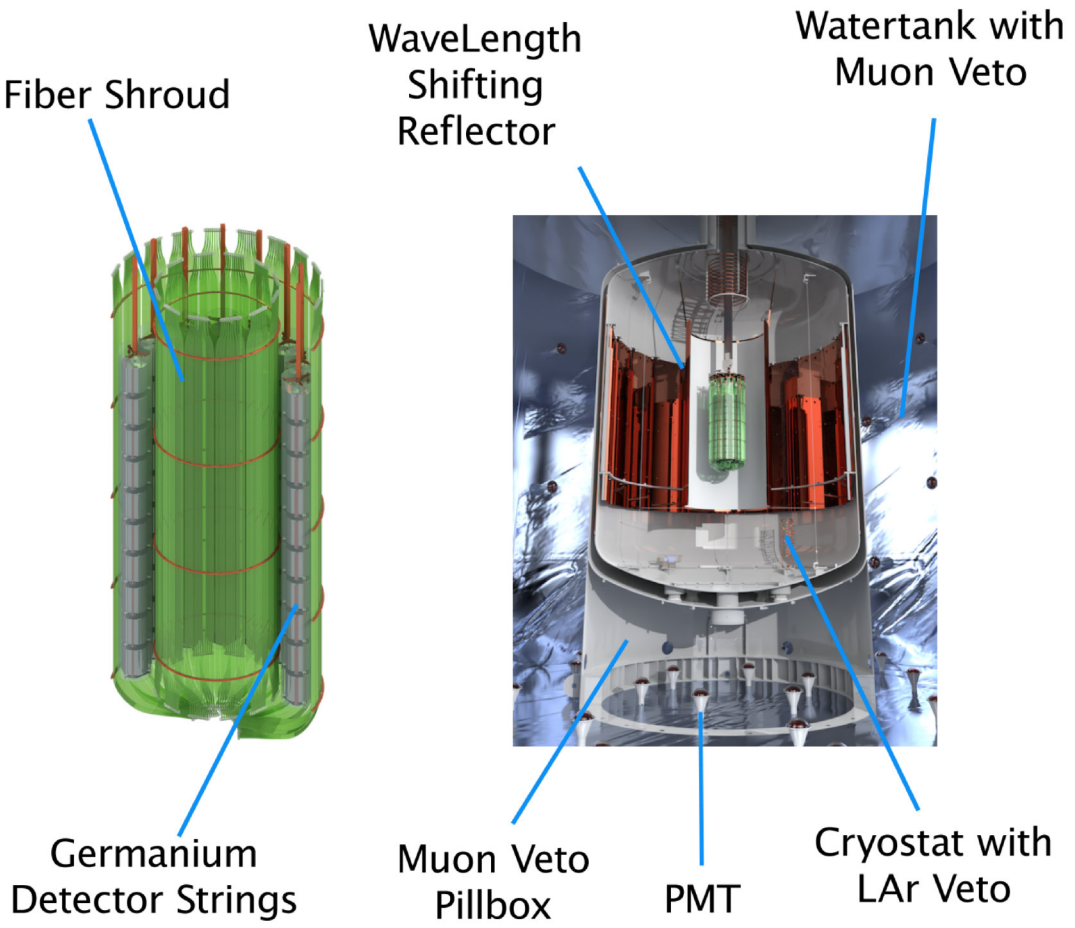


Figure 3.1.: On the left, just the germanium string array surrounded by the wavelength shifting fibers of the LAr Veto can be seen. On the right, the full L200 setup is drawn, including the outer watertank and muon veto, the cyrostat and LAr Veto and inside the Ge strings. (From [24])

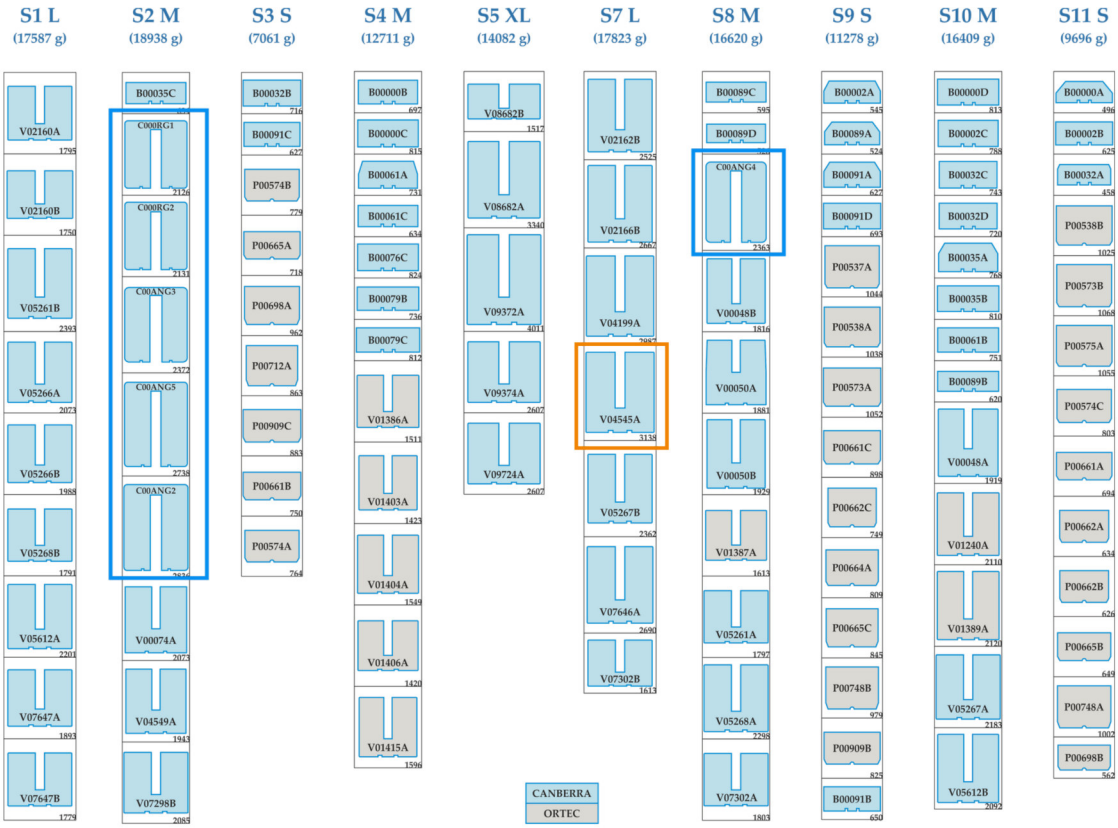


Figure 3.2.: Map of Detector strings, also showing the different types of detectors used in LEGEND, as they are described in 3.3 (From [28]). The detectors used in this work are marked blue (Coax) and orange (ICPC).

part of the characterisation campaign are explained in more detail.

The HADES facility, located 225 meters underground (corresponding to 500m.w.e.), is in proximity to Mirion Technologies, a manufacturer of some HPGe components. This allows detectors to be reworked in the event of anomalies without excessive transport distances, which would not be possible if characterisation were carried out directly at the LGNS. During the characterization phase, tests were conducted using ^{241}Am, ^{60}Co, ^{133}Ba, and ^{228}Th sources. These tests assessed the performance of the later used analysis parameter as well as different detector characteristics like the best operating voltage. [31]).
The scans were done for the new Inverted coaxial p-Type point contact detector (ICPC) detectors, but also for some Broad Energy High-purity Germanium Detector (BEGe) and one Coax from GERDA. The different type of detectores are described further in section 3.3.2.

To investigate the PSA performance by A/E (see section 3.4.2), a ^{228}Th source with an activation of 87 kBq (2013-11-01) was used. Also for the development of a new, ML-based PSA technique, this scans are of importance. For every detector, a scan with the source on top as well as a latitudinal position of the ^{228}Th source was performed. Normaly, the measurement time for the latitudinal scans was 17h and for the top scans 5h respectively. In both cases, the setup was surrounded by a lead castle of 16cm thickness to shield against background radiation.

3.3. High Purity Germanium Detectors

LEGEND, following the legacy of GERDA and MAJORANA, utilizes HPGe, which will be described in the following. These semiconductor diode detectors offer a significant advantage in terms of high energy resolution, attributed to the large number of charge carriers generated per energy pulse.
In addition, there are different types of detectors in LEGEND (Section 3.3.2). Also the reaction of the detector to different event types is described in Subsection 3.3.3.

3.3.1. HPGe operating mode and Signal Formation

Since Germanium is a semiconductor, meaning it has an energy gap between its valence and conduction bands. Unlike in the use of an insulator, this gap can be overcome if sufficient external energy is provided, allowing the creation of free electron-hole pairs. Semiconductor detectors, like those made of Germanium, exploit this effect to detect energy depositions within the detector material.
Every detector has an n^+ and p^+ contact at its surface. The n^+ contact is formed via lithium diffusion, while the composition of the p^+ contact is implemented by a boron beam. During operation, a high voltage is applied to the electrodes on this surfaces, creating an electric field within. The exact position and size of the contacts vary for the different detector types (see Section 3.3.2) and determine the geometry of the field inside.
When an electron-hole pair forms, the holes move along the electric field lines toward the readout (which is the p^+ contact), generating the signal, while the electrons move towards the n^+ surface.
To minimize thermal noise, HPGes are cooled to approximately -200°C, typically using liquid nitrogen. This cooling is crucial because, at higher temperatures, thermally generated electron-hole pairs would interfere with the detection of actual signals, degrading the detector's performance.

Signal Formation

Important for the work with HPGe are the details of signal formation and the different effects on the pulse shapes.
First of all, the output pulse begins to form immediately after energy is deposited inside the detector and carriers start to move towards the electrodes. After the last carrier is collected by the electrode, the end point of the output pulse is reached, normally in a range of $1000 - 2000$ns.

Shockley-Ramo-Theorem

This instantaneous current induced on a specific electrode i is described via the Shockley-Ramo-Theorem ([32], [33]). This theorem tells that

$$i = q\vec{v} \cdot \vec{E_0} \tag{3.3}$$

while q is the carried charge, $\vec{v}$ is the charge velocity and $\vec{E_0}$ is the weighting field. The weighting potential and the corresponding weighting field are not the same as real electric potentials or fields inside the detector. To calculate the weighting potential φ, the following boundary conditions are applied:

1. $V = V_0$ at one electrode
2. $V = 0$ at all other electrodes
3. No free charges inside the field, leading to $\vec{\nabla}^2\varphi_0 = -\vec{\nabla} \cdot \vec{E_0} = 0$.

With these boundary conditions and the Shockley-Ramo-Theorem, the charge pulse can be calculated, corresponding to the weighting field inside the detector.

For detectors with higher mass, boundary condition 3 is an inaccurate approximation. In addition to the pure signal formation, the risetime of the pulse is affected by the charge drift collection. Since an event with 1MeV produces 10^6 electron-hole pairs [34], it is not a single, point-like charge carrier moving through the detector. Instead, it is a charge cloud with a distinct size, and therefore the electric field will be slightly different at different parts of this cloud, leading to a deformation. Moreover, there are diffusion and self-interaction effects, leading to an increased size of the charge cloud during its drift through the detector. This increased size influences the measured pulse shape of the event. More details about HPGe can be found in [35].
In L200, this has to be taken into account for the analysis of the larger ICPC detectors due to their higher mass and greater resolution. Therefore, this detectors require additional adjustments in the analysis, described in Section 3.4.2.

Important for the work at hand is, that the Shockley-Ramo-Theorem shows the correlation between field gradient in the detector and the returned pulse shape. The stronger the gradient of the electric field, the higher is the spacial resolution and its impact to the pulse shape. While the field geometry depends directly on the geometry of the detector, the pulses of the different detector types in L200 can vary strongly.

3.3.2. The LEGEND HPGes

Beside of the general explanation of HPGes, the following Section will now discuss the characteristics of the detectors used in LEGEND.

In the L200 setup, four distinct types of detectors are employed, each differing in geometry and their applied contacts as depicted in figure 3.3. This differences also result in a different field inside the detector as shown in figure 3.4. Following the theoretical aspects

Figure 3.3.: Contact geometries of different detector types. The p^+ contact is shown in orange, while the n^+ contact is grey, separated by a white groove or passivated surface, depending on the detector type. From left to right: PPC, BEGe, ICPC, Coax. (From [36])

of signal formation as described above, this leads to a different behavior in therms of signal formation and the following possibilities and challenges for PSA.

Semi-Coaxial HPGe (Coax)

The first type, Coax, are between 2-3 kg in weight. The p^+ contact, which is used for the readout, is large and covers the full borehole. This results in a field inside the detector depending on the radius, but is homogenous otherwise, similar to a capacitor. By following the Shockley-Ramo-Theorem in section 3.3.1, a mostly location-independent pulse shape is produced. This requires the use of advanced Machine Learning Algorithms for PSA (see Section 3.4.2).

Broad Energy HPGe (BEGe) and P-Type Point Contact Detector (PPC)

The second type includes BEGes from GERDA and PPCs from MAJORANA. Despite minor differences, these detectors are often considered a single category due to their similar advantages and limitations. They are lighter, with a maximum mass of 1kg, and "point-like" p^+ contact with a diameter < 1.5 mm, in contrast to the n^+ contact which covers the rest of the surface, expect from a small passivation layer. This contact geometry facilitates effective PSA due to the high field gradient inside the detector and the following differences in the pulse shapes depending on their location.
The difference between PPC and BEGe is the shape of the p^+ contact. For PPCs, this contact is really point-like and surrounded by a passivation surface. For BEGes, the p^+ contact is a bit larger and surrounded by a groove.

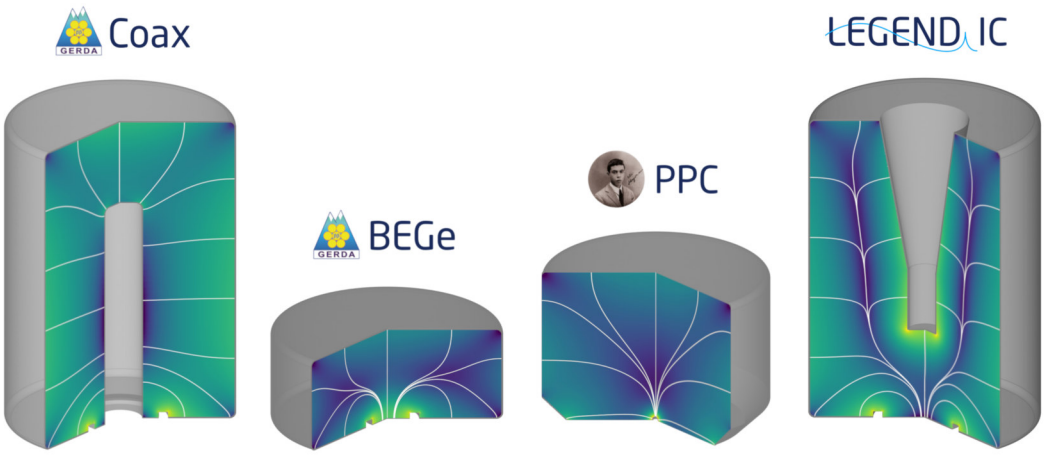


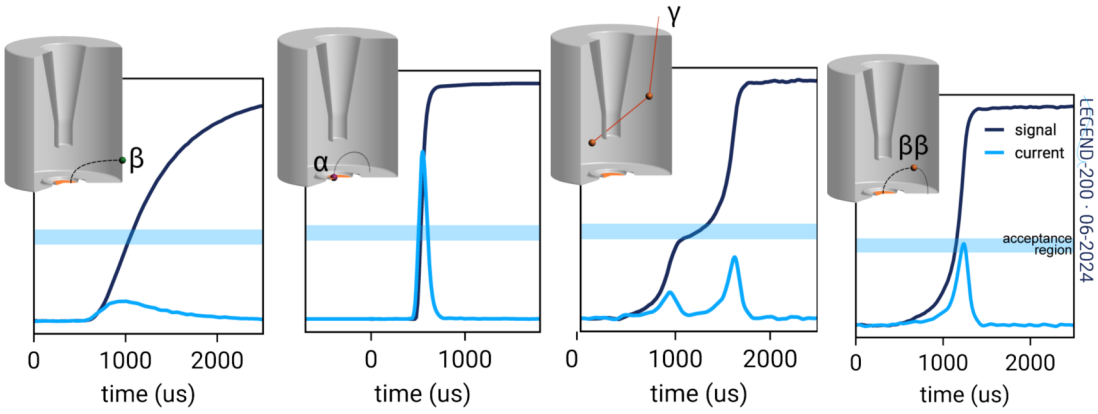
Figure 3.4.: Geometry and internal field configurations of Coax (left), BEGe (middle left), PPC (middle right), and ICPC (right) detectors. (From [24])

p-type Inverted Coaxial Point Contact (ICPC)
The last type, ICPCs, merges the high mass of semi-coaxial detectors with the precise PSA capabilities of BEGes or PPCs. The key difference between Coax and ICPC detectors lies in the placement of the p^+ contact: within the bore hole for semi-coaxial and on the opposite side as a point contact for ICPCs. This distinction significantly affects the internal field of the detector, as illustrated in figure 3.4.

3.3.3. Different types of events

A critical aspect of high-sensitivity experiments is effective background reduction. This can be achieved through a combination of experimental setup strategies, such as active or passive shielding, and rigorous analytical techniques. An integral part of this analysis is the PSA of detected Germanium pulses.

As illustrated in figure 3.5, we categorize waveforms based on the location of energy deposition within the detector, which correlates with the nature of the underlying radiation.
The readout of the HPGes return a signal proportional to the collected charge in the de-



(a) n^+ surface event from β background (b) p^+ surface event from an α particle (c) MSE from γ radiation (d) SSE from a $\beta\beta$ decay

Figure 3.5.: Correlation of Pulse Shapes with specific event types. ([37])

tector. Out of this charge pulse, for some applications a current pulse is calculated over the derivative of the measured signal. If not mentioned otherwise, pulse and waveform both describe the directly measured charge pulse in the following, while its derivative is called current pulse.

Single Site Events

An $0\nu\beta\beta$ decay event is expected to deposit energy in the volume of 100 μm^3, which is nearly single point in case of the HPGe. Events with similar characteristics like $2\nu\beta\beta$ decay events are termed **Single-Site-Event (SSE)** and are generally classified as signal-like.
Their pulse shape acts like a reference for all other type of events and pulses detected in L200.

Multi Site Events

Conversely, events where energy is deposited at multiple locations within the detector are called **Multi-Site-Event (MSE)**. They are mainly produced by γ radiation and can be seen as the superposition of multiple SSEs. This leads to a pulse with a step-like shape and a lower height of the corresponding current pulse.

n^+ surface Events

Events at the n^+ surface, known as slow pulses or **n^+ surface events**, deposit energy in the dead layer, a result of lithium diffusion at the surface. The deadlayer is a region of 1mm close to the surface where the electric field is zero. Due to the missing electric field inside the deadlayer, produced charges can not be measured. The only possibility

to leave the deadlayer is via diffusion. This leads to a slow pulse rise and, therefore, a lower amplitude of the current pulse compared to signal-like events with the same energy. n^+ surface events are mostly influenced by β radiation, which penetrates to the depth of the transition layer. However, given the large surface area of the n^+ contact, n^+ surface events play an important role, especially since they potentially leading to signal-like events as well.

p^+ surface Events

On the other hand, **p^+ surface events** are located on the p^+ contact. These are mainly α-particles. With their limited penetration depth, they are predominantly shielded by the n^+ surface and can only penetrate at the thin p^+ contact. Events at this contact exhibit significantly shorter risetimes, since they are located directly at the readout. This leads to a steeper charge rise and, if calculated out of the charge pulse, a higher current pulse.

3.4. LEGEND Analysis

The following section gives first a short overview of the analysis in LEGEND in general. In section 3.4.2 the classical PSA is explained in more detail.

3.4.1. Overview over full LEGEND Analysis chain

The L200 analysis process is a sophisticated, multi-step procedure designed to ensure high-precision results by integrating multiple subsystems. Each subsystem plays a critical role in data preparation, calibration, and event selection, contributing to the overall analysis.

Muon and Liquid Argon Veto Systems

The LEGEND experiment employs two dedicated veto systems to further reduce background contamination, which has to be analysed as well:
Muon Veto: This subsystem identifies events coincident with cosmic muons, which are flagged and excluded from the analysis.
Liquid Argon (LAr) Veto: Events coincident with signals from the liquid argon detector are similarly flagged and excluded.

These veto systems operate independently. Both feature their own, independent analysis framework. Their outputs result in muon and LAr flags in the Germanium dataset.

Only events without coincident signals in either the muon or LAr veto are accepted for further analysis.

Event Multiplicity

Another key criterion in the analysis is the requirement of event multiplicity. Only events with a single coincident germanium signal are considered valid. This ensures that multi-site events, which are more likely associated with background processes, are excluded.

Data Cleaning and Energy Calibration

Following the data acquisition, an essential first step is data cleaning, which involves the removal of unphysical events. This step ensures that only physically meaningful and analyzable data are retained for subsequent processing.
To achieve accurate energy measurements, energy calibration is performed by dedicated calibration runs. These runs provide reference points that allow the conversion of raw detector signals into precise energy values.

Pulse Shape Discrimination

Additionally, a cut for Pulse Shape Discrimination (PSD) is determined and fine-tuned during calibration and applied to the data from physical runs afterward. PSD techniques enable the differentiation between signal types based on the characteristic shapes of their waveforms. This step is crucial for rejecting background events.

3.4.2. Pulse Shape Analysis

The PSA methods used in LEGEND to distinguish the different types of Germanium events described in 3.3.3, will be described in this Section. How the PSA will perform and which methods can be used depends on the detector type, since the technical opportunities differ a lot. Especially the opportunities for PSA on Coax are different from the other types of detectors. Beside of the detector type, different types of background events can be analysed by varying analysis methods.
The following section will focus on the different analysis parts of PSA on BEGe, PPC and ICPC, as they are used but not enhanced in this work. The PSD for Coax are described and investigated in Chapter 5.

Necessary Data for PSA

Fundamental for every PSA is an appropriate measurement containing the possibility to select event populations with known event type. Requirement in general is a suitable

energy range and a certain amount of peaks containing the types of events which shall be analysed. It is possible to get these measurements from the characterisation campaign or the calibration runs in L200 directly. Thereby is a general PSA or the determination of a PSD cut done with a calibration source, while the PSD cut can be applied later on to the low-background physics data.

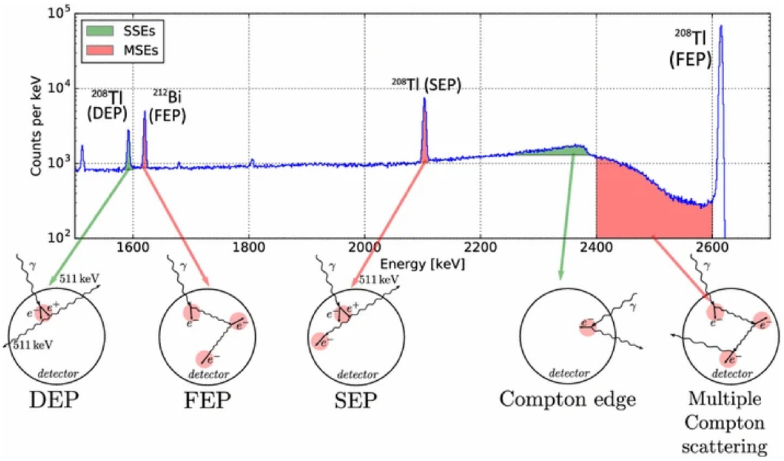


Figure 3.6.: Calibration spectrum with different peaks important for PSA [38]. For PSA, the different amount of MSE and SSE in DEP, SEP and FEP is used.

Usage of Th-Spectrum

In the HADES characterization campaign as well as in the L200 calibration runs, a ^{228}Th source is used to take data. In figure 3.6 a ^{228}Th spectrum is shown, including the important peaks for PSA.
During the measurement time, the spectrum in the region of 1-2 MeV consists mostly of decays of ^{208}Tl in the decay chain of ^{228}Th. The decay

$$^{208}\mathrm{Tl} \rightarrow {}^{208}\mathrm{Pb} + e^- + \overline{\nu_e} + \gamma, \tag{3.4}$$

creates a γ peak at 2614 keV, known as Full Energy Peak (FEP), since it contains the full enery of the γ.
The possibility of pair production induced by the emitted γ leads to the Double Escape Peak (DEP) at 1592keV and Single Escape Peak (SEP) at 2103keV. In case of the DEP, both produced 511keV photons leave the detector and deposit their energy outside the detector. In case of the SEP, just on photon escapes the detector, while the second deposites its energy inside. If no photon escapes, both contribute to the energy deposition

in the detector and therefore again the FEP.
Driven by this nature of the peaks, the DEP mainly consist of SSE, while SEP and FEP contain $5-10\%$ more MSE, depending on the detector size and the supposed definition of MSE. The detailed expectations regarding the amount of MSE in the peaks are described further in the last part of this section.

Beside this peaks from ^{208}Tl, the FEP of ^{212}Bi is worth mentioning, since its γ with 1621keV is very close to the DEP of ^{208}Tl at 1592keV. This will also be used in Chapter 5 for the PSA on Coax.

A/E Cut

The PSD in BEGe and ICPC detectors is performed via the so called A/E analysis. Due to the high field gradient inside these detectors, the different types of events can be differentiated by the current amplitude A and the energy E of an event.
As shown in figure 3.5, the current pulse of a MSE event has a lower amplitude as a SSE event with the same energy. This is due to the nature of MSE as an event depositing energy at two or more sites in the detector volume, resulting in a superposition of multiple SSE.
Since p^+ events are located close to the readout, the pulses of these events are really fast and therefore feature a higher A/E than bulk events. The n^+ surface events, or slow pulses, have a low A/E, since the charge has to diffuse through the deadlayer until it reaches the electric field inside the detector.

energy correction

Since not only A/E depends on energy, but also A itself is energy dependent, it is necessary to do an energy correction. This is done by selecting several slices of 20keV energy ranges on the compton continuum. For each of this slices, a histogram of the A/E distribution of all events in this energy range is generated. Afterwards, a fit of this distribution is performed with a gaussian function plus tail towards lower A/E to determine the mean $\mu_{A/E}$ and sigma $\sigma_{A/E}$. After this is done for all energy ranges, $\mu_{A/E}$ can be fitted by

$$\mu_{A/E}(E) = a + b \cdot E \tag{3.5}$$

If a, b are determined, $\mu_{A/E}(E)$ can be used to calculate an energy corrected A/E. Similar, $\sigma_{A/E}(E)$ follows this behaviour

$$\sigma_{A/E}(E) = \sqrt{a + b \cdot E^c} \tag{3.6}$$

With a, b, c determined, $\sigma_{A/E}(E)$ is later on used to define the width of the so-called SSE band containing the signal-like events. The result of such an energy correction including the determination of with SSE band width is shown in Figure 3.7. This is later used for the definition of an A/E classifier.

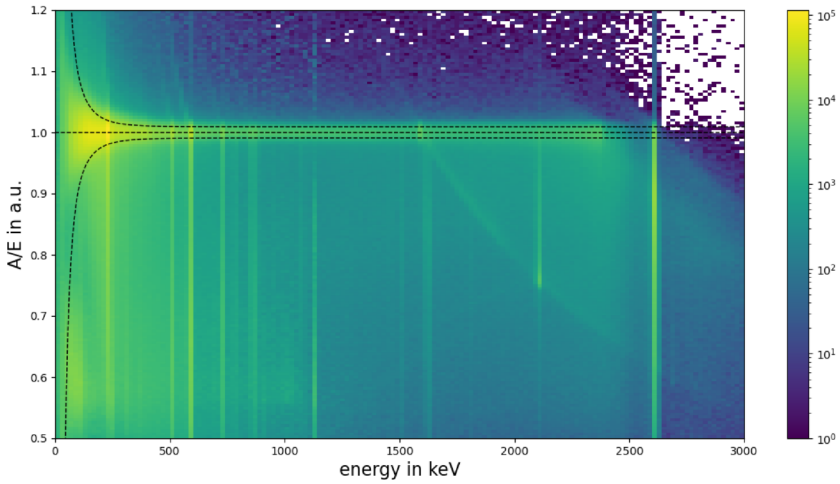


Figure 3.7.: AoE vs Energy plot with the characteristic SSE Band. This band is normalized to one and the A/E cut (black dotted line) is set relative to the width of it. The upper A/E cut is the high cut, while the lower one is the low cut. SEP and the two FEP can be seen in the lines for lower AoE with a higher population. DEP is harder to recognize, since it results in a higher population of SSE inside the SSE Band.

risetime correction

As a characteristic of the ICPC, a risetime correction is necessary. This is due to the large size of these detectors, which leads to charge collection effects [39].
In each HPGe detector, the produced charge clouds increase due to self-interaction and diffusion while they are traveling through the bulk of the detector, potential resulting in an modified risetime. Since BEGes and PPCs are comparably small, they remain unaf-

fected by this effect. The risetime distribution in the DEP but also a simulated $0\nu\beta\beta$ distribution shows a double peak, although with less pronounced characteristics. They are due to the detector geometry and appear also in simulations ([39]), the topology of risetime and A/E leading to this effect is shown in Figure 3.8b. This figure shows also the higher impact of this effect in outer regions of the detector to the double peak structure. In general, events from an external source, like in the DEP, are located closer to the surface of the detector, than events from $0\nu\beta\beta$ decay, which are homogeneously distributed. With this, the difference in the double peak structure between DEP and $0\nu\beta\beta$ decay can be explained, but also the fact, that it appears in both cases.
The risetime correction is used to bring the A/E of these two peaks in alignment.

As part of the HADES characterization campaign, measurements with a ^{228}Th source located at the top and at a latitudinal position where done. The top scan contains more events with a long risetime due to more events in the upper part of the bulk and misses the double peak structure. This can be explained by the highly asymmetrical distribution of different risetimes over the detector. All events in the upper half of the detector have a similar risetime, while the gradient increases strongly in the lower half.
The latitutdinal scan features two populations, one with larger and one with shorter risetimes, respectively. Therefore it is used to perform a risetime correction.
Normally, the risetime is defined as the difference between t_{10} and t_{90}, which define the timepoint of either 10% or 90% pulse height. A value which shows the same effect as the risetime is the Charge Drift (QDrift) value. It labels the area over the full rising edge of the pulse, starting from t_0 up to t_{max}. t_0 is the time, where the pulse start to rise and t_{max} is the time at maximum height of each single pulse.

QDrift

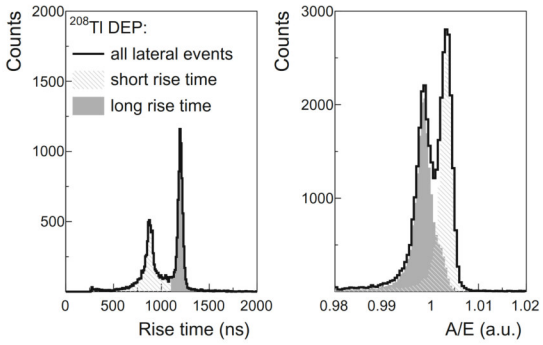
Since the risetime is based on only two points of the waveform, it is more susceptible to fluctuations in the pulse shape, whereas the QDrift is calculated over the full range of data points. Therefore, the QDrift parameter has a higher sensitivity to pulse shape variations and a higher robustness to noise, it is used in the analysis at hand in Section 6 and 7 instead of a risetime parameter.

A/E classifier

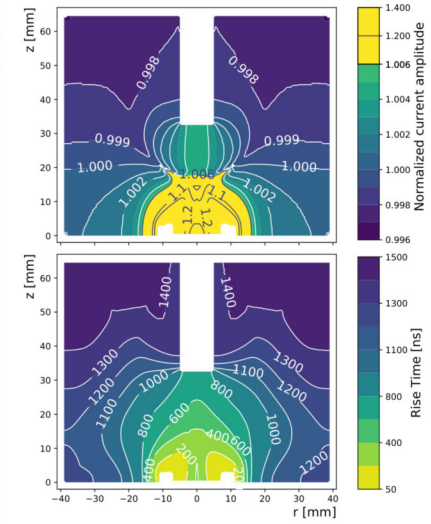
After all corrections are done, the A/E classifier $class_{A/E}$ is defined by including the sigma of the SSE band as following

$$class_{A/E} = \frac{A/E_{\text{corr}} - 1}{\sigma(E)} \tag{3.7}$$

with the risetime corrected A/E_{corr} and the width of the SSE band $\sigma(E)$.



(a) Risetime distribution in DEP for a latitudinal scan. The events in the upper parts of the detector have a longer risetime, while the population with the lower risetimes are from the bottom part of the detector. At risetime correction, the A/E of this regions is aligned to one single peak. (From [31])



(b) Topology of A/E (top) and rise time (bottom) values in an ICPC. (From [39])

γ background reduction

To reduce the γ-dominated background using PSD, the knowledge about the fraction of SSE and MSE in the different peaks is used. In the DEP a ratio of 90% SSE is expected and used to define a cut value. After applying this cut the ratio in SEP and FEP is be expected to be around 10% from previous measurements and simulations [31].
In this section the calculation of the cut value for A/E is discussed, while the general procedure is explained in 3.4.2, together with the underlying physical expectations and the uncertainty calculations.

The cut value is determined in two steps. Every event with an $class_{A/E} > 3$, is cut out by default. This cut value of 3 is called the high cut, since it cuts out all events above the SSE band. To determine the low cut value, which cuts events below the SSE band, the background is subtracted from the peaks and the A/E low cut value is set to 90% acceptance inside the DEP.
This low cut has to be calculated for all detectors individually and monitored over the different runs, since fluctuations of the cut value of ≈ 0.5 over the different periods in L200 were observed over the operation time up to know. Not adjusting the cut value would cause a high impact to the efficiency of this cut. Due to this fact, the low cut

value is often simply named cut value, since it is the crucial parameter. This naming will also be used in the work at hand.

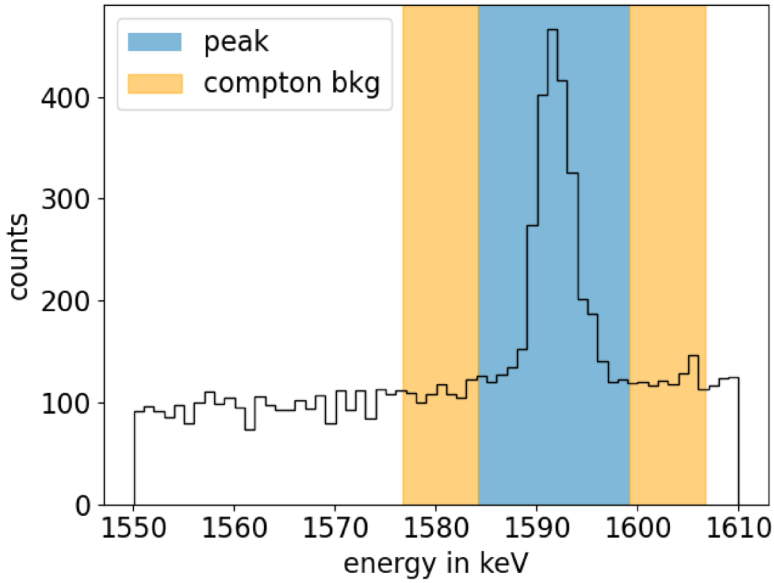


Figure 3.9.: Defined peak width of $\pm 4.5\sigma$ and the regions from Compton background with the same width.

Late Charge Cut

The Late Charge (LQ) Value is defined similar to the QDrift, by using the area above the pulse from t_{80} up to t_{max}. Its main benefit is the sensitivity to MSE close to the p^+ surface. Since these events are a superposition of one event with high A/E and one event with low A/E, it can result in a A/E close to 1 and therefore survive the A/E cut. These events feature a steep rise in the beginning, but the most important feature located in the last 20% of the rise. This is the part of the pulse, the LQ is sensitive to.

LQ Classifier

The LQ cut is adjusted similar to the A/E cut by adding energy and drifttime corrections as well as normalisation. Afterwards, an LQ-classifier $class_{LQ}$ is definded as

$$class_{LQ} = \frac{LQ(E) - 1}{\sigma(E)} \tag{3.8}$$

using that the LQ of SSE form a band with width $\sigma(E)$. This results in a distribution around zero. The cut is then set to 3[40]. In contrast to A/E, all events below are accepted and everything above the LQ cut value is rejected.

Efficiency Estimation and Survival Probabilities

For the analysis of a possible PSD classifier like the A/E, a parameter to estimate the efficiency estimation is necessary. Therefore, the knowledge about the expected ratio of MSE and SSE in DEP, FEP and SEP plays an important role. The exact ratio of SSE/MSE in these peaks depends strongly on the definition of a MSE. Depending on this assumtions, the survival fractions are between 2% to 20% in SEP and 5% to 25% in FEP[41][42]. In addition, the fraction depends on the detector mass, which was shown by [42]. Also their simulations predict a ratio of up to 95% SSE in DEP, making the A/E a relatively conservative cut.
Former experiences from GERDA, MAJORANA and characterisation measurements shows survival fractions from A/E of <10% in SEP and <15% in FEP [31].
The slightly higher amout of SSE in FEP is expected due to the possibility of photoelectric absorption. If the photon is absorbed directly, the full energy is deposited in form of a SSE, which contributes to the FEP and increase the ratio of SSE vs MSE in this peak. In addition, a general higher ratio of MSE is expected for higher energies, which also affect the survival probability.

To calculate such a survival fraction f, the peak width has to be determined. For the survival fractions, a range of $\pm 4.5\sigma$ around the mean is chosen as the "full peak" [42]. To get rid of the influence of the compton background, above and below the peak a window of 4.5σ each is choosen (see Figure 3.9) and subtracted as the following

$$f = \frac{N_c - B_c}{N - B} \tag{3.9}$$

where N is the number of events in the peak before the cut and N_c the number of events after the cut. B and B_c are the equivalent values in the background region. The corresponding uncertainties Δf are calculated via

$$\Delta f = \left| f \cdot \sqrt{\frac{N+B}{(N-B)^2} + \frac{N_c + B_c}{(N_c - B_c)^2} - 2\frac{N_c + B_c}{(N-B)(N_c - B_c)}} \right| \tag{3.10}$$

In addition to the survival fractions of the peaks, also the fraction in the Compton Continuum at $Q_{\beta\beta}$ is calculated to check, if there is a meaningful result at this value. In this case, the subtraction of the Compton background does not make sence, since the region around $Q_{\beta\beta}$ is nothing else than the Compton Continuum. Therefore, the calculation is

done simply with $B = 0, B_c = 0$.

By executing the analysis described this section, a full PSA for most of the HPGe signals is possible, as it is also implemented in L200. Especially the A/E is a powerful tool, but contain some challenges and at least for the Coax, it is not possible to apply an A/E cut. Therefore, the PSD in Coax is done with ML and further improvements could be possible with ML also for the other type of detectors.

4. Common Principles of Machine Learning

Since Machine Learning is an upcoming and highly efficient technique for manifold tasks, this work will examine its impact on the PSA in LEGEND by using characterisation measurements from HADES as well as calibration measurements from L200.
All Machine Learning methods used are explained in this chapter. The core of the developed model for PSA in LEGEND is the FIS, described in Section 6.2. This method can be applied to multiple Machine Learning techniques, such as Fully Connected Neural Network (FCNet), Convolutional Neural Network (CNN), or Recurrent Neural Network with attention score (RNN+att).
In the following sections, these common ML architectures are explained. Section 4.1 explains the general principle of Feedforward Neural Network (FNN), while Section 4.2 explains the principles of Recurrent Neural Network (RNN). For a comprehensive overview of this topic, [43] provides a detailed explanation.
Machine Learning and Artificial Intelligence have a wide range of applications, from classification to the generation of new content. Since this work focuses solely on the classification of pulse shapes, the following will concentrate on the classification techniques and applications of ML that are used later on. Furthermore, it only deals with labelled data and supervised learning, while other methods are not discussed.

4.1. Feedforward Neural Networks

FNN are the quintessence of deep learning models. The main idea is to approximate a function f, which can map a given input x to a classifier $\bar{y}$. This results in a mapping of $\bar{y} = f(x;\theta)$, where θ represents the parameters of the model. By training the Multilayer Perceptron (MLP), θ is adjusted to provide the best approximation. To demonstrate a good approximation, a training dataset with given labels y is necessary. During training, $\bar{y}$ approaches y as close as possible to ensure meaningful outputs when applying the

model to unlabelled data later.
In the terminology of machine learning, a FNN consist of multiple neurons. The word "neuron" is used due to the analogy to neurons in a brain. Those artificial neurons are a mathematical, non-linear function, combining different inputs and weights to a new output. Multiple neurons build a so-called layer and every network consist of at least one input layer, one output layer and one hidden layer in between. The composition of different layers and its neurons build the function f mentioned above.
The number of layer as well as their size and other fixed parameters (inside the model but also every fixed parameter connected to the training) are the **hyperparameters**. These hyperparameters have to be choosen as an essential part of the fine-tuning of the training. This can be done either by hand, or with a hyperparameter optimisation method like the Adaptive Experimentation Platform (Ax) [44], which was also used for this work.

All layers are connected only in a forward direction, giving rise to the name Feedforward Neural Networks. A graphical representation of such a network can be found in Figure 4.1. In contrast to an FNN, the neurons inside an RNN are also connected to earlier parts of the model. More details can be found in Section 4.2.
In the special case of a FCNet, every neuron in one layer is connected with every neuron of the next layer. This leads to the denomination "'Fully Connected"'.

Linear Layer

In the most basic case, each hidden layer is a linear function of the shape

$$\mathbf{h}_i = g_i(\mathbf{W}_i \cdot x + \mathbf{b}_i) \tag{4.1}$$

with an input x, weights W, biases $\mathbf{b}$ and the activation function g. By choosing the shape of W, it is possible to change the shape of the different layers $\mathbf{h}$.

Activation Function

To introduce non-linearity into the model, each layer, is fed into an activation function g. In general, all non-linear functions are possible, but in modern models, a Rectified Linear Unit (ReLU) or LeakyReLU is the default choice. LeakyReLU is defined as

$$g(x) = \begin{cases} x, & \text{if } x \geq 0 \\ s \cdot x, & \text{if } x < 0 \end{cases} \tag{4.2}$$

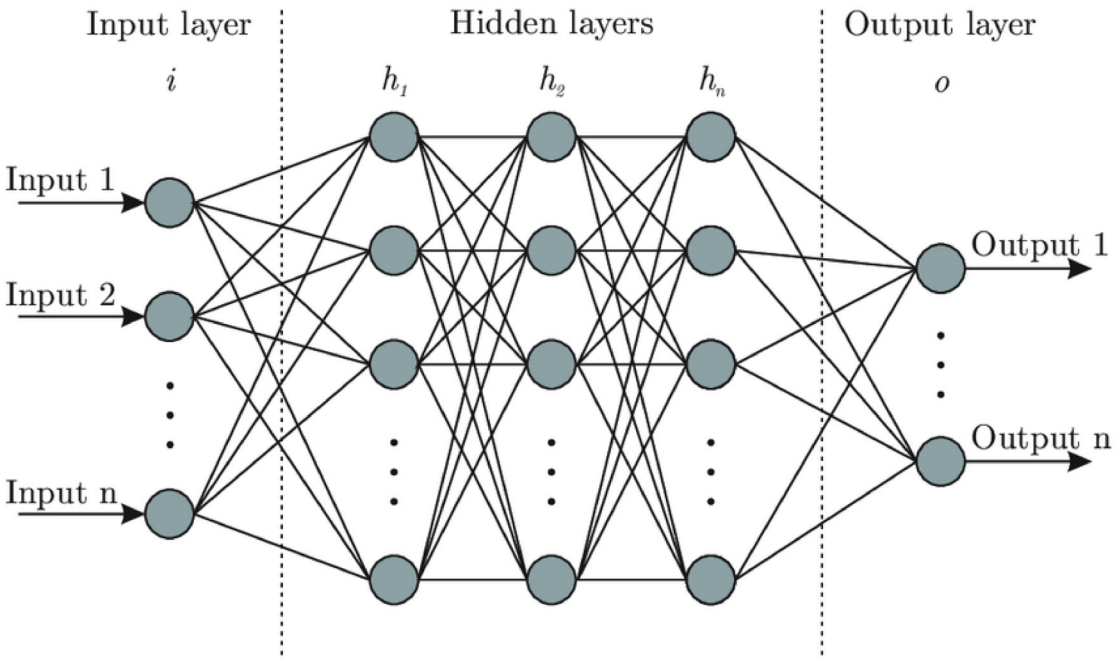


Figure 4.1.: Structure of FNN. The input is fed into the model through multiple hidden layers and results in an output layer. In the case of binary classification, as this work will later deal with, the output layer has size $n = 1$. The graphic shows the special case of a fully connected network, where every neuron is connected to all other neurons. (Graphic from [45])

with the slope $s = 10^{-2}$. The (non-leaky) ReLU is defined as the special case of $s = 0$. After the activation function is applied, the layer can build the input of the next layer

$$\mathbf{h}_{i+1} = g_{i+1}(\mathbf{W}_{i+1} \cdot \mathbf{h}_i + \mathbf{b}_{i+1}). \tag{4.3}$$

By applying this for all hidden layers, it results in $f(x, \theta(W, b))$ for the whole model. Before training, W and $\mathbf{b}$ are randomly chosen parameters, which are optimised during the training process.

Since we are dealing with supervised learning, it is necessary to have a labelled dataset to train the model. For each input x, the correct label y must be provided.

Loss Function

The next step of training is to calculate the distance between the predicted output from the model $\bar{y}$ and the expected output y. Therefore, a loss function L is used. Depending on factors such as the shape of the output or the labels parameter space, the loss function must be chosen accordingly.

For multi-class classification problems, both the label and prediction will be vectors. For binary classification, the label will be either zero or one, and the prediction is typically reduced to a value in [0,1]. These different task types require different loss functions due to the differing output formats.

There are several possible loss functions, depending on the output dimension or the complexity of the model. However, each loss function calculates a distance between the values $\bar{y}$ and y. For binary classification, for example, a basic loss function is the Mean Square Error (MSE):

$$L_{\text{MSE}}(y, \bar{y}) = \frac{1}{N} \sum (y - \bar{y})^2 \tag{4.4}$$

A slightly more complex but frequently used alternative is the Binary Cross Entropy Loss (BCE), given by

$$L_{\text{BCE}}(y, \bar{y}) = -(y \cdot \log(\bar{y}) + (1 - y) \cdot \log(1 - \bar{y})) \tag{4.5}$$

Optimiser

The loss function is then minimised using an optimiser, with multiple options available. Most optimisers are based on gradient descent but incorporate improvements to speed up the process and reduce the risk of getting stuck in a local minimum instead of reaching the global minimum.

A common and well-performing optimiser is, for example, the Adam Optimiser [46]. The name comes from Adaptive Moment Estimation, and it uses a dynamic learning rate by

scaling it with the current gradient. This leads to a decrease in the learning rate as it approaches a (local) minimum. Additionally, it uses momentum to handle local variations better and move past local minima. The combination of both often results in a fast convergence to the global minimum.

Backpropagation

By using the optimiser to minimise the loss function, the training process updates the parameters θ while propagating backwards through the entire model. This approach is known as backpropagation and is based on gradient-based optimisation.
To successfully train a model, multiple so-called epochs of backpropagation are performed. During each epoch, all training data are split into different batches and fed into the model one batch at a time. This process is repeated multiple times to get a well-trained model at the end.

Validation

Afterwards, the current state of the model is evaluated using a test dataset with known labels. This is necessary to monitor the training process and check the performance on an unseen dataset. The first step is to verify whether the model is learning and performing the intended task correctly. However, it is also important to check for overfitting. Overfitting may initially appear to yield excellent performance, but it indicates that the model has memorised the exact data points from the training dataset rather than generalising well to new data.

Dropout Layer

To prevent overfitting, different possibilities exist, but a common approach is to add a Dropout layer in between the model. Such a layer requires one hyperparameter $\phi \in (0, 1)$. The Dropout Layer sets input values to 0 with a probability of ϕ, while the remaining values stay unchanged.
Even with this basic version, good classification performance is possible for simple tasks.

4.1.1. Convolutional Neural Networks

A CNN is a subtype of FNN that is very common in image classification. It uses so-called convolutional layers to compress large images but can also be used to classify 1D, 3D, and higher-dimensional inputs. Such a layer uses a convolution operation to produce the input for the next layer. The key idea is to transform one volume of the input layer into another volume in the next layer rather than processing neuron by neuron, as in a fully connected network.

Convolutional Layer

The eponymous convolutional layer reduces the given input using a kernel of variable size. This kernel is a small matrix (often 3x3 or 5x5) sliding over the input data and

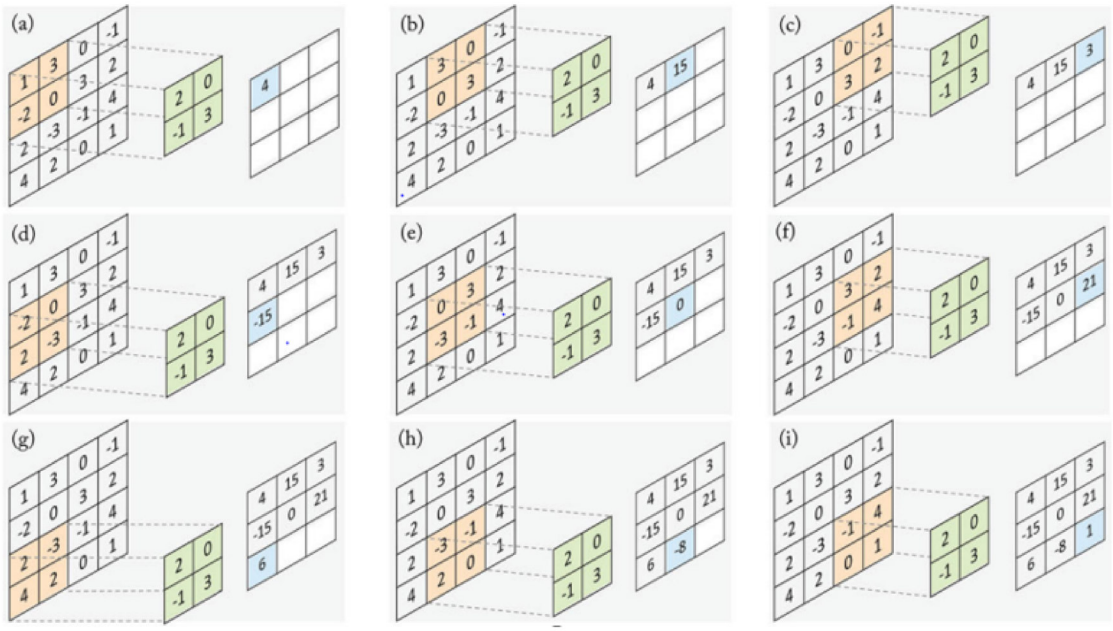


Figure 4.2.: Main concept behind a convolutional layer, exemplarily shown with a 4x4 matrix, a kernel size of 2, and a corresponding output shape of 3x3. [47]

perform the following operation:

$$y[i,j] = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} x[i+m, j+n] \cdot w[m,n] \tag{4.6}$$

with the input x and kernel w of height k_h and width k_w. Similar to the linear layer, the parameters of the kernel must be optimised as part of the training process. This method is graphically represented in Figure 4.2

Pooling Layer

In addition to convolutional layers, it is possible to reduce the layer size by adding a pooling layer. Like the convolutional layer, a pooling layer requires a kernel with a specified size. The full layer is divided into several subregions with a size similar to the kernel size. For example, in max pooling, only the maximum element of each subregion is retained in the next layer. In mean pooling, the mean value of each subregion is used to form the next layer.

These pooling layers provide an efficient way to reduce the model's dimensionality and speed up training. Unlike convolutional layers, they do not introduce additional parameters that must be optimised during training.

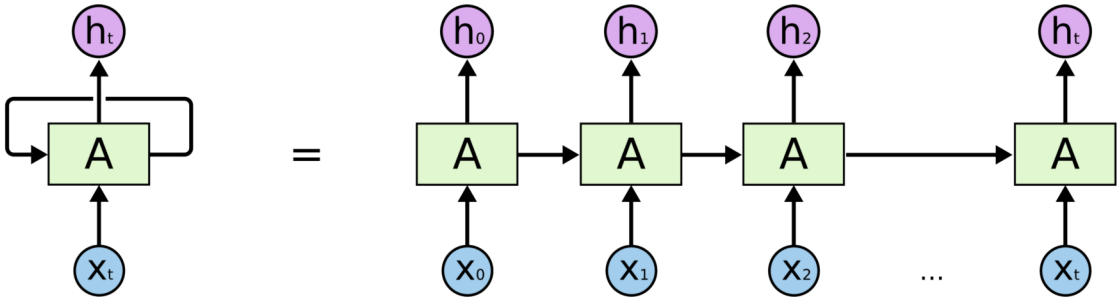


Figure 4.3.: RNNs are built from multiple cells. Depending on the type of RNN, the intern architecture A of the cells varies in complexity, but all of them account for the order of the input. (Graphic from [48])

Besides these specific layers of a CNN, it is also possible to incorporate components of a FCNet into a CNN.

4.2. Recurrent Neural Networks

RNNs are mostly used for problems that contain time series in various forms. A popular application is, for example, in Natural Language Processing and all types of language translation, where the order of words is important—for instance, the end of a sentence can refer to its beginning. In the case of LEGEND, the HPGe signals are time-ordered, making it logical to use an ML model that accounts for this characteristic.
A RNN does not only propagate forward (as an FNN), but its hidden layers are also connected to previous layers, making it possible to reference earlier states. This improves the model's ability to capture time dependencies but results in a large model requiring significant computational power. The following section explains the functionality of the RNN in more detail.
A RNN is built from multiple cells, one for each element of the input. These cells are recurrently connected to each other, as shown in Figure 4.3. The input $\mathbf{x} = \mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_n$ is fed in element-wise (with each element generally being a vector itself), resulting in a hidden state $\mathbf{h}_i$. Each hidden state $\mathbf{h}_i$ depends on $\mathbf{h}_{i-1}$ and $\mathbf{x}_i$.
Depending on the type of RNN, the cells have different internal compositions and may include, in addition to the hidden state, a cell state. Figure 4.4 illustrates three types of RNN, which are described later.

Basic RNN

A basic RNN simply feeds $\mathbf{x}^t$ and the previous hidden state $\mathbf{h}^{t-1}$ into an activation function (typically tanh) to compute the next hidden state $\mathbf{h}^i$. Usually, the first hidden

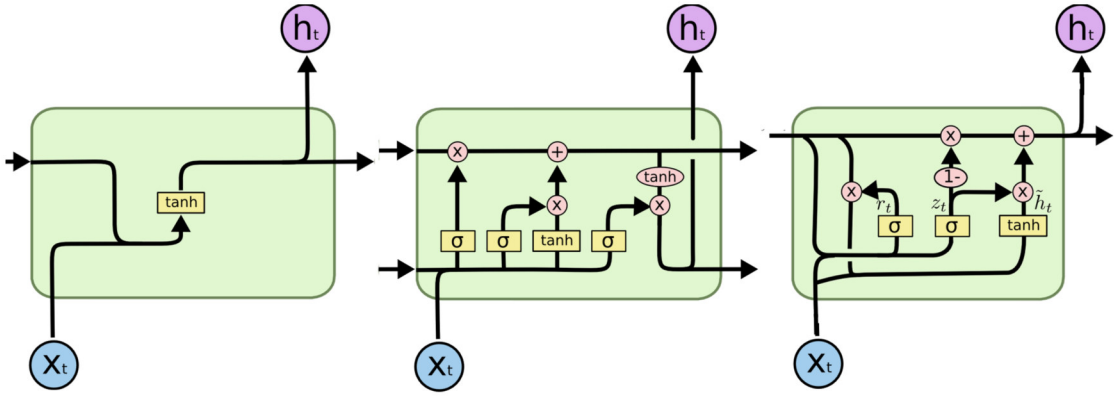


Figure 4.4.: Graphical representation of a basic RNN cell and two more complex variations, Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). The LSTM improves long-term memory by incorporating a forget gate. However, since LSTM results in a very large and computationally intensive model, the GRU offers a compromise between training speed and long-term memory. (Graphics from [48])

state $\mathbf{h}^0$ is set randomly. The last hidden state $\mathbf{h}^n$ forms the output of the RNN, and its dimension is controlled by the chosen length m of $\mathbf{h}^0$. The output can be used directly or passed through an FNN.

Such an RNN performs well for short-term tasks but struggles with long-term dependencies due to the Vanishing Gradient Problem ([49]), making it difficult to retain and utilise information from earlier timestamps over extended periods. However, since it forms the foundation for all other types of RNN, understanding its mechanics is crucial.

To address this problem, gated RNNs provide an efficient solution. They introduce pathways through time that prevent vanishing or exploding gradients. These networks are designed not only to retain important information over long periods but also to effectively discard less relevant data.

These networks are called "gated" because the weights controlling long-term information retention are regulated by additional hidden units known as gates.

Long Short-Term Memory (LSTM)

The first version of these gated RNNs is the so-called LSTM, illustrated in Figure 4.4. In addition to the **hidden state** $\mathbf{h}^{(t)}$, they include a **cell state** $C^{(t)}$ in each cell, allowing for long-term information retention, accumulation, and forgetting.

The cell state is trained alongside the hidden state, and both interact as schematically depicted in Figure 4.4.

To examine an LSTM cell in detail, the first step is to merge the input $\mathbf{x}^{(t)}$ with the

previous hidden state $\mathbf{h}^{(t-1)}$ and pass it through the **forget gate f**, defined as

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right) \in (0,1) \tag{4.7}$$

where $\mathbf{b}^f, \mathbf{W}^f, \mathbf{U}^f$ are the weights and biases of the forget gate. This part of the network determines which information should be discarded. Mathematically, this is achieved by multiplying it with the cell state.
The next part of the cell is the **input gate i**, which decides which information from the current input should be retained:

$$i_i^{(t)} = \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right) \tag{4.8}$$

$$\tilde{C}_i^{(t)} = \sigma\left(b_i^c + \sum_j W_{i,j}^c h_j^{(t-1)} + \sum_j U_{i,j}^c x_j^{(t)}\right) \tag{4.9}$$

The updated cell state is then calculated as

$$C_i^{(t)} = f_i^{(t)} \cdot C_i^{(t-1)} + i_i^{(t)} \cdot \tilde{C}_i^{(t)}. \tag{4.10}$$

The final step is to compute the new hidden state $\mathbf{h}^{(t)}$ using the **output gate o**:

$$\vec{o}_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}\right) \tag{4.11}$$

$$\vec{h}_i^{(t)} = \vec{o}_i^{(t)} \cdot \tanh\left(C_i^{(t)}\right) \tag{4.12}$$

Gated Recurrent Unit (GRU)

Besides the classical LSTM, there are different variations. One of them is the GRU, shown in Figure 4.4. A GRU simplifies the LSTM by integrating the additional cell state pipeline directly into the regular cell, eliminating the need for an extra input.
A GRU results in a less complex model that balances the full power of an LSTM with the efficiency of a simple RNN. It is faster in training while still being able to manage long-term dependencies in most cases.

Bidirectional RNN

Another variation is the bidirectional RNN. Since RNNs process data sequentially, they have a directional bias. A bidirectional RNN processes the data both forward and backward, combining both results into one output. This enhances the model's flexibility.

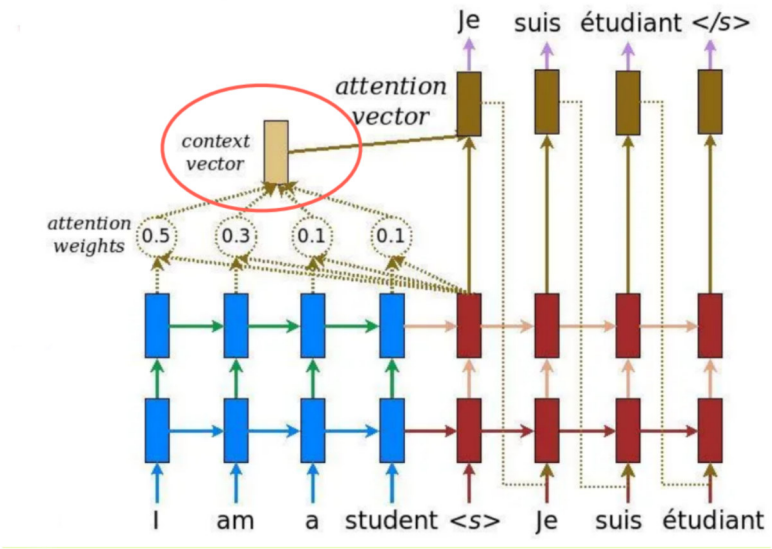


Figure 4.5.: Schematic drawing of an attention mechanism, showing the determination of context and attention vector. (From [51])

4.2.1. Self-Attention

The attention mechanism ([50]) is an additional feature that can be integrated into an RNN. It is trained as part of the model to optimize performance by directing focus toward specific input segments, thereby enabling the monitoring of these crucial parts. This makes it possible to cross-check whether the model's results are reasonable and trustworthy. As illustrated in figure 4.5, the first step in the attention mechanism is computing the similarity between each intermediate hidden state $\mathbf{h}^{(t)}$ for $t = \{0, ..., n-1\}$ and the final hidden state $\mathbf{h}^{(n)}$. All vectors $\mathbf{h}$ share the same length m. Depending on the similarity metric, different types of attention mechanisms can be implemented. However, they all ultimately compute a scalar quantity $s_{i,t}$ to measure the difference between the current part of the model and its endpoint. These quantities are calculated for each element i of the hidden state outputs $\mathbf{h}^{(t)}$. Using these quantities, an attention weight $a_i^{(t)}$ is computed as:

$$a_i^{(t)} = c_i^{(t)} \cdot s_i \quad \text{with} \quad \sum_t a_i^{(t)} \stackrel{!}{=} 1 \quad \text{for every } i \in [0, m] \tag{4.13}$$

From these attention weights, a context vector $\vec{C}$ is generated as follows:

$$C_i = \sum_{t=0}^{n} a_i^{(t)} \cdot h_i^{(t)} \tag{4.14}$$

Subsequently, an attention vector $\mathbf{A}$ is formed by concatenating $\mathbf{C}$ with the last hidden state $\mathbf{h}^{(n)}$:

$$\mathbf{A} = \mathbf{C} \oplus \mathbf{h}^{(n)} \tag{4.15}$$

In the final step of training, the context vector and the last hidden state are fed into a simple fully connected network, which merges both sources of information into a single output. In addition to influencing the model's predictions, the attention scores serve as a tool for assessing the model's reliability. If high attention weights are assigned to input regions that a human would deem unimportant, it may indicate a need to reassess or refine the model.

5. Artificial Neural Network for Pulse Shape Analysis with Semi-Coaxial Detectors

Since the GERDA experiment, Machine Learning has been used for Pulse Shape Analysis (PSA) of coaxial detectors. These six detectors are the oldest in LEGEND, and due to their design and the resulting low field gradient inside the detector, it is not possible to apply the A/E parameter. As a result, a ML model was already employed in GERDA. The first part of this chapter provides an explanation of the GERDA-like model, followed by an analysis of its application to the data from L200. Since an alternative approach did not perform reliably (see Section 6.6), the well-established methods from GERDA for Coax were adapted and slightly modified for the L200 setup. The traditional analysis consists of an ANN to reject MSE events and a risetime cut to remove alpha events.
This section focuses on the adapted ANN for LEGEND, which is designed to separate MSE events from gamma background. A detailed explanation of the ML model used in GERDA can be found in [52]. For LEGEND, several modifications were made, including the transition from ROOT to Python-based software and updates in ML techniques. However, the fundamental model structure, based on a MLP, remained unchanged, with adjustments in layer sizes and the choice of optimiser.
Since there was insufficient time for an internal review process at the depth expected by the collaboration, the ANN model was not included in the unblinding process in June 2024.

5.1. Training with Calibration Data

To train the ANN, data from the weekly calibration runs were used, specifically from periods 3, 4, and 6 to 9. These cover 48 weeks of measurement time and provide between 18,000 and 84,000 training events per detector. The exact number of events in

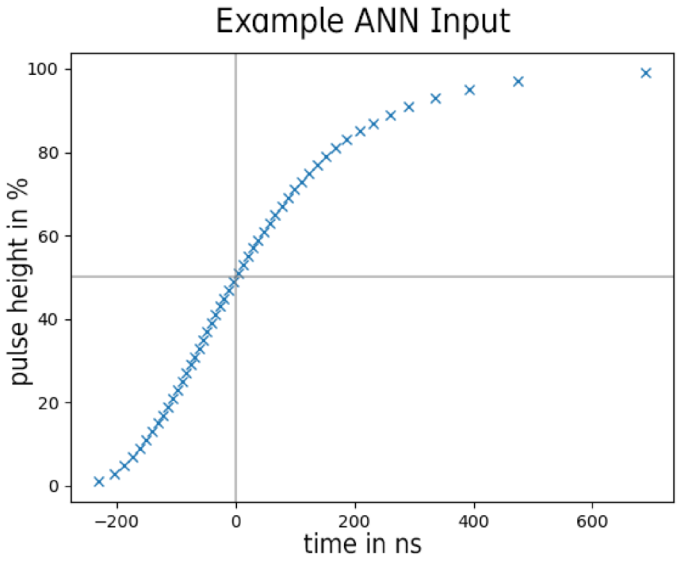


Figure 5.1.: Input of the ANN, normalised to one, consisting of 50 timepoints at 1%, 3%, 5%, ... 97%, 99% height, centred around the timepoint at 50% height after applying a moving window average.

the training dataset is presented in Table 5.1.
Events within the DEP were selected as signal-like, while events within the FEP of bismuth were labelled as background. Although these labels are not entirely pure, they are the best available. The FEP of Bi was chosen due to its energy of 1621 keV, which is close to the DEP at 1592 keV. The small energy difference between the signal and background training peaks helps ensure an energy-independent model. The corresponding energy spectrum is shown in Figure 3.6 and explained closer in Section 3.4.2.

To prepare the input data for the ANN, following the approach used in GERDA, a moving window average of 10 ns was applied three times. Next, baseline subtraction was performed, and the waveform height was normalised to 1. Each waveform was centred around the 50% height timepoint. The input array for the model consists of timepoints at 1%, 3%, 5%, ... 97%, 99% height, as shown in Figure 5.1.
The ANN consists of four linear layers with sizes 256, 128, 32, and 1. The first three layers are followed by LayerNorm1D and a LeakyReLU activation function. The final layer is followed by a sigmoid activation function to produce an output between 0 and 1.
Training was conducted using a binary cross-entropy (BCE) loss function and an Adam

optimiser with a weight decay of 10^{-4}. The model was trained with a learning rate of 0.00005 and a batch size of 32. To ensure sufficient statistics, data from all calibration runs for each detector were combined to create a training and a test dataset. Graphics showing the loss functions during training, used to monitor overtraining, are included in the appendix B.1.

Most Coax detectors were trained over 300 epochs. However, training was halted after 200 epochs for C00ANG4 and C000RG1, as both began to show signs of overtraining. Ideally, all detectors should have larger datasets, so 200 epochs should generally be sufficient for training. However, during the analysed periods, the calibration sources became unavailable for all coaxial detectors except C00ANG4. As a result, C00ANG4 benefited from higher statistics, significantly improving its ANN performance. It is expected that the performance of the ANN for the other Coax detectors will improve once calibration sources are available again.
The overtraining observed in C000RG1 after more than 200 epochs is attributed to the detector's poor performance. This was also a key factor in the decision to dismount and reprocess this detector after period 11.

5.2. Evaluation with Calibration Data

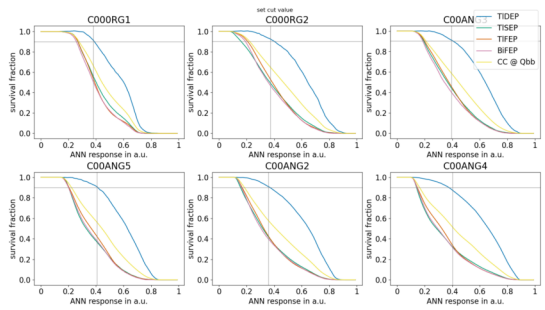
After training, several tests were conducted to assess the model's performance using calibration data. These evaluations include checks for energy dependence, timing stability, and efficiency across different peaks of the spectrum.

Set Cut Value

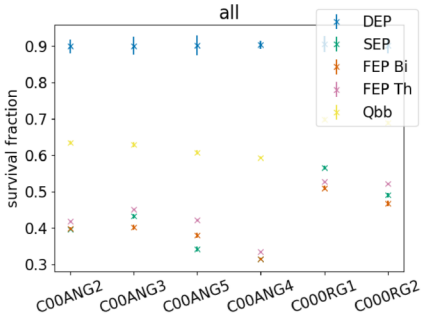
Initially, a cut value was determined for each detector based on the dataset from all calibration runs. This value was selected to ensure a 90% survival rate in the DEP, as illustrated in Figure 5.2 a). To compute this threshold, the Compton background was first subtracted from the peak. The full peak width was defined as $\pm 4.5\sigma$, and two slices of width 4.5σ each, before and after the peak, were selected as background (similar to the analysis procedure for A/E as described in 3.4.2). Due to background subtraction and the associated statistical fluctuations, survival fractions exceeding one are possible in Figure 5.2 a), as are values below zero in the distributions shown in Figure 5.2 c).
Further investigations examine the behaviour of SEP and FEP. While DEP primarily consists of SSE events, FEP and SEP predominantly contain MSE. In Figure 5.2 c), the

distribution of ANN outputs for different peaks is shown after background subtraction. A higher ANN response indicates SSE, whereas a lower response is associated with MSE. The distribution peaks at higher ANN values for events in the DEP and at lower values for SEP and FEP, demonstrating that the ANN effectively distinguishes between SSE and MSE.

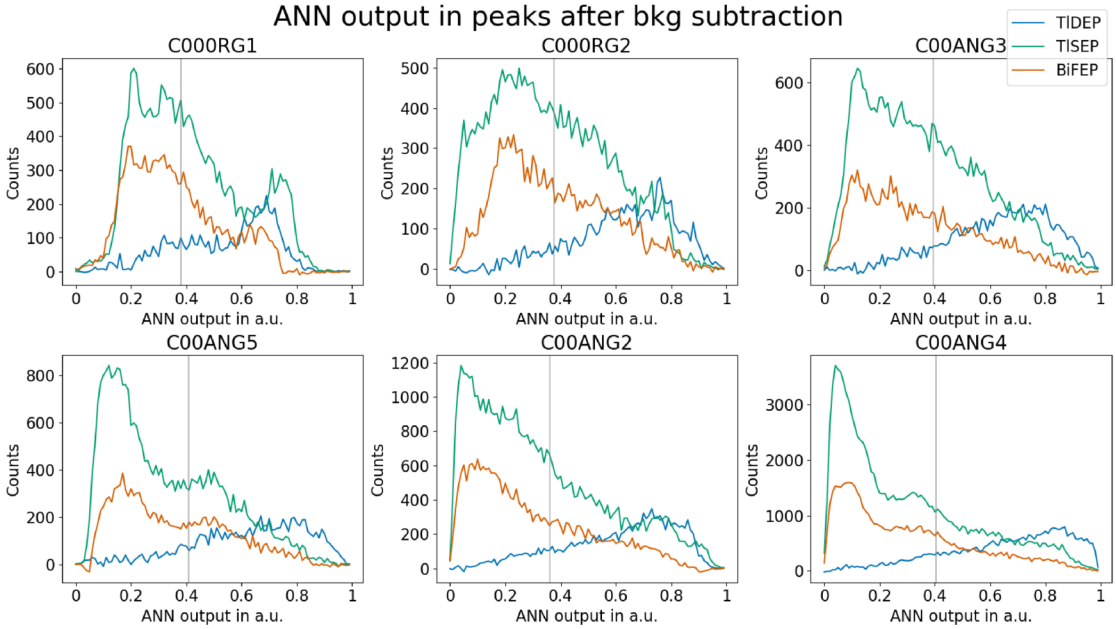
Survival fractions for different peaks and the Compton continuum around $Q_{\beta\beta}$ are presented in Figure 5.2 b), with exact values listed in Table 5.1. Applying this cut value to events within the SEP or FEP should result in <20% MSE survival from a physical point of view. However, due to the resolution of coaxial detectors, expected values from GERDA were around 30–40%, while the survival fraction in the Compton continuum at $Q_{\beta\beta}$ was approximately 60% [52].
For L200, survival fractions in SEP and FEP range from 30% to 57%, with Compton continuum survival around 60%, except for C000RG1 and C000RG2, where it approaches 70%. These two detectors also exhibit the highest survival fractions in SEP and FEP and have the smallest number of training events. This suggests that increasing the training dataset size could enhance performance. Additional evidence supporting this hypothesis is the strong performance of C00ANG4, which benefited from four times more training events due to its proximity to a calibration source. This detector exhibits survival fractions between 31% and 34% in the peaks and 59% at $Q_{\beta\beta}$. Future measurements will allow further validation of this hypothesis, and improvements are expected when calibration sources become available for the majority of coaxial detectors.
Since not all performance variations can be attributed to training dataset size, detector characteristics also play a role. This was a key factor in the decision to decommission C000RG1 during the 2024 summer reconstruction. The corresponding uncertainties follow the distribution described in Equation 3.10.



(a) Possible cut values for survival fractions in the different peaks and the Compton continuum (CC) at $Q_{\beta\beta}$



(b) Survival fraction for all detectors and the combined calibration dataset.



(c) Classifier response in DEP and SEP/FEP, illustrating the higher population of SSE in DEP and the dominance of MSE in SEP/FEP. This aligns with physical expectations, confirming that the model functions as intended.

Figure 5.2.: The cut value is set to 90% survival fraction in the DEP, with data from all calibration runs combined to determine the threshold for each detector.

Detector	N	SEP	FEP Bi	FEP Th	CC@Qbb
C00ANG2	33,536	$39.7\% \pm 0.5\%$	$39.8\% \pm 0.6\%$	$41.9\% \pm 0.8\%$	$63.5\% \pm 0.5\%$
C00ANG3	24,126	$43.2\% \pm 0.7\%$	$40.3\% \pm 0.8\%$	$45.1\% \pm 1.1\%$	$63.0\% \pm 0.6\%$
C00ANG4	83,462	$31.5\% \pm 0.4\%$	$31.5\% \pm 0.3\%$	$33.5\% \pm 0.5\%$	$59.3\% \pm 0.6\%$
C00ANG5	22,300	$34.2\% \pm 0.7\%$	$38.0\% \pm 0.8\%$	$42.2\% \pm 1.1\%$	$60.9\% \pm 0.3\%$
C000RG1	18,148	$56.6\% \pm 0.7\%$	$51.0\% \pm 0.7\%$	$52.7\% \pm 1.1\%$	$69.9\% \pm 0.5\%$
C000RG2	19,830	$49.1\% \pm 0.7\%$	$46.8\% \pm 0.7\%$	$52.3\% \pm 1.1\%$	$69.0\% \pm 0.6\%$

Table 5.1.: Number of training events N and survival fractions for different peaks in the Th-spectrum for each coaxial detector. Since the fraction is set to 90% in the DEP, values around 40% were expected from GERDA for coaxial detectors in other peaks, and around 60% in the Compton continuum near $Q_{\beta\beta}$.

Energy Dependence

Even if efficiency in different peaks is informative, it is essential to assess energy dependence in the ANN response. To monitor the energy dependence, a z-score is computed from the cut spectrum and defined as

$$z = \frac{x - x'}{\sigma} \tag{5.1}$$

for each bin in a spectrum, where x is the expected bin height, x' is the measured bin height and σ is the standard deviation of $x - x'$.

No detector exhibited fluctuations in the z-score exceeding 2σ, which is within an acceptable range. The areas around DEP, SEP, and FEP were excluded from the calculation, as their survival fraction variations are physically expected. Figure 5.3 illustrates this for C00ANG3, while plots for all other detectors are provided in the appendix.

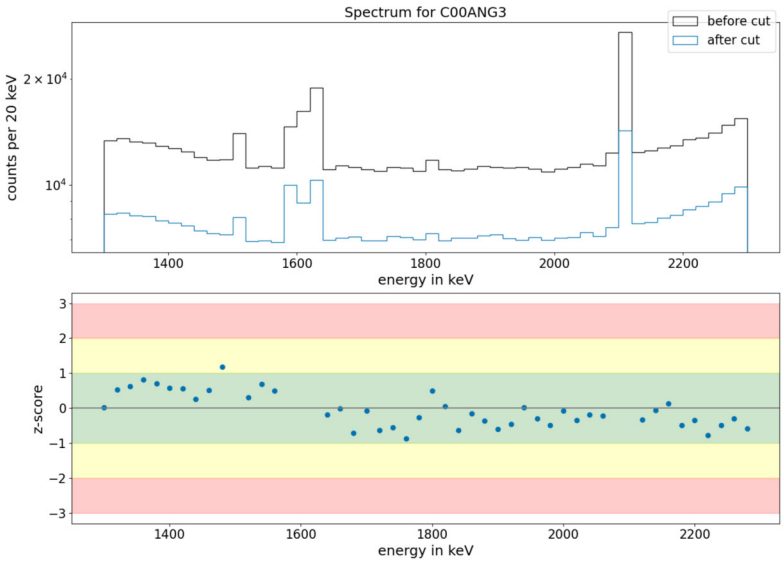


Figure 5.3.: Spectrum with and without cut as well as the corresponding z-score. It can be seen, that it is not totally flat, but in a good range inside 1 to 2 σ.

To assess the model's timing stability, the ANN was applied separately to calibration data from each run. The observed fluctuations in survival fractions are within statistical uncertainties. The timing stability check was performed without background subtraction, considering only the fraction of events inside the DEP before and after the cut (see Figure 5.4). Background subtraction would introduce higher statistical fluctuations without providing essential insights into timing stability.

Timing Stability

Observing survival fractions across different runs reveals a slight downward trend in some detectors, but these variations are within the statistical uncertainties as it can be seen in Figure 5.4. Overall, this confirms that a single ANN output and cut value can be applied consistently across all runs.

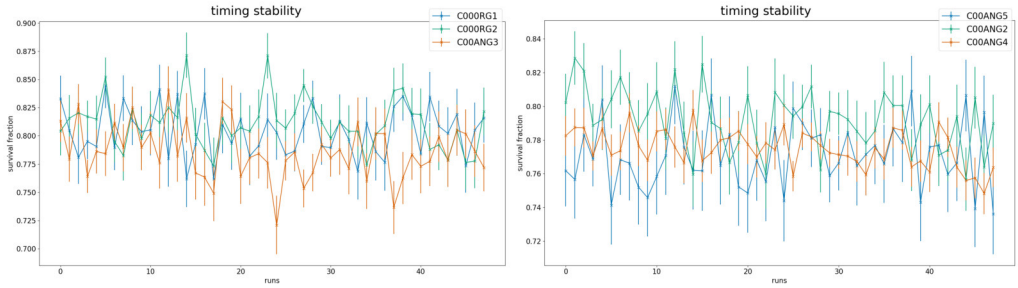


Figure 5.4.: Observation of timing stability of the survival fraction in DEP before background subtraction over different runs. Aside from statistical fluctuations, the performance remains stable across all runs.

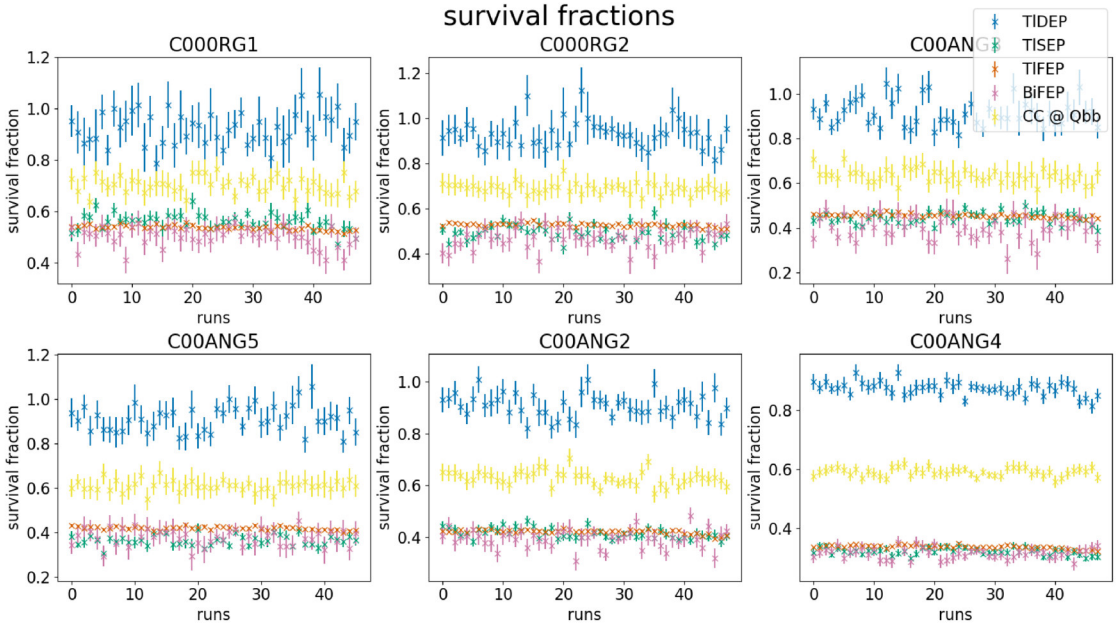


Figure 5.5.: Survival fraction for all runs after background subtraction. Since C00ANG4 has higher statistics, its survival fractions are more stable. For all other detectors, fluctuations remain within statistical uncertainties. The fractions are within the expected range: around 90% in DEP, 40% in SEP and FEP, and 60% in the Compton continuum at $Q_{\beta\beta}$. Values exceeding one are possible due to background subtraction.

While Figure 5.2 presents survival fractions for the combined dataset, Figure 5.5 displays values for individual runs. Due to the lower event count in each run, statistical fluctuations are significantly larger. Background subtraction can lead to survival fractions exceeding one. Even though the cut value was set to ensure 90% survival in DEP,

it fluctuates between runs. As shown in the timing stability analysis, no strong deviations occur over time. Notably, C00ANG4 exhibits much smaller uncertainties and consistently lower fractions for SEP, FEP, and Compton Continuum at $Q_{\beta\beta}$.

5.3. Application on Low-Background Physics Data

After evaluating the ANN performance on calibration data, its behaviour on low-background physics data, where no additional sources are present, must be analysed. This includes examining the remaining events in the $2\nu\beta\beta$ spectrum and those rejected around the Region of Interest (ROI) at 2039 keV.

$2\nu\beta\beta$

Between 1000 and 1300 keV, the spectrum is dominated by the $2\nu\beta\beta$ decay, which primarily consists of SSE, similar to $0\nu\beta\beta$ events. The known background contribution in this region is low, and the event count is relatively high.
In this $2\nu\beta\beta$ range, the survival fraction is approximately 80%, as shown in Table 5.2. This is comparable to GERDA, where the fraction ranged between 75% and 85% and varied between Phases I and II.

Detector	$N_{2\nu\beta\beta}$	Survival Fraction	$N_{0\nu\beta\beta}$	After Cut
C000RG1	454	$82\% \pm 1.9\%$	3	1
C000RG2	422	$84\% \pm 1.8\%$	1	0
C00ANG2	611	$86\% \pm 1.4\%$	4	4
C00ANG3	507	$81\% \pm 1.8\%$	5	2
C00ANG4	573	$77\% \pm 1.8\%$	4	2
C00ANG5	550	$86\% \pm 1.5\%$	2	1

Table 5.2.: Survival fractions for $2\nu\beta\beta$ between 1000 and 1300 keV, along with the total event count N within this energy range.

Overall, these values align closely with results from GERDA and match both simulation results and physical expectations.

Close to ROI

In the energy region closer to the Region of Interest (ROI) around 2039 keV, specifically between 1900 and 2200 keV (excluding the blinded window), a total of 19 events were detected, of which 10 were removed by the ANN.
As illustrated in Figure 5.6, some events are closer to the cut value than others, but none are exactly at the threshold. This provides confidence in the validity of rejected events. However, to determine a reliable efficiency for $0\nu\beta\beta$, cross-checks with energy-

rescaled events shifted to 2039 keV would be necessary. Due to time constraints, these calculations are not included here but will be performed for further LEGEND analyses.

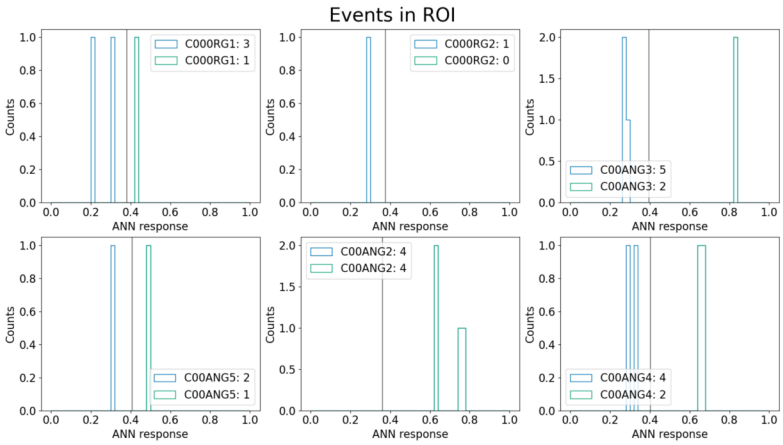


Figure 5.6.: Events before and after cuts near $Q_{\beta\beta}$ in low-background physics data (1900–2190 keV). No event is extremely close to the cut value.

5.4. Conclusion

The application of FIS on the semi-coaxial detectors is challenging due to the poor resolution of this detector type. While it is possible to achieve similar results to the ANN, the training process remains unstable. Additionally, there is no clear advantage in further pursuing this approach as long as FIS is still under investigation and development. Given the need for a classifier for the coaxial detectors in the near future, the well-established ANN from GERDA was chosen.

By implementing and adapting this model, we achieved the best current performance for the coaxial detectors. This was demonstrated using calibration data for training, and further validated by its application to low-background data, where the ANN produced meaningful and reasonable results. It was observed that the performance improves significantly when a calibration source is placed near individual detectors, as seen with C00ANG4. This suggests that further improvements are expected once SIS4 is available again. Additionally, C000RG1 was removed in later periods due to its poor performance. Both of these factors are expected to enhance the ANN performance and positively impact its results.

Performance metrics—including energy dependence, timing stability, survival fractions, and successful classification of $2\nu\beta\beta$ events in low-background data—demonstrate that

the ANN is a reliable and consistent classifier. It meets all necessary requirements for use in the ongoing L200 analysis.

To further refine the model and estimate efficiency at $Q_{\beta\beta}$, additional investigations with rescaled events will be necessary in the future. Ideally, simulated events would also be included in the study, but their generation requires significant effort. However, given that the ANN is a well-understood model, the current analysis provides a solid foundation. Moving forward, the presented investigations cover the most critical aspects of the analysis and will simply need to be repeated periodically as additional L200 data is collected.

6. Feature Importance Supervision

While the previous chapter provided an overview of the established ML techniques in L200, this chapter focuses on a new method called FIS. The core idea of the FIS model is to integrate physical knowledge about the important parts of the pulse for different types of events into the model. This approach should enable the model to classify pulses correctly for the right reasons and, ideally, consolidate various "'classical"' cuts and corrections into a single, more effective cut. The investigation of FIS is conducted on a single ICPC detector, as this type of detector offers the best resolution and the broadest range of pulse shapes.

At the beginning of this chapter, knowledge about different event types and their corresponding pulse shapes is compiled (Section 6.1), and the critical regions of the pulses are identified. In Section 6.2, the FIS concept is explained and connected to the information from Section 6.1 to develop an understanding of a model driven by physical knowledge and its potential applications.
The first part of the FIS application examines the method's ability to create an energy-independent model (Section 6.4.1). Subsequently, the performance of three different model architectures — FCNet, CNN, and RNN+att — is analysed (Section 6.4.2). Since RNN+att performs best in this comparison, it is then used to investigate the effect of different maskings in Section 6.5.
The key questions addressed are whether the approach is generally effective, how different architectures and maskings influence performance, and how FIS compares to the classical A/E analysis.

6.1. Human Knowledge about Germanium Pulses

Humans can observe various characteristics in the measured waveforms of germanium detectors, which are described in this section. There are two main characteristics that are crucial for developing a well-performing ML model:

6.1.1. Energy Dependence in Baseline Noise

Due to the process of signal formation, the height of each pulse is proportional to the energy of the underlying event. To mitigate the impact of this effect and enable the model to focus on pulse shape rather than pulse height, each pulse is normalized. During this process, the baseline noise, which is generally energy-independent, still introduces a slight energy dependence. Although this dependence is minimal, it becomes significant when the model requires high sensitivity.

In Figure 6.1, this energy dependence is shown by plotting energy against the standard deviation of the waveform baseline. The baseline is defined as the first 50 samples of a windowed waveform (size 512) centered around tp_{50}, the time point corresponding to 50% pulse height. To quantify this dependence, the maximum value of the baseline standard deviation distribution over small energy ranges (20 keV width) was determined. After fitting a linear function to these maxima, the resulting slope of

$$-9.6 \times 10^{-5}\ \text{std/keV} \pm 5 \times 10^{-11}\text{std/keV} \tag{6.1}$$

indicates a minor but clear dependence.

This observation suggests that energy dependence affects each pulse in its baseline and tail, as these regions are dominated by the detector's electronic noise rather than the induced charge in the detector. Near the start and end of the rising edge — specifically before t_5 (time at 5% pulse height) and after t_{95} (time at 95% pulse height) — this effect can also influence pulse shape itself, although it may already be mixed with meaningful features (see kinked waveforms in the next section).
Other approaches exist to address this energy dependence, such as ML-based denoising methods or scaling the ground noise with energy and adding it back to the pulse. However, both methods carry the risk of losing sensitivity, while FIS retains additional benefits, as described in the next section.
It is important to mention at this point, that there is also an energy dependence caused by the increasing amount of MSE at higher energies. This effect is physical meaningful and can influence the PSA especially in the survival fractions of SEP and FEP.

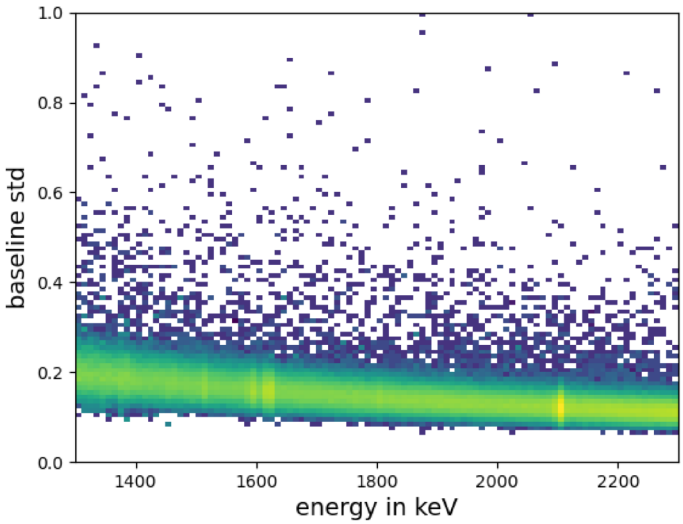


Figure 6.1.: Plot of standard deviation vs. energy after normalization. A small but noticeable energy dependence is evident, which proves to be crucial in the application of machine learning models.

6.1.2. Pulse Shape Effects

The various pulse shapes in germanium detectors are well characterized through signal formation theory, simulations, and previous PSA analyses in different types of detectors. The following section focuses on different pulse shape classes, how they can be distinguished, and the relevant analysis parameters.

In the classical PSA for L200 ICPC detectors, several corrections and cuts are combined. Ideally, a ML model should integrate all of these into a single step:

1. **Energy Correction**: This correction is necessary to achieve an energy-independent classifier. The rationale for training an energy-independent model has already been discussed in Section 6.1.1.

2. **QDrift Correction**: This correction compensates for charge drift collection effects, which are non-negligible in large-volume detectors like ICPC.

3. A/E **Parameter**: This parameter is used to distinguish SSE from other event types. It relies solely on the event energy E and pulse amplitude A. Despite its simplicity, it performs well but is not flawless and requires constant fine-tuning.

4. **LQ Parameter**: Defined as the area above the pulse between 80% and 100% height, this parameter is used alongside A/E to filter out events that are a superposition of p^+-surface events and MSE.

Each of these aspects is elaborated further in Section 3.4.2. Table 6.1 presents six types of waveforms along with their descriptions. While Section 3.3.3 covered the basic principles of signal formation and underlying radiation, the following section focuses on the analytical parameters used to distinguish different event types. In addition to SSE, MSE, p^+ surface event, and n^+ surface event, two additional classes are considered: kinked waveforms and MSE events depositing energy at the p+ contact. Both exhibit unique features and can complicate classical analysis.

- **SSE** deposit energy in a small volume of the detector ($< 1\,\mu\text{m}^3$) and are the desired signal-like events.

- **Kinked waveforms** are generally SSE (and therefore signal-like) occurring approximately within the first centimetre of the active detector volume. They exhibit a low QDrift and a small characteristic kink near t_0 but are mostly classified as SSE by A/E and LQ. This waveform shape arises due to the large detector volume. Their A/E is slightly higher than that of bulk SSE, an effect managed by applying a QDrift correction.
  Since efforts are being made to develop even larger germanium detectors, it is important to consider these events and ensure the classifier correctly identifies them as SSE.

- Most background events consist of **MSE** and can be identified using A/E. It is a well-performing standard; however, for L1000, it requires regular fine-tuning for each individual detector.

- n^+ **surface events** cannot currently be discriminated from MSE using A/E or LQ. Their major feature is the long risetime. Therefore, they are also called slow pulses. Additionally, a distinguishing feature of n^+ surface events compared to MSE is the presence of a single peak in the current pulse, whereas MSE pulses have two or more maxima.
  Even though n^+ surface event events could theoretically arise from $0\nu\beta\beta$ decay, they exhibit a shifted energy due to partial energy deposition in the dead or transition layer, preventing them from being fully measured. Thus, such events are

unlikely to be accepted as valid $0\nu\beta\beta$ decay candidates. However, if n^+ surface event events can be selected, they could provide valuable opportunities for further analysis.

- p^+ **surface events** exhibit high A/E values and can be selectively identified. Some misclassification may occur with MSE on the p^+ surface (see next point), but this can be effectively resolved using the LQ cut.
- **MSE events depositing energy close to the p^+-surface** result in a superposition of two pulse types. This can lead to an A/E value resembling that of an SSE or p^+ surface event, requiring LQ for proper distinction and rejection.
  Furthermore, superpositions of n^+ surface event and MSE are possible, but these do not pose a risk of being misclassified as SSE by any implemented cut parameter.

The goal of FIS is to systematically provide this information to the model in a structured manner. Ideally, it should lead to a model that integrates all described corrections and cuts, effectively distinguishing all pulse shapes listed in Table 6.1. While achieving this goal may not be straightforward, a general proof of principle would serve as a strong foundation for further research. Since FIS introduces several new possibilities, a systematic investigation of these options is warranted. Before discussing the various methodologies, the next section explains the technical implementation of FIS and how human knowledge can be effectively incorporated into the model.

6.2. Feature Importance Supervision

The idea of FIS is to utilise the knowledge described above and integrate it into the model, enabling it to classify pulses while considering this information. The detailed workings of this method are described below; it is adapted and customised from [53].
The main challenge in using Machine Learning for PSA is the hidden energy information in the baseline and tail of the waveform. Since the model should classify waveforms based on their pulse shape rather than their energy, FIS appears to be a reliable way to embed human knowledge about important and unimportant parts of the waveform into the model.
From a broader perspective, this approach can also benefit other applications where providing additional information about relevant and irrelevant features improves the model's performance.

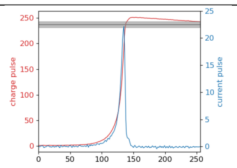
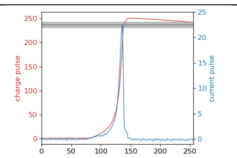
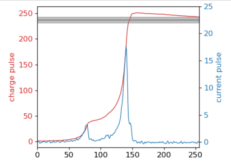
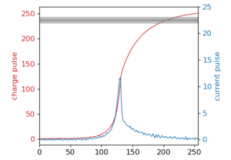
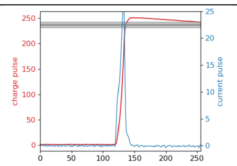
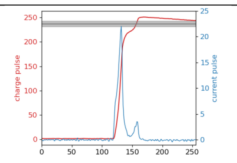
Event Type	**Category**	**Example Pulse**	**A/E**	**LQ**	**QDrift**	**Important Differences**
Classic SSE	Signal		≈ 1	<LQ cut	High QDrift	A/E peak at 0.9975
Kinked waveform	Signal		≈ 1	Low LQ	Low QDrift	A/E peak at 1.0065, QDrift
MSE	Background (mostly γ)		Low A/E	Varies	Varies	tp_{amax}, multiple local maxima in A
n^+ surface event	Background (including β)		Low A/E	High LQ	Tendency to higher QDrift	Later part of rising edge, single local max in A
p^+ surface event	Background (including α)		High A/E	Low LQ	Very low QDrift	tp_{amax}
MSE with p+ Surface Events	Background (mostly γ)		$A/E \geq 1$	High LQ	Increased low QDrift	tp_{80} to tp_{100} (LQ)

Table 6.1.: Overview of different waveform types with corresponding characteristics. The red curve represents the directly measured charge pulse, while the blue curve corresponds to the current. The grey area indicates the accepted current height from A/E as SSE.

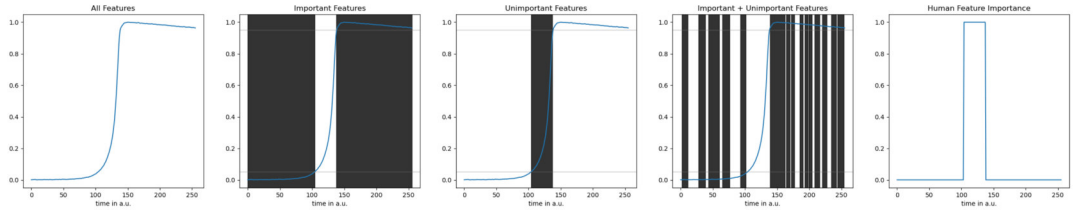


Figure 6.2.: Different inputs for FIS containing varying information: 1. all information, 2. only important information, 3. only unimportant information, 4. random selection of unimportant parts + all important information, 5. human knowledge about important parts for the explanation metric. The grey parts of the pulses are selected by a reference point (5% and 95% height of the pulse in this case) and ignored.

As with any supervised learning method, it is necessary to provide a label y for the training dataset. (More about the basics of ML see Section 4.) In addition, the input to the model consists of five different parts, as shown in Figure 6.2. The first four inputs are fed into a neural network, which provides the model with information from different perspectives. Additionally, the last input provides information about the expected explanation metric for the model.

1. All Features

The first input is the full waveform x and used to calculate a first Loss L_{task}.

$$L_{\text{task}} = \text{BCE}(y, f(x)) \tag{6.2}$$

$$= -(y \cdot \log(f(x)) + (1 - y) \cdot \log(1 - f(x))) \tag{6.3}$$

After the model is trained, just this full waveform builds the input for the utilisation of the model. The other inputs are solely used to enhance training, particularly to ensure energy stability, but are not used as inputs during the final application.

2. Important Features

The second input, x_{suff}, contains only the important features of the waveform, as determined by human knowledge. It is expected to yield an accurate output. All values below 5% height and above 95% height of the waveform are masked, leaving only the rising edge, which should be sufficient for the model to produce a

meaningful classification. The corresponding loss function for training is

$$L_{\text{suff}} = \text{BCE}(y, f(x_{\text{suff}})) \tag{6.4}$$
$$= -(y \cdot \log(f(x_{\text{suff}})) + (1-y) \cdot \log(1 - f(x_{\text{suff}}))) \tag{6.5}$$

Technically, the masking is achieved by setting all values in the unimportant regions to either the minimum or maximum value (or slightly lower/higher).

3. Unimportant Features

Conversely, the model should return an uncertain output when provided only with unimportant information from the baseline and tail. For this reason, the rising edge of the waveform is masked, producing x_{unc}. The model is trained to return a random output in this case using an arbitrary number $U \in (0, 1)$ with the loss function:

$$L_{\text{unc}} = \text{KL}(f(x_{\text{unc}}), U) \tag{6.6}$$
$$= f(x_{\text{unc}}) \cdot (\log(f(x_{\text{unc}})) - \log(U)) \tag{6.7}$$

while KL is the Kullback-Leibler Divergence, for an exact definition see Appendix A.1.

4. Invariance to Unimportant Features

At this stage, two approaches can be applied: a data augmentation approach and a Feature Importance supervision approach, as described in [54] and [55]. In both cases, the goal is to ensure that variations in the unimportant features selected in Step 3 do not affect the model's output.

1. The data augmentation approach combines all important features with random samples of unimportant features, which change during training. The outputs from the combined input $f(x_{\text{inv}})$ and the input containing only the important features $f(x_{\text{suff}})$ are compared. By applying a KL loss, the model is forced to make

the two outputs as similar as possible, following the equation:

$$L_{\text{inv-aug}} = \text{KL}(f(x_{\text{inv}}), f(x_{\text{suff}})) \tag{6.8}$$
$$= f(x_{\text{inv}}) \cdot (\log(f(x_{\text{inv}})) - \log(f(x_{\text{suff}}))) \tag{6.9}$$

2. The Feature Importance (FI) supervision approach uses the model's explanation metric at the datapoint level, e_{model}. Here, the model is directly penalised for focusing on unimportant features using:

$$L_{\text{inv-FI}} = \text{SmoothL1}(e_{\text{model}}) \tag{6.10}$$
$$= \{l_1, ..., l_N\}^T \tag{6.11}$$
$$\text{with} \quad l_n = 0.5(1 - e_{\text{model}}) \tag{6.12}$$

where SmoothL1 denotes the SmoothL1 Loss (A.1). An explanation metric gives the possibility to picture on which part of the input the model focus to produce its output. Two versions of such an metric is described in 5.
In the following analysis, only the FI supervision approach is used, as it provides better model stability. Additionally, there is no advantage in comparing both approaches, as this was already done in [53].

5. Human Feature Importance

In addition to the masked waveform variations, the last input, e_{human}, represents our human expectation of the important waveform features. The model is trained to produce an explanation metric e_{model} that closely matches e_{human} by minimising:

$$L_{\text{align}} = \text{CosEmb}(e_{\text{human}}, e_{\text{model}}) \tag{6.13}$$
$$= 1 - \cos(e_{\text{human}}, e_{\text{model}}) \tag{6.14}$$

while CosEmb is the Cosine Embedding Loss (A.1). CosEmb has the advantage, that it supports tensors as input and can therefore deal also with the n-dimensional explanation metrics, while n equal the input length.

The explanation metric can take one of two forms:

1. A gradient-based explanation, such as a vanilla gradient. The gradient through the full model from input to output provides structural information about the model and serves as a good foundation for an explanation metric. In the case of the vanilla gradient, the gradients through the full model are summed for each input, highlighting the most influential parts of the input.

2. An attention-based explanation, such as the attention score described in Section 4.2.1. The attention score guides the model to focus on particular parts of the input while simultaneously providing an explanation for its output.

Finally, the overall loss function is constructed as follows:

$$L_{\text{FIS}} = w_0 L_{\text{task}} + w_1 L_{\text{suff}} + w_2 L_{\text{unc}} + w_3 L_{\text{inv}} + w_4 L_{\text{align}} \tag{6.15}$$

with hyperparameters w_i, and it is used to train the model like any other ML approach without FIS.

6.2.1. Idea of Different Maskings

Based on the fundamental principles of pulse shape analysis and knowledge of waveform characteristics, different masking versions are introduced and discussed in this section. These maskings can be applied by leveraging different characteristic points of the pulse. First, a distinction must be made between the classic masking for important and unimportant features (used to train the model on meaningful parts of the waveform) and the Human Feature Importance (HFI), which is not a masking in the classical sense.

For important and unimportant features, different parts of the input are either masked or retained—there is no intermediate state. This mechanism is particularly useful in marking the baseline and tail of the waveform as unimportant, thereby suppressing energy dependence. The only modifiable aspect is the scope of the masked region. By considering the masking in the important and unimportant features, two versions are proposed:

1. The **95-Model** masks the range from t_5 (time point at 5% pulse height) to t_{95} (time point at 95% pulse height) as illustrated in Figure 6.3 for different types of waveforms. This captures most of the rising edge while excluding the exact start and endpoint, which

are difficult to determine precisely.

$$HFI(t) = \begin{cases} 0, & \text{if } t_5 < t < t_{95} \\ 1, & \text{otherwise} \end{cases} \tag{6.16}$$

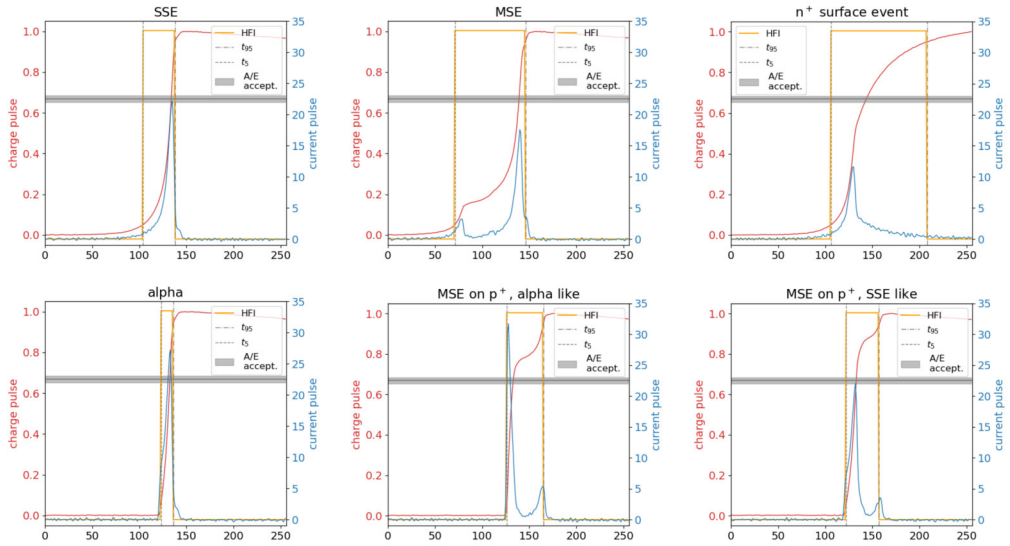


Figure 6.3.: Exemplary waveforms for the 95-model. The measured charge pulse (red), the current pulse (blue), and the Human Feature Importance (HFI, orange) are displayed. The HFI is applied over t_5 to t_{95}, similar to the masking for important and unimportant features. This covers most of the rising edge, providing a broad range of information, but fewer details and less guidance compared to the following models.

2. The **amax-Model** masks the range of ±5 samples around the time point corresponding to the maximum current pulse height, t_{amax}. It is depicted in Figure 6.4, showing the implications for different pulses. This model aims to replicate the behavior of A/E, which considers only the maximum slope point.

$$HFI(t) = \begin{cases} 0, & \text{if } t_{\text{amax}} - 5 < t < t_{\text{amax}} + 5 \\ 1, & \text{otherwise} \end{cases} \tag{6.17}$$

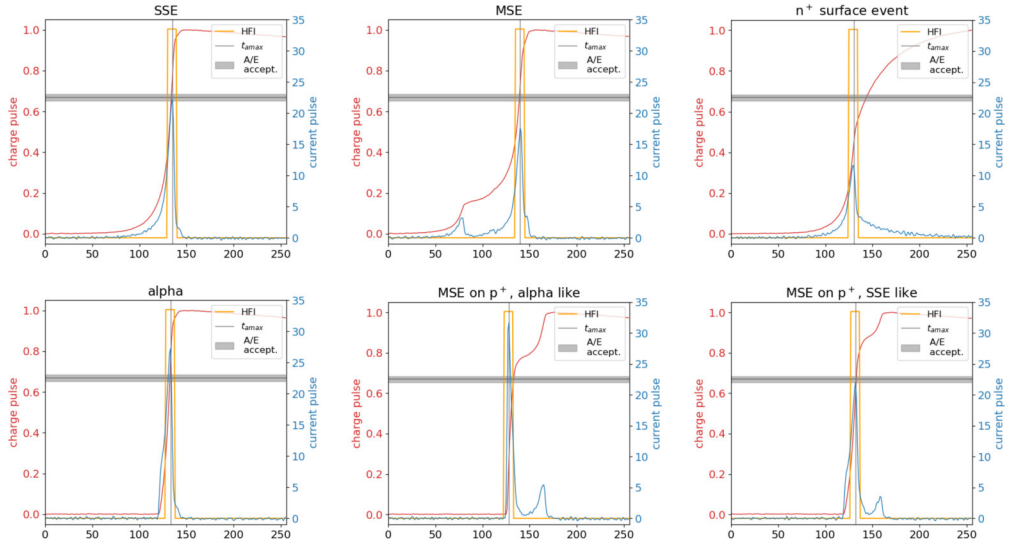


Figure 6.4.: Similar to Figure 6.3, the HFI for the amax-model is shown here for different event types. The amax-model mimics A/E by focusing only on the region surrounding the maximum of the current pulse. This results in a limited amount of information, but it is known from A/E that this region contains the most crucial features. A close agreement with A/E is expected, though this model does not address its shortcomings. An example is the last waveform, where the crucial region that identifies the event as MSE is not covered by HFI.

For Human Feature Importance (HFI), values can be assigned variably within the important range. The simplest approach is to set the entire important region to one. In addition to this, additional alternative versions can be proposed:

3. A masking later referred to as the **Fork-Model** combines the knowledge from A/E with information about risetime by highlighting t_5 and t_{amax} as follows:

$$HFI(t) = \begin{cases} 1, & \text{if } t_0 - 5 \leq t \leq t_0 + 5 \\ 1, & \text{if } t_{\text{amax}} - 5 \leq t \leq t_{\text{amax}} + 5 \\ 0, & \text{if } t < t_5 - 5 \\ 0, & \text{if } t > t_{95} + 5 \\ 0.5, & \text{otherwise} \end{cases} \tag{6.18}$$

This is illustrated in Figure 6.5 for different waveform types. By taking the regions

around t_0 and t_{amax} into account with a higher impact, a higher sensitivity to the different types of pulses is expected. Furthermore, these two peaks in the HFI contain an information similar to the QDrift (see Section 3.4.2). This is expected to be energy dependent, but contains interesting possibilities in terms of classification of n^+ surface events or MSE on p^+ surface.

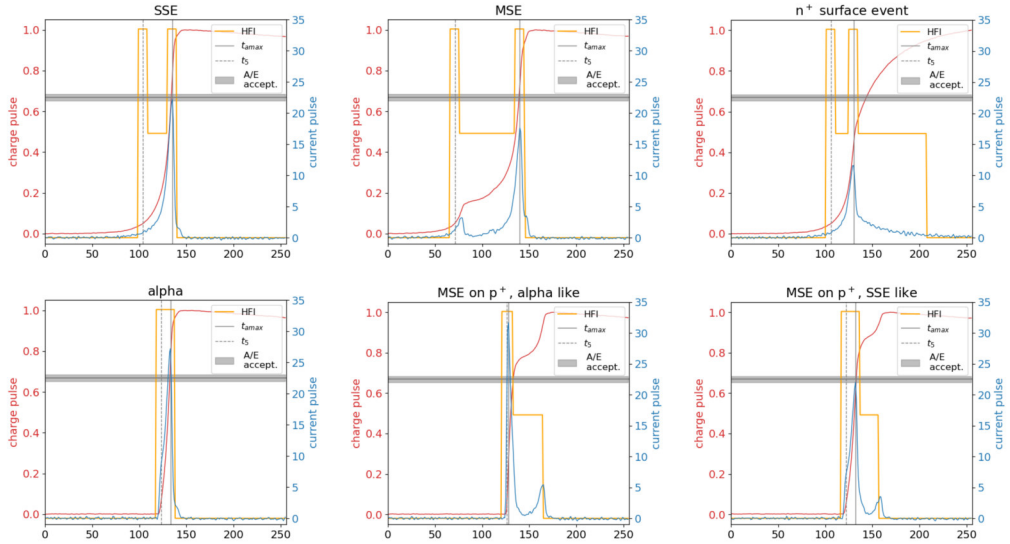


Figure 6.5.: Example waveforms for the Fork-Model, illustrating how HFI is applied to different event types. Different to the previous versions, the full rising edge is taken into account, but the region of t_0 and t_{amax} are weighted more than the rest of the rising edge. This leads to a higher focus on these two points and giving also some kind of a risetime information. With this information, the Model could be sensitive to pulse shapes like n^+ surface event or MSE on p^+-surface. On the other hand, the risetime is in general energy dependent, so the fork masking can result in an energy dependent model.

4. The next version incorporates a combination of A/E, QDrift, and additional information from local maxima in the current pulse. This is achieved by including the corresponding time points $t_{\text{amax, local}}$ in addition to t_{amax}, which represents the global maximum, as depicted in 6.6. This approach enhances the model's ability to distinguish between slow pulses with a single peak in the current pulse and MSE, which typically

exhibits multiple peaks.

$$HFI(t) = \begin{cases} 1, & \text{if } t_0 - 5 \leq t \leq t_0 + 5 \\ 1, & \text{if } t_{\text{amax}} - 5 \leq t \leq t_{\text{amax}} + 5 \\ 0.75, & \text{if } t_{\text{amax, local}} - 5 \leq t \leq t_{\text{amax, local}} + 5 \\ 0, & \text{if } t < t_5 - 5 \\ 0, & \text{if } t > t_{95} + 5 \\ 0.5, & \text{otherwise} \end{cases} \tag{6.19}$$

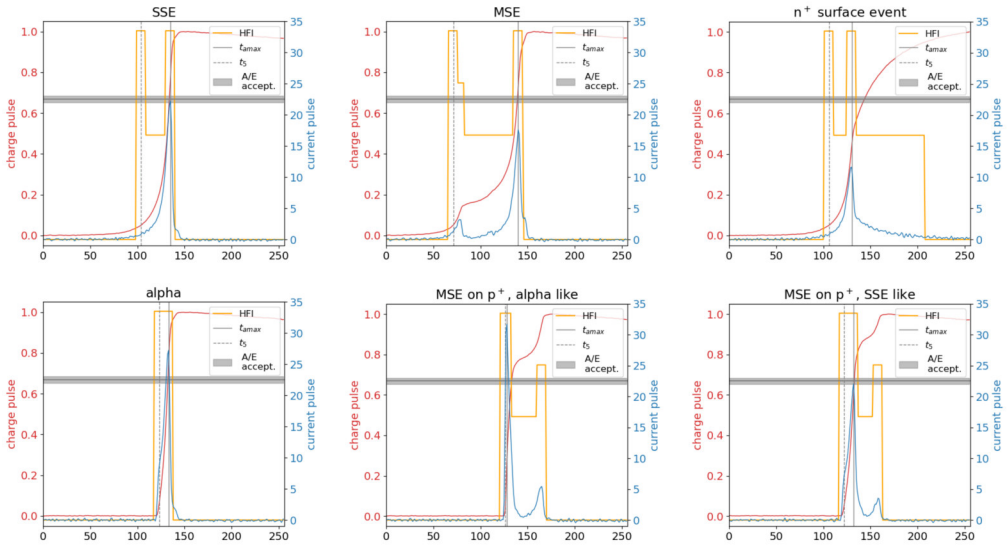


Figure 6.6.: Example waveforms for the local-maxima-enhanced model, showing how HFI is applied for different types of events.

These examples illustrate just a few possibilities for applying FIS. Another approach could involve using the normalized current pulse directly as HFI or incorporating t_{80}, similar to the LQ cut. But there are several different combinations or modifications possible, giving additional knowledge to the model.
On the other hand, if too much information is included at once, this can lead to an uncertain result. A potential solution could be to train separately for different event types (e.g., n^+ surface event events) using a distinct masking approach. All of this are starting points for future investigations.

The following sections will begin with the simplest masking approach, covering the range from 5% to 95% of pulse height, as a proof of principle. In Section 6.5, the 95-, amax-, and Fork-Model maskings are evaluated. These three versions are sufficient to demonstrate the general principle of FIS and the influence of different masking strategies, either through changes in full masking or modifications in HFI.

6.3. General Training of the FIS Model

At the beginning, the first question is how to train the model, what data is available, and what the boundary conditions are. These aspects will be investigated and clarified in this section.
The first step in any ML approach is selecting a training dataset. This dataset must include two distinct sets of events: one set of clear signal-like events and another set of clear background events. In this work, the focus is on distinguishing between SSE and MSE, while the selection of surface events is not addressed here.

One of the key challenges in this process is the difficulty of correctly labelling events. In theory, simulated pulses of signal and background events could be used for this purpose. However, in practice, this approach is unreliable due to the lack of accurate models for electronic response, noise, the effects of dead and transition layers, and the underlying field geometry. While most of these effects can be incorporated into a simulation, it is time-consuming, and achieving the necessary precision remains a challenge. Although ongoing efforts aim to simulate realistic pulse shapes, achieving full control is difficult.

A more practical approach is to use measurements from a thorium source. In the resulting spectrum, peaks such as the DEP or SEP can be identified based on energy. While these peaks do not represent pure datasets, the DEP contains a higher proportion of SSE compared to the SEP or FEP, which have a higher fraction of MSE. This fact provides a sufficient basis for training the model to distinguish between SSE and MSE events.

In previous work ([52]), ML was used for PSA in semi-coaxial detectors. There, an Artificial Neural Network (ANN) was trained using the DEP from thorium and the FEP from bismuth due to the small energy difference between these peaks. However, this approach is limited because the proportion of MSE in FEP is lower than in SEP. Simply using pulses from SEP as MSE causes the model to differentiate between background

and signal-like datasets based on energy rather than pulse shape, leading to energy-dependent model outputs.
This energy dependence is due to the energy dependence of the noise in baseline and tail and is particularly relevant since the training peaks differ in energy by 511 keV between signal and background-like events.

Additionally, the difference between coaxial detectors and ICPC detectors must be considered when adapting the training process. ICPC detectors offer higher resolution and more details in the measured pulse shapes. This allows the usage of a more precise model which is sensitive to this detailed structures in the pulse. However, they also provide a better capacity for resolving energy dependence compared to semi-coaxial detectors.
In general, we expect a stronger impact of energy differences between background and signal-labelled events when using ICPC.

By applying the FIS method, it should be possible to minimise the energy dependence by focusing on the rising edge of the pulse while masking the baseline and tail as unimportant features. This should enable the use of SEP as an MSE label without introducing energy dependence in the model. Applying a hard cut rather than a soft masking approach ensures that baseline information is fully suppressed, allowing the model to focus on important waveform features, particularly in the initial percentage of the rising edge.

All models are trained with 10,000 events each from DEP and SEP over 30 epochs using an Adam optimiser and the combined loss L_{FIS} from Equation 6.15. Momentum and learning rate vary depending on the model architecture and are shown in the corresponding sections.
For all architectures, the region between t_5 and t_{95} is selected as the important features, as described in Section 6.2.1 and illustrated in Figure 6.3.
The exact configurations and hyperparameters of each model are described in Table 6.2. Examples of the resulting plots with the loss function over training can be found in the Appendix (Table C.2), together with different waveforms, their attention score and the HFI (Table C.1). In this plot, also the alignment, but also slight differences between HFI and trained attention score are visible.

6.4. General Performance of FIS

6.4.1. Creating an Energy-Independent Model with FIS

Energy Dependence

To demonstrate the ability to eliminate energy dependence, a model is trained both with and without the FIS mechanism. As explained above, the baseline and tail contain energy information accessible to the ML model. Due to the significant energy difference between background and signal-labelled events, a model trained without the FIS mechanism learns to discriminate events based on energy rather than pulse shape. This effect is clearly visible in Figure 6.7.
On the other hand, the fraction of MSE increases at higher energies. Thus, there is a slight energy dependence due to a physical effect, and, additionally, an energy dependence may occur due to incorrect training. For this reason, a slight energy dependence is expected, making it challenging to demonstrate that the model with FIS is entirely energy-independent.

z-score

To analyse energy dependence, the z-score is used to evaluate the spectral behaviour after applying a cut. In general, the z-score is defined as:

$$z = x - x' \tag{6.20}$$

for each bin in a spectrum, where x is the expected number of events in a bin and x' is the measured number of events in a bin. In this analysis, an expected spectrum is approximated by randomly selecting N events from the full spectrum, where N is the total number of events surviving the cut in the observed energy range.

A plot of the z-score can be found in the lower panel of Figure 6.7 is obtained. A z-score close to zero across all energy levels indicates good agreement between expectation and result, confirming low energy dependence. If a clear gradient is observed, this suggests energy dependence in the model.

Since the spectra exhibit a linear energy dependence, a parameter Δz is defined by fitting z with:

$$z(E) = \Delta z \cdot E + c \tag{6.21}$$

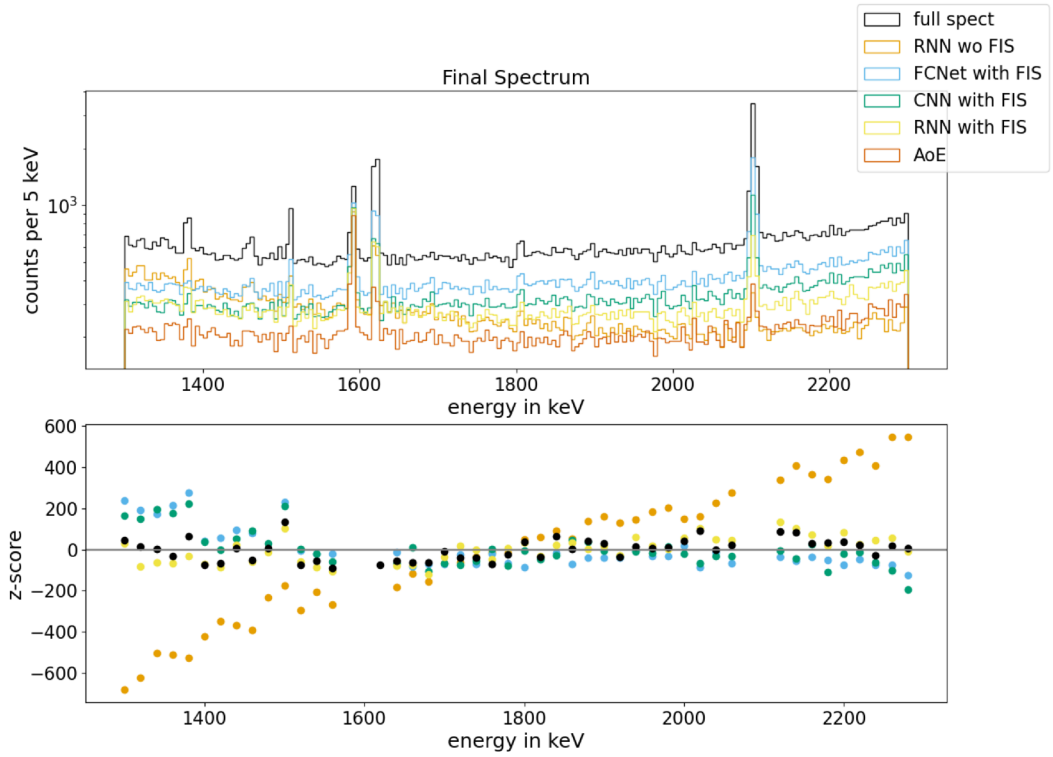


Figure 6.7.: Energy spectrum for different models with corresponding z-score. Since the ratio of SSE to MSE differs in DEP, FEP, and SEP, these peaks are excluded, and only the Compton continuum is considered. In the model without FIS, a clear energy dependence is observed, while the other models and A/E do not exhibit this behaviour.

An absolute higher Δz value indicates stronger energy dependence, while a value close to zero suggests an energy-independent model. The uncertainty in Δz provides additional information about the general noise in z. Combining both parameters offers an effective way to compare different models and training cycles.

The calculated Δz values for models trained with and without FIS can be found in Table 6.3, where a clear distinction between the two approaches is evident. For this first investigations, the masking from 5% to 95% is used, since it is enough for a proof of energy independence. By using $z(E = 2039\text{keV})$, it is possible to interpolate the influence of the energy dependence on $Q_{\beta\beta}$ and get an ideal value for the number of events cut out by the classifier $N_{c,id}$. At least, if the real value of cut events N_c is

known.

$$N_{c,id} = N_c + z(E) = N_c + \Delta z \cdot E + c \tag{6.22}$$

Out of this, the ratio between real and ideal can be determined and observed. Out of this, the parameter δ_E is defined as

$$\delta_E = 100 - \frac{N_{c,id}(E_{\text{Qbb}})}{N_c(E_{\text{Qbb}})} \cdot 100 = \frac{z(E_{\text{Qbb}})}{N_c(E_{\text{Qbb}})} \cdot 100 \tag{6.23}$$

to monitor the energy dependence, with $E_{\text{Qbb}} = 2039\text{keV}$. This gives the fluctuation caused by energy dependence of the survival fraction at $Q_{\beta\beta}$ in percent. The overall survival fraction in this region is usually around $30\% \pm 2\%$. These 2% in total are 6.7% relative to the 30%. We considered a deviation smaller than 5% of the uncertainties to be a normal fluctuation, while deviations below 10% were classified as small but acceptable effects, at least for the current state of development.

6.4.2. Different Architectures Combined with FIS

Even though FIS is a new method, it is built upon the use of neural networks and merely extends the training process. Therefore, a general network architecture must be selected. As a first step, the FIS method is tested with different types of architectures: a simple FCNet (technical similar to the ANN from Chapter 5), a slightly more complex CNN, and an RNN+att. An overview of the technical details is provided in Section 4.
Every choice of a network architecture involves a trade-off between performance, which is often directly linked to model complexity, and computational efficiency. A natural approach is to begin with a simple model with a minimal number of layers and increase complexity in subsequent steps. Due to this reasoning, the initial approach is a FCNet consisting solely of linear layers and dropout. If this model performs well enough, it would be the preferred choice, as such networks are computationally efficient. However, fully connected networks often struggle with complex tasks. Therefore, the next approach is a CNN, which improves performance on intricate problems by incorporating convolutional layers. Finally, RNN+att is selected as the third architecture since it is particularly well-suited for this task. Pulse shapes form time series data, and RNNs are specifically designed for such inputs, making RNN+att a logical choice. However, training an RNN+att is computationally intensive and takes more time, due to the higher complexity of this architecture (see Chapter 4).

Since it is necessary to provide some form of explanation metric, a vanilla gradient method is implemented for FCNet and CNN. This method computes the total gradient through the model for each input element, highlighting elements that have a greater impact on the output. For RNN+att, this additional implementation is unnecessary, as it is designed with an RNN+attention mechanism, which inherently provides an explanation metric (for further details, see Section 4.2.1).
The hyperparameters and details for the different architectures are summarized in Table 6.2. For the energy dependence evaluation, only the basic masking approach from 5% to 95% pulse height is applied.

	FCNet + FIS	**CNN + FIS**	**RNN + att + FIS**	**RNN + att**
Epochs	30	30	30	30
Optimiser	Adam	Adam	Adam	Adam
Loss	L_{FIS}	L_{FIS}	L_{FIS}	L_{BCE}
Learning Rate	0.00008	0.002	0.007	0.007
Momentum of Adam Optimiser	0.0009	0.0008	$2 \cdot 10^{-5}$	$2 \cdot 10^{-5}$
FIS Weights	1, 20.43, 54.56, 0.44, 8.23	1, 47.68, 46.73, 48.82, 40.0	1, 20, 15, 1.5, 20	0, 0, 0, 0

Table 6.2.: Additional information about the different hyperparameters for the applied architectures. The lower learning rate of FCNet+FIS slows down the training, but reduces the risk of skipping the global minimum. On the other hand, the probability that a local minimum will be found instead of the global minimum increases. This is compensated by the higher momentum of the optimiser. Since the CNN and RNN models are in general slower than an FCNet, a higher learning rate is used. The exact fine-tuning of this parameter was done with Ax [44]. For the exact implementation of the different architectures, see Appendix A.2

Table 6.3 presents the survival fractions for different architectures, giving an impression of the performance and plausibility of the applied cut from the different models, as

	SEP Tl [%]	FEP Bi [%]	FEP Tl [%]	CC@$Q_{\beta\beta}$[%]	Δz std/keV	δ_E[%]
Energy	2103 keV	1621 keV	2614 keV	2039 keV	-	-
A/E	6.7 ± 1.1	12.0 ± 1.5	7.4 ± 0.12	34.3 ± 1.0	0.024 ± 0.0006	0.7
Simulations	$< 20\%$	< 25 %	< 25 %	30 - 50%		
RNN w/o FIS	12 ±11	33 ±11	15 ±11	44 ±10	0.76 ±0.27	18 ± 11
FCNet+FIS	44 ±14	45 ±13	47 ±12	66 ±9	-0.15 ±0.06	0.2 ±0.06
CNN+FIS	32 ±6	35 ±5	38 ±6	59 ±5	-0.2 ±0.03	0.3 ±0.03
RNN+att+FIS	16 ±11	27 ±10	21 ±11	43 ±8	0.11 ±0.07	2.6 ±3

Table 6.3.: Different survival fractions for the various architectures. The survival fraction in the DEP is set to 90% in every case. While Δz gives the slope of the z-value, δ_E equates to the corresponding energy dependence at $Q_{\beta\beta}$ (see Equation 6.23). The statistical fluctuations result from 100 training cycles. Since ML models contain randomly chosen starting parameters for training, these fluctuations are expected. The lower survival fractions for the RNN+attention model are closer to the physically expected values, while the CNN and FCNet models perform worse.

discussed below. The general idea of calculating survival fractions in different peaks of the Th-spectrum is explained in 3.4.2. In addition, Figure 6.7 shows the corresponding energy spectra before and after applying the cut, along with the z-score for energy dependence.

In contrast to the RNN+att model without FIS, none of these models exhibit significant energy dependence. However, as will be discussed later (Section 6.5), certain training cycles can still lead to energy dependence due to the influence of the random seed. This is also why the uncertainties in the survival fractions presented in Table 6.3 are relatively high.

Beyond energy independence, all three models successfully train and are able to distinguish signal and background events. If this were not the case, the survival fractions across all peaks (DEP, SEP, FEP, $Q_{\beta\beta}$) would be similar. The expected trend is also maintained: the DEP fraction is fixed at 90%, while SEP should have the lowest proportion of SSE, followed by the FEP, while the Compton continuum at $Q_{\beta\beta}$ contains slightly more MSE than SSE.

One notable discrepancy is the ordering of survival fractions in the two FEP peaks. Using the classical A/E method, the expected order is SEP with the lowest survival

fraction, followed by FEP Th, and FEP Bi with the highest survival fraction. This is because higher energy events generally contain a greater fraction of MSE (higher energy → more MSE → lower survival fraction) and because SEP naturally has fewer MSE. However, in the FCNet+FIS and CNN+FIS models, the two FEP fractions appear more in the same order, whereas the RNN+att+FIS model correctly follows the expected trend.
Despite this anomaly, the RNN+att+FIS model significantly outperforms both the FCNet+FIS and CNN+FIS, which gives two reason for the choice of RNN+att as the fundamental model and against FCNet or CNN. The only reason against RNN+att+FIS are the high uncertainties, which will be investigated further in Chapter 7. But even if the uncertainties are smaller for CNN+FIS, the overall performance is much worse and the RNN+att+FIS therefore preferred.

6.5. Detailed Influence of Special Masking

In the second step, the primary architecture is fixed to an RNN+att+FIS with an attention mechanism, based on the previous comparison. The exact technical details of the RNN+att+FIS implementation can be found in Section A.3. Instead, this section focuses on analysing the impact of different FIS maskings to understand the extent to which FIS modifications influence performance.
Different ideas for this maskings were already explained and introduced in Section 6.2.1. In the following analysis, the 95-, amax- and fork-Model are investigated. Shortly summarised is the expectation, that the 95-model will give a first rough result, which will probably lack detail, since the model does not have very detailed information. For the amax-model, a result closer to A/E is expected. But as the A/E, it is not expected to be sensitive to features like QDrift or risetime, which needs information over a broader range of the rising edge. Last but not least, the fork-model is expected to give the most precise output, since it contains most information. But also, this amount of information can increase the uncertainties, since the model has to deal with a more complex input.

The training conditions remain the same across all three versions: Training is conducted over 30 epochs using an Adam optimiser with a weight decay of $2 \cdot 10^{-5}$ and a learning rate of 0.007. The loss functions are weighted using $w = [1, 20, 15, 1.5, 20]$.

Statistical Analysis

To conduct a statistical analysis, the training process is repeated 100 times without any

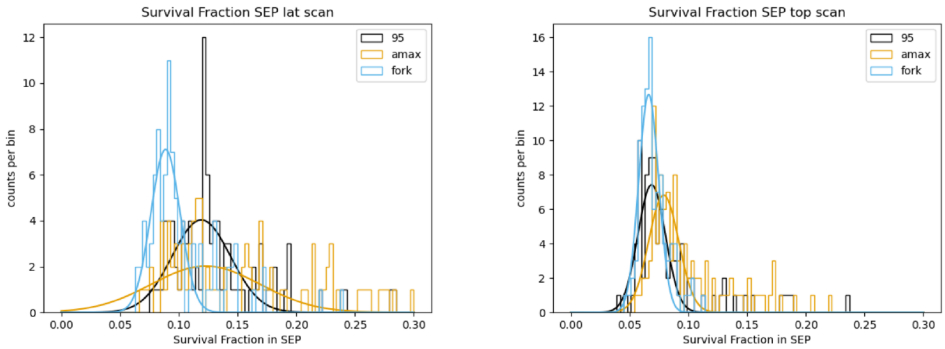


Figure 6.8.: Distribution and fit of the survival fraction in SEP for the latitudinal scan (left) and the top scan (right). Significant differences between the two scans are visible, along with some outliers towards higher fractions. By applying a Gaussian fit, these outliers carry less weight. The plots for all survival fractions can be found in the Appendix, Table C.5.

modifications. Since the model is complex and contains a large number of parameters, it is not entirely stable, and some variation in results across training runs is expected. The possible causes of this uncertainty are explored and discussed in detail in Section 7. At this stage, the primary aim is to observe general trends and behaviours across different maskings and scan configurations.

6.5.1. General Description of the Analysis

Since QDrift is expected to influence the results, the effect of different scan configurations is of interest. From HADES measurements, two distinct scan types are available: a source positioned on top (top scan) and a source placed on the side (latitudinal scan). These scans allow for an analysis of the effects of risetime and QDrift.
In the top scan, a larger proportion of events deposit energy in the upper part of the detector, leading to a higher risetime. In contrast, the latitudinal scan results in a more evenly distributed set of interaction points throughout the detector's active volume. Since energy depositions in the upper part of the detector produce electron-hole pairs where the holes collect additional charge as they travel through the detector, the risetime and QDrift tend to be higher. This wider spatial distribution of interactions results in two distinct peaks in the QDrift distribution (see Section 3.4.2).

While both risetime and QDrift reflect these effects, the following analysis focuses specifically on QDrift due to the higher stability and sensitivity of this value.

Observed Parameter

Alongside the different survival fractions, δ_E is calculated to assess the energy dependence of each model, as described in Section 6.4.1. Additionally, the performance of the RNN+att+FIS model is compared to the traditional A/E classifier, as shown in Figure 6.9. The ratios R_{diag} is determined as the ratio of events, where A/E and RNN+att+FIS are in alignment. The reverse parameter would be R_{offdiag}, but is split into R_{top} and R_{bottom}. Corresponding to Figure 6.9, R_{top} show the ratio of events selected as SSE by RNN+att+FIS, but as MSE by A/E. While R_{bottom} is defined the other way around. These calculations, as well as the δ_E evaluation, consider only events with energies between 1300 keV and 2300 keV.

All calculated parameters for different masking strategies and scan configurations are listed in 6.4. The key patterns, differences, and performance trends are discussed in the following sections. To select the "best" model, the one with the lowest SEP survival fraction while maintaining an energy dependence below a defined threshold $\delta_E < 5\%$ is chosen for the latitudinal scans, while the limit is $\delta_E < 10\%$ for top scans. This is acceptable, since the top scans have a slight energy dependence, also in A/E, which is due to the broader SSE band and therefore higher fluctuations in the A/E energy correction.

Additionally, two types of mean values are reported:

- Overall mean value $\varnothing$ – A direct average of all training runs.
- Gaussian fit mean μ – A mean value obtained by fitting a Gaussian function to the distribution of results. This method reduces the impact of models that exhibit strong energy dependence or poor classification performance.

Due to this approach, the overall mean tends to be higher than the fitted mean for all survival fractions.

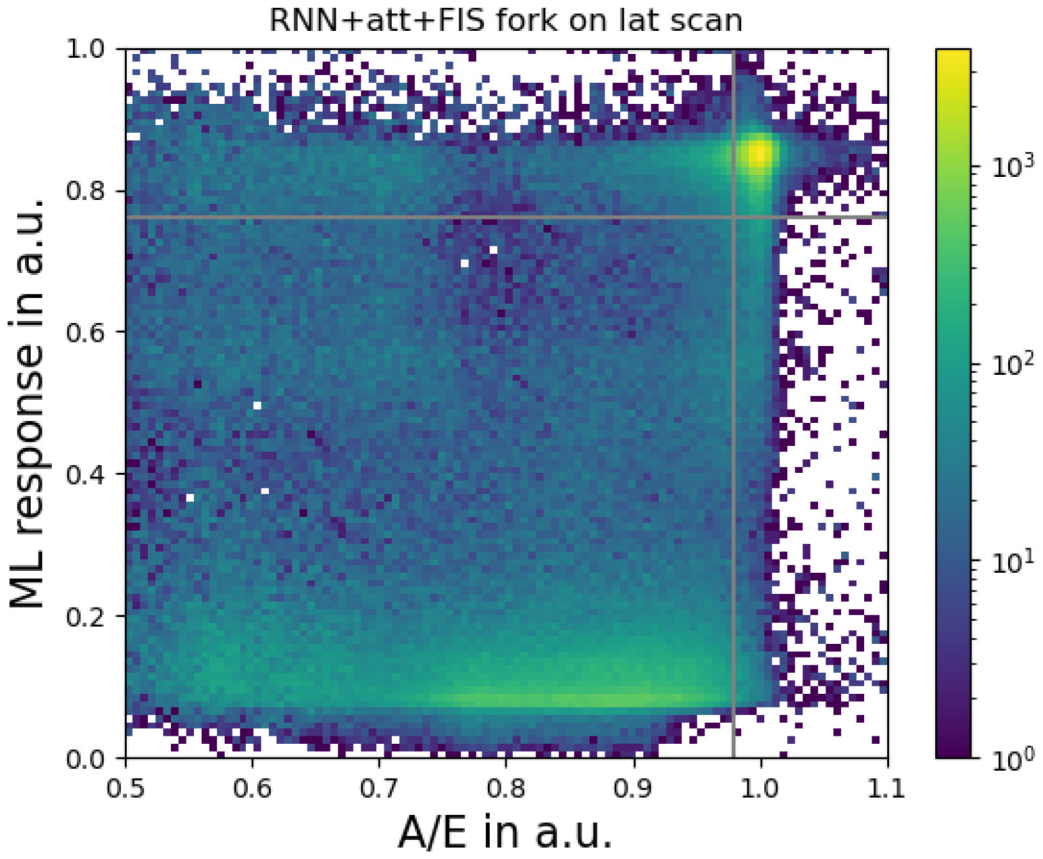


Figure 6.9.: Comparison of the A/E classifier with the RNN+att+FIS model. The both lines show either the RNN+att+FIS or the A/E cut, in both cases the events below the cut value are rejected. Therefore, for the diagonal events in this plot, both cuts are in alignment. For the upper left part, the events are accepted by RNN+att+FIS, but rejected by A/E and vice versa for the lower right part.

Table 6.4.: Table with all calculated values for different scans and maskings.

Model	**Type**	**SEP ^{208}Tl [%]**	**FEP ^{214}Bi [%]**	**FEP ^{208}Tl [%]**	**CC@Qbb [%]**	**Δz [std/keV]**	**δ_E[%]**	**R_{top} [%]**	**R_{bottom} [%]**	**diag [%]**
energy		2103 keV	1621 keV	2614 keV	2039 keV	-	-	-	-	
Top										
A/E		4.9 ± 0.8	8.3 ± 1.1	4.40 ± 0.09	29.8 ± 0.9	0.32 ± 0.0009	9.0	-	-	
A/E rt		5.4 ± 0.8	8.3 ± 1.1	5.4 ± 0.1	30.9 ± 0.9	0.24 ± 0.0009	6.5	-	-	
95	⌀	14±19	22±17	18±18	42±13	0.31±0.14	7.4±3.3	12±12	4.7±1.3	83±11
95	μ	6.9±1.2	15.8±3.5	10.5±2.9	36±4	0.31±0.08	8.3±2.6	6.4±2.7	5.0±1.2	88.0±3.2
95	best	4.8±0.8	12.4±1.1	6.75±0.11	31.4±0.9	0.2953 ±0.0009	8.61	3.16	5.79	91.05
amax	⌀	10±4	17±5	11±5	41±7	0.42±0.10	9.8±2.6	11±6	4.6±0.9	84±6
amax	μ	7.9±1.2	13.9±1.8	9.0±2.1	38±5	0.42±0.11	9.6±2.4	8±4	4.5±1.0	87.0±3.4
amax	best	6.5±0.8	13.9±1.1	8.77±0.12	35.2±0.9	0.317 ±0.001	8.2	5.92	4.9	89.18
fork	⌀	7.0±1.2	15.8±3.0	10.3±2.9	36±4	0.42±0.10	10.9±2.2	7.8±3.3	5.5±0.8	86.7±3.0
fork	μ	6.6±0.9	15.0±2.7	9.3±2.2	34.7±2.5	0.41±0.11	10.6±2.2	6.7±2.8	5.5±0.9	87.5±3.2
fork	best	5.0±0.8	16.0±1.1	8.76±0.12	34.7±0.9	0.3634 ±0.001	9.69	5.41	5.38	89.2
Lat										
A/E		7.8 ± 2.3	12.9 ± 2.6	9.0 ± 2.4	36.3 ± 2.5	$(-9 \pm 67) \cdot 10^{-5}$	0.03	-	-	
A/E rt		6.7 ± 1.1	12.0 ± 1.5	7.4 ± 0.12	34.3 ± 1.0	0.024 ± 0.0006	0.7	-	-	
95	⌀	16±11	27±10	21±11	43±8	0.10±0.08	2.7±1.8	14±7	3.6±1.1	83±7
95	μ	11.9±2.4	23±4	17±4	40±4	0.11±0.07	2.7±1.7	11.4±3.1	3.6±1.1	85.1±3.0
95	best	7.9±1.0	13.7±1.5	9.66±0.14	33.3±1.0	0.111 ±0.0004	3.52	5.08	4.6	90.33
amax	⌀	18±10	25±11	22±11	47±11	0.14+/-0.12	3.2±2.4	17±10	2.9±1.0	80±9
amax	μ	12±4	17.6±3.1	15±4	41±7	-0.00±0.21	2.7±2.4	11±4	2.8±1.1	86.5±3.2
amax	best	6.8±1.0	13.2±1.5	8.98±0.14	33.2±1.0	0.1422 ±0.0006	4.57	5.13	4.51	90.36
fork	⌀	12±9	26±9	20±9	42±7	0.22±0.08	5.8±2.3	14±6	4.2±1.1	82±6
fork	μ	8.9±1.3	24±5	17±4	39±4	0.21±0.08 ±	5.6±2.2	11.9±3.4	4.3±1.0	83±4
fork	best	6.9±1.1	16.9±1.5	12.22±0.16	36.2±1.0	0.0829 ±0.0004	2.39	7.13	3.59	89.28

6.5.2. Analysis Output

The detailed investigations of the different maskings at different scans show at many points reasonable results, showing that FIS is in general a promising method. The following section will explain the observations over different training runs in detail, since there are many interesting characteristics.
All in all, a striking observation is the considerable difference in results between the top and latitudinal scans, indicating that QDrift or risetime plays a significant role in influencing model performance.
Beside this point, the performance of individual training runs varies significantly. While some runs produce exceptional results, others deviate strongly from the expected behaviour, forming notable outliers. This behaviour is a general risk and will be discussed in Section 7.

Latitudinal Scan Energy Dependence

An analysis of the energy dependence of different models for the latitudinal scan reveals that most models exhibit $\delta_E < 5\%$ and are therefore below our definition of an energy independent model. Amax- and 95-model are both in the same range, just the fork models are in μ and $\varnothing$ slightly above 5%, but within the uncertainties (5.6% $\pm$2.2% respectively 5.8% $\pm$2.3%). In total, 9% of the trained 95-models are energy dependent, 21% of the amax- models and 59% of the fork-models. This exception can be attributed to residual energy dependence within QDrift, which persists in the pulse shapes.
Eliminating this effect would likely require normalisation of the risetime on the waveform level with respect to the energy. Such a normalisation can have additional side effects and does not correspond to the idea of a waveform that is as unprocessed as possible.
By looking at the training run selected as best fork model, it appears that the fork model can also be definitively energy independent with a great performance, there are just other training runs with a clear energy dependence.

Top Scan Energy Dependence

One of the most prominent observations from the top scan is that it closely approximates the A/E survival fractions, much more so than the latitudinal scan. However, this scan exhibits a higher δ_E value, a broader SSE band, and larger fluctuations in the z-score. This effect is appears for both, A/E and RNN+att+FIS.

As an general effect of this measurement type it is acceptable and the threshold is therefore increased to $\delta_E = 10\%$. Within this new range, 18% of the trained 95-models are energy dependent, 43% of the amax- models and 55% of the fork-models, showing a higher impact for 95- and amax model, but not for the fork-model.

Latitudinal Scan Survival Fraction

As explained in 3.4.2, the survival fractions in different peaks of the spectrum give a good insight in the performance of a PSD cut. To determine the cut, the survival fraction is set to 90% in the DEP, and expected to be low in SEP and FEP. As mentioned in 3.4.2, the exact value from simulation depend on the definition of the maximum volume in which the event has to deposit its energy to count as a SSE. Especially how the distance between the different points of energy deposition is defined can make a hugh difference. Simulations were not done as part of this work, but this effect can be seen in previous analysis on this topic. [41][42]
The known A/E survival fractions could be used as a reference point, however it is not the goal of this work to reproduce A/E. Even if it is a good tool for PSD, it is not perfect. Likewise, the improvement of a working ML tool for PSD would be in containing different or additional features.

Examining the survival fractions sf derived from μ, the following pattern emerges:

$$\begin{aligned}
\text{SEP:}\quad & sf^{\mu}_{\text{SEP, amax}} \gtrapprox sf^{\mu}_{\text{SEP, 95}} > sf^{\mu}_{\text{SEP, fork}} \\
& \Delta sf^{\mu}_{\text{SEP, amax}} > \Delta sf^{\mu}_{\text{SEP, 95}} > \Delta sf^{\mu}_{\text{SEP, fork}} \\
\text{FEP:}\quad & sf^{\mu}_{\text{FEP, fork}} \approx sf^{\mu}_{\text{FEP, 95}} > sf^{\mu}_{\text{FEP, amax}} \\
& \Delta sf^{\mu}_{\text{FEP, fork}} \gtrapprox \Delta sf^{\mu}_{\text{FEP, 95}} \gtrapprox \Delta sf^{\mu}_{\text{FEP, amax}}
\end{aligned}$$

This provides strong evidence that incorporating QDrift information, as in the fork model, impacts classification performance. Interesting is the fact, that it seem to be an positiv impact by looking at the survival fraction in the SEP, but a negative impact on the FEP. With the amax-model, it is the other way around, it gives the best results on the FEP.
The reason for this is not clear and should be investigated further, especially since it is possible to find an explanation for a good performance of both, amax and fork. This

is due to the fact, that the knowledge about QDrift, at it is implemented in the fork model, keeps high advantages, but also risks.

Top Scan Survival Fractions

Following nearly the same pattern as for the latitudinal scan, the fitted survival fractions exhibit the following order:

$$\begin{aligned}
\text{SEP:} \quad & sf^{\mu}_{\text{SEP, amax}} > sf^{\mu}_{\text{SEP, 95}} \approx sf^{\mu}_{\text{SEP, fork}} \\
& \Delta sf^{\mu}_{\text{SEP, amax}} \approx \Delta sf^{\mu}_{\text{SEP, 95}} > \Delta sf^{\mu}_{\text{SEP, fork}} \\
\text{FEP:} \quad & sf^{\mu}_{\text{FEP, 95}} > sf^{\mu}_{\text{FEP, fork}} \gtrapprox sf^{\mu}_{\text{FEP, amax}} \\
& \Delta sf^{\mu}_{\text{FEP, 95}} > \Delta sf^{\mu}_{\text{FEP, fork}} > \Delta sf^{\mu}_{\text{FEP, amax}}
\end{aligned}$$

One difference is the fact, that the fraction in FEP is worse for 95-model than for the fork model. This can be explained by the low influence of QDrift in the top scan. Since most events in this scan originate from the upper part of the detector, where QDrift is naturally high, it does not serve as a discriminating feature. In such a scenario, the point where A is maximal contains nearly all the relevant information, making it sufficient for an effective classification model. Both the amax and fork configurations retain this information, while the 95 model appears to capture too much surrounding information, which may dilute its effectiveness.
Another difference are the general smaller uncertainties than for the latitudinal scan. The most reasonable explanation for this point is the smaller variance between the different waveforms due to the higher population of events in the upper part of the detector.

Off-Diagonal Elements

The overall trend is that more events are accepted by RNN+att+FIS but rejected by A/E than vice versa ($R_{\text{top}} > R_{\text{bottom}}$). This aligns with the generally higher survival rate of events passing the RNN+att+FIS cut compared to A/E, reflected in the different survival fractions, especially at $Q_{\beta\beta}$, which is part of the compton continuum.
Closer investigations about the events in the offdiagonal elements were not successful, it is currently not possible to see any significant structure of this events.

Best Performing Model

By selecting the best performing models, it is possible for all scans and maskings to get a well performing model. In direct comparison, 95- or amax- model seem to perform better, for both, top and latitudinal scan, while for the fork-model especially the high survival fraction in the FEP of ^{214}Bi is eye-catching. But since the fork-Model takes QDrift into account, open different possibilities, it is worth to investigate further. Especially, since the model is very unstable for the moment and it would be methodically questionable to pick one well performing model out and use it as base for futher analysis.

General Conclusion

Several factors influence the model's performance, and in some cases, these factors interact in contradictory ways. A particularly challenging issue is the effect of QDrift (especially in the fork-Model) and its inherent energy dependence, which cannot be entirely eliminated without significant additional effort.

Due to the high uncertainties in the results, a definitive analysis is not feasible. However, the general trends in performance across different masking techniques can be reasonably explained by the information each masking method retains. The possibilities to get a stable training and therefore the base for future refinement and closer analysis are discussed later on in Section 7.

6.6. Feature Importance Supervision on Semi-Coaxial Detectors

As part of this work, the FIS was tested as a new possibility for PSA on the Coax detectors. But the resolution of the pulses of the coax detectors are much lower than for the ICPCs. This leads to much higher uncertainties by using the FIS at they appear in the ICPCs already.
In general, the RNN turned out to be a to complex model for this type of task and it worked better with a CNN approach. Especially by using a CNN+FIS model with slightly different masking, which include the tail of the pulse to the important part. Even with this adjustments, the training was instable and it will have to be further investigated, before it can be used in the official analysis. Figure 6.10 gives a short impression of what the FIS would be capable of. The model is not adjusted to the higher amount

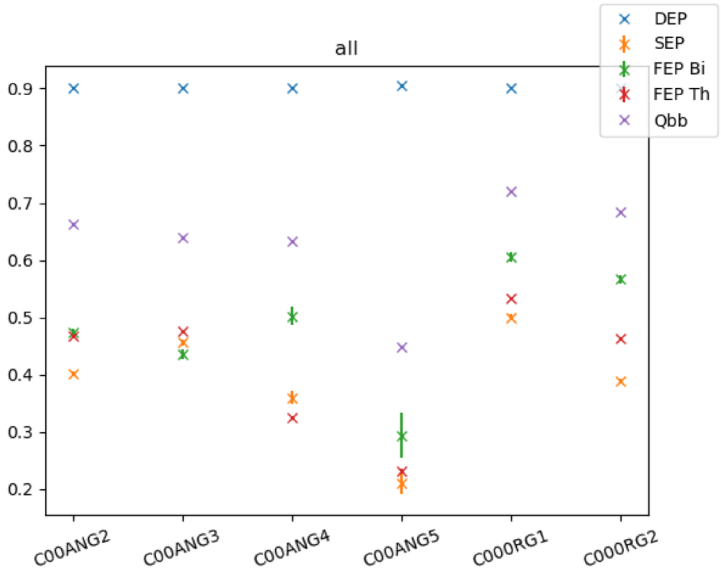


Figure 6.10.: Survival fractions of different detectors for one of the better training cycle with CNN. The uncertainties just include the binomial fluctuations for the survival fraction calculation, but not the training uncertainties.

of events in the training dataset for C00ANG4, which leads to an energy dependence in this detector. But for the other detectors, the survival fractions are in the right range and there is no or just a small energy dependence observed (see figure 6.11.
Especially C00ANG5 has an incredible low survival fraction and seems to outperform the ANN in this trainingcycle.

The FIS was not investigated further for the coaxial detectors, especially since we expect a worse performance on this type of detectors compared to the ICPC in general.

6.7. Conclusion

An energy-independent model can be effectively developed using a FIS approach, shown by a value of $\delta_E < 5$ for the tested FCNet, CNN and RNN+att approaches. Given the complexity and fine details of pulse shapes, a combination of a RNN+att with FIS has proven to be the most effective in balancing energy independence and detailed analysis of the pulses.

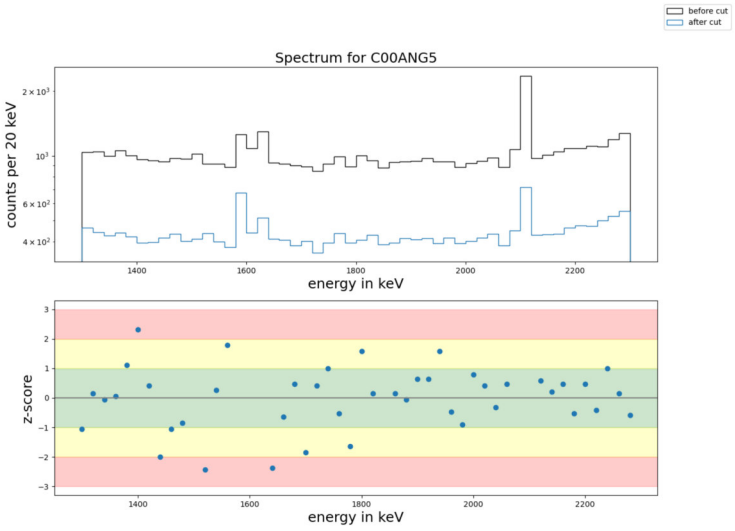


Figure 6.11.: An exemplary energy dependence investigation for C00ANG5 with the corresponding spectrum.

The use of different masking techniques within the FIS framework allows for the integration of significant physical knowledge into the model. This influences its general predictions in a controlled manner and different, but from a physical point of view reasonable, behaviours can be seen for different maskings. From the tested maskings, the fork model contains most physical knowledge and would be therefore preferred, but have flaws in its performance.

However, the current training process exhibits uncertainties. They are especially high in the fork model, since this masking contain most information and therefore offer the highest freedom to the model. The possible reasons for this uncertainty in general will be discussed in the next chapter.
If this instabilities can be resolved, FIS keep the possibility to combine different cuts and corrections since it is sensitive to QDrift, a sensitivity to LQ can not seen until know due to the high fluctuation, but is also reasonable.
Also it is possible, that the fork model perform way better, if the reason for the uncertainties can be eliminated. But this has to be investigated further, especially since the fork model incorporates some information similar to an uncorrected QDrift, which is in general energy dependent.

As an additional point, the performance of FIS was tested on the semi-coaxial detectors, but it lacks precision, mainly due to the low resolution of this type of detector.

7. Mislabelled Data in the Training Set

As shown in the last chapter, the principle of FIS works in general for PSA of HPGe pulses. This chapter focuses on possible explanations for the significant variation in outcomes depending on the random seed used during training. This chapter will now focus on the possible reason for this fluctuations and discuss, how to deal with it.

One possibility for the high fluctuations in the different training runs are the mislabelled data. DEP and SEP are naturally no clear selection of SSE or MSE, in addition the compton background add more impurities to the trainingdata. In previous ML approaches at LEGEND or its precursors MAJORANA and GERDA, this impurities were seen as acceptable. That is why the approach was not questioned at first.
A closer look shows around 25% of the events in DEP are labeled as MSE by the A/E cut (see figure 7.1). Studies show, that most ML models can handle mislabeled data between 10% and 20% in a acceptable manner. But even with this percentage, the portion of wrong classified data increase and can be critical for a low background experiment, where efficiency and precision is crucial.
In any case, the amount of mislabelled data would be an obvious reason for the training instabilities.

ML Methods to suppress mislabelled data

Since a completely clear training dataset does not correspond to reality, there are several possibilities to deal with impurities in the training data.
A simple possibility is a mixup augmentation [56], where a linear combination of two events, one background and one signal-like is used as an input, as well as a linear combination of its labels as input label.

In addition, an Attentive Feature Mixup (AFM) was tested, a method provided by [57] to suppress mislabelled data. It uses a grouping mechanism and training with attention

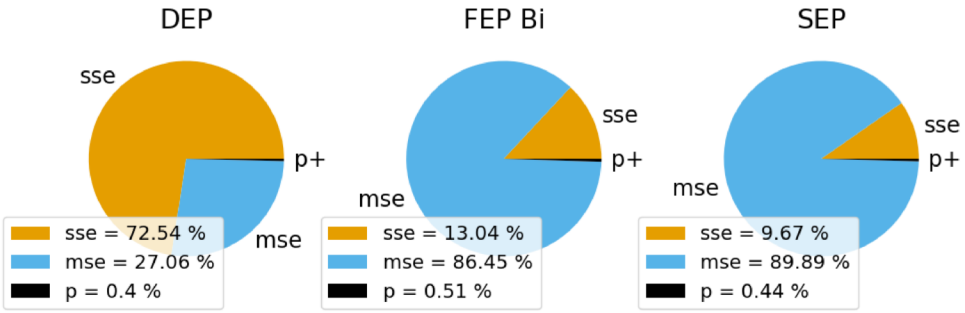


Figure 7.1.: Amount of events selected by A/E as SSE, MSE or p^+ surface events in DEP, SEP and FEP.

score, to determine a weight for each data point of the training data, which leads to a lower impact of wrong labelled data. The results in applying this mechanism to the FIS model can be found in Appendix Table D.1.

This methods where tested, but non of them lead to any improvement. They just slow down the training, but give no better performance. This is a clear sign, that there is a limit reached and it would be a better approach to find a way to clean the training data before.

7.1. Preselection via A/E cut

An obvious method to get a cleaner dataset is to do a preselection with the A/E cut. To get an improvement from FIS to A/E, it is not meaningful to use the exact same cut, but it is a good starting point to test the theory, that the instability mostly result from the mislabelled data.

Training with correct A/E cut

As a first check, some training cycles were done with an additional A/E cut on the training data. So just events in the DEP with A/E above the low cut value are accepted as SSE, and vice versa for the MSE labels. Events with A/E above the high cut are ignored completely for the moment.
This labels lead to a high agreement between A/E and ML response, as shown in table

7.1. The high agreement can be seen in the low fraction of off diagonal elements as well as the survival fractions. Since the training was also really fast, it was interupted after 5 epochs. But even with this short training, some runs ended in overtraining, as can be seen in the high uncertainty of ⌀ for 95-model or of the fork-model for latitudinal scan.

Table 7.1.: Table with all calculated values for different scans and maskings for a training with an additional A/E cut to the training dataset. The training was repeated 10 times. The results are in high agreement to the A/E values.

Model	**Type**	**SEP ^{208}Tl [%]**	**FEP ^{214}Bi [%]**	**FEP ^{208}Tl [%]**	**CC@Qbb [%]**	**Δz**	**δ_E**	**R_{top}**	**R_{bottom}**	**diag**
energy		2103 keV	1621 keV	2614 keV	2039 keV	-	-	-	-	
Top										
A/E		4.9 ± 0.8	8.3 ± 1.1	4.40 ± 0.09	29.8 ± 0.9	0.32 ± 0.0009 z	9.0	-	-	
A/E rt		5.4 ± 0.8	8.3 ± 1.1	5.4 ± 0.1	30.9 ± 0.9	0.24 ± 0.0009 z	6.5	-	-	
95	⌀	15±23	20±21	15±22	41±16	0.21±0.10	5.3±2.0	9±16	3.2±0.4	88±16
95	μ	5.9±0.5	10.75±0.30	6.2±0.6	33.4±0.6	0.227±0.018	6.0±0.6	2.04±0.19	3.21±0.06	-28±12
95	best	5.5±0.8	10.4±1.1	5.63±0.10	32.9±0.9	0.2391 ±0.0009	6.52	1.66	3.38	94.96
amax	⌀	7.5±1.8	12.8±2.1	8.3±2.4	36.2±2.4	0.20±0.05	4.9±1.3	3.8±1.6	3.1±0.4	93.2±1.4
amax	μ	6.7±1.5	12.2±1.8	7.1±1.6	35.3±1.6	0.205±0.008	5.27±0.07	3.1±1.0	3.2±0.4	93.8±0.5
amax	best	5.9±0.8	11.1±1.1	6.34±0.11	34.2±0.9	0.223 ±0.0009	5.81	2.39	3.54	94.08
fork	⌀	5.98±0.35	10.9±0.7	6.0±0.4	33.3±1.0	0.232±0.032	6.3±1.1	1.81±0.33	3.3±0.7	94.9±0.5
fork	μ	5.9±0.4	10.9±0.7	5.8±0.6	33.71±0.31	0.23±0.05	6.3±1.3	1.7±0.4	3.4±0.8	94.9±0.4
fork	best	5.3±0.8	10.8±1.1	5.28±0.10	32.2±0.9	0.2518 ±0.001	7.06	1.42	3.96	94.62
Lat										
A/E		7.8 ± 2.3	12.9 ± 2.6	9.0 ± 2.4	36.3 ± 2.5	$(-9 \pm 67) \cdot 10^{-5}$	-	-	-	-
A/E rt		6.7 ± 1.1	12.0 ± 1.5	7.4 ± 0.12	34.3 ± 1.0	0.024 ± 0.0006	-	-	-	-
95	⌀	11±6	16±5	13±6	39±5	-0.00±0.06	0.1±1.2	7±5	1.02±0.33	92±4
95	μ	8.5±1.2	13.3±0.8	9.9±1.5	36.2±1.6	0.027±0.029	0.5±1.1	4.6±1.4	1.1±0.5	94.1±1.2
95	best	7.0±1.0	12.9±1.5	8.21±0.13	34.9±1.0	0.0275 ±0.0005	0.81	3.17	1.41	95.42
amax	⌀	8.6±1.6	14.2±2.1	10.1±2.3	36.1±3.3	0.09±0.07	2.8±1.6	5.7±2.1	2.3±0.7	92.0±1.7
amax	μ	8.4±1.7	14.3±2.6	9.9±2.3	35.7±2.9	0.10±0.06	3.2±1.6	5.8±2.5	2.5±0.5	91.8±2.0
amax	best	6.5±1.0	12.3±1.5	7.93±0.13	32.8±1.0	0.1032 ±0.0005	3.32	3.45	2.87	93.68
fork	⌀	12±10	18±10	13±10	39±9	0.03±0.08	1.3±2.0	8±7	1.6±1.0	91±7
fork	μ	7.1±1.3	12.53±0.16	8.3±2.5	35±4	0.07±0.05	1.6±2.1	3.5±2.9	1.3±0.8	94.3±1.7
fork	best	5.6±1.0	11.1±1.5	6.05±0.12	31.0±1.0	0.1022 ±0.0004	3.48	1.53	2.66	95.81

Effective monitoring and adjustment of the A/E parameter throughout the different measurement periods in L200 are critical for ensuring a well performing A/E cut. Preliminary observations from current L200 data indicate that the A/E cut value fluctuates within a range of 0.89σ across different detectors.
To get an balance between using the A/E cut for data cleaning and still get an improvement in manpower and effort through the usage of ML, a coarse cut is proposed. In the best case, it is possible to apply such a cut without any fine-tuning to the training data and it will be enough for the FIS to outperform this rough cut by handling the remaining mislabelled datapoints.

Varying applied A/E cut value

In contrast to the previous training with applied A/E labels, an abortion condition was implemented to prevent overfitting, which mostly terminates the training after 4-7 epochs.
Training with varying A/E cuts demonstrated a clear dependence of model performance on the cut value, as illustrated in Figure 7.2. In general, this behaviour was expected, but it was hoped for a broader distribution. The current valley, associated with the best performance, barely cover a area of $\leq 1\sigma$. This range is approximately the same as the observed fluctuations over the different periods in L200, but to use a coarse cut without A/E fine-tuning and confidential results, a larger range would be required.

In general, the fast training indicates that the model can quickly learn the relationship between A/E values and their corresponding labels, resulting more in a imitation of A/E than an independent model. These findings imply that the inclusion of an initial rough A/E cut does not improve the ML performance in a sensible manner.

7.2. Analysis of the Training Data

Since applying a rough A/E cut for training does not yield satisfactory results, an alternative approach must be considered. Before exploring other methods to address the mislabeling in the training data, a more detailed examination of the dataset is conducted. A thorough understanding of the different pulse types can aid in proposing a solution for the impure training data while also providing a better overview of the general possibilities and challenges.
To analyse the training data more closely, a Principle Component Analysis (PCA) ([58]) and subsequently a t-distributed Stochastic Neighbour Embedding (t-SNE) ([59]) were

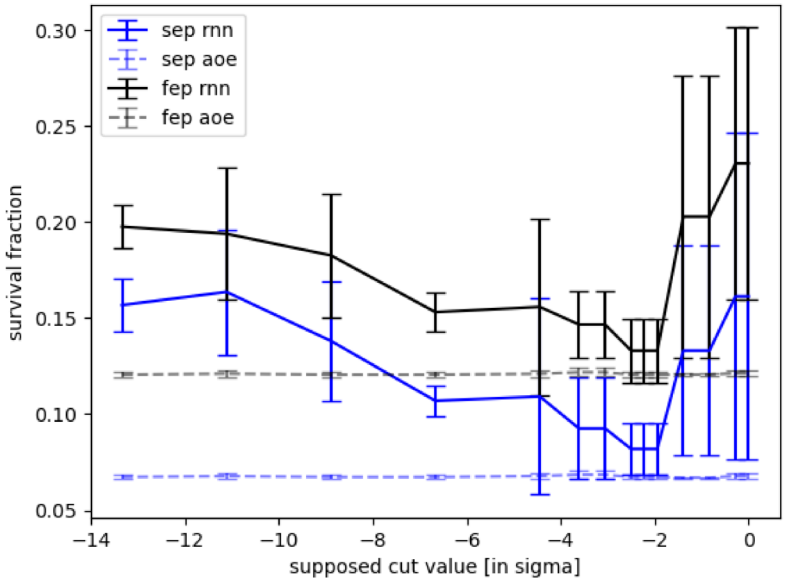


Figure 7.2.: This plot shows the behaviour of the survival fractions from a RNN+att+FIS output under different pre-selected training datasets. This selection is done by applying an A/E cut but changing the cut value stepwise to the values on the x-axis. The correct A/E cut value is at 0.98, exactly where the optimal results appear and the RNN+att+FIS is closest to the A/E fractions.

used to categorize the data into different clusters. PCA is a linear dimensionality reduction technique that reduces pulses of 256 samples to a size of 10. The fundamental principle of PCA is to determine a transformation matrix $\mathbf{W}$ that reduces the dimensionality of an input matrix $\mathbf{X}$ via $\mathbf{W} \cdot \mathbf{X}$ while preserving as much variance as possible. For more information, check also [60].
t-SNE in contrast, is a nonlinear dimensionality reduction technique designed for data visualization. Unlike PCA, which applies a linear transformation, t-SNE calculates the probability of two points being neighbors using a Gaussian or Student's t-distribution.

The resulting t-SNE plot is presented in Figure 7.3. Based on the distribution of events, seven clusters were manually defined, and the corresponding waveforms are depicted in Figure 7.2.
Several notable characteristics emerge within these categories:

- Label 1 (26.1%) and Label 2 (15.8%): Both are in general SSE, but PCA successfully differentiates between events occurring in the upper and lower regions of the detector. From simulations, it is known that pulses originating in the lowest centimeter of the detector exhibit a characteristic kink at their starting point, which can be observed in the waveforms of Label 2 events. Additionally, their QDrift values align with the lower peak of the QDrift distribution observed in DEP, whereas events in Label 1 correspond to the higher peak.

- Label 3 (23.1%): This category contains events with high LQ and low A/E, consistent with expectations for n^+ surface event events and certain MSE.

- Label 4 (6.5%) and Label 5 (15.2%): These labels correspond to clear MSE events. Label 4 consists exclusively of MSE with an $A/E \leq 0.8$. A closer inspection of the current pulse for Label 4 events reveals that many contain three local maxima—an effect correlated with bremsstrahlung peaks in the SEP.

- Label 6 (1.8%): The events in this category exhibit distinct SSE pulses, yet they are incorrectly classified as MSE by A/E. Although a small fraction, this confirms that some events are mislabeled by A/E.

- Label 7 (11.6%): Events in this category are ambiguous and do not fit clearly into any classification. To further analyze these events, a separate PCA and t-SNE analysis was performed, as illustrated in Table 7.3.

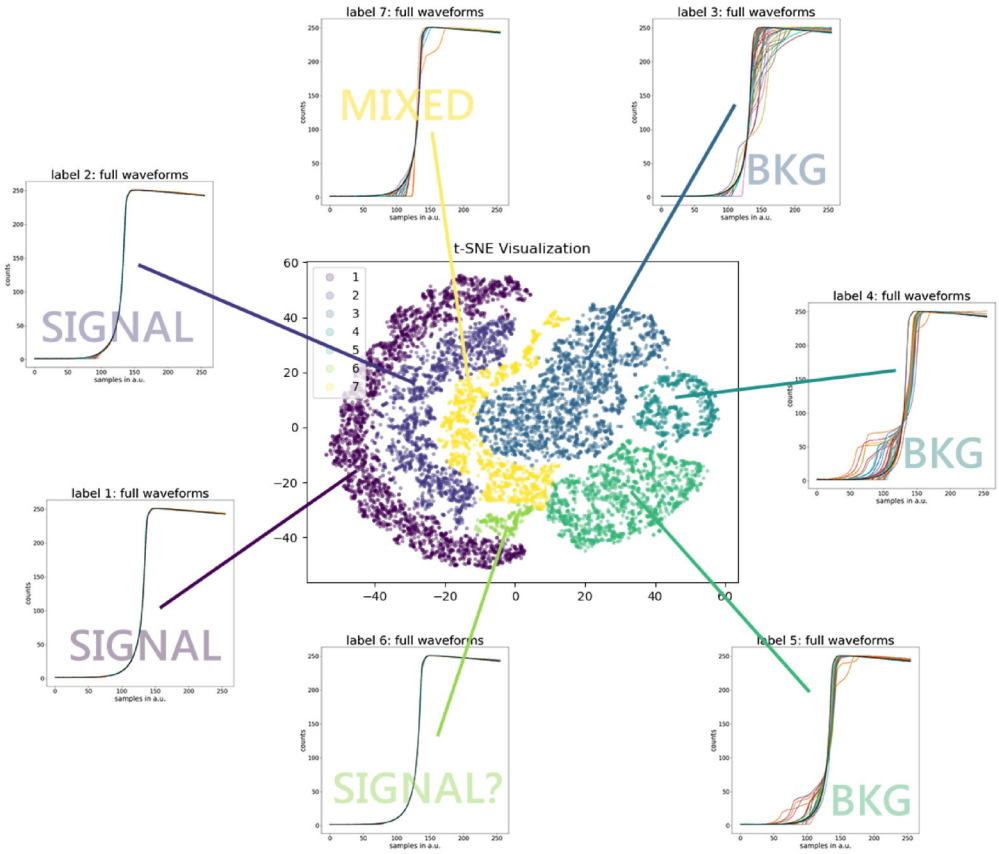
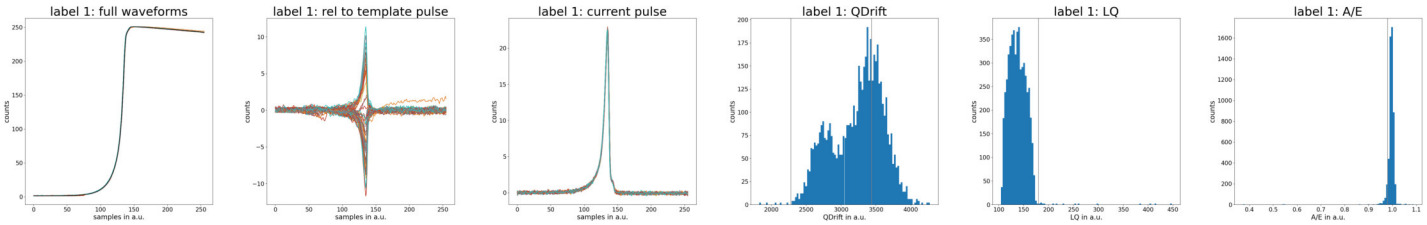
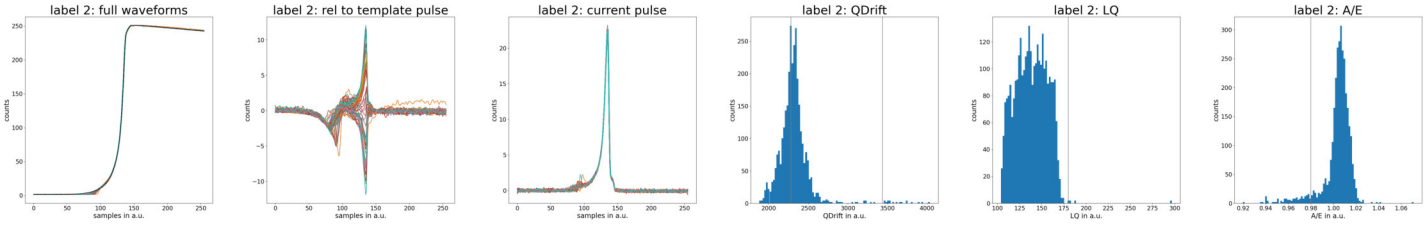


Figure 7.3.: 2D representation of events inside the training dataset after PCA and t-SNE. The highlighted clusters are handselected by using the structures from t-SNE, representing distinct waveform types as plotted on the side. The exact boundaries between these clusters are not rigid and could be adjusted. In Table 7.2, the corresponding pulse shapes, as well as their A/E, QDrift, and LQ values, are presented to demonstrate the meaningfulness of the selected clusters.

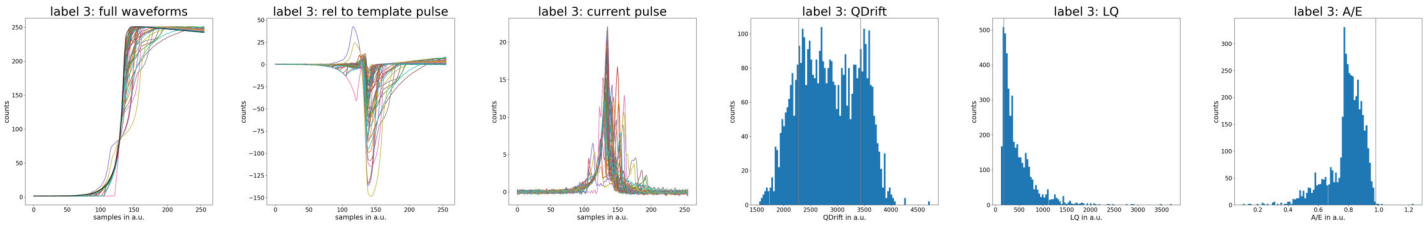
Label 1 (26.1%): SSE pulses (upper part of the detector)



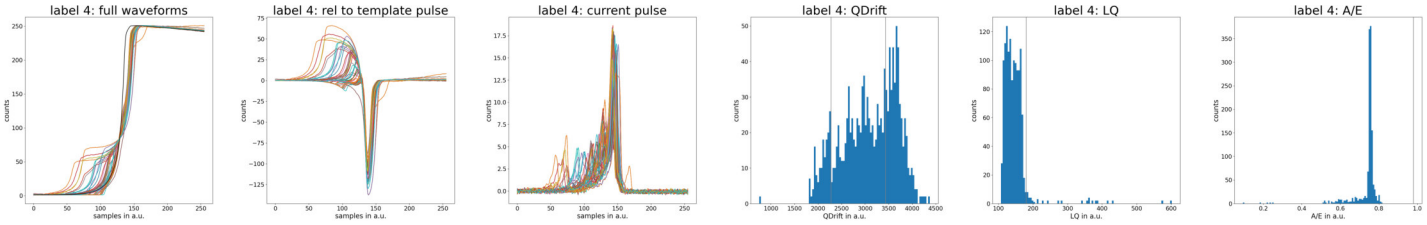
Label 2 (15.8%): kinked SSE pulses (lower part of the detector)



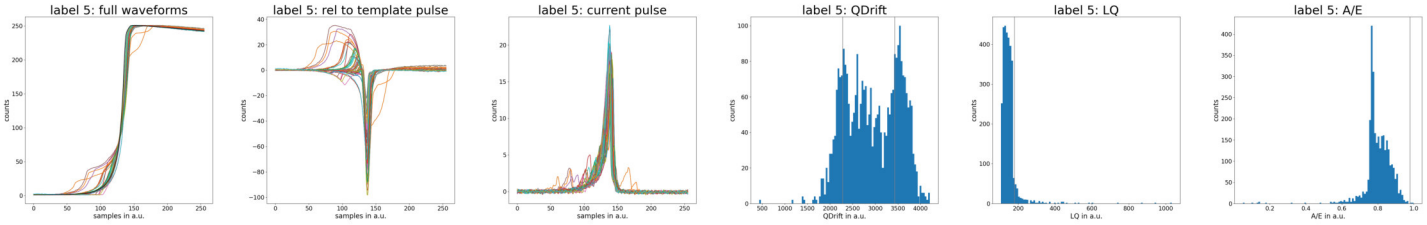
Label 3 (23.1%): background, cut out by A/E, maybe n^+ surface events?



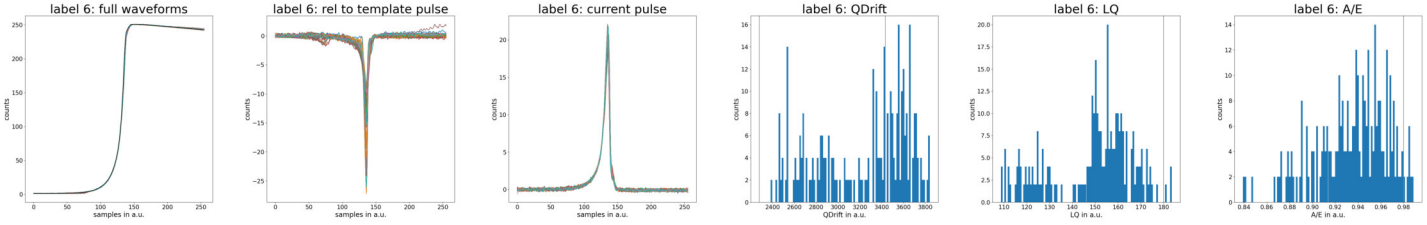
Label 4 (6.5%): background, strong MSE, cut out by A/E



Label 5 (15.2%): background, cut out by A/E



Label 6 (1.8%): SSE? but cut of by A/E



Label 7 (11.6%): Mixed population

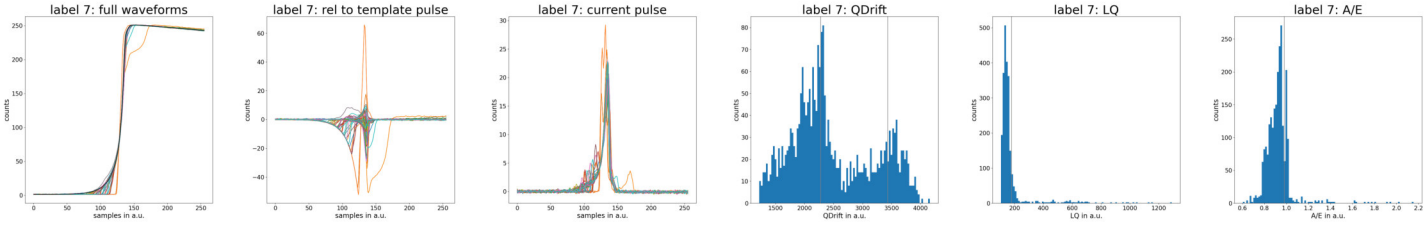


Table 7.2.: Different parameter corresponding to the different identified categories. From left to right: 100 waveforms, distance between waveform and template pulse, current pulse, QDrift, LQ and A/E

A preliminary attempt to refine the training labels by using only events from Labels 1 to 6, while discarding the ambiguous events in Label 7, was unsuccessful. This outcome is unsurprising, as the indistinct events in Label 7 are likely the primary contributors to the instability observed during training.
Against this suggestion, it was not possible to find a coincidence between the events classified with Label 7 here and the off-diagonal events from Section 6.5. But independent of this, they can induce a general instable training.
To gain a better understanding of these events, they were reclassified, revealing the following characteristics:

- Label 7.1 (3.0%): These events resemble SSE waveforms that were rejected by A/E, yet they do not necessarily exhibit MSE characteristics. Some of them display slight artifacts, but none deviate significantly from the template pulse ($< \pm 10$). The probability of these events being falsely rejected is high, as the A/E cut is set conservatively—90% of DEP events were accepted, whereas simulations suggest an expected acceptance of approximately 93-95%.

- Label 7.2 (2.5%): These events are similar to SSE, but with a slower rising edge—possibly indicating proximity to the n^+ surface. The deviation from the template pulse is larger than in Label 7.1, approximately ± 30. This raises the question of what level of deviation from the template pulse is acceptable, particularly in the upper region: at what point does an event transition into an n^+ surface event?

- Label 7.3 (1.9%): These events closely resemble those in Label 2, but all exhibit kinked waveforms (low QDrift and a visible kink).

- Label 7.4 (0.6%): These events correspond to MSE that deposited energy near the p^+ surface, as described in Section 6.1.2. Some were accepted by A/E, while others were rejected by LQ, as expected.

- Label 7.5 (1.4%): These events resemble those in Label 4 but do not appear to be as clear-cut MSE. This could make them particularly important for classification. Further separating them using A/E, reveals that they are primarily kinked

waveforms located near the p^+ contact. Those rejected as MSE by A/E exhibit an additional artifact in the upper portion of the rising edge, resembling MSE or n^+ surface event. The from A/E accepted events have a pronounced kink, while the remainder are clear p^+ surface event.

- Label 7.6 (0.8%): These events appear to be kinked waveforms, but they were mostly accepted as SSE by A/E.

- Label 7.7 (1.5%): These events also exhibit kinked waveforms, but with a particularly pronounced kink, leading to 70% of them being rejected by A/E. Further investigation is needed to determine whether they originate from SSE or another source. If they are indeed SSE, this effect warrants further study, as the presence of kinked waveforms is expected to increase with larger detector masses, making them crucial for future detector development.

Several of these observations require methods to classify pulses not only by A/E, like the events with Label 7.1 to 7.3 are cut out by A/E, but seem to be SSE anyway. Instead, additional parameters such as deviation from a template pulse, LQ, QDrift, and current pulse characteristics should be incorporated. Some categories appear to represent superpositions of waveforms influenced by the n^+ and/or p^+ surfaces, or different bulk volume regions with varying electric fields.
For a more thorough investigation and a clearer classification of background versus signal-like events, detailed simulations of these mixed waveform types could be valuable. These simulations should then be compared against the mentioned parameters. Once a classification decision is reached for each category, techniques such as PCA combined with t-SNE (or similar clustering methods) could be used to preselect a training dataset and improve label quality. Even if this method does not yield perfect labels, it would be a significant step toward reducing misclassification in training data.

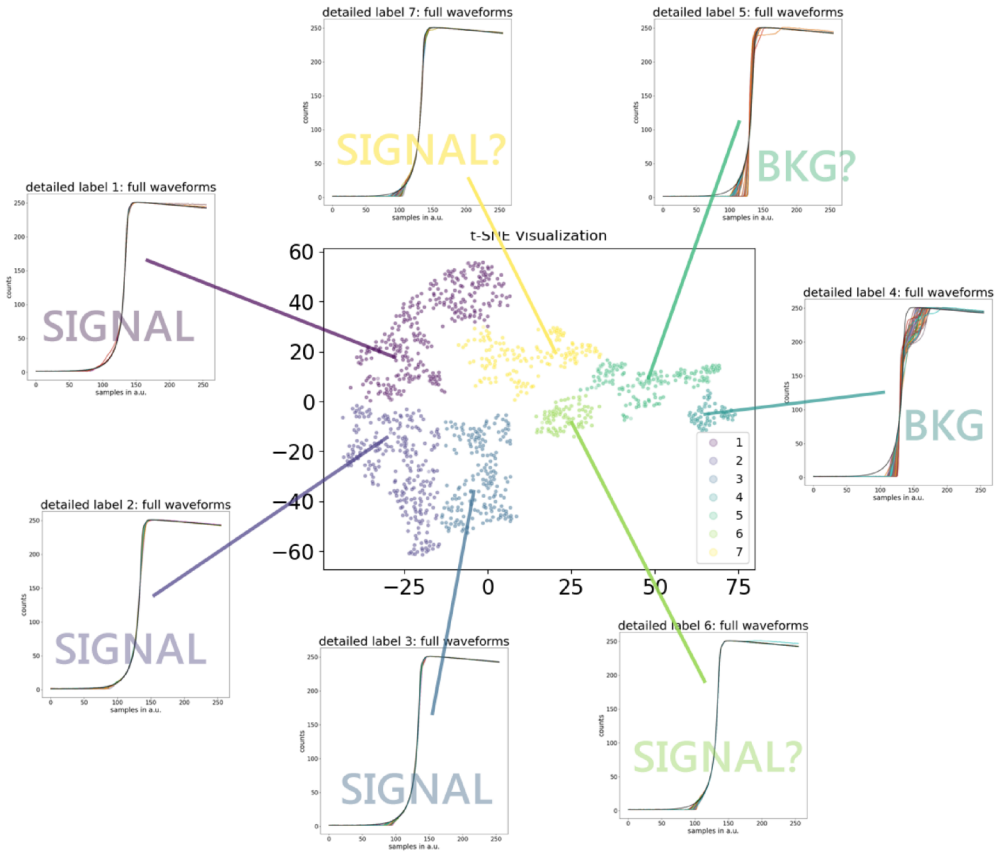
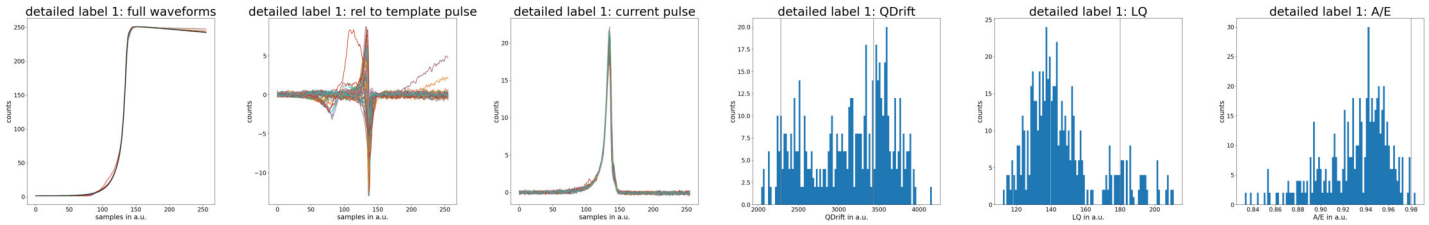
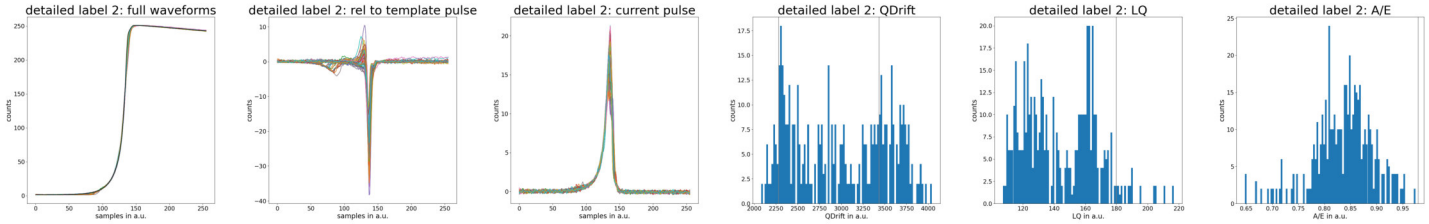


Figure 7.4.: Representation of the different classes from PCA, similar to Figure 7.3, but focused only on events that were not classified meaningfully in the initial run. In Table 7.3, the corresponding pulse shapes, as well as their A/E, QDrift, and LQ values, are presented to demonstrate the meaningfulness of the selected clusters.

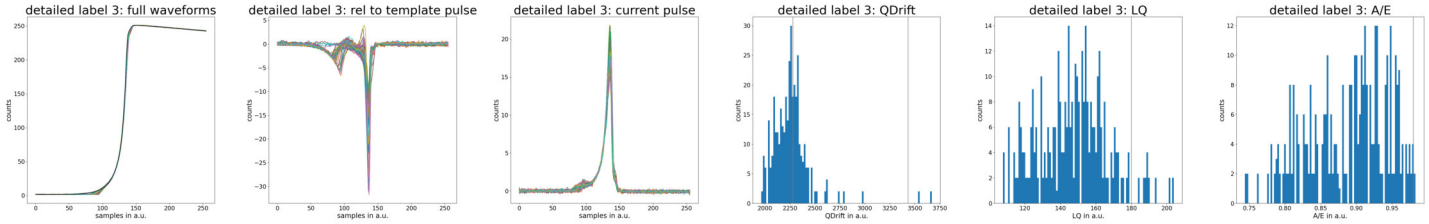
Label 7.1 (3.0%): SSE? but cut of by A/E



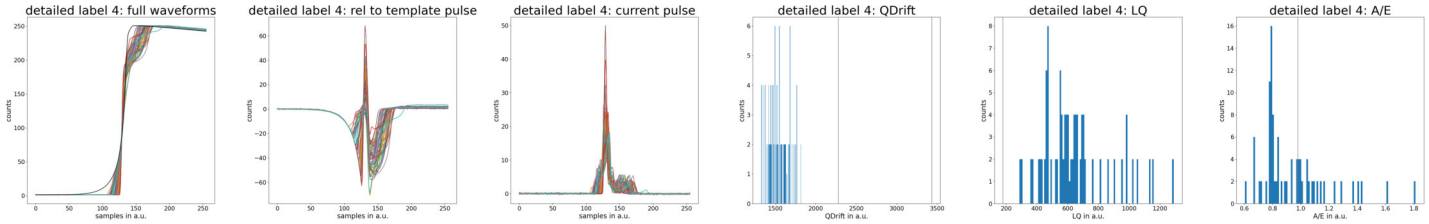
Label 7.2 (2.5%): SSE? but cut of by A/E



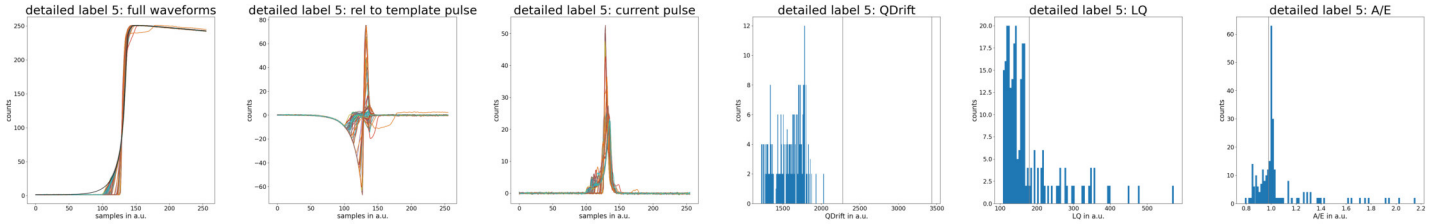
Label 7.3 (1.9%): SSE? but cut of by A/E



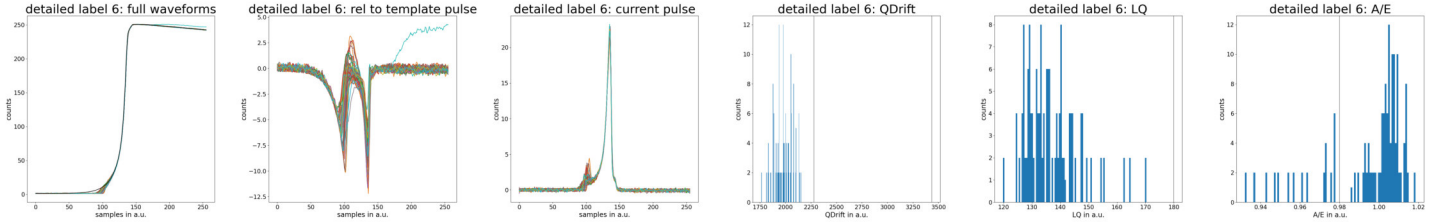
Label 7.4 (0.6%): background (MSE on p^+, cut out by LQ)



Label 7.5 (1.4%): mixed?



Label 7.6 (0.8%): SSE?



Label 7.7 (1.5%): SSE?

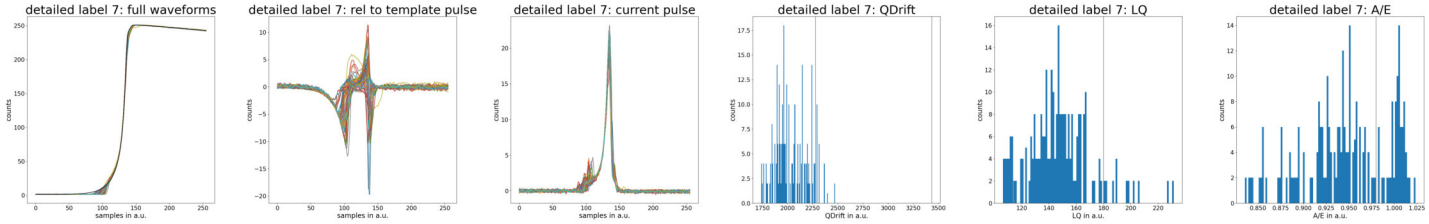


Table 7.3.: Different parameter from the different categories previously grouped under the ambiguous Label 7. From left to right: 100 waveforms, distance between waveform and template pulse, current pulse, QDrift, LQ and A/E.

7.3. Conclusion

The instability observed during training, as discussed in the FIS chapter, is most likely due to mislabeled data within the training dataset. Existing machine learning methods to reduce the impact of mislabeled data, such as linear mixup or AFM, have proven ineffective. This is due to the inherent reduction in information within these approaches, which suggests that alternative methods should be explored.

Preselection using A/E enables the model to replicate A/E outputs effectively; however, this approach does not yield significant improvements. A rough cut selection is not feasible without fine-tuning the A/E, resulting in no improvement through the RNN+att+FIS compared to the existing A/E cut.
A fundamental challenge remains that the model may learn to exploit the underlying parameters used for preselection,
A rough preselection solely based on PCA is not viable unless the most challenging 12% of cases are excluded. The expected approach to preselection involves a hybrid method combining machine learning clustering/visualization algorithms with manual selection. This process would integrate multiple factors, including A/E outputs, LQ, Q-drift, current pulse characteristics, deviations from template pulses, and general pulse shape analysis. A more detailed investigation using simulations is necessary, particularly for critical waveform classes.

Future efforts will focus on refining the training dataset, as previously attempted with PCA, and evaluating the effectiveness of a robust preselection strategy. If this proves insufficient, further testing of additional methods to suppress mislabeled data will be pursued.

8. Conclusion and Outlook

High sensitivity and a low background is crucial for the LEGEND experiment, as explained in Chapter 3. To achieve this background, alongside active and passive shielding, a robust data analysis is essential. A key component of this data analysis for the HPGe detectors is PSA, which is currently performed using the A/E analysis, combined with an additional LQ cut.
The only exception is the Coax detectors, where the PSD is handled by an ANN. This ANN is already implemented in GERDA and is now successfully adapted to the new LEGEND software as part of this work (see Chapter 5).
For the Coax detectors, any type of PSA is limited by their field geometry, whereas the pulses from BEGe, PPC, and ICPC detectors have better resolution and contain more information about the underlying event. This raises the question of how ML methods can be used to improve the PSA in these detector types—particularly in the ICPCs, which are intended for use in L1000.
Especially considering the large number of detectors in L1000, a potential PSD approach that does not require detector-wise fine-tuning would offer a significant advantage and might be achievable using ML. Additionally, an alternative PSA method based on ML could serve as a crosscheck for the existing analysis chain and provide greater flexibility in the overall analysis, including adaptability to possible future detector geometries.

To investigate a viable method for such an alternative PSA, FIS is explored in Chapter 6. It represents a promising concept that allows prior physical knowledge to be passed to the model so it can produce more faithful results.
The first major benefit of FIS is greater flexibility in selecting training data, without introducing energy dependence in the model. This opens up the possibility of using events in the SEP as background data instead of the FEP. Since the SEP contains more background events than the FEP, this approach is preferable — but it was not feasible with commonly used models like FCNet, CNN or RNN due to the larger energy difference between DEP and SEP.

This concept is tested using different model architectures: FCNet+FIS, CNN+FIS, and RNN+att+FIS, with RNN+att+FIS emerging as the best-performing variant.
The second advantage of FIS lies in the idea of highlighting specific features known to be important for distinguishing between different pulse types. By applying different masking strategies, targeted information can be provided to the FIS model, and this leads to varying results. These differences are interpretable based on the amount and nature of the information provided to the model, demonstrating that the overall conceptual idea. However, a remaining challenge is the high variation between multiple training runs with fixed hyperparameters. While minor fluctuations are normal in ML, the level of variation observed in RNN+att+FIS is critical, especially given the high sensitivity requirements in LEGEND. For this reason, it is currently not possible to favour one specific masking approach over another.

Additionally, it is necessary to investigate the reasons for the uncertainties observed across different training runs. This is addressed in Chapter 7. The most probable reason for the varying training outcome is the incorrect assumption that plain data labelling using DEP and SEP/FEP is sufficient. These peaks can be easily selected by their energy, but the necessary labels are SSE (25% of the events in DEP) and MSE (10% in SEP). The extent to which a model can handle mislabelled data in the training dataset varies, but in general, a ratio above 20% of mislabelled data noticeably impacts model performance. For high-sensitivity applications, this threshold can be even lower.
Different methods designed to handle mislabelled data are tested, but they do not yield improvements. As a result, it becomes necessary to remove at least part of the mislabelled data through alternative approaches.
A test using A/E to preselect the training data leads to replication of A/E, as the model appears to simply mimic the A/E parameter. This results in a fast training process but fails to offer any improvement over the classical A/E cut. The idea of using a less fine-tuned A/E cut for preselection does not lead to better outcomes, since the fluctuations in the cut value over different periods are comparable to the fluctuations seen when a rough A/E cut is applied directly to the training data—which still leads to acceptable results.
A more in-depth analysis of the training dataset using a clustering algorithm is possible and results in several distinct classes of pulses. These classes exhibit different characteristics in parameters such as A/E, LQ, and QDrift, as well as differences in their overall pulse shape and in the deviation between the pulse shape and a reference tem-

plate pulse. This observation opens future possibilities for improving training datasets through preselection strategies or through a hybrid approach that combines preselection with ML-based methods designed to handle mislabelled data.

In the future, the studies carried out on the composition of the training data can be used to find out whether this improves FIS training and reduce the uncertainties. Therefore, various techniques are possible, like ML clustering algorithms, a preselection with combined analysis parameters or methods to suppress mislabelled data in the training dataset. In order to make a decision for one masking, a new investigation of the different maskings in view of all cuts and corrections of the classical PSA is proposed afterwards. Especially with respect to critical events which are difficult to select by A/E. Also it can be followed by the examination of other maskings or a combination of different maskings, each for another type of event.

9. Acknowledgements

First of all, I want to thank Prof. Dr. Josef Jochum for the supervision during the whole time of my PhD. Even in chaotic times, you provided my the security to show up and ask for advise. Also I am grateful for your support in terms of alternative means of transportation to several meetings.
Thanks to Prof. Dr. Tobias Lachenmaier for being my second supervisor, but - sometimes way more important - a great coffee collaboration and the diverse discussions in the lunch break.

During my PhD, I had a great time at the University of North Carolina for several months. In this connection, I want to thank asst. Prof. Dr. Julieta Gruszko for the support she provided me, as well as asst. Prof. Dr. Aobo Li for the great collaboration during my time at UNC, but also the assistance over the full time of my PhD.
I would like to mention the Reinhard-Frank-Stiftung in this context, who offered my this opportunity by founding my stay in North Carolina.

Special thanks goes to the whole LEGEND Group in Tübingen. Especially Dr. Lukas Rauscher, the best office mate I ever had and Dr. Ann-Kathrin Schütz, for all your advice and encouragement.
But also to my colleagues Gina, Jessi, Dr. Katja, Lukas, Marc, Dr. Tobi^2 - not all LEGEND, but everyday social interaction and often necessary distraction or moral support.

Danke an meine Familie, die beste Unterstützung, die man sich denken kann. Dafür, dass ich mich immer auf euch verlassen kann, egal ob die Sonne scheint oder die Zeichen gerade auf Sturm stehen. Ich liebe euch.

All die anderen Menschen in meinem Leben aufzuzählen, die mich über die Jahre hinweg unterstützt und begleitet haben, würde den Rahmen sprengen - aber fühlt euch alle gedrückt, ohne euch wäre ich nie so weit gekommen.

List of Abbreviations

Symbols

$0\nu\beta\beta$ **decay** Neutrinoless double beta decay. 11–14, 16–19, 21, 23, 32, 38, 70

$2\nu\beta\beta$ **decay** two neutrino double beta decay. 16, 17, 19, 32

$Q_{\beta\beta}$ Q value of $2\nu\beta\beta$. 22, 58–60, 62, 64, 65, 85, 87, 95

A

AFM Attentive Feature Mixup. 101

ANN Artificial Neural Network. 12, 55–58, 60, 61, 63–65, 97, 117

B

BEGe Broad Energy High-purity Germanium Detector. 27, 30, 31, 34, 36, 38, 117

C

CNN Convolutional Neural Network. 43, 47, 49, 67, 85–88, 97, 117

CNN+FIS Convolutional Neural Network with Feature Importance Supervision. 87, 88, 118

Coax Semi-coaxial detector. 24, 27, 30, 31, 34, 36, 42, 55, 57, 96, 117, 140

D

DEP Double Escape Peak. 35, 36, 38, 39, 41, 56–62, 81, 82, 84, 87, 94, 101, 102, 107, 111, 117, 118

F

FCNet Fully Connected Neural Network. 43, 44, 49, 67, 85–88, 97, 117

FCNet+FIS Fully Connected Neural Network with Feature Importance Supervision. 86–88, 118

FEP Full Energy Peak. 35, 36, 39, 41, 56–60, 62, 63, 81, 84, 87, 88, 94–96, 102, 117, 118

FIS Feature Importance Supervision. 11, 12, 43, 64, 67, 68, 71, 76, 80–85, 87, 88, 93, 96–99, 101, 102, 105, 117–119

FNN Feedforward Neural Network. 43–45, 47, 49, 50

G

GERDA Germanium Detector Array. 12, 19, 21, 23, 27, 41, 55, 56, 58, 60, 63, 117

GRU Gated Recurrent Unit. 50, 51

H

HADES High Activity Disposal Experimental Site. 24, 27, 35, 38, 43

HPGe High Purity Germanium Detector. 21–24, 27–32, 37, 42, 49, 101, 117

I

ICPC Inverted coaxial p-Type point contact detector. 27, 29, 31, 34, 36, 37, 39, 67, 69, 82, 96, 97, 117

L

L1000 LEGEND 1000. 11, 21, 70, 117

L200 LEGEND 200. 12, 21, 22, 24, 29, 32, 35, 39, 42, 55, 58, 65, 67, 69, 105

LAr Liquid Argon. 22, 23

LEGEND Large Enriched Germanium Experiment for Neutrinoless $\beta\beta$ Decay. 11, 12, 17, 19, 21, 22, 26, 27, 29, 34, 49, 55, 64, 117, 118

LQ Late Charge. 40, 69–71, 98, 107, 109, 111–113, 117, 118

LSTM Long-Short Term Memory. 50, 51

M

MAJORANA MAJORANA Demonstrator. 23, 41

ML Machine Learning. 11, 12, 27, 42, 43, 49, 55, 67–69, 73, 76, 81, 83, 87, 94, 101, 102, 105, 117–119

MLP Multilayer Perceptron. 43, 55

MSE Multi-Site-Event. 32, 35, 36, 39–41, 55, 57–59, 69–72, 78, 79, 81–84, 87, 90, 101, 102, 107, 111, 112, 118

O

ORNL Oak Ridge National Laboratory. 24

P

PCA Principle Component Analysis. 105, 115

PPC P-Type Point Contact Detector. 30, 31, 34, 38, 117

PSA Pulse Shape Analysis. 11, 12, 24, 27, 30, 31, 33–36, 42, 43, 68, 69, 71, 81, 96, 101, 117, 119

PSD Pulse Shape Discrimination. 34–36, 39, 41, 42, 94, 117

Q

QDrift Charge Drift. 38, 40, 69, 70, 72, 79, 88, 89, 93–96, 98, 107, 109, 111–113, 118

R

ReLU Rectified Linear Unit. 44, 46

RNN Recurrent Neural Network. 43, 44, 49–52, 85, 86, 117

RNN+att Recurrent Neural Network with attention score. 43, 67, 85–88, 97

RNN+att+FIS Recurrent Neural Network with attention score and Feature Importance Supervision. 87, 88, 90, 91, 93, 95, 106, 115, 118

S

SEP Single Escape Peak. 35, 36, 39, 41, 57–60, 62, 63, 81, 82, 84, 87, 94, 101, 102, 107, 117, 118

SiPM Silicon Photomultiplier. 23

SSE Single-Site-Event. 32, 35–41, 57–59, 63, 69–72, 81, 84, 87, 90, 94, 101, 102, 107, 111, 112, 118

T

t-SNE t-distributed Stochastic Neighbour Embedding. 105, 107

Bibliography

[1] E. C. G. Sudarshan and R. E. Marshak. "Chirality Invariance and the Universal Fermi Interaction". In: *Physical Review* 109.5 (Mar. 1958), pp. 1860–1862. ISSN: 0031-899X. DOI: 10.1103/physrev.109.1860.2.

[2] R. P. Feynman and M. Gell-Mann. "Theory of the Fermi Interaction". In: *Physical Review* 109.1 (Jan. 1958), pp. 193–198. ISSN: 0031-899X. DOI: 10.1103/physrev.109.193.

[3] M. Goldhaber, L. Grodzins, and A. W. Sunyar. "Helicity of Neutrinos". In: *Physical Review* 109.3 (Feb. 1958), pp. 1015–1017. ISSN: 0031-899X. DOI: 10.1103/physrev.109.1015.

[4] Y. Fukuda et al. "Evidence for Oscillation of Atmospheric Neutrinos". In: *Physical Review Letters* 81.8 (Aug. 1998), pp. 1562–1567. ISSN: 1079-7114. DOI: 10.1103/physrevlett.81.1562.

[5] Q. R. Ahmad et al. "Direct Evidence for Neutrino Flavor Transformation from Neutral-Current Interactions in the Sudbury Neutrino Observatory". In: *Physical Review Letters* 89.1 (June 2002), p. 011301. ISSN: 1079-7114. DOI: 10.1103/physrevlett.89.011301.

[6] K. Eguchi et al. "First Results from KamLAND: Evidence for Reactor Antineutrino Disappearance". In: *Physical Review Letters* 90.2 (Jan. 2003), p. 021802. ISSN: 1079-7114. DOI: 10.1103/physrevlett.90.021802.

[7] M. Fukugita and T. Yanagida. "Barygenesis without grand unification". In: *Physics Letters B* 174.1 (June 1986), pp. 45–47. ISSN: 0370-2693. DOI: 10.1016/0370-2693(86)91126-3.

[8] B. Pontecorvo. "Inverse beta processes and nonconservation of lepton charge". In: *Zh. Eksp. Teor. Fiz.* 34 (1957), p. 247.

[9] B. Pontecorvo. "Mesonium and anti-mesonium". In: *Sov. Phys. JETP* 6 (1957), p. 429.

[10] Ziro Maki, Masami Nakagawa, and Shoichi Sakata. "Remarks on the Unified Model of Elementary Particles". In: *Progress of Theoretical Physics* 28.5 (Nov. 1962), pp. 870–880. ISSN: 0033-068X. DOI: 10.1143/ptp.28.870.

[11] Max Aker et al. "Direct neutrino-mass measurement based on 259 days of KATRIN data". In: *Science* 388.6743 (Apr. 2025), pp. 180–185. ISSN: 1095-9203. DOI: 10.1126/science.adq9592.

[12] X. Qian and P. Vogel. "Neutrino mass hierarchy". In: *Progress in Particle and Nuclear Physics* 83 (July 2015), pp. 1–30. ISSN: 0146-6410. DOI: 10.1016/j.ppnp.2015.05.002.

[13] S. Gariazzo. "Neutrino Masses in Cosmology". In: *Moscow University Physics Bulletin* 79.S1 (Dec. 2024), pp. 202–209. ISSN: 1934-8460. DOI: 10.3103/s0027134924701662.

[14] Peter W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Physical Review Letters* 13.16 (Oct. 1964), pp. 508–509. ISSN: 0031-9007. DOI: 10.1103/physrevlett.13.508.

[15] Matteo Agostini et al. "Toward the discovery of matter creation with neutrinoless $\beta\beta$ decay". In: *Reviews of Modern Physics* 95.2 (May 2023), p. 025002. ISSN: 1539-0756. DOI: 10.1103/revmodphys.95.025002.

[16] M. Agostini et al. "Final Results of GERDA on the Search for Neutrinoless Double Beta Decay". In: *Physical Review Letters* 125.25 (Dec. 2020). ISSN: 1079-7114. DOI: 10.1103/physrevlett.125.252502. URL: http://dx.doi.org/10.1103/PhysRevLett.125.252502.

[17] G. Anton et al. "Search for Neutrinoless Double-βDecay with the Complete EXO-200 Dataset". In: *Physical Review Letters* 123.16 (Oct. 2019), p. 161802. ISSN: 1079-7114. DOI: 10.1103/physrevlett.123.161802.

[18] Koichi Ichimura. "Recent results from KamLAND-Zen". In: *Proceedings of Neutrino Oscillation Workshop — PoS(NOW2022)*. NOW2022. Sissa Medialab, Dec. 2022, p. 067. DOI: 10.22323/1.421.0067.

[19] D. Q. Adams et al. "Search for Majorana neutrinos exploiting millikelvin cryogenics with CUORE". In: *Nature* 604.7904 (Apr. 2022), pp. 53–58. ISSN: 1476-4687. DOI: 10.1038/s41586-022-04497-4.

[20] A. Agrawal et al. "Improved Limit on Neutrinoless Double β Decay of ^{100}Mo from AMoRE-I". In: *Physical Review Letters* 134.8 (Feb. 2025), p. 082501. ISSN: 1079-7114. DOI: 10.1103/physrevlett.134.082501.

[21] A.S. Barabash et al. "Calorimeter development for the SuperNEMO double beta decay experiment". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 868 (Oct. 2017), pp. 98–108. ISSN: 0168-9002. DOI: 10.1016/j.nima.2017.06.044.

[22] C. Adams et al. "Neutrinoless Double Beta Decay". In: (Dec. 2022). DOI: 10.48550/ARXIV.2212.11099. arXiv: 2212.11099 [nucl-ex].

[23] Michelle J. Dolinski, Alan W. P. Poon, and Werner Rodejohann. "Neutrinoless Double-Beta Decay: Status and Prospects". In: *Annual Review of Nuclear and Particle Science* 69.1 (Feb. 11, 2019), pp. 219–251. ISSN: 1545-4134. DOI: 10.1146/annurev-nucl-101918-023407. arXiv: 1902.04097 [nucl-ex].

[24] LEGEND Collaboration et al. "LEGEND-1000 Preconceptual Design Report". In: (2021). arXiv: 2107.11462 [physics.ins-det].

[25] M. Agostini et al. "Upgrade for Phase II of the Gerda experiment". In: *The European Physical Journal C* 78.5 (May 2018). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-018-5812-2.

[26] K. Freund et al. "The performance of the Muon Veto of the Gerda experiment". In: *The European Physical Journal C* 76.5 (2016), p. 298. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-016-4140-7. URL: https://doi.org/10.1140/epjc/s10052-016-4140-7.

[27] Nina Burlac. "Optimizing the Search for Neutrinoless Double Beta Decay - Liquid Argon Instrumentation for Background Suppression in LEGEND-200 Experiment". PhD thesis. Università degli studi Roma Tre, 2024.

[28] Konstantin Gusev. LEGEND presentation material.

[29] Luigi Pertoldi. "The first year of LEGEND-200 physics data in the quest for $0\nu\beta\beta$ decay". In: *Neutrino 2024*. 2024. URL: https://zenodo.org/records/12706010.

[30] Yannick Müller. "Calibration of the LEGEND-200 Experimentto Search for Neutrinoless Double Beta DecayandSearches for Signatures of New Physicswith the GERDA Experiment". PhD thesis. Universität Zürich, 2023.

[31] M. Agostini et al. "Characterization of inverted coaxial ^{76}Ge detectors in GERDA for future double-βdecay experiments". In: *The European Physical Journal C* 81.6 (2021), p. 505. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-021-09184-8. URL: https://doi.org/10.1140/epjc/s10052-021-09184-8.

[32] W. Shockley. "Currents to Conductors Induced by a Moving Point Charge". In: *Journal of Applied Physics* 9.10 (Apr. 1938), pp. 635–636. ISSN: 0021-8979. DOI: 10.1063/1.1710367. eprint: https://pubs.aip.org/aip/jap/article-

pdf/9/10/635/8059000/635_1_online.pdf. URL: https://doi.org/10.1063/1.1710367.

[33] S. Ramo. "Currents Induced by Electron Motion". In: *Proceedings of the IRE* 27.9 (1939), pp. 584–585. DOI: 10.1109/JRPROC.1939.228757.

[34] Tommaso Comellato. "Inverted Coaxial Detectors for Legend". en. PhD thesis. Technische Universität München, 2022, p. 188. URL: https://mediatum.ub.tum.de/?id=1662747.

[35] Glenn F. Knoll. *Radiation Detection and Measurement*. Wiley, 2010.

[36] David Hervas Aguilar. "Characterizing bulk Signals in an Inverted Coaxial Point-Contact Detector to inform rare-event Searches". PhD thesis. University of North Carolina at Chapel Hill, 2023.

[37] Luigi Pertoldi George Marshall. LEGEND presentation material.

[38] M. Misiaszek et al. "Improving sensitivity of a BEGe-based high-purity germanium spectrometer through pulse shape analysis". In: *The European Physical Journal C* 78.5 (2018), p. 392. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-018-5852-7. URL: https://doi.org/10.1140/epjc/s10052-018-5852-7.

[39] Tommaso Comellato, Matteo Agostini, and Stefan Schönert. "Charge-carrier collective motion in germanium detectors for $\beta\beta$-decay searches". In: *The European Physical Journal C* 81.1 (Jan. 2021). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-021-08889-0. URL: http://dx.doi.org/10.1140/epjc/s10052-021-08889-0.

[40] I. Guinn et al. "Pulse Shape Discrimination for the 2024 Unblinded Analysis of LEGEND-200". Internal Document.

[41] H. V. Klapdor-Kleingrothaus, I. V. Krivosheina, and I. V. Titkova. "Theoretical investigation of the dependence of double beta decay tracks in a Ge detector on particle and nuclear physics parameters and separation from gamma ray events".

In: *Physical Review D* 73.1 (Jan. 2006), p. 013010. ISSN: 1550-2368. DOI: 10.1103/physrevd.73.013010.

[42] Victoria Wagner. "Pulse Shape Analysis for the GERDA Experiment to Set a New Limit on the Half-life of $0\nu\beta\beta$ Decay of 76Ge". PhD thesis. Ruperto-Carola-University of Heidelberg, 2017.

[43] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[44] *Ax - Adaptive Experimentation Platform*. URL: https://ax.dev/.

[45] Facundo Bre, Nadia Roman, and Víctor D. Fachinotti. "An efficient metamodel-based method to carry out multi-objective building performance optimizations". In: *Energy and Buildings* 206 (Jan. 2020), p. 109576. ISSN: 0378-7788. DOI: 10.1016/j.enbuild.2019.109576.

[46] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[47] Salman Khan et al. *A Guide to Convolutional Neural Networks for Computer Vision*. Springer International Publishing, 2018. ISBN: 9783031018213. DOI: 10.1007/978-3-031-01821-3.

[48] Christopher Olah. *Understanding LSTM Networks*. Blog Article. Aug. 2015. URL: https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[49] Sepp Hochreiter. "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 06.02 (Apr. 1998), pp. 107–116. ISSN: 1793-6411. DOI: 10.1142/s0218488598000094.

[50] Ashish Vaswani et al. *Attention Is All You Need*. 2017. DOI: 10.48550/ARXIV.1706.03762. URL: https://arxiv.org/abs/1706.03762.

[51] Synced. *A Brief Overview of Attention Mechanism*. Online Article. Sept. 2017. URL: https://medium.com/syncedreview/a-brief-overview-of-attention-mechanism-13c578ba9129.

[52] Andrea Kirsch. "Search for the neutrinoless double β-decay in Gerda Phase I using a Pulse Shape Discrimination technique". PhD thesis. Ruperto-Carola University of Heidelberg, 2014.

[53] Zhuofan Ying, Peter Hase, and Mohit Bansal. "VisFIS: Visual Feature Importance Supervision with Right-for-the-Right-Reason Objectives". In: (2022). DOI: 10.48550/ARXIV.2206.11212. URL: https://arxiv.org/abs/2206.11212.

[54] Chun-Hao Chang, George Alexandru Adam, and Anna Goldenberg. *Towards Robust Classification Model by Counterfactual and Invariant Data Generation*. 2021. DOI: 10.48550/ARXIV.2106.01127.

[55] Becks Simpson et al. *GradMask: Reduce Overfitting by Regularizing Saliency*. 2019. DOI: 10.48550/ARXIV.1904.07478.

[56] Hongyi Zhang et al. *mixup: Beyond Empirical Risk Minimization*. 2017. DOI: 10.48550/ARXIV.1710.09412.

[57] Xiaojiang Peng et al. *Suppressing Mislabeled Data via Grouping and Self-Attention*. 2020. DOI: 10.48550/ARXIV.2010.15603.

[58] H. Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of Educational Psychology* 24.6 (Sept. 1933), pp. 417–441. ISSN: 0022-0663. DOI: 10.1037/h0071325.

[59] Laurens van der Maaten and Geoffrey Hinton. "Viualizing data using t-SNE". In: *Journal of Machine Learning Research* 9 (Nov. 2008), pp. 2579–2605.

[60] Jonathon Shlens. *A Tutorial on Principal Component Analysis*. 2014. DOI: 10.48550/ARXIV.1404.1100.

A. Technical Machine Learning Details

A.1. Definition of Loss functions

The exact definition of the Cosine Embedding Loss is:

$$\text{CosEmb}(x_1, x_2, y) = \begin{cases} 1 - \cos(x_1, x_2), & \text{if } y = 1 \\ \max(0, cos(x_1, x_2) - margin), & \text{if } y = -1 \end{cases} \tag{A.1}$$

while x_1, x_2 are tensors

The exact definition of the SmoothL1Loss is:

$$\text{SmoothL1} = \{l_1, ...l_N\}^T \tag{A.2}$$

$$\text{with} \quad l_n(x_n, y_n) = \begin{cases} 0.5 \cdot \frac{(x_n - y_n)^2}{\beta}, & \text{if } |x_n - y_n| < \beta \\ \max(|x_n - y_n| - 0.5 \cdot \beta, & \text{else} \end{cases} \tag{A.3}$$

While the Kullback-Leibler Divergence is defined as:

$$\text{KL} = \sum y_{\text{true}} \cdot \log\left(\frac{y_{\text{true}}}{y_{\text{pred}}}\right) \tag{A.4}$$

A.2. Detailed Architecture of FCNet, CNN and RNN+att

Listing A.1: FCNet

```
class FCNet(nn.Module):
    def __init__(self, first_unit, last_unit):
        super(FCNet, self).__init__()

        #Number of channels in each fully connected layers
        fc1, fc2 = (first_unit, int(first_unit*0.25))
        do = 0.5
        self.fcnet = nn.Sequential(
            torch.nn.Linear(fc1, fc2),
            torch.nn.LeakyReLU(),
            torch.nn.Dropout(do),
            torch.nn.Linear(fc2, last_unit),
            torch.nn.Sigmoid()
        )
    def forward(self, x):
        return self.fcnet(x)
```

Listing A.2: CNN

```
class CNNDiscriminator(nn.Module):
    def __init__(self):
        super(CNNDiscriminator, self).__init__()
        params = {'cnn1': 3, 'cnn2': 30, 'cnn3': 26, 'cnn4': 79, 'w1':
            17.681005079076083, 'w2': 36.72950484280138, 'w3':
            18.823737570046184, 'w4': 1.0, 'f1': 8, 'f2': 8, 'f3': 9, 'f4':
            2, 'inv_loss': 'aug', 'distance': 'l2'}
        self.seq_len = 256
```

```
        conv1, conv2, conv3, conv4 = (params["cnn1"]*2,params["cnn2"]*2,
            params["cnn3"]*2,params["cnn4"]*2)
        flatten_num = ((((self.seq_len - (params["f1"] - 1))//2 - (params["f2
            "] - 1))//2 - (params["f3"] - 1))//2 - (params["f4"] - 1))*params
            ["cnn4"]*2
        self.CNNBackbone = nn.Sequential(
            torch.nn.Conv1d(1,conv1,params["f1"]),
            torch.nn.MaxPool1d(kernel_size=2),
            torch.nn.LeakyReLU(),
            torch.nn.Conv1d(conv1,conv2,params["f2"]),
            torch.nn.MaxPool1d(kernel_size=2),
            torch.nn.LeakyReLU(),
            torch.nn.Conv1d(conv2,conv3,params["f3"]),
            torch.nn.MaxPool1d(kernel_size=2),
            torch.nn.LeakyReLU(),
            torch.nn.Conv1d(conv3,conv4,params["f4"]),
            torch.nn.LeakyReLU(),
        )
        self.fcnet = nn.Sequential(
            torch.nn.Linear(flatten_num , 1),
            torch.nn.Sigmoid()
        )


    def forward(self, x):
        batch = x.size(0)
        x = self.CNNBackbone(x)
        x = x.view(batch,-1)
        x = self.fcnet(x)
        return x
```

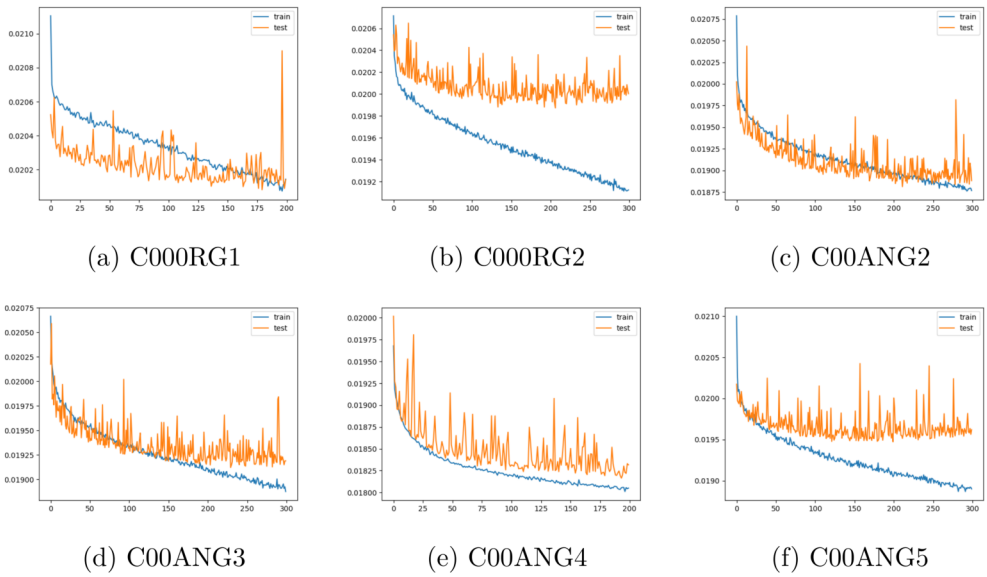
Listing A.3: RNN

```
class RNN(nn.Module):
    def __init__(self,get_attention = False):
        super(RNN, self).__init__()

        bidirec = False    #Whether to use a bidirectional RNN
        self.bidirec = bidirec
        feed_in_dim = 64
        self.seg = 1      #Segment waveform to reduce its length. If the
            original waveform is (2000,1), then segment it with self.seg=5
            can reduce its length to (400,5)
        self.emb_dim = 32
        self.embedding = nn.Embedding(MASK_TOKEN+10,self.emb_dim)
        self.seq_len = (256)//self.seg
        if bidirec:
            self.RNNLayer = torch.nn.GRU(input_size = self.emb_dim,
                hidden_size = feed_in_dim//2,num_layers=3, batch_first=True,
                bidirectional=True,dropout=0.5)
            feed_in_dim *= 2
        else:
            self.RNNLayer = torch.nn.GRU(input_size = self.emb_dim,
                hidden_size = feed_in_dim//2,num_layers=3, batch_first=True,
                bidirectional=False,dropout=0.5)
        self.fcnet = FCNet(feed_in_dim,1)
        self.attention_weight = nn.Linear(feed_in_dim//2, feed_in_dim//2,
            bias=False)
        self.norm = torch.nn.BatchNorm1d(feed_in_dim//2)
        self.get_attention = get_attention

    #@torchsnooper.snoop()
    def forward(self, x):
        x = x.squeeze()
        x = self.embedding(x.long())
        bsize = x.size(0)
```

```
        output, hidden = self.RNNLayer(x)
        if self.bidirec:
            hidden =  hidden[-2:]
            hidden = hidden.transpose(0,1).reshape(bsize,-1)
        else:
            hidden =  hidden[-1]



        #Attention Mechanism
        hidden_attention = hidden.unsqueeze(-1)
        w_attention = self.attention_weight(output)
        w_attention = torch.einsum("ijl,ilm->ijm",w_attention,
            hidden_attention).squeeze(-1)
        attention_score = torch.softmax(w_attention,dim=-1)

        context = torch.sum(attention_score.unsqueeze(-1).expand(*output.size
            ()) * output,dim=1)
        x = self.fcnet(torch.cat([context,hidden],dim=-1))
        return x, attention_score

#The fully connected part of neural network
class FCNet(nn.Module):
    def __init__(self, first_unit, last_unit):
        super(FCNet, self).__init__()

        #Number of channels in each fully connected layers
        fc1, fc2 = (first_unit, int(first_unit*0.25))
        do = 0.5
        self.fcnet = nn.Sequential(
            torch.nn.Linear(fc1, fc2),
            torch.nn.LeakyReLU(),
            torch.nn.Dropout(do),
            torch.nn.Linear(fc2, last_unit),
```

```
            torch.nn.Sigmoid()
        )
    def forward(self, x):
        return self.fcnet(x)
```

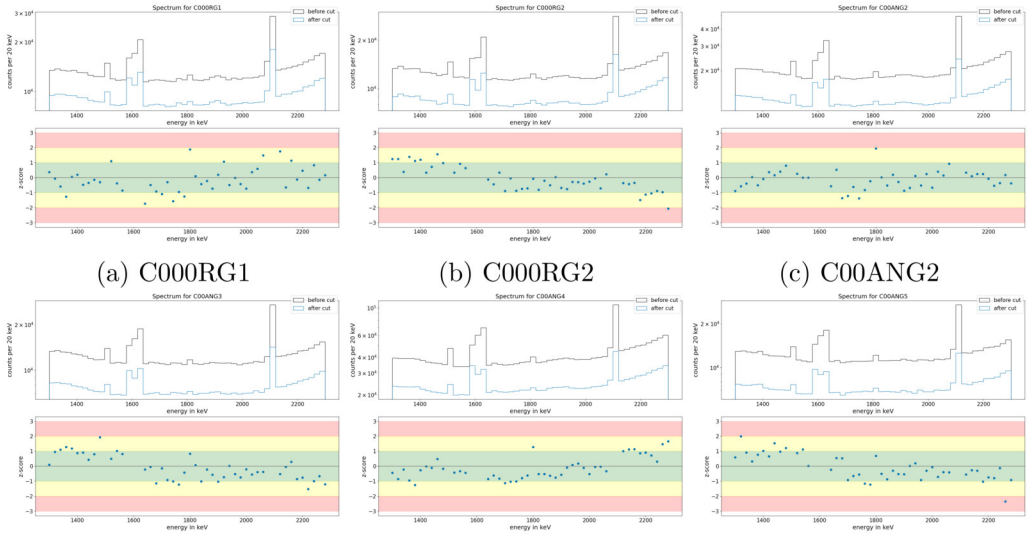
B. Additional Figures for the Application of ANN



(a) C000RG1 (b) C000RG2 (c) C00ANG2

(d) C00ANG3 (e) C00ANG4 (f) C00ANG5

Figure B.1.: Loss during the training for different coaxial detectors (blue). The orange curve is the loss for the test sample to monitor a potential overfitting.



(a) C000RG1 (b) C000RG2 (c) C00ANG2

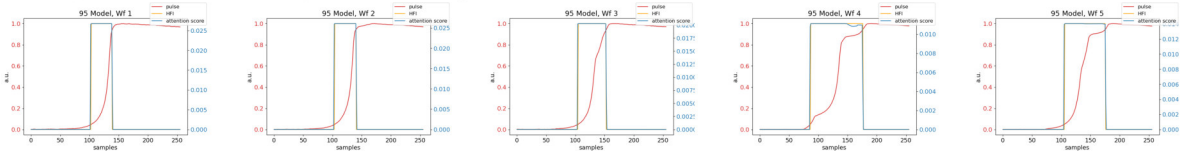
(d) C00ANG3 (e) C00ANG4 (f) C00ANG5

Figure B.2.: Analysis of the energy dependence for the different Coax. It can be seen that none of them has a fluctuation of the z-score from more than 2σ, while most are even below 1σ.

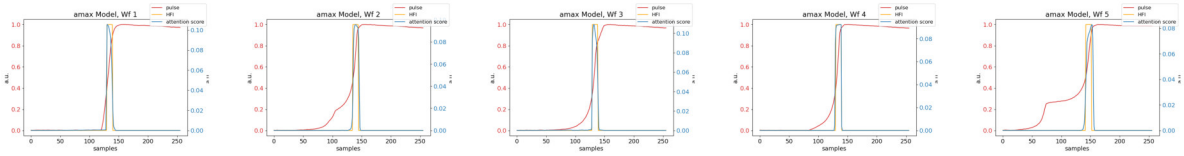
C. Additional Figures for the Analysis of FIS

Table C.1.: Attention Score and Human Feature Importance

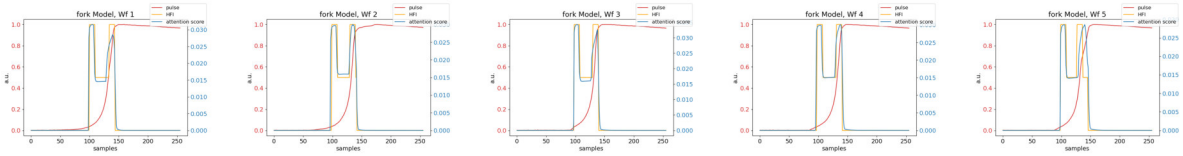
Run 1, Latitudinal Scan, 95-Masking



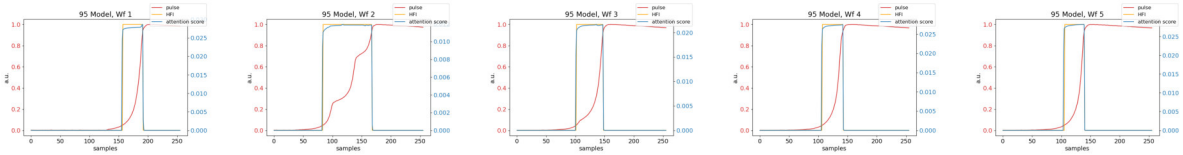
Run 1, Latitudinal Scan, amax-Masking



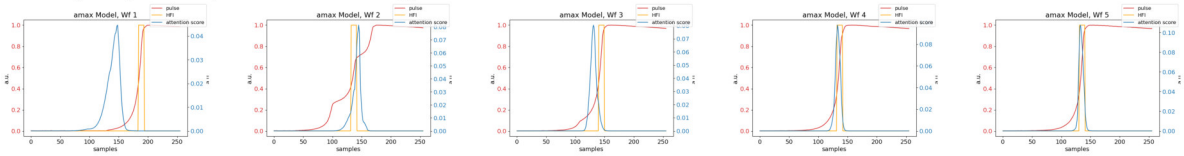
Run 1, Latitudinal Scan, fork-Masking



Run 1, Top Scan, 95-Masking

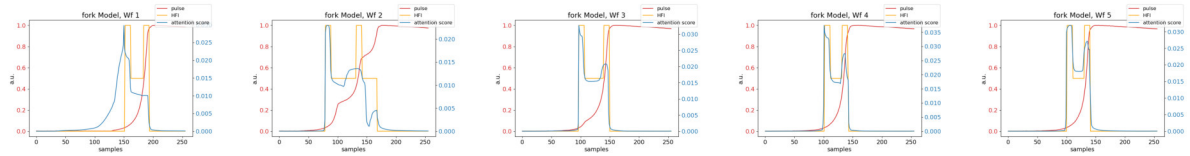


Run 1, Top Scan, amax-Masking

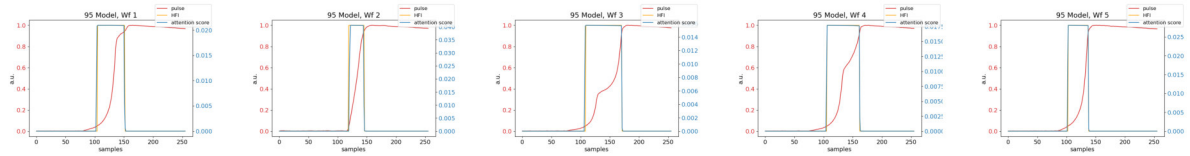


Run 1, Top Scan, fork-Masking

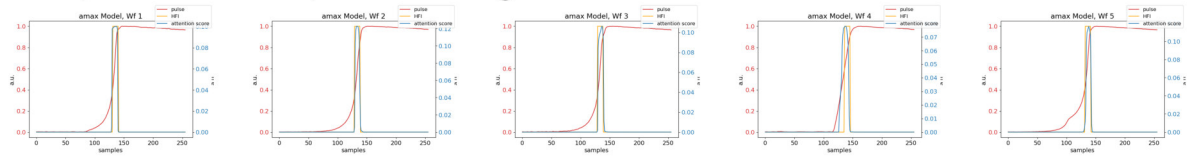
Table C.1.: (Continued)



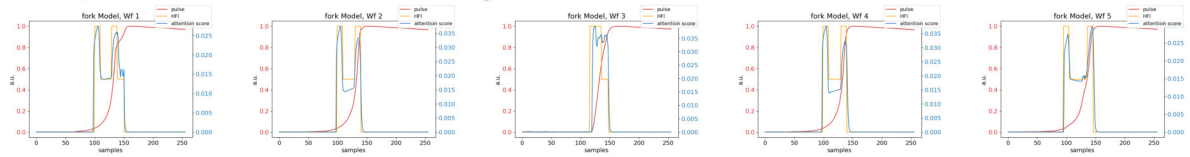
Run 2, Latitudinal Scan, 95-Masking



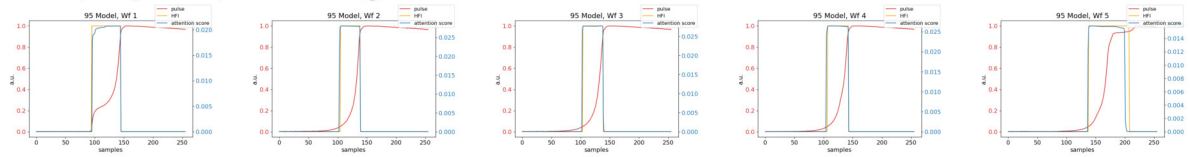
Run 2, Latitudinal Scan, amax-Masking



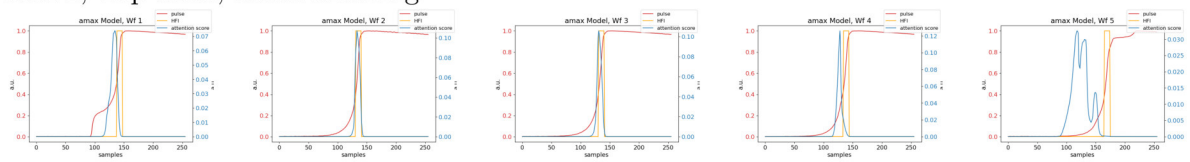
Run 2, Latitudinal Scan, fork-Masking



Run 2, Top Scan, 95-Masking



Run 2, Top Scan, amax-Masking



Run 2, Top Scan, fork-Masking

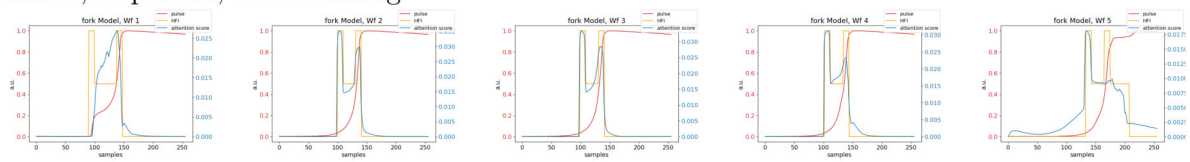
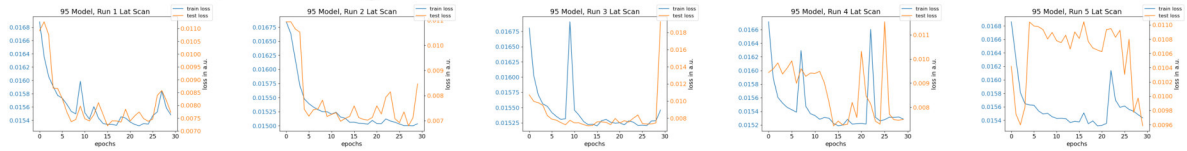
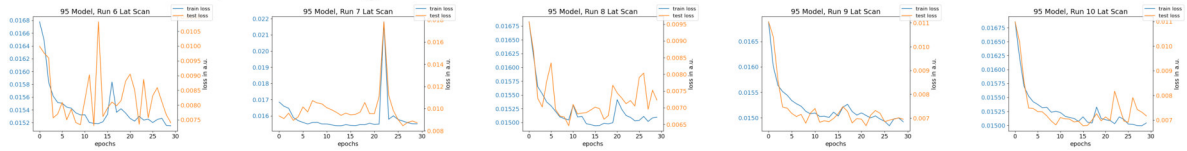


Table C.2.: Training (blue) and validation (orange) loss developement during the training, 10 random samples

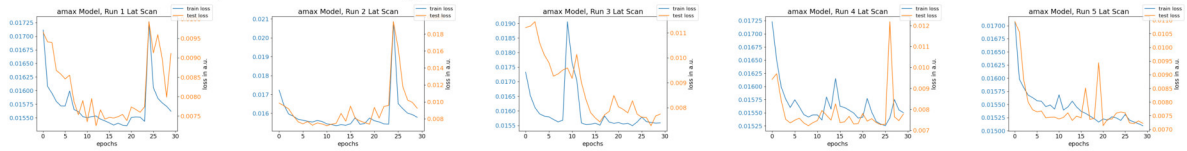
Latitudinal Scan, 95-Masking



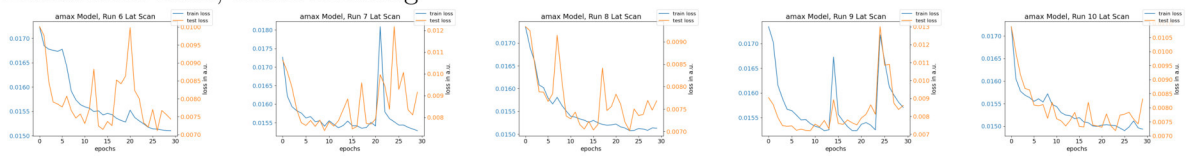
Latitudinal Scan, 95-Masking



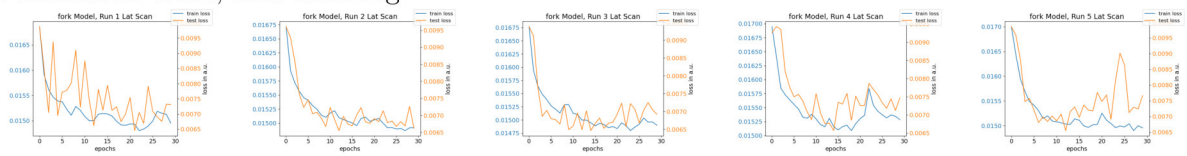
Latitudinal Scan, amax-Masking



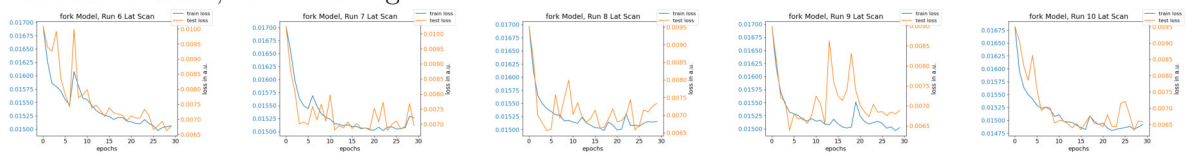
Latitudinal Scan, amax-Masking



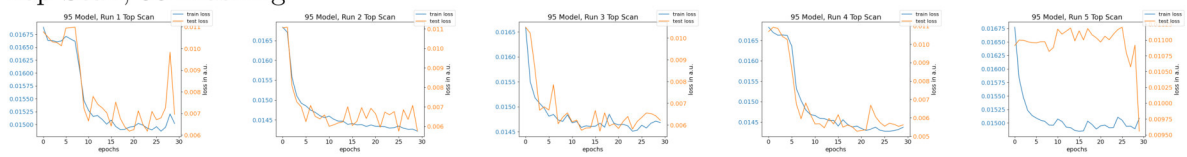
Latitudinal Scan, fork-Masking



Latitudinal Scan, fork-Masking



Top Scan, 95-Masking



Top Scan, 95-Masking

Table C.2.: (Continued)

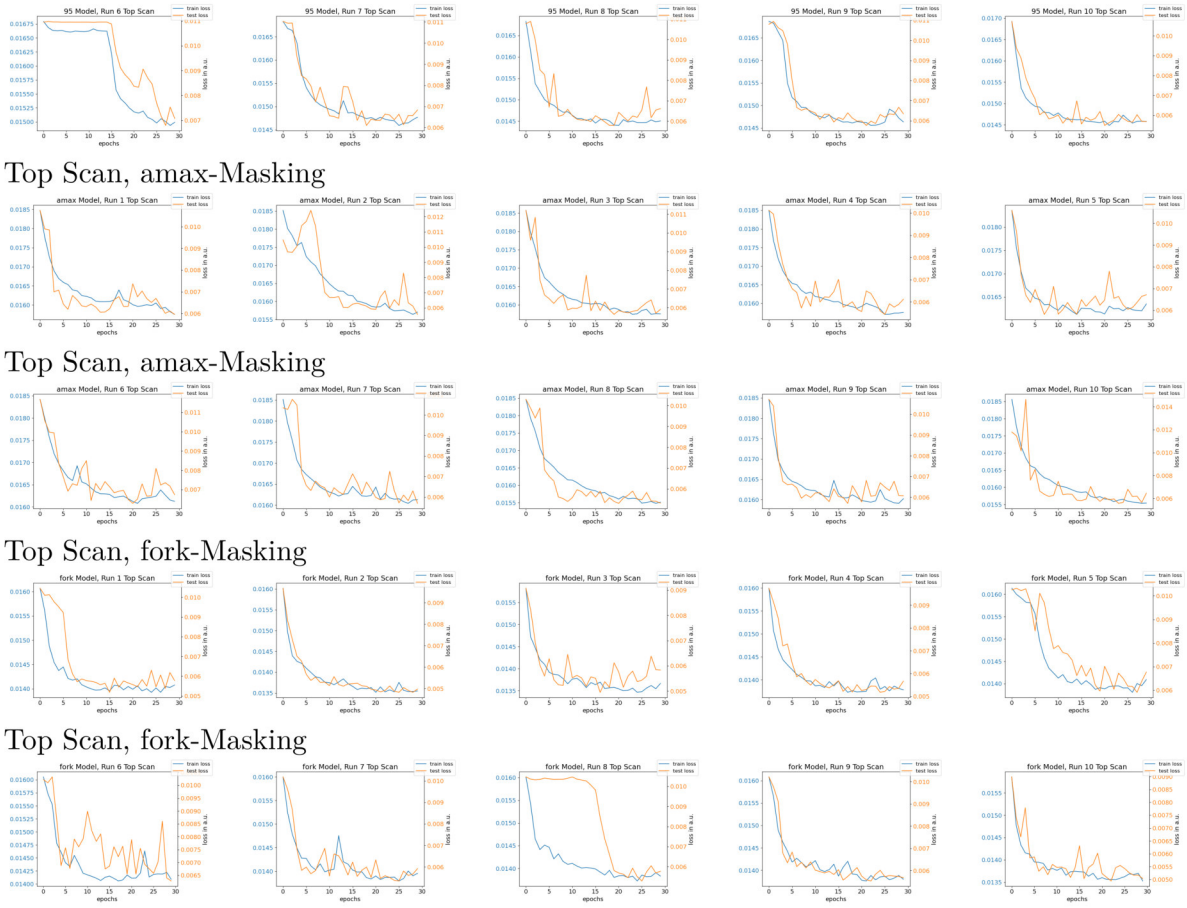
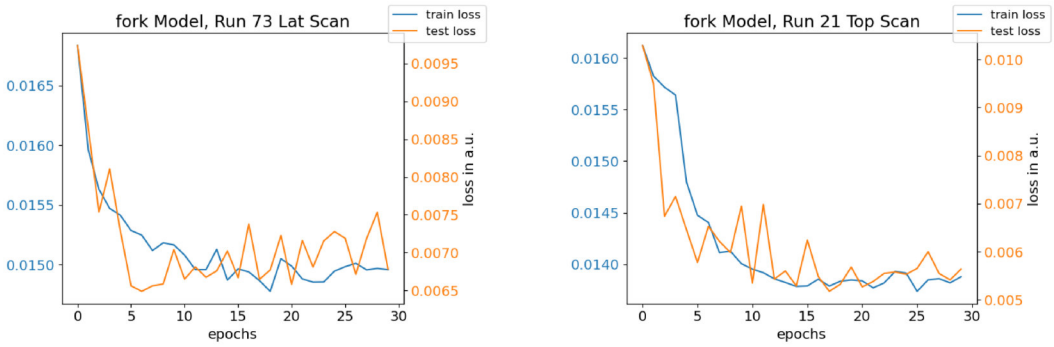
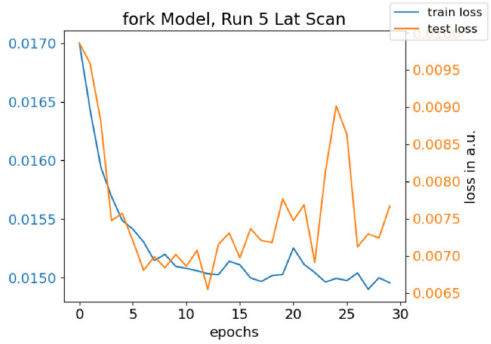


Table C.3.: Training (blue) and validation (orange) loss developement during the training for the best performing models.

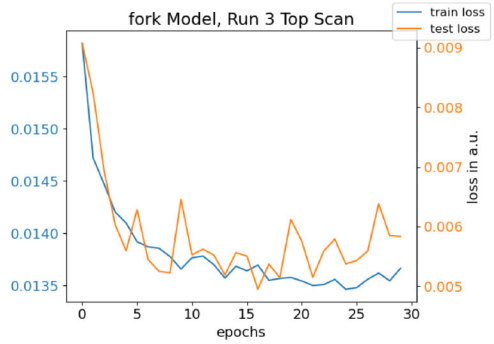


(a) loss evaluation of best 95-model lat **(b)** loss evaluation of best 95-model top

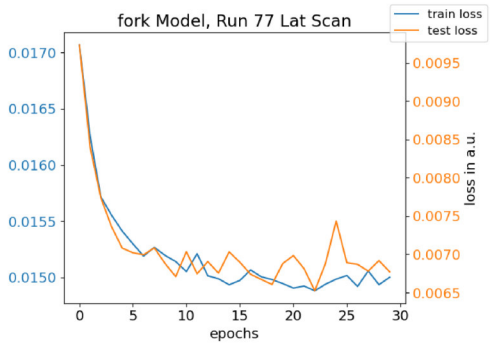
Table C.3.: (Continued)



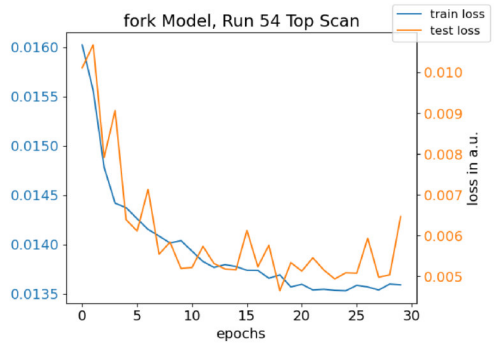
(c) loss evaluation of best amax-model lat



(d) A/E vs best amax-model top

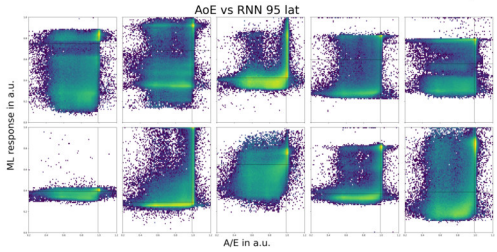


(e) loss evaluation of best fork-model lat

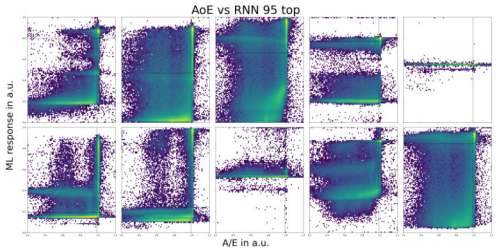


(f) loss evaluation of best fork-model top

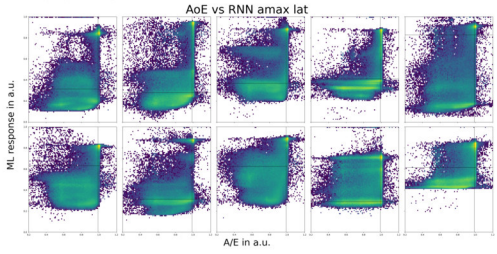
Table C.4.: RNN vs A/E plots 10 random models.



(a) A/E vs random 95-model lat



(b) A/E vs random 95-model top



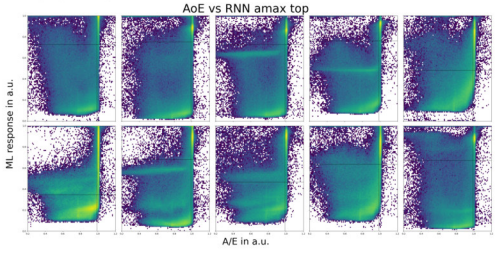
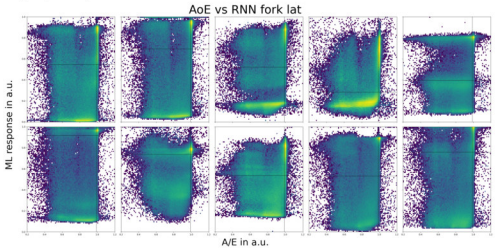
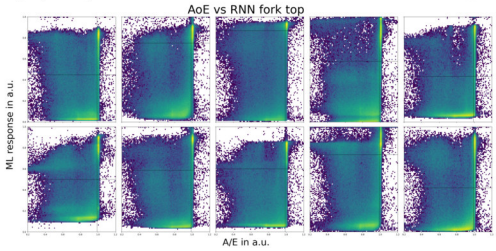


Table C.4.: (Continued)



(c) A/E vs random amax-model lat

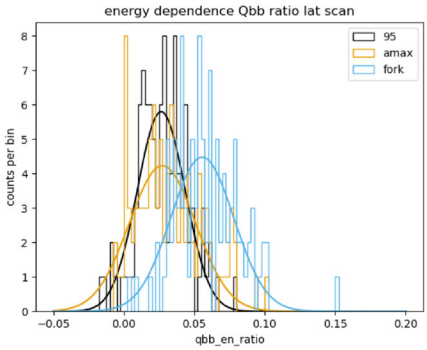


(d) A/E vs random amax-model top

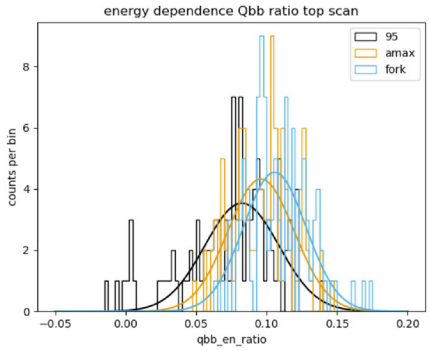
(e) A/E vs random fork-model lat

(f) A/E vs random fork-model top

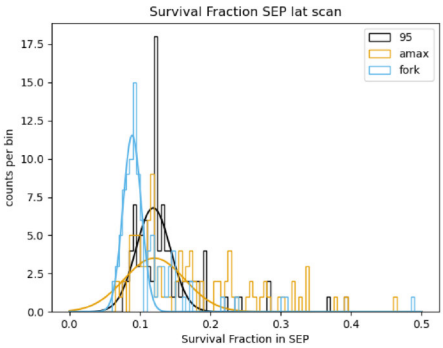
Table C.5.: Performance of FIS over 100 training cycles on different parameters with a pure energy-defined label.



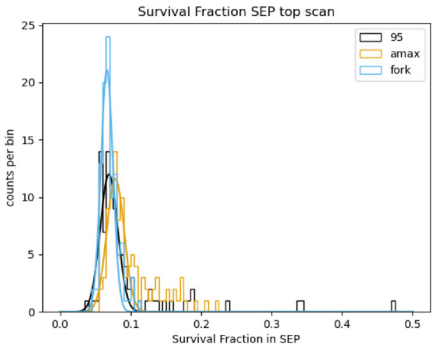
(a) Energy dependence lat



(b) Energy dependence top

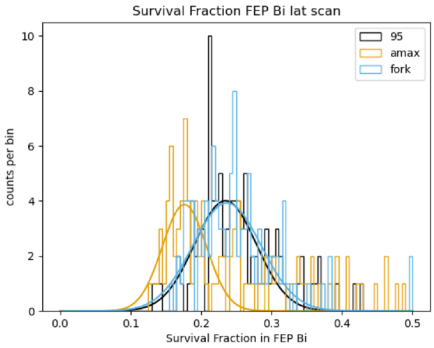


(c) SEP lat

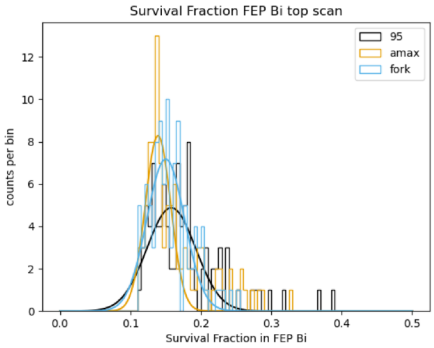


(d) SEP top

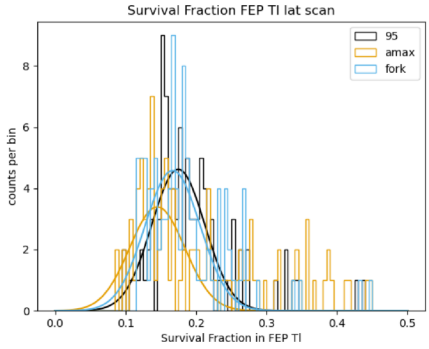
Table C.5.: (Continued)



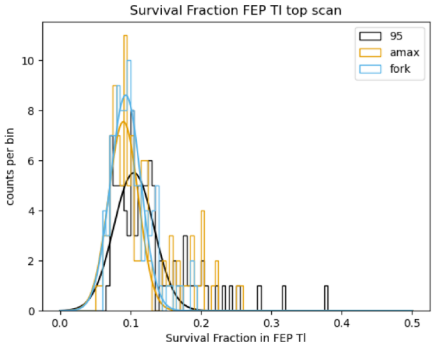
(e) FEP Bi lat



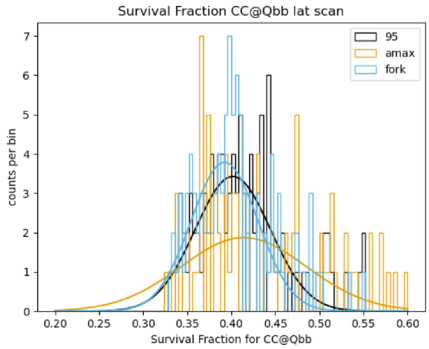
(f) FEP Bi top



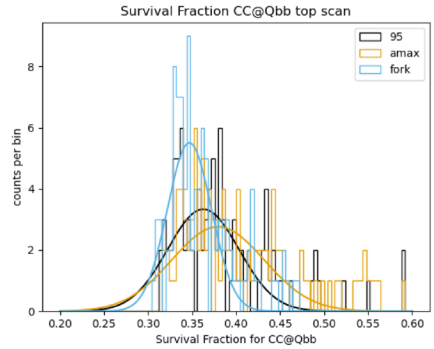
(g) FEP Tl lat



(h) FEP Tl top

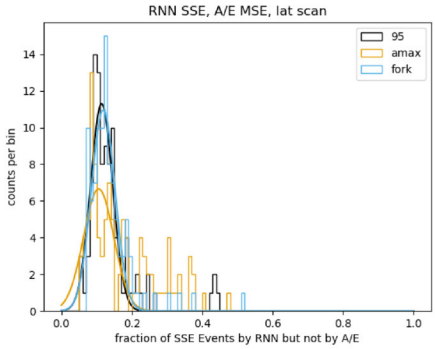


(i) Qbb lat

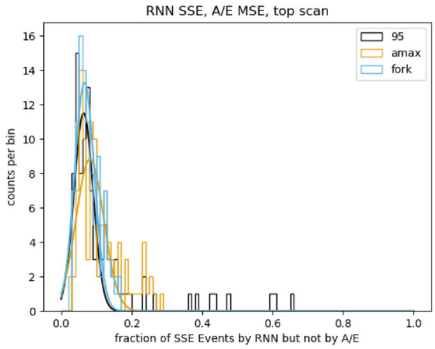


(j) Qbb top

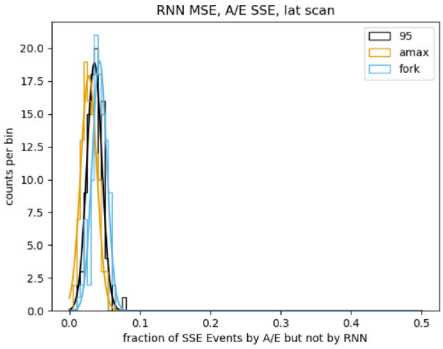
Table C.5.: (Continued)



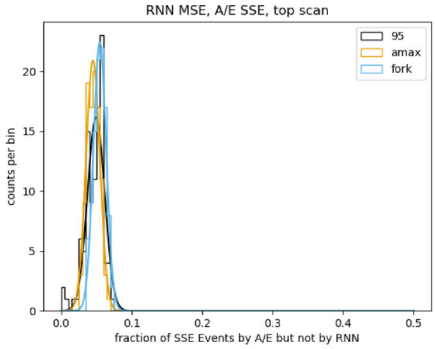
(e) R_{top} lat



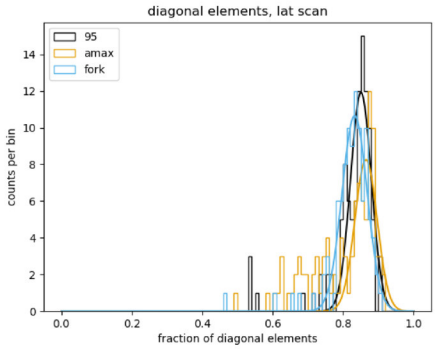
(f) R_{top} top



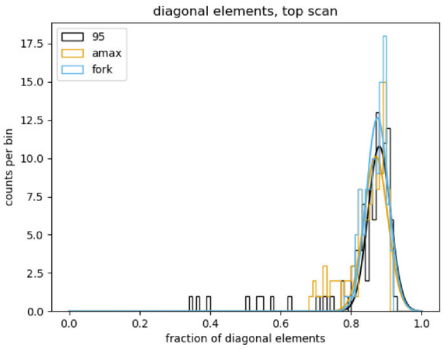
(g) R_{bottom} lat



(h) R_{bottom} top

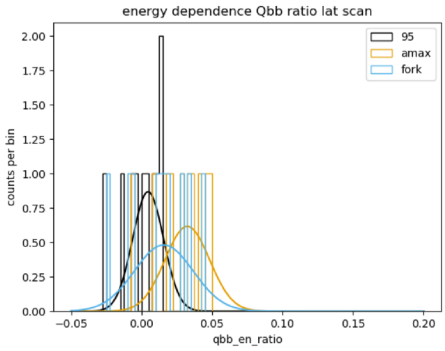


(i) Diagonal Elements lat

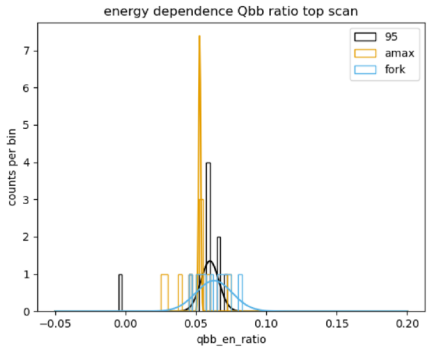


(j) Diagonal Elements top

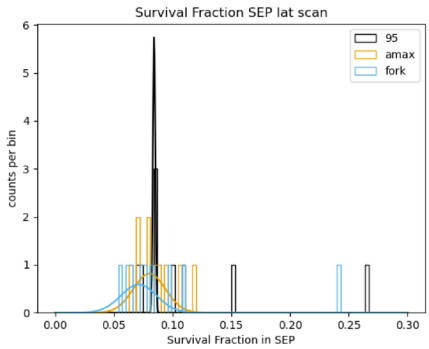
Table C.6.: Performance of FIS over 10 training cycles on different parameters with a label set by a combination of energy and A/E.



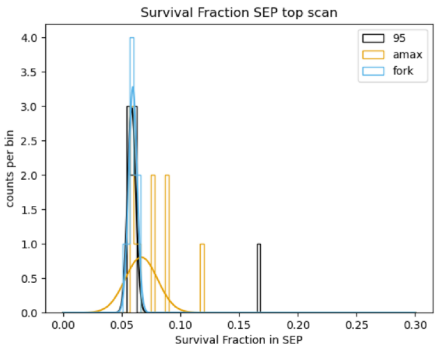
(a) Energy dependence lat



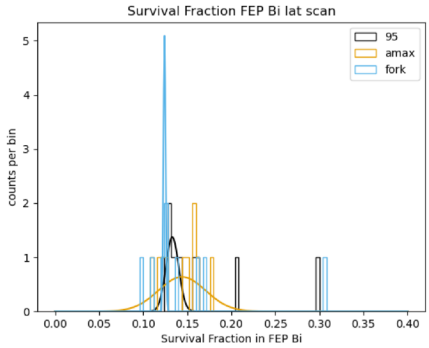
(b) Energy dependence top



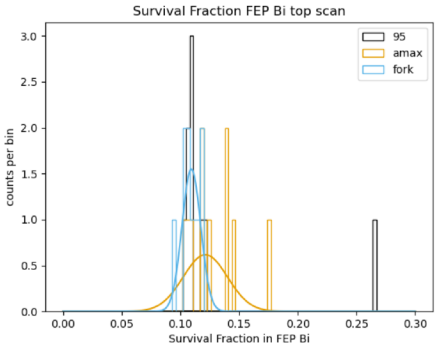
(c) SEP lat



(d) SEP top

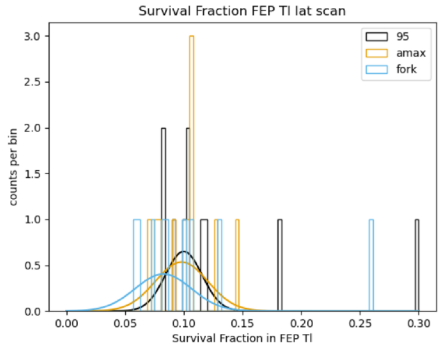


(e) FEP Bi lat

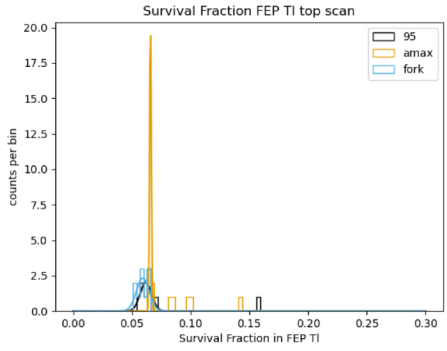


(f) FEP Bi top

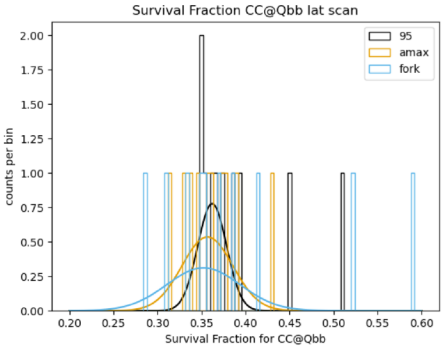
Table C.6.: (with A/E label Continued)



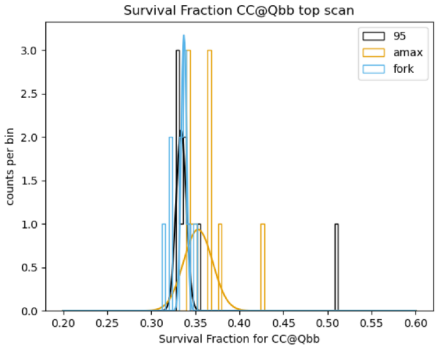
(g) FEP Tl lat



(h) FEP Tl top

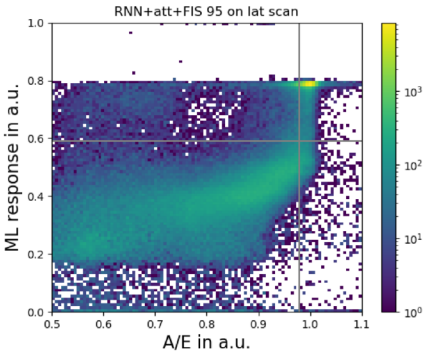


(i) Qbb lat

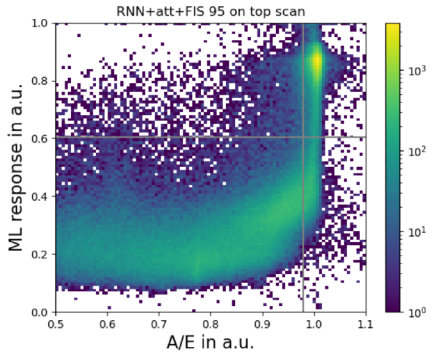


(j) Qbb top

Table C.7.: RNN vs A/E plots for the best performing models.

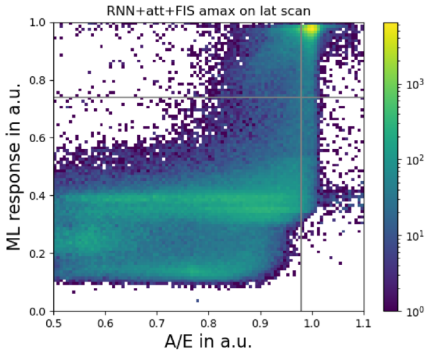


(a) A/E vs best 95-model lat

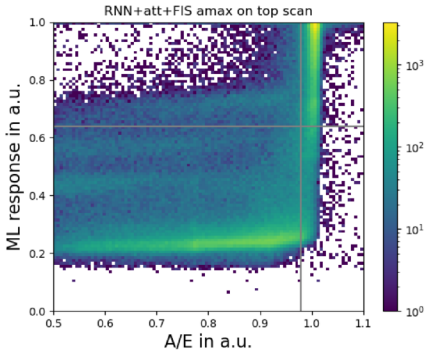


(b) A/E vs best 95-model top

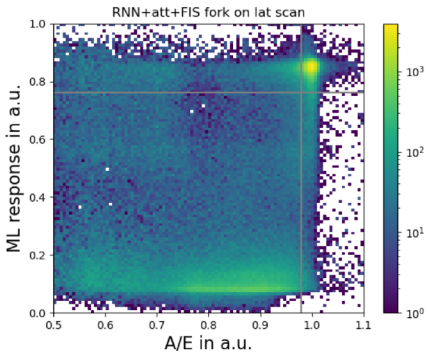
Table C.7.: (Continued)



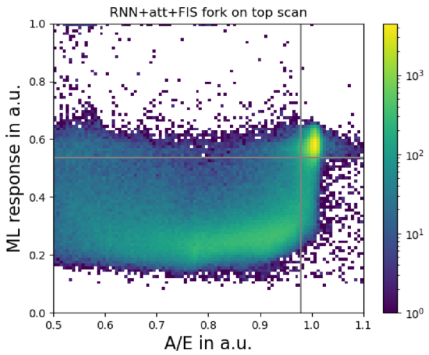
(c) A/E vs best amax-model lat



(d) A/E vs best amax-model top



(e) A/E vs best fork-model lat



(f) A/E vs best fork-model top

D. Results with Attentive Feature Mixup

Table D.1.: Table for using RNN+att+FIS+AFM with all calculated values for different scans and maskings. Worse than without AFM.

Model	**Type**	**SEP ^{208}Tl[%]**	**FEP ^{214}Bi [%]**	**FEP ^{208}Tl[%]**	**Qbb [%]**	**Δz**	**δ_E**	**off diag [%]**	**diag [%]**
energy		2103 keV	1621 keV	2614 keV	2039 keV	-	-	-	-
Top									
A/E		4.9 ± 0.8	8.3 ± 1.1	4.40 ± 0.09	29.8 ± 0.9	0.32 ± 0.0009	9.0	-	-
A/E rt		5.4 ± 0.8	8.3 ± 1.1	5.4 ± 0.1	30.9 ± 0.9	0.24 ± 0.0009	6.5	-	-
95	⌀	16±15	26±14	21±15	45±11	2.5±0.9	6.4±2.7	19±9	81±9
95	μ	10±4	21±6	16±6	41±6	2.5±0.9	6.6±2.7	15±5	85±5
95	best	5.7±0.8	15.6±1.1	9.48±0.13	33.7±0.9	2.88	9.33	5.29	59.87
amax	⌀	12±8	20±9	15±9	43±8	3.4±0.8	8.6±2.0	18±7	82±7
amax	μ	9.6±1.5	16.7±2.3	11.7±2.4	40±6	3.4±0.8	9.2±1.4	15±4	85±4
amax	best	7.5±0.8	14.9±1.2	9.07±0.13	35.3±0.9	3.09	9.61	6.43	58.72
fork	⌀	9.4±3.2	19±5	13±5	41±7	3.3±0.9	8.9±2.1	17±6	83±6
fork	μ	7.8±1.2	16.9±2.8	11.2±2.2	37±4	3.1±0.9	8.8±2.0	13.4±3.5	86.6±3.5
fork	best	6.0±0.8	14.5±1.1	9.13±0.13	33.1±0.9	2.8	9.24	5.44	59.72
Lat									
A/E		7.8 ± 2.3	12.9 ± 2.6	9.0 ± 2.4	36.3 ± 2.5	$(-9 \pm 67) \cdot 10^{-5}$	0.03	-	-
A/E rt		6.7 ± 1.1	12.0 ± 1.5	7.4 ± 0.12	34.3 ± 1.0	0.024 ± 0.0006	0.7	-	-
95	⌀	23±18	36±16	29±16	50±12	1.2±0.6	2.6±1.7	24±11	76±11
95	μ	13±4	28±7	20±6	43±7	1.2±0.5	2.7±1.4	18±4	82±4
95	best	8.3±1.0	21.9±1.5	15.24±0.18	36.5±1.0	1.6	4.59	10.07	56.91
amax	⌀	24±11	33±12	30±11	53±9	1.5±0.8	2.9±1.6	26±8	74±8
amax	μ	21±10	30±11	27±10	54±10	1.3±0.7	2.7±1.3	25±9	75±9
amax	best	9.0±1.0	14.8±1.5	10.74±0.15	34.9±1.0	0.83	2.44	5.85	61.13
fork	⌀	22±18	37±16	31±17	51±12	2.0±1.0	4.4±2.3	26±11	74±11
fork	μ	11.6±3.3	28±8	22±8	44±6	2.2±0.7	5.1±1.3	21±6	79±6
fork	best	8.2±1.0	17.0±1.5	12.30±0.16	35.9±1.0	1.51	4.39	7.4	59.58