University of Zurich[UZH]

Data-Driven Analysis for Improved Signal Detection in LEGEND-200

Master Thesis in Physics

Yuri van der Burg

Supervised by
Prof. Dr. Laura Baudis
Dr. Marta Babicz

August 22, 2025

Abstract

Neutrinoless double beta decay is a hypothesized nuclear process whose observation would have far-reaching implications for particle physics, providing insight into the fundamental nature of neutrinos and the possible violation of lepton number conservation. The Large Enriched Germanium Experiment for Neutrinoless double beta Decay (LEGEND) experiment aims to detect this rare decay using high-purity germanium detectors enriched in ^{76}Ge, where pulse shape discrimination plays a critical role in suppressing background events.

This thesis presents the development and application of a novel pulse shape discrimination algorithm using a Transformer-based neural network trained on time-series charge waveforms. The pulse shape discrimination efficiency at the energy corresponding to the neutrinoless double beta decay signal was evaluated using data from Inverted Coaxial Point Contact detectors produced by Mirion Technologies. The Transformer-based classifier achieved an efficiency of $(86.7 \pm 1.3)\%$, which is consistent with the $(84.3 \pm 0.5)\%$ obtained using the conventional A/E method, where A/E is the ratio of the maximum current amplitude to the reconstructed energy. A two-sided p-value of $p \approx 0.08$ indicates no significant difference between the two approaches at the 5% level. The total detection efficiency achieved with the Transformer-based approach is $(66.8 \pm 2.1)\%$, compared to $(65.0 \pm 2.0)\%$ for the A/E method. A Bayesian statistical framework is employed to propagate uncertainties in the efficiency and evaluate the resulting impact on LEGEND's neutrinoless double beta decay discovery sensitivity.

If LEGEND-200 – the first of two stages of the LEGEND experiment, with a total target mass of 200 kg – reaches its design goals of a background index of 2×10^{-4} counts/(keV·kg·yr) and a total exposure of 1000 kg $\cdot$ yr, the projected sensitivity to the neutrinoless double beta decay half-life is $T^{0\nu}_{1/2} = 1.22^{+0.33}_{-0.37} \times 10^{27}$ yr for A/E and $T^{0\nu}_{1/2} = 1.25^{+0.34}_{-0.37} \times 10^{27}$ yr for the Transformer-based method.

While the Transformer-based classification does not yet significantly enhance the experimental sensitivity, it demonstrates the potential of advanced machine learning techniques in improving background rejection and sensitivity. Further development and validation, such as through refined data cleaning procedures, better balanced training sets, and enhanced detector-specific modeling, are required to consistently outperform conventional methods. Continued progress along these lines could pave the way for the robust integration of machine learning models into the LEGEND analysis workflow, ultimately helping to maximize discovery sensitivity.

Contents

1 Introduction

Neutrino physics lies at the forefront of modern particle physics, addressing some of the most fundamental questions about the nature of matter and the evolution of the universe. One of the most compelling processes in this context is neutrinoless double beta ($0\nu\beta\beta$) decay. Its observation would confirm that neutrinos are Majorana particles, with profound implications for our understanding of fermion mass generation. As a direct consequence, it would imply the violation of lepton number – a necessary condition for baryogenesis via leptogenesis, which may explain the observed matter-antimatter asymmetry in the universe. While the sensitivity of $0\nu\beta\beta$ decay to the absolute neutrino mass scale is model-dependent, a null result at high exposure would still provide strong and complementary constraints, especially when combined with limits from cosmology and beta-decay experiments. The process would appear as a monoenergetic peak at $Q_{\beta\beta}$. To search for this rare process with the required sensitivity, the LEGEND experiment was conceived as a phased program, with LEGEND-200 now operational and using modular arrays of high-purity germanium (HPGe) detectors enriched in ^{76}Ge. To reach its background goal of 10^{-5} counts/(keV·kg·yr), LEGEND combines ultra-clean detector materials with powerful analysis techniques. A crucial element in this effort is pulse shape discrimination (PSD), which enables the rejection of multi-site and surface background events by analyzing the shape of recorded waveforms. Traditionally, PSD relies on hand-engineered features such as the ratio between area and energy (A/E). However, such approaches can be sensitive to electronic noise and calibration drifts, and may not fully exploit all the discriminative information in the waveform. This motivates the use of machine learning techniques – particularly Transformer architectures – which employ self-attention mechanisms to model long-range dependencies in sequential data, enabling the automatic extraction of rich, high-dimensional correlations from raw waveform data. Such models may improve discrimination performance beyond conventional methods.
The structure of this thesis is as follows:

Chapter 2 establishes the theoretical motivation for this work. It introduces neutrinos within and beyond the Standard Model, motivating the search for neutrinoless double beta decay as a probe of the Majorana nature of neutrinos. The chapter concludes by summarizing the landscape of current and future neutrino experiments.

Chapter 3 addresses the experimental realization of such searches in the LEGEND experiment. It explains how HPGe detectors are used to achieve the required sensitivity, focusing on signal formation and the challenge of

background suppression. Pulse shape discrimination is introduced as a key strategy, and the chapter concludes with an overview of the experiment's calibration procedures.

Chapter 4 covers the machine learning concepts relevant to this work, focusing on deep learning. It justifies applying deep learning to PSD and motivates the use of Transformers as particularly well-suited for waveform data. The chapter concludes with a discussion of the Transformer architecture and the specific model used in this thesis.

Chapter 5 describes the practical implementation of a Transformer-based PSD method. It details the LEGEND-200 data processing, the preparation of training data for the Transformer models, their configurations, and their performance. The extraction of PSD efficiencies at $Q_{\beta\beta}$ is also described, and an overview of the current status of pulse shape simulation in LEGEND-200 is presented. The chapter concludes with a brief summary of the key results.

Chapter 6 presents the Bayesian sensitivity study used to quantify the impact of the PSD methods on the attainable $0\nu\beta\beta$ half-life. It introduces the statistical framework and outlines the procedure used in this work to translate detector performance into projected half-life sensitivities. It concludes with a concise summary of the results.

The thesis concludes with a discussion of the results, including limitations and possible directions for further improving PSD methods.

2 Neutrino physics

This chapter covers relevant concepts for $0\nu\beta\beta$ decay searches. It begins with an introduction to the Standard Model, explaining why its minimal formulation cannot account for neutrino masses. It then presents evidence that neutrinos can change flavor during propagation, implying that neutrinos can not be massless. Next, it discusses two different mechanisms for generating neutrino masses, considering neutrinos as Dirac or Majorana particles. After a brief theoretical introduction to $0\nu\beta\beta$ decay, several experimental results are discussed. The chapter concludes with the introduction of the next-generation $0\nu\beta\beta$ decay experiments.

2.1 Neutrinos in the Standard Model of particle physics

The last few centuries of physics research have concluded that there are four fundamental interactions in nature: gravity, described by the theory of general relativity, and the electromagnetic, weak, and strong interactions. The latter three are unified in the framework of the Standard Model of particle physics (SM), a gauge quantum field theory in which particles arise as quantized excitations of the underlying fields [1].

The total gauge symmetry of the SM is $SU(3)_C \times SU(2)_L \times U(1)_Y$. The $SU(3)$ component of the SM is related to quantum chromodynamics, the theory of the strong force. It describes how quarks, carrying color charge, are confined to form hadrons. The remaining part of the SM gauge symmetry corresponds to the unification of the electromagnetic and the weak interaction. The L indicates that the weak interaction couples to left-handed particles, as only left-handed fermions (and right-handed anti-fermions) participate in weak-isospin doublets. The group $U(1)_Y$ is associated with the weak hypercharge Y.

In the SM, there are three generations of leptons: The three charged leptons e, μ, and τ and their corresponding neutrinos $(\nu_e, \nu_\mu, \nu_\tau)$. These leptons are grouped into $SU(2)_L$ doublets for the left-handed components. Usually, l_L denotes an unspecified $SU(2)_L$ doublet while l_R refers to the corresponding right-handed singlet, which does not participate in weak-isospin interactions:

$$l_L = \begin{pmatrix} \nu_\mathrm{l} \\ \mathrm{l} \end{pmatrix}, \; l_R = \mathrm{l}_R \,. \tag{2.1}$$

Here, l represents one of the charged leptons and ν_l its corresponding neutrino. The four generators of the electroweak gauge symmetry, correspond-

ing to $SU(2)_L$ and $U(1)_Y$, do not directly correspond to physical particles. Instead, specific linear combinations of these generators give rise to the physical gauge bosons of the weak and electromagnetic interactions. Two $SU(2)_L$ gauge bosons form the charged weak gauge bosons $W^{\pm}_{\mu}$, which mediate charged current (CC) interactions such as β-decays. The Lagrangian is given in (2.2), where γ^{μ} represents the Dirac matrices and g denotes the weak coupling constant. The notation $\overline{\nu_{l,L}} = (\nu_{l,L})^{\dagger}\gamma^0$ and $\overline{l^-_L} = (l_L)^{\dagger}\gamma^0$ denotes the Dirac adjoint of the left-handed neutrino and lepton fields, respectively.

$$\mathcal{L}_{CC} = -\frac{g}{\sqrt{2}} \sum_l \left(\overline{\nu_{l,L}} \gamma^{\mu} l^-_L W^+_{\mu} + \overline{l^-_L} \gamma^{\mu} \nu_{l,L} W^-_{\mu} \right) . \tag{2.2}$$

The remaining $SU(2)_L$ gauge boson, W^3_{μ} mixes with the $U(1)_Y$ gauge boson to form the electrically neutral Z_{μ} boson and the photon, A_{μ}, according to the following relation

$$\begin{pmatrix} Z^0_{\mu} \\ A_{\mu} \end{pmatrix} = \begin{pmatrix} \cos\theta_W & -\sin\theta_W \\ \sin\theta_W & +\cos\theta_W \end{pmatrix} \begin{pmatrix} W^3_{\mu} \\ B_{\mu} \end{pmatrix} , \tag{2.3}$$

where W^3_{μ} is the third gauge boson of the $SU(2)_L$ group, B_{μ} is the gauge boson of the $U(1)_Y$ group, and θ_W is the weak mixing angle [2, 3]. The Z^0_{μ} boson mediates weak neutral current (NC) interactions, with the corresponding Lagrangian given in equation (2.4). Weak NC interactions depend on the weak hypercharge and therefore couple also to right-handed fermions (and left-handed anti-fermions). However, in the minimal Standard Model, right-handed neutrinos do not exist, as they are completely uncharged under the gauge interactions.

$$\mathcal{L}_{NC} = -\frac{g}{2\cos\theta_W} \sum_l \overline{\nu_{l,L}} \gamma^{\mu} \nu_{l,L} Z^0_{\mu} . \tag{2.4}$$

In the unbroken electroweak theory, all gauge bosons and fermions are massless. However, the introduction of the Higgs field ϕ, a complex scalar doublet, leads to spontaneous symmetry breaking of the $SU(2)_L \times U(1)_Y$ gauge symmetry. This mechanism generates masses for $W^{\pm}$ and Z^0 bosons, while leaving the photon massless. When the Higgs field acquires a non-zero vacuum expectation value $\langle\phi\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}$, the symmetry is spontaneously broken and the weak gauge bosons acquire mass. The mass of the charged weak bosons is given by

$$m_W = \frac{1}{2} g v \,, \tag{2.5}$$

where g is weak coupling constant and $v \approx 246$ GeV is the Higgs vacuum expectation value [3]. At energies much lower than m_W, the exchange of a virtual W boson can be approximated by a point-like interaction. The corresponding coupling strength is the Fermi constant G_F, defined as

$$G_F = \frac{\sqrt{2}}{8} \frac{g^2}{m_W^2} \approx 10^{-5} \mathrm{GeV}^{-2} \,. \tag{2.6}$$

This low-energy limit reproduces the structure of Fermi's original theory of beta decay, in which weak interactions are modeled as point-like four-fermion interactions with a universal coupling constant [1].

Fermion masses are generated through interactions with the Higgs field via the Yukawa coupling. The Yukawa coupling for the electron is described by equation (2.7), where the Hermitian conjugate of the Higgs field is indicated by $\phi^\dagger$ and the Yukawa coupling constant for the electron by Y_e [1]:

$$\mathcal{L}_Y^e = -Y_e \left(\overline{l_L} \phi l_R + \overline{l_R} \phi^\dagger l_L \right) \,. \tag{2.7}$$

When the Higgs field acquires a non-zero vacuum expectation value, the symmetry is spontaneously broken and the electron acquires a mass term proportional to both the Yukawa coupling constant and the vacuum expectation value of the Higgs field [3, 4]:

$$m_e = Y_e \cdot \frac{v}{\sqrt{2}} \,. \tag{2.8}$$

This mechanism applies analogously to other fermions, with each particle's mass depending on its specific Yukawa coupling constant. However, since right-handed neutrinos are completely uncharged in the electroweak theory, it is not possible to construct a Yukawa term analogous to (2.7). As a result, the Standard Model predicts that neutrinos are massless [3].

2.2 Neutrino oscillation

Already in 1967, Pontecorvo proposed a mechanism where lepton flavor is not conserved during neutrino propagation, leading to the phenomenon known as neutrino oscillation [5]. Since then, several experiments have observed deficits in neutrino fluxes from the sun (solar neutrinos), the atmosphere (atmospheric neutrinos), nuclear reactors and particle accelerators,

providing strong evidence for neutrino oscillation. In addition to these disappearance signals, later experiments such as OPERA and NOvA have directly observed the appearance of different neutrino flavors, confirming the oscillation mechanism through both disappearance and appearance channels [6, 7].

Atmospheric neutrinos are produced in collisions of cosmic rays with nuclei from the upper atmosphere. The flux is dominated by leptonic pion decays, yielding an expected ratio of:

$$\frac{\Phi_{\nu_\mu} + \Phi_{\overline{\nu}_\mu}}{\Phi_{\nu_e} + \Phi_{\overline{\nu}_e}} \approx 2\,. \tag{2.9}$$

The first precise measurement for neutrino oscillation was presented in 1998 by the Super-Kamiokande experiment, a 50 kiloton water Cherenkov detector equipped with over 11'000 photomultiplier tubes (PMTs), located 1000 meters below ground in Japan. Neutrinos are detected via charged current neutrino scattering and subsequent Cherenkov radiation of the leptons [8]. Super-Kamiokande also detected a smaller than expected solar ν_e flux originating from different nuclear fusion processes, indicating neutrino oscillation in solar neutrinos [9]. This discrepancy between theoretically predicted and experimentally observed ν_e flux is called the solar neutrino problem. It was solved by the Sudbury Neutrino Observatory (SNO), which provided a crucial breakthrough by detecting both charged-current and neutral-current interactions. By comparing these interaction rates, SNO showed that the total flux of solar neutrinos agreed with theoretical predictions from the Standard Solar Model, but a substantial fraction had transformed into ν_μ and ν_τ flavors [10].

In general, a neutrino ν_α is produced in a definitive flavor eigenstate (ν_e, ν_μ, ν_τ) via the weak interaction. This neutrino then propagates through spacetime as a linear superposition of the three mass eigenstates (ν_1, ν_2, ν_3), which are the actual fundamental particles. When a neutrino is later detected, it is again projected onto a specific flavor eigenstate, which may differ from its original production state, leading to the possibility of flavor transitions [11, 12]. The flavor eigenstates ν_α and mass eigenstates ν_i are connected through the unitary Pontecorvo–Maki–Nakagawa–Sakata (PMNS) matrix $\mathcal{U}$:

$$|\nu_\alpha\rangle = \sum_i^3 \mathcal{U}\,|\nu_i\rangle\ . \tag{2.10}$$

This mixing leads to the phenomenon of neutrino oscillation, where a

neutrino can change its flavor as it propagates through space. The PMNS matrix is typically parametrized by three mixing angles (θ_{12}, θ_{13}, θ_{23}) and a complex phase δ, which can lead to charge-parity (CP) violation in the lepton sector. CP violation in neutrino oscillations is a key area of current research, as it could provide insights into the matter-antimatter asymmetry of the universe. The parametrization, shown in (2.11), uses the shorthand $s_{ij} = \sin\theta_{ij}$ and $c_{ij} = \cos\theta_{ij}$. Although the full PMNS matrix contains 18 real parameters, only four of them are physically observable in neutrino oscillation experiments [3, 11].

$$\mathcal{U} = \begin{pmatrix} c_{12}\, c_{13} & s_{12}\, c_{13} & s_{13}\, e^{-i\delta_{CP}} \\ -s_{12}\, c_{23} - c_{12}\, s_{13}\, s_{23}\, e^{i\delta_{CP}} & c_{12}\, c_{23} - s_{12}\, s_{13}\, s_{23}\, e^{i\delta_{CP}} & c_{13}\, s_{23} \\ s_{12}\, s_{23} - c_{12}\, s_{13}\, c_{23}\, e^{i\delta_{CP}} & -c_{12}\, s_{23} - s_{12}\, s_{13}\, c_{23}\, e^{i\delta_{CP}} & c_{13}\, c_{23} \end{pmatrix} \tag{2.11}$$

For two-flavor mixing, the probability of neutrino oscillation is given by:

$$P_{\alpha\to\beta}(t) = |\langle\nu_\beta|\nu_\alpha(t)\rangle|^2 = \sin^2(2\theta)\cdot\sin^2\left(\Delta m_{ij}^2\cdot\frac{L}{4E}\right), \tag{2.12}$$

where, $\Delta m_{ij}^2 = m_i^2 - m_j^2$ is the squared mass difference between mass eigenstates i and j, L the distance between source and detection, and E the neutrino energy [12]. Equation (2.12) illustrates that the oscillation probability depends on the squared mass differences, not the absolute masses. Consequently, neutrino oscillations require $\Delta m_{ij}^2 \neq 0$, and oscillation experiments cannot determine the absolute neutrino mass scale. The mixing angle θ in the two-flavor oscillation probability arises from the underlying unitary transformation that relates flavor and mass eigenstates, analogous to equation (2.10). In the absence of CP violation, this transformation reduces to a real, orthogonal matrix with a single parameter (an element of $SO(2)$). The angle θ determines the amplitude of the oscillation probability [12].

While the two-flavor approximation is useful for illustrative purposes, a complete description of neutrino oscillations involves three-flavor mixing, incorporating all three mass eigenstates and the full PMNS matrix. This more general framework leads to a richer phenomenology, including interference effects between multiple oscillation modes and potential CP-violating effects.

2.3 Dirac and Majorana neutrinos

The discovery of neutrino oscillation – and therefore the existence of non-zero masses for the neutrinos – raises the fundamental question of how neutrinos acquire mass. Two popular mechanisms exist to generate neutrino mass terms: If neutrinos are Dirac particles, masses can be generated by introducing right-handed neutrinos, which are often called sterile because they do not participate in SM interactions. This allows the construction of a Yukawa coupling term analogous to that of the electron, as shown in equation (2.7), leading to a mass term:

$$m_\nu = \frac{Y_\nu v}{\sqrt{2}}\,. \tag{2.13}$$

However, the SM offers no satisfactory explanation for why the neutrino Yukawa couplings (and therefore the neutrino masses) are many orders of magnitude smaller than those of other fermions [13].

The second possibility is that neutrinos are Majorana particles, meaning they are their own antiparticles: $\nu^C = \nu$. In this case, right-handed neutrinos are not required, since the charge-conjugated left-handed neutrino field ν_L^C transforms as a right-handed field. The Majorana condition for a general field is expressed as

$$\psi = C\overline{\psi}^T\,, \tag{2.14}$$

where C is the charge-conjugation operator and $\overline{\psi}^T = (\psi^\dagger\gamma^0)^T$ is the transposed Dirac adjoint of ψ. If neutrinos are Majorana particles, the total lepton number $L = L_e + L_\mu + L_\tau$ is not conserved, violating a fundamental symmetry of the Standard Model. This has profound implications for physics beyond the Standard Model, including potential mechanisms such as leptogenesis, which could potentially explain the matter-antimatter asymmetry of the universe [13]. Moreover, Majorana neutrinos introduce two additional CP-violating phases, known as Majorana phases, to the PMNS matrix. The full Majorana mixing matrix can be written as

$$\mathcal{U}^M = \mathcal{U}^D \cdot \mathrm{diag}(1, e^{i\alpha}, e^{i\beta})\,, \tag{2.15}$$

where $\mathcal{U}^D$ is the standard Dirac neutrino mixing matrix, and α, β are Majorana phases [12]. These phases, however, do not affect neutrino oscillation probabilities since they cancel out in the oscillation formula (equation (2.12)). The Majorana mass term for the neutrino is given by

$$\mathcal{L}^M_{mass} = -\frac{1}{2} m \left(\overline{\nu_L^C} \nu_L + \overline{\nu_L} \nu_L^C \right) , \tag{2.16}$$

where m is the Majorana mass term.

While this discussion assumes three active neutrino species, extensions with sterile neutrinos are under investigation but are not addressed here. A compelling framework that incorporates both Dirac and Majorana mass terms is the Type-1 see-saw mechanism [14]. In this scenario, heavy right-handed neutrino fields are introduced as Standard Model singlets. These fields allow for both Dirac mass terms m_D, arising from the Higgs mechanism as in the Dirac sense, and a large Majorana mass term M_R for the right-handed neutrinos. The resulting mass Lagrangian takes the form:

$$\mathcal{L}_{mass} = -\frac{1}{2} \begin{pmatrix} \overline{\nu_L} & \overline{(\nu_R)^C} \end{pmatrix} \begin{pmatrix} 0 & m_D \\ m_D^T & M_R \end{pmatrix} \begin{pmatrix} (\nu_L)^C \\ \nu_R \end{pmatrix} + \text{h.c.} , \tag{2.17}$$

where the Hermitian conjugate is not explicitly written. Diagonalizing this mass matrix yields two mass eigenstates: a very light neutrino with mass approximately $m_\nu \approx \frac{m_D^2}{M_R}$ and a heavy neutrino with mass $\sim M_R$. The light state predominately corresponds to the observed active neutrino, while the heavy state is mostly sterile and lies far beyond the electroweak scale.

The see-saw mechanism not only naturally explains the smallness of neutrino masses but also implies that neutrinos are Majorana particles. Neutrino oscillation experiments indicate a mass hierarchy characterized by $\Delta m^2_{21} \ll |\Delta m^2_{31}| \simeq |\Delta m^2_{32}|$, leading to two possible mass orderings [3, 11, 15], which are illustrated in figure 1:

- Normal ordering: $m_1 < m_2 \ll m_3$
- Inverted ordering: $m_3 \ll m_1 < m_2$

2.4 Neutrinoless double beta decay

The most promising method to determine whether neutrinos are Dirac or Majorana particles is to search for processes that violate the total lepton number. If such a process is discovered, it would imply that neutrinos are Majorana particles. The most sensitive experimental probe for the violation of L is neutrinoless double beta decay, a simultaneous beta decay of two neutrons without emission of neutrinos [3]. Beta decays describe the decay of a neutron into a proton, $(A, Z) \to (A, Z+1)$ inside a nucleus (β^--decay). If $m_{(A,Z+2)} < m_{(A,Z)}$, double beta decay ($2\nu\beta^-\beta^-$) can occur:

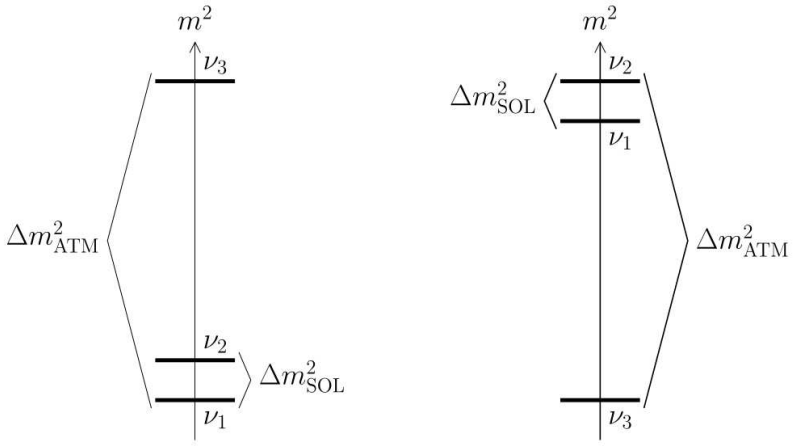

Figure 1: The two possible neutrino mass hierarchies. Normal ordering is shown on the left and inverted ordering on the right [13].

$$(A, Z) \to (A, Z+2) + 2\, e^- + 2\, \bar{\nu}_e\,. \tag{2.18}$$

This is a second-order weak interaction. As such, the decay rate scales as $\Gamma^{2\nu} \propto G_F^4$, leading to a highly suppressed transition probability. The resulting half-lives are typically of order $T^{2\nu}_{1/2} \sim 10^{20}$ yr, depending on the nuclear matrix element and available phase space. To be experimentally feasible, normal β^- decay should be energetically forbidden ($m_{(A,Z+1)} > m_{(A,Z)}$) [12]. Only if neutrinos are Majorana particles, L is not conserved, and neutrinoless double beta decay can occur:

$$(A, Z) \to (A, Z+2) + 2\, e^-\,. \tag{2.19}$$

The Feynman diagram for $0\nu\beta\beta$ decay is shown in figure 2 and illustrates why this process is not allowed in the SM: the upper vertex represents $W^- \to e^- + \bar{\nu}_e$ and the lower vertex corresponds to the decay $W^- + \nu_e \to e^-$. Since the exchanged neutrino is emitted as an antineutrino and absorbed as a neutrino, the process is only possible if $\nu_e = \bar{\nu}_e$.

For $0\nu\beta\beta$ decay, the decay rate $\Gamma^{0\nu}$ is given by Fermi's golden rule:

$$\Gamma^{0\nu} = \frac{\ln 2}{T^{0\nu}_{1/2}} = \ln 2 \cdot G^{0\nu}(Q, Z) \cdot \left|M^{0\nu}\right|^2 \cdot \left(\frac{m_{\beta\beta}}{m_e}\right)^2\,. \tag{2.20}$$

Here, $T^{0\nu}_{1/2}$ is the half-life, $G^{0\nu}(Q, Z)$ the phase space factor, and Q the

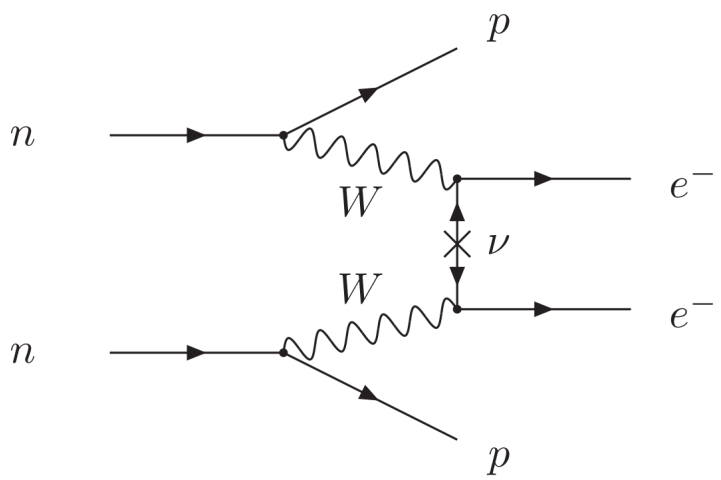

Figure 2: Feynman diagram of neutrinoless double beta decay: Two neutrons decay into protons and electrons. Only if the neutrino is its own antiparticle, it can be emitted and reabsorbed by a virtual intermediate state, enabling this process [3].

total kinetic energy released. The quantity $M^{0\nu}$ is the nuclear matrix element, $m_{\beta\beta}$ the effective Majorana mass and m_e the electron mass. The nuclear matrix element $M^{0\nu}$ carries a model-dependent uncertainty of a factor 2-3 [16], and the effective Majorana mass is given by:

$$|m_{\beta\beta}| = \left| \sum_i \mathcal{U}_{ei}^2 m_i \right| \tag{2.21}$$

where $\mathcal{U}$ is the PMNS matrix and m_i are the mass eigenstates. A full derivation of (2.20) is presented in [17].

In this work, we refer to the Q-value of $0\nu\beta\beta$ decay as $Q_{\beta\beta}$. Furthermore, we define the signal half-rate as

$$\mathcal{S} = \frac{1}{T_{1/2}^{0\nu}} \,. \tag{2.22}$$

This definition of $\mathcal{S}$ corresponds to the half-rate, as it is the inverse of the half-life. This follows the common usage in the field (see [18], eq. (8)), where the phase-space factor $G^{0\nu}$ is defined to include a factor of $1/\ln 2$:

$$G^{0\nu}(Q, Z) \propto \frac{G_F^4}{\ln 2} \cdot Q_{\beta\beta}^5 \tag{2.23}$$

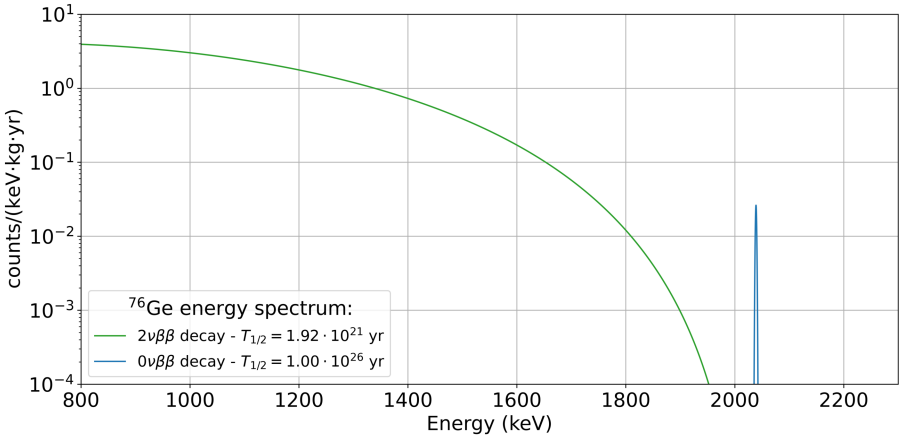

Figure 3: Expected energy spectra for $2\nu\beta\beta$ and $0\nu\beta\beta$ decay in ^{76}Ge. The $2\nu\beta\beta$ decay spectrum is computed using the phase space distribution from [22], scaled to a half-life of $1.92 \cdot 10^{21}$ yr. The $0\nu\beta\beta$ decay signal is modeled as a Gaussian at $Q_{\beta\beta} = 2039$ keV, assuming a detector resolution of FWHM = 2.5 keV and a half-life of 10^{26} yr, following the estimates from the GERDA collaboration [23].

2.4.1 Experimental searches

Double beta decay should be observable in 35 isotopes, but due to practical reasons, only 9 are of experimental interest: ^{48}Ca, ^{76}Ge, ^{82}Se, ^{96}Zr, ^{100}Mo, ^{116}Cd, ^{130}Te, ^{136}Xe and ^{150}Nd [19]. The detection of (neutrinoless) double beta decay is characterized by measuring the kinematic parameters of the emitted electrons [20]. In the case of $0\nu\beta\beta$ decay, the total energy carried by the emitted electrons equals the Q-value of the decay (neglecting nuclear recoil), resulting in a monoenergetic peak. In practice, this peak is broadened due to the finite energy resolution of the detector. In contrast, $2\nu\beta\beta$ decay produces a continuous energy spectrum, as the undetected neutrinos carry away varying amounts of energy. Figure 3 illustrates the expected energy depositions for both decay modes in the transition $^{76}\text{Ge} \rightarrow {}^{76}\text{Se}$. The Q-value for this process is $Q_{\beta\beta} = 2039.0612 \pm 0.0075$ keV [21].

The sensitivity to the half-life depends heavily on the amount of background:

$$T_{1/2}^{0\nu} \propto \begin{cases} a \cdot M \cdot \epsilon \cdot t & \text{background free} \\ a \cdot \epsilon \cdot \sqrt{\frac{M \cdot t}{B \cdot \Delta E}} & \text{with background} \end{cases} \tag{2.24}$$

Here, a is the isotopic abundance, ϵ the efficiency of the detection, M the total mass used in kg, t the time measured in years, B the background in counts/(keV·kg·yr), and ΔE is the energy resolution in keV [12, 20]. Equation (2.24) shows why it is of utmost importance to achieve a quasi-free background: If there are many background events, the sensitivity scales $(M{\cdot}t)^{1/2}$ instead of $M{\cdot}t$. Moreover, the energy resolution must be sufficiently good, otherwise the sharp peak at $Q_{\beta\beta}$ is indistinguishable from the $2\nu\beta\beta$ decay tail [24].

Neutrinoless double beta decay has not been observed yet. However, several different experiments can constrain the $0\nu\beta\beta$ decay half-life: The KamLAND-Zen consists of a tank filled with a liquid scintillator that contains a balloon of Xenon-loaded liquid scintillator as the $0\nu\beta\beta$ decay source. Inside the tank are almost 2000 PMTs alongside the walls. KamLAND-Zen achieved $T^{0\nu}_{1/2} > 2.3 \times 10^{26}$yr with ^{136}Xe [25].

The CUORE experiment is a cryogenic calorimeter that consists of 988 TeO_2 crystals. Energy depositions in the crystals will lead to a temperature rise, which is transformed into an electrical signal. The experiment measured a half-life of $T^{0\nu}_{1/2} > 2.2 \times 10^{25}$yr using over 200 kg of ^{130}Te [26]. Reported values are at 90% C.L.

CUPID-Mo, which served as a demonstrator for the upgrade of CUORE, contains enriched Li_2MoO_4 scintillating calorimeters, obtaining a half-life of $T^{0\nu}_{1/2} > 1.8 \times 10^{24}$ for ^{100}Mo [27].

A very promising approach is ionizing radiation detection, as shown by the GERDA and MAJORANA DEMONSTRATOR (MJD) experiments. While not exclusively to these setups, both experiments showed that a HPGe-based approach is highly efficient in terms of energy resolution and intrinsic background suppression due to the detector-source identity [23]. The GERDA collaboration achieved an unprecedentedly low background of 5.2×10^{-5} counts/(keV·kg·yr) and an exposure-weighted energy resolution of (2.9 ± 0.1) keV full width half maximum (FWHM) at $Q_{\beta\beta}$ for the detector geometry used in this work . The final value reported by the GERDA collaboration was $T^{0\nu}_{1/2} > 1.8 \times 10^{26}$yr at 90% C.L. [28]. MJD reached an impressive energy resolution of (2.55 ± 0.09) keV FWHM at $Q_{\beta\beta}$, with a final half-life of $T^{0\nu}_{1/2} > 8.3 \times 10^{25}$ yr at 90% C.L. [29].

In addition to direct searches, complementary constraints on the absolute neutrino mass scale arise from cosmological observations and kinematic measurements. The Planck satellite has placed an upper limit on the sum of neutrino masses, currently constrained to $\sum m_\nu < 0.16$ eV at 95% confidence level under the ΛCDM model [30].

On the experimental side, the KATRIN experiment probes the kinematic endpoint of tritium beta decay, directly measuring the effective electron neutrino mass: $m_\nu = \sqrt{\sum |U_{ei}|^2 m_i^2}$. Its latest results set an upper limit of $m_\nu < 0.8\,\mathrm{eV}c^{-2}$ at 90% CL [31]. While not sensitive to the Majorana or Dirac nature, this result further constrains the possible values of $m_{\beta\beta}$, especially in the case of inverted mass ordering.

2.5 Current challenges and future prospects

Assuming three neutrinos, oscillation experiments constrain six parameters: Two mass differences, three mixing angles, and a CP-violating phase. A global analysis of neutrino oscillation data, such as provided by the NuFIT collaboration, yields ever-tighter bounds on these parameters [32].

However, oscillation experiments are insensitive to the absolute neutrino mass scale. In this regard, cosmological observation provides a complementary probe: Relic neutrinos affect the large-scale structure of the universe and leave an imprint on the cosmic microwave background through their contribution to the total energy density. To determine the neutrino mass hierarchy, neutrinoless double beta decay experiments remain a promising approach, as the measured half-life can be translated into upper limits on the effective Majorana mass – assuming that the decay is dominated by the exchange of light Majorana neutrinos [28]. The sensitivity difference between normal and inverted ordering arises because in the latter case, the two heaviest neutrino mass eigenstates contribute significantly to $m_{\beta\beta}$, leading to a non-zero lower bound.

For normal ordering, destructive interference between the contributions from the three mass eigenstates can suppress $m_{\beta\beta}$ values close to zero, potentially beyond the reach of upcoming experiments. Current limits include $m_{\beta\beta} < 79 - 180$ meV from GERDA and $m_{\beta\beta} < 36 - 156$ meV from KamLAND-Zen [28, 33]. The relatively large range in values reflects uncertainties in the nuclear matrix element involved in the decay process.

Figure 4 shows the allowed regions for the effective Majorana masses $m_{\beta\beta}$ as a function of the lightest neutrino mass m_{min}, for both normal and inverted mass ordering. For the inverted hierarchy, next-generation $0\nu\beta\beta$ decay experiments such as SNO+ (using Cherenkov detectors), NEXT (a xenon time projection chamber), or LEGEND (using HPGe detectors) are expected to reach sensitivities that fully probe the relevant parameter space [34].
Among these, the LEGEND experiment stands out for its combination of

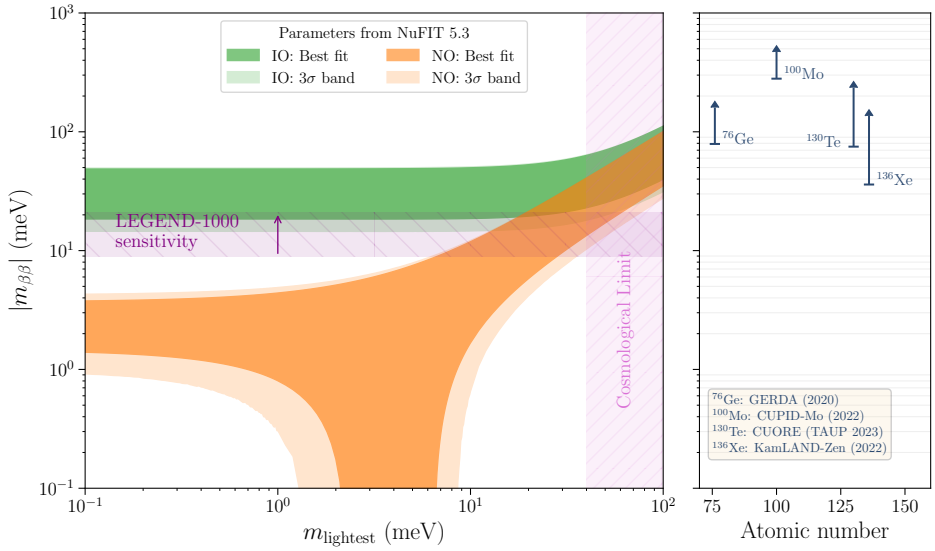

Figure 4: Left: Effective Majorana mass $m_{\beta\beta}$ for inverted and normal ordering as a function of the lightest neutrino mass m_{min}, including the projected sensitivity of the LEGEND-1000 experiment. Right: Effective Majorana masses obtained for different isotopes. The figure is created with code from [35].

excellent energy resolution, ultra-low background techniques and scalability. With its staged approach from LEGEND-200 to LEGEND-1000, the experiment is designed to explore half-lives beyond 10^{28} years and approach the full coverage of inverted ordering. The next chapter introduces the LEGEND experiment in detail, with an emphasis on the detector technology.

3 The LEGEND experiment

Following the discussion of neutrino physics and the significance of neutrinoless double beta decay, this chapter introduces the LEGEND experiment, a next-generation effort to search for neutrinoless double beta decay. It builds on the advancements of its predecessors, MJD and GERDA. LEGEND reuses the infrastructure at the Laboratori Nazionali del Gran Sasso (LNGS), previously used by GERDA, and upgrades it to improve both background suppression and detector performance. The chapter begins with a short overview of how LEGEND evolved from GERDA and MAJORANA, focusing on key improvements in detector design and instrumentation. It then introduces the high-purity germanium (HPGe) detectors used in the experiment, and the principles of semiconductor physics, including doping and charge collection. Finally, it covers how these detectors are used to identify neutrinoless double beta decay events, how pulse shape discrimination (PSD) helps to reject background, and how the detectors are calibrated in LEGEND-200.

3.1 From GERDA to LEGEND

The GERmanium Detector Array (GERDA) aimed to search for $0\nu\beta\beta$ decays using high-purity germanium detectors enriched to 87% in ^{76}Ge. Located at LNGS, the setup was shielded by 3500 meters of water equivalent of rock, significantly reducing the cosmogenic background. GERDA operated in two phases. Phase I reached a background index of 10^{-2} counts/(keV·kg·yr) with a total exposure of 21.6 kg $\cdot$ yr. The experiment set a new lower limit on the half-life of $T^{0\nu}_{1/2} > 2.1 \times 10^{25}$yr (90% CL). In phase II, the goal was to improve sensitivity by an order of magnitude. This was addressed by introducing a new detector geometry with better energy resolution and improved pulse shape discrimination (BEGe detectors), as well as a new liquid argon (LAr) veto system and a denser detector array [36].

In GERDA, the HPGe detectors were attached on strings and submerged in a cryostat filled with 64 m^3 LAr, which served both as a coolant and as an active shield against background radiation, including cosmic and environmental radiation. Surrounding the cryostat was a 590 m^3 ultra-pure water tank equipped with 66 PMTs providing Cherenkov veto against muons. The clean room was located above the cryostat [36]. Within the LAr volume, an additional veto system detected energy deposits in both the germanium detectors and the LAr, with scintillation light being measured through wavelength-shifting fibers. These fibers converted the 128 nm scintillation

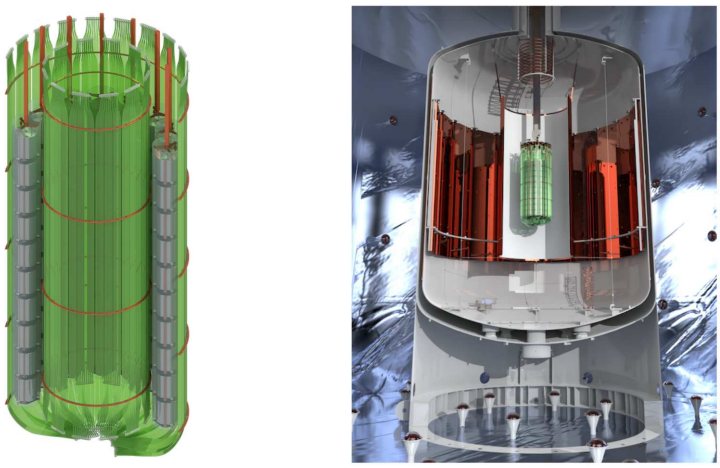

Figure 5: LEGEND-200 design. The left image shows the germanium detectors mounted on the strings, surrounded by optical fibers that guide the LAr scintillation light to silicon photomultipliers. The right image shows the cryostat located in the water Cherenkov tank, which is equipped with PMTs to detect coincident muons [38].

light into green light, directing the photons to silicon photomultipliers for precise detection [36].

The LEGEND experiment continues from both GERDA and MJD. The idea is to combine GERDA's low background techniques with the high energy resolution of MJD. The final stage, LEGEND-1000, is planned to consist of 1000 kg of germanium and aims for a discovery sensitivity on the order of 10^{28} years at 99.7% C.L. [37]. In a first step, LEGEND-200 aims to employ 200 kg of HPGe detectors inside the existing infrastructure of GERDA, combining it with the low-noise electronics of the MAJORANA experiment, which was achieved by using low-mass front-end electronics close to the detector. The goal is to enter the background-free regime, which is defined as having < 1 expected background counts in the region of interest over the experiment's exposure [38]. For LEGEND-200, this corresponds to a background index of $2 \times 10^{-4}\text{counts}/(\text{keV} \cdot \text{kg} \cdot \text{yr})$. The design of the LEGEND-200 detector setup is shown in figure 5.

3.2 Germanium detectors

High-purity germanium detectors are at the core of the LEGEND experiment. Their excellent energy resolution and ability to reconstruct interaction topologies make them particularly well suited for $0\nu\beta\beta$ decay searches.

3.2.1 Germanium semiconductors

In crystalline solids like germanium, the periodic atomic structure gives rise to a quasi-continuum of energy levels known as energy bands. The valence band is the highest band fully occupied by electrons, while the conduction band is the lowest unoccupied band. The energy difference between these two bands is called the band gap, and its value determines the material's electrical properties. In conductors, the bands overlap; in insulators, the band gap is large. Semiconductors fall in between with band gaps typically below < 3 eV. In the case of germanium, the band gap is 0.67 eV [39]. Each atom contributes four electrons that form covalent bonds with neighboring atoms. At zero temperature, the valence band is completely filled and the conduction band is empty, resulting in a very low electrical conductivity. However, at non-zero temperatures, thermal excitation can promote electrons into the conduction band, creating electron-hole pairs. Holes represent the absence of electrons in the valence band and act as positively charged carriers. Although the ions themselves are stationary, the movement of holes contributes to electrical conduction [39, 40]. The law of mass action relates the concentration of electrons n and holes h in a semiconductor at thermal equilibrium:

$$n \cdot p = \text{constant}\,. \tag{3.1}$$

For a perfectly pure semiconductor at room temperature, the charge carrier concentrations are very low, resulting in correspondingly low electrical conductivity. However, in practice, there are always small amounts of impurities present, which increase the number of free charge carriers and thereby the conductivity. There are two important types of electronically active impurities: Atoms that act as electron donors are called n-type, while atoms that readily accept electrons (and thus effectively donate holes) are referred to as p-type. Semiconductors are typically doped with either type to increase the conductivity to a reasonable value. For tetravalent elements like germanium, n-type doping is achieved by introducing pentavalent atoms like phosphorus. The additional outer shell electron cannot participate in a covalent bond and remains only loosely bound to the positively charged

nucleus. Its energy level lies just below the conduction band, and only a small amount of thermal energy is required to excite it into the conduction band [39, 41, 42]. P-type doping is achieved by adding trivalent atoms such as boron or aluminium, which act as electron acceptors with energy levels just above the valence band. At high enough temperatures, where the electrons are no longer bound to the dopant atoms, n-type doping will lead to free electrons and p-type doping leads to free holes, thus increasing the free charge carrier concentration in either case. Regions with very high doping concentrations are particularly well suited for electrical contacts due to their high conductivity. These regions are referred to n^+ and p^+, depending on the type of impurity introduced [40, 42].

3.2.2 Semiconductors as particle detectors

If p-type and n-type semiconductors are brought into direct contact, the concentration gradient causes electrons to diffuse into the p-doped region, while holes diffuse towards the n-doped region. In the overlapping region, known as the depletion zone, electrons and holes recombine. For each electron that diffuses away from the n-doped side, a net positive charge remains due to the immobile donor impurity. Analogously, for every hole that diffuses out of the p-doped side, an immobile acceptor impurity remains. This charge imbalance creates an electric field across the depletion region, which acts to counter further diffusion until equilibrium is reached [40, 42]. The depletion region contains almost no free charge carriers and therefore exhibits very high resistivity.

An external voltage can be applied across the junction in either direction: When the voltage is applied in the same direction as the concentration difference (i.e., opposite the existing electric field), the junction is said to be in forward bias. In this case, electrons can move easily from the n-doped to the p-doped region, and the conductivity increases.

The reverse case is more relevant for particle detection. In reverse bias, which is shown in figure 6, the external voltage is applied in the same direction as the existing electric field, opposing the natural diffusion of charge carriers. This pulls electrons from the p-doped region toward the n-side. Since the carrier concentrations are already low, the current is greatly suppressed, and the depletion region widens [39, 40, 42].

When ionizing radiation interacts in the depletion zone of a reverse-bias semiconductor, it deposits energy and generates a number of electron-hole pairs within a few picoseconds [44]. These charge carriers are separated by the electric field and drift towards their respective electrodes, inducing

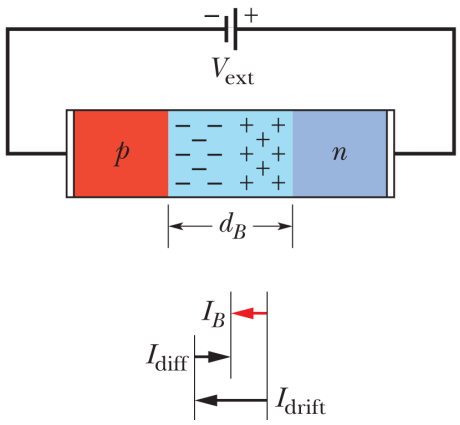

Figure 6: Schematic view of a p-n junction in reverse bias. Since the applied voltage opposes the direction of the diffusion I_{diff}, only a small current I_B remains. As a result, the depletion zone d_B widens and the conductivity decreases [43].

a measurable signal. Due to the low ionization energy in semiconductors, detectors of this type achieve a high signal-to-noise ratio. The probability for such interactions, and its dependence on photon energy in germanium, is illustrated in figure 7, which shows the mass attenuation coefficient for germanium, including prominent K- and L-shell absorption edges.

In HPGe detectors, the applied reverse bias is large enough to deplete almost the entire detector volume, but still below the breakdown voltage. The small region that remains undepleted is referred to as dead layer, where charge collection is incomplete and signals are suppressed [42].

The electron-hole pairs created during an interaction inside the detector drift toward the electrodes in a few microseconds. While in principle signals can be read out from either p^+ or n^+ electrode of the detector, the point-contact detectors in LEGEND-200 always read from the p^+ electrode. In LEGEND-200, the HPGe diodes are operated at 77 K, where the typical charge carrier drift velocities are on the order of 10^7 cm/s [46].

The measured signal does not result from the collection of discrete charges at the electrode itself, but rather from the movement of the charge carriers, which induces a current in the electrodes. The induced charge can be determined by integrating the normal component of the electric field $\vec{E}$ over the surface $\vec{S}$ surrounding the electrode:

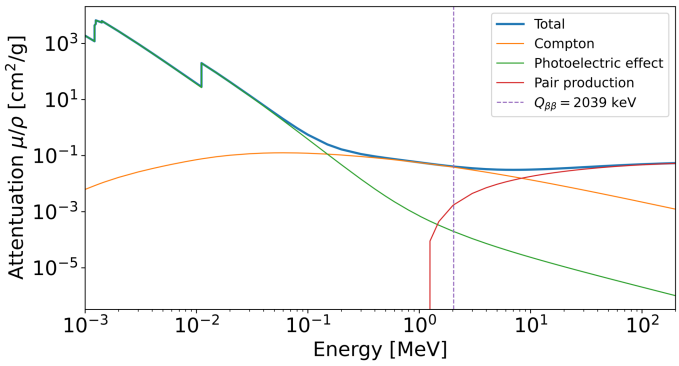

Figure 7: Mass attenuation coefficients for germanium, showing contributions from photoelectric absorption, incoherent (Compton) scattering, and pair production, along with the total attenuation. The region of interest around $Q_{\beta\beta}$ (horizontal line) is dominated by Compton scattering. Data from NIST XCOM [45].

$$Q = \oint_S \epsilon \vec{E} \cdot d\vec{S}\,, \tag{3.2}$$

where ϵ is the dielectric constant of the material. To achieve precise charge reconstruction, the electric field must, in principle, be calculated for many points along the carriers' trajectories [47]. However, this can be simplified using the Ramo-Shockley theorem [48, 49], which allows for the induced charge to be calculated more efficiently. In this approach, the readout electrode is set to 1 volt while all other electrodes are grounded. The induced charge from a single moving charge q at position $x(t)$ is then:

$$Q_{ind}(x(t)) = -q \cdot \varphi_0(x(t))\,. \tag{3.3}$$

The electric potential φ_0 is the weighting potential, and it only needs to be computed once for a given geometry. In most applications, it is assumed that all charge carriers are created at a single position in the active volume and that the electric field is strong enough for the charge carriers to reach saturated drift velocity. The total induced charge includes contributions from both electron and hole movement. For a planar detector geometry with thickness d, the induced charge can be expressed as

$$Q_{ind} = \frac{q_0}{d}\,(\text{electron drift distance} + \text{hole drift distance})\,, \tag{3.4}$$

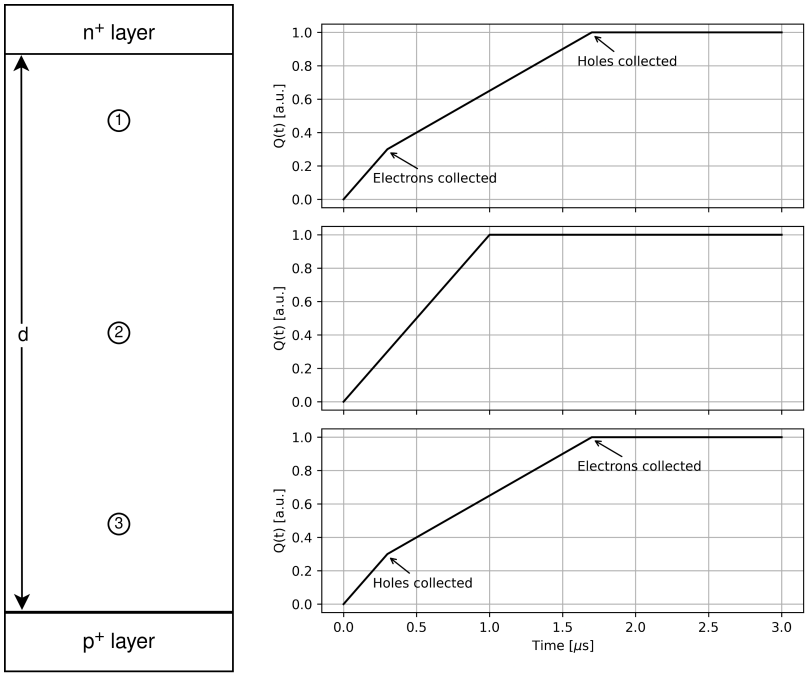

Figure 8: Shape of the leading edge for three energy depositions inside a planar HPGe detector. Charge depositions close to the p^+ layer (bottom panel) are characterized by a very fast hole-induced signal, followed by a slower electron-induced signal, as the latter must drift across the entire detector volume. Similarly, charge depositions close to the n^+ layer are characterized by fast signals from the electrons, followed by a slower hole-induced signal. Here, we assume a detector of height $d = 20$ cm and drift velocities of 10^7 cm/s.

where q_0 determines the maximum induced charge. It is defined as $q_0 = N \cdot e$ with N being the number of electron-hole pairs and e the electronic charge. Depending on the location of the energy deposition, either electrons or holes may dominate the signal. This results in characteristic pulse shapes, as illustrated in figure 8. The general concept also applies to more complex detector geometries [42].

In practice, however, signals recorded in HPGe detectors deviate from these idealized shapes due to additional effects that smooth out the signal. Crystal defects such as vacancies, dislocations, or impurities can locally trap charge carriers during their drift, preventing a fraction of the charge from reaching the electrodes. This reduces the signal amplitude and degrades

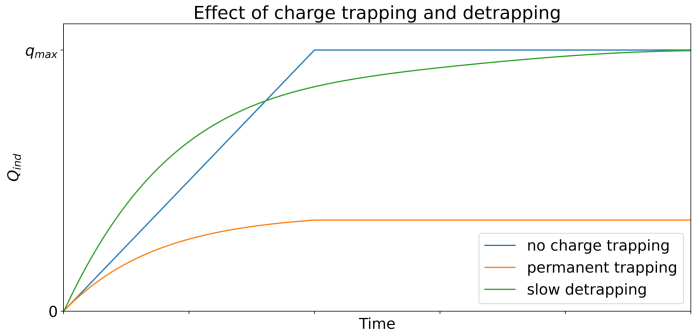

Figure 9: Effect of charge trapping and de-trapping on the rising edge of a HPGe waveform. Assuming a perfect semiconductor where no charge trapping occurs (blue line), all charges would arrive in a very short period of time, leading to a sharp edge when q_{max} is reached. On the other hand, permanent trapping would reduce the maximum amplitude. The combined effect of trapping and detrapping results in a smoothed waveform that rises quickly at first, then more slowly over time. For simplicity, only one type of charge carrier is shown. Adapted from [42].

energy resolution, as a variable amount of charge is lost. Not all trapped carriers remain trapped permanently. Some are released after a time delay through de-trapping, which partially restores the signal but leads to a longer rise time. Both effects are illustrated in figure 9 [42].

Other physical effects further modify the waveform in similar ways: Self-repulsion and diffusion cause the charge carriers to spread out into a charge cloud, rather than moving as point-like. As a result, not all charges arrive at the electrode simultaneously, and the sharp edge at q_{max} becomes smoothed out. Crystal anisotropy leads to non-uniform drift velocities, since the mobility of charge carriers depends on the orientation of the crystal axes. Table 1 shows electron and hole mobilities for different crystallographic axes, indicated by Miller indices. The charge carrier drift velocity $v_{e/h}$ is related to the mobility as:

$$v_{e/h} = \mu_{e/h} \cdot E \,, \tag{3.5}$$

where the subscript indicates electrons and holes, respectively.

Table 1: Electron and hole mobilities for different crystal axes in germanium at 77 K. The axes are indicated by Miller indices [50].

Crystal axis	μ_e **[cm^2/(Vs)]**	μ_h **[cm^2/(Vs)]**
$\langle 100 \rangle$	38609	61824
$\langle 111 \rangle$	38536	61215

3.2.3 HPGe detectors in LEGEND-200

The LEGEND-200 experiment initially employed the majority of the enriched HPGe detectors that were previously manufactured, characterized, and operated in the GERDA and MJD experiments. GERDA primarily used broad energy germanium (BEGe) detectors, while MJD mainly used p-type point contact (PPC) detectors, where the main germanium crystal is p-doped. Both detector types are optimized for $0\nu\beta\beta$ decay searches, offering excellent energy resolution, low background levels, and precise event reconstruction capabilities [38]. In later phases of both experiments, a new type of detector was introduced. The p-type inverted coaxial point contact (ICPC) detector allows for larger individual detector sizes while simultaneously increasing the pulse shape discrimination and energy resolution. Typical masses for these types are 1 kg for PPC, 0.7 kg for BEGe and 2-4 kg for ICPC detectors. For LEGEND-200, new ICPC detectors were developed and installed. Using larger detectors provides multiple advantages. Most importantly, it reduces the number of required cables, readout electronics, and mechanical support structures, all of which contribute to the overall background. Furthermore, a lower surface-to-volume ratio reduces the detector's sensitivity to surface-related backgrounds [38].

Traditional coaxial detectors, which were used in GERDA and have been installed in LEGEND-200 as well, do not offer the same level of performance as ICPC detectors. Consequently, they are not planned for use in future phases of LEGEND. In fact, the LEGEND collaboration will primarily rely on ICPC detectors moving forward. Figure 10 illustrates the three detector geometries. The layout of the LEGEND-200 detector array, as it was at the point of this work, is shown in figure 11.

The detectors used in LEGEND-200 are all p-type HPGe diodes, equipped with a relatively thin ($\sim$ 300 nm) p^+ electrode that is formed via boron implantation. The n^+ electrode, doped with lithium, is significantly thicker ($\sim$ 1-2 mm) and covers most of the detector surface [44]. The signal is read out from the small p^+ electrode, because the charge signal is dominated by hole drift in this geometry.

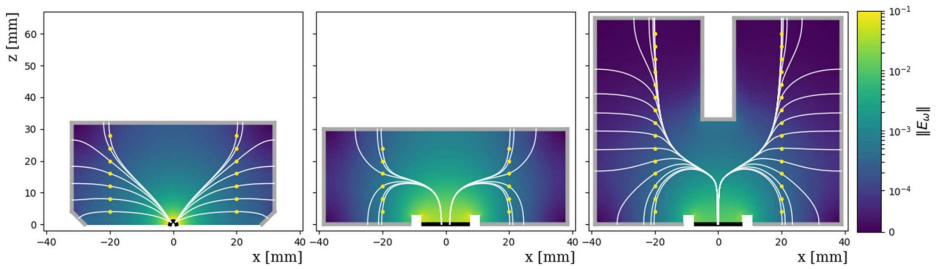

Figure 10: Cross section of three detector geometries used in LEGEND-200, including their respective weighting fields: PPC (left), BEGe (middle), and ICPC (right). The black line at the bottom indicates the p^+ contact, and the grey line surrounding the detector represents the n^+ electrode. Yellow dots mark exemplary interaction points, and white lines show the corresponding charge carrier trajectories. Plot from [38].

In LEGEND-200, detector signals are processed by a two-stage resistive-feedback charge-sensitive amplifier (CSA) operated in liquid argon for cooling and radiopurity. The Low-Mass Front End is mounted only millimeters from each HPGe detector, using ultra-radiopure materials and a high-value amorphous-germanium feedback resistor. A differential amplifier, positioned 30–150 cm away, boosts the signal for transmission to the data acquisition (DAQ) system. This design achieves electronic noise equivalent to < 1 keV FWHM, energy resolution of ≤ 2.5 keV at $Q_{\beta\beta}$, and fast (≤ 100 ns) rise times, enabling precise energy reconstruction and effective pulse-shape discrimination [51].

The DAQ system reads out the signals at a rate of 62.5 MHz, corresponding to a sampling interval of 16 ns. This is sufficiently fast to resolve the characteristic signal rise times. The digitized waveform is a voltage step with an exponential decay from the feedback circuit and superimposed electronic noise, and therefore requires further processing.

The digital processing begins with a baseline subtraction and pole-zero cancellation to remove the exponential decay, yielding an idealized step. The resulting waveform is then step-like. The signal is then shaped using a trapezoidal (trap) filter, implemented by subtracting two moving averages of different widths to produce a waveform with a defined rise, flat top, and fall. The flat top is rounded via additional moving averages to yield a single, well-defined maximum. The amplitude is taken as this maximum, and its value is proportional to the collected charge [51, 52].

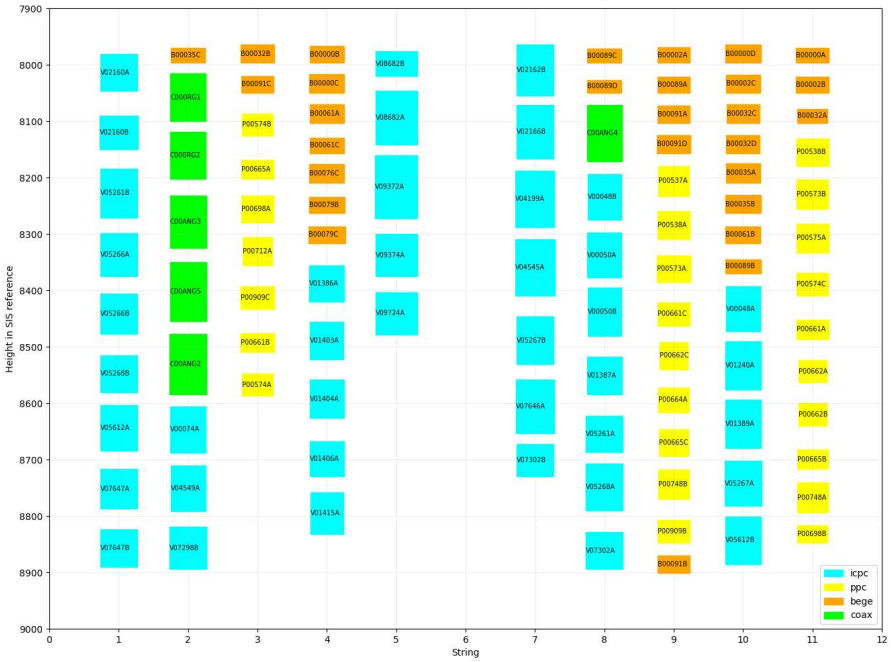

Figure 11: Detector map of the HPGe array installed in LEGEND-200. The germanium diodes are mounted on 11 strings. The y-axis indicates the vertical position relative to the top of the detector, where the source insertion system is located. The colors indicate the different detector geometries. ICPC detectors are blue, PPC detectors are yellow, BEGe detectors are orange and the coaxial detectors are indicated in green. Image credits of Sandro Gaelli.

3.3 Calibration of the LEGEND-200 experiment

A successful search for $0\nu\beta\beta$ decay requires excellent energy resolution and a stable energy scale, which in turn requires regular calibration of every detector. After digital signal processing, the waveform remains in units of ADC counts. To convert these into physical energy, radioactive sources with a well-known γ spectrum are periodically inserted into the cryostat using an automated source insertion system (SIS). This is necessary because the isotope used for LEGEND-200, as well as GERDA and MJD, ^{228}Th, has a relatively short half-life of 1.9 years [38]. The SIS allows sources to be inserted for calibration runs and removed during physics data-taking. Sources are

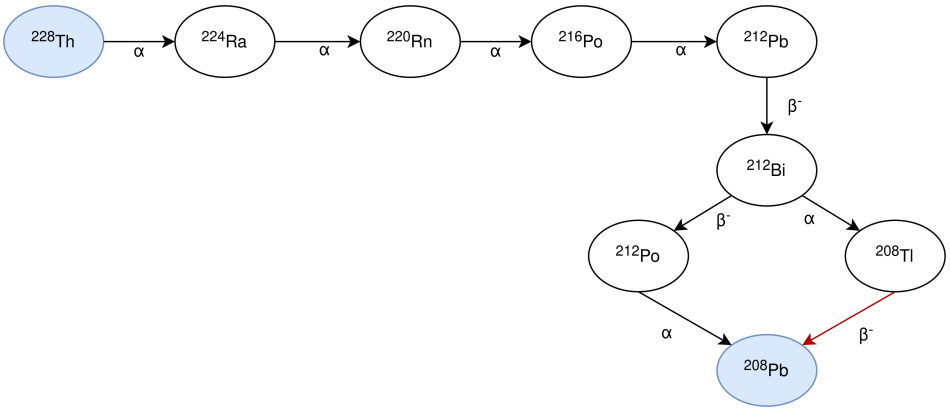

Figure 12: Decay chain of ^{228}Th to ^{208}Pb. For energy calibration and pulse shape studies, the β^- decay of ^{208}Tl is most important, because it produces several γ lines used for calibration. This decay is indicated by the red arrow.

guided near the detector strings through dedicated copper funnels, which are required to navigate the sources to their calibration positions next to the detectors. LEGEND-200 comprises four SIS, each accommodating four sources [53].

Once sufficient statistics are collected, the resulting γ lines appear in the energy histogram; matching their known energies to the measured ADC positions yields the calibration curve. In addition to determining the absolute energy scale, calibration is essential for evaluating the detector's energy resolution. This is critical for $0\nu\beta\beta$ searches, as a precise determination of the peak at $Q_{\beta\beta}$ with a high resolution is necessary for effective background discrimination and therefore maximizing signal sensitivity.

^{228}Th decays through a series of α and β transitions to ^{208}Pb, emitting several characteristic γ rays along the way [54]. The decay scheme is shown in figure 12.

The most important line is the ^{208}Tl 2614.5 keV γ. When this photon undergoes pair production inside a HPGe detector, the resulting electron deposits its kinetic energy, and the positron annihilates into two 511 keV photons. Depending on whether none, one, or both annihilation photons escape, three distinct peaks appear:

Full-energy peak (2614.5 keV): No photons escape, and the waveform contains multiple events.

Single-escape peak (2103.5 keV): One photon escapes, there are still multiple events per waveform due to the time separation between deposits.

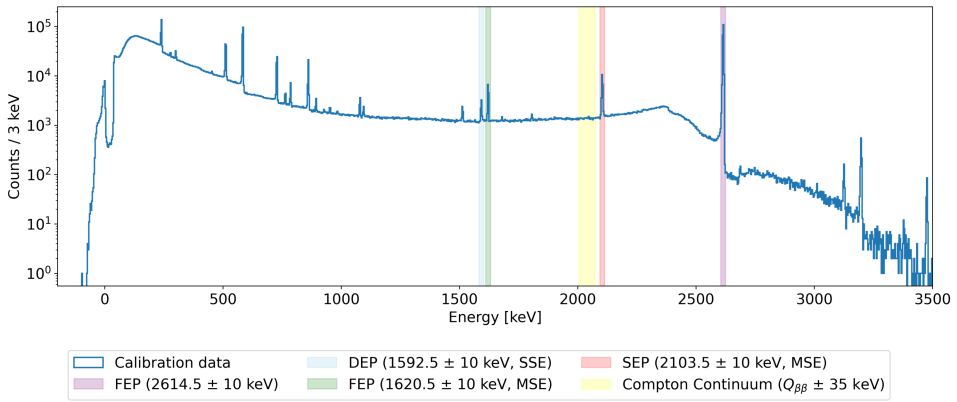

Figure 13: Energy spectrum from a ^{228}Th calibration run (period 9, run 3) in LEGEND-200. The spectrum corresponds to a single IC detector (V09372A) and a 5-hour data acquisition. Prominent γ-ray lines used for calibration and PSD studies are indicated.

Double-escape peak (1592.5 keV): Both photons escape, and the remaining energy deposition is highly localized and thus single-site [44, 54].

A further useful line is the ^{212}Bi full-energy peak at 1621 keV. Its proximity to the double-escape peak makes it convenient for energy-independent comparisons of single-site and multi-site populations [44]. Figure 13 shows a representative ^{228}Th calibration spectrum recorded with a single ICPC detector in LEGEND-200, with the key peaks highlighted.

In addition to ^{228}Th, LEGEND also recorded data with a ^{56}Co source. It is not used for detector calibration, but since it emits several γ lines at energies that complement those of ^{228}Th, it is a valuable source for studying the energy dependence of the PSD efficiency.

3.4 Background rejection: pulse shape discrimination

Even with large detector masses and multi-year exposures, only a handful of $0\nu\beta\beta$ decay events are to be expected. At the same time, numerous background processes can deposit energy in the region of interest around $Q_{\beta\beta} = 2039$ keV. Without effective background rejection methods, these events would obscure the $0\nu\beta\beta$ signal entirely. Moreover, as shown in equation (2.24), the experimental sensitivity to the half-life scales inversely with the square root of the background index, meaning that lowering the background significantly improves sensitivity. Background rejection in LEGEND

is achieved through a combination of material selection, active veto systems, and signal-based analysis techniques. Of particular importance for this work is pulse shape discrimination, which is possible because HPGe detectors record ionization signals with high temporal and spatial precision. In LEGEND, we distinguish between four characteristic pulse shapes:

Single-site events (SSE) are characterized by very localized energy depositions. This applies to both $0\nu\beta\beta$ and $2\nu\beta\beta$ decays, where the electrons deposit their energy within a small volume (typically ~ 1 mm^3). The resulting charges drift to the electrodes nearly simultaneously, and the waveform resembles that of a single interaction.

Multi-site events (MSE) involve energy deposited at multiple locations, typically due to multiple Compton scatterings of high-energy γ rays from natural radioactivity. These events are classified as background in the context of neutrinoless double beta decay.

Surface events occur near the detector boundaries and are also associated with the background. **P-contact events**, such as α interactions near the p$^+$ electrode, produce a fast-rising signal, since the drift path is short. Such surface events are particularly dangerous because their localized energy deposition can mimic the topology of SSEs, but their distorted charge collection leads to subtle differences in the pulse shape that must be carefully identified and rejected. In contrast, **n-contact events**, which originate near the n^+ electrode, involve long drift paths for the holes through the entire detector. These signals tend to rise more slowly and may show distortions due to trapping, de-trapping, or charge loss in the dead layer.

Figure 14 illustrates these event types. The standard PSD parameter is the amplitude over energy ratio (A/E), also referred to as AoE. Here, A denotes the maximum of the waveform and E is the reconstructed event energy, expressed in keV after energy calibration. The A/E parameter is sensitive to the event topology: in SSE, all the energy is deposited in a localized volume, leading to a short charge collection time and thus a sharply peaked current pulse. In MSE, the energy is deposited in several locations, resulting in less peaked current pulses and smaller A/E ratios.

The A/E distribution is calibrated by performing peak fits on ^{228}Th calibration data. In the Compton bands, the A/E distribution is characterized by a Gaussian single-site band with a low-side tail from MSE and n^+ surface events. A high-energy tail accounts for p^+ surface events. The tails are modelled as an exponential distribution convolved with a Gaussian.

Two corrections are applied. The A/E parameter depends on the drift time and the energy. Charge clouds drifting through the detector are subject to diffusion, which broadens the current signal. The A/E energy dependence

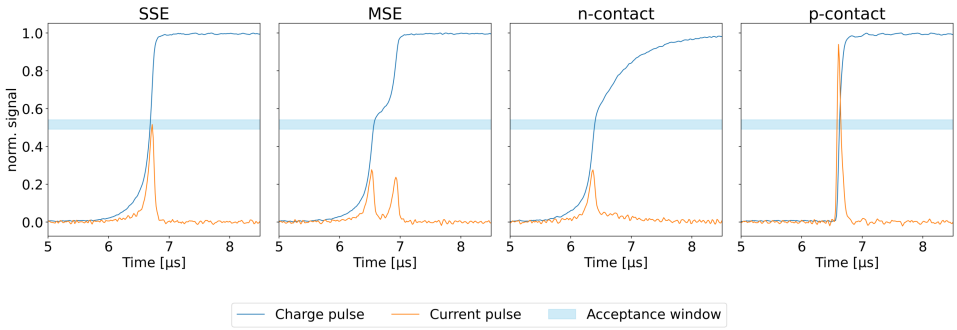

Figure 14: Example waveforms from a ^{228}Th calibration run in LEGEND-200, showing the normalized induced charge and corresponding current signal for four different event types. All examples have the same total deposited energy. The blue band illustrates, schematically, the A/E acceptance window for SSE-like signals; it is not derived from a quantitative cut but serves to illustrate the idea.

is modelled as:

$$\mu_{\mathrm{A/E}}(E) = a + b \cdot E \tag{3.6}$$

$$\sigma_{\mathrm{A/E}}(E) = \sqrt{c + \frac{d}{E^2}}\,, \tag{3.7}$$

where a, b, c, d are determined from calibration data in the Compton region from 900 keV to 2300 keV. For uniformity across detectors and calibration periods, a normalized A/E classifier is defined:

$$A/E_{\mathrm{classifier}} = \frac{\frac{A/E}{\mu_{A/E}(E)-1}}{\sigma_{A/E}(E)}\,. \tag{3.8}$$

This normalization accounts for detector-specific and time-dependent variations in the A/E response, enabling a unified classifier scale that allows consistent pulse shape discrimination across all detectors and calibration periods [55].

To isolate single-site events, two cut values are applied. The lower cut, which removes multi-site and n-contact events, is chosen such that 90% of events in the ^{208}Tl DEP survive. The upper cut is fixed to exclude high-amplitude surface events, in particular, α interactions [44].

The Late Charge (LQ) is a PSD parameter that measures the slowness of the final 20% of charge carriers for an event. It is defined as the area

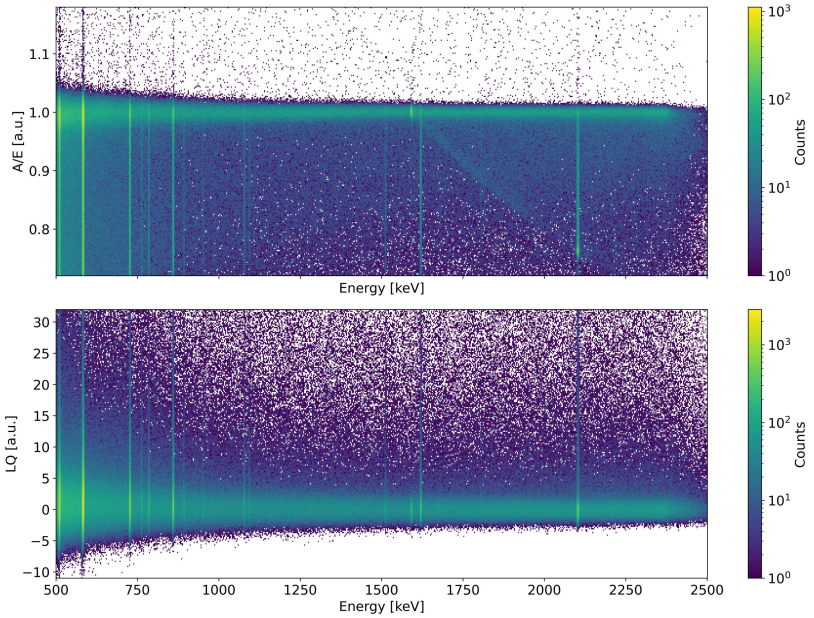

Figure 15: Distributions of A/E (top) and LQ (bottom) as a function of energy for the same dataset as in figure 13 (^{228}Th, period 9, run 3, detector V09372A). Both PSD parameters show a narrow single-site band, centered around 1 for A/E and around 0 for LQ, which is successfully corrected for energy dependence.

above the rising edge of the waveform after it reaches 80% of its maximum value. It is a powerful discriminator to identify events with slow charge collection, such as p^+ contact surface α interactions or events affected by significant charge trapping. Like the A/E parameter, the LQ parameter is both drift-time and energy dependent, and needs to be corrected for [55]. The A/E and LQ distributions as functions of energy for a single IC detector are shown in figure 15.

4 Machine learning principles and Transformer architecture

This chapter introduces fundamental concepts of machine learning and deep learning that underpin the waveform classification task addressed in this thesis. It begins with linear regression, a foundational method familiar to most physicists, to illustrate basic supervised learning principles. Next, the multilayer perceptron (MLP) is presented as a foundational neural network model to introduce core deep learning principles, such as optimization techniques and the backpropagation algorithm. Finally, the chapter explores the Transformer architecture, focusing on the encoder component, which is most relevant for waveform classification tasks like those in LEGEND-200. Core components of Transformer networks, such as positional encoding and the attention mechanism are explained, as they are essential for modeling the temporal structure of waveform data.

4.1 Machine learning and deep learning

There is no fixed definition of Artificial Intelligence (AI), but it usually describes systems that simulate intelligent behavior. Machine learning (ML) is a subfield of AI concerned with constructing mathematical models that learn patterns from data to perform tasks such as classifying events or estimating continuous quantities [56, 57]. ML algorithms are commonly grouped into several paradigms. Two of the most widely used are:

a) Unsupervised learning: Only the input data is available, and the goal is to discover structure without labeling targets. Typical techniques are clustering (partitioning data into similar subsets) and dimensionality reduction (compressing data while preserving salient information) [56].

b) Supervised learning: Models learn a mapping $f(x, \boldsymbol{\theta})$ from inputs x to known targets y. If y takes discrete values (e.g., classifying a waveform as signal or background), the problem is one of classification; if y is continuous, it is a regression task [56, 57].

To illustrate these concepts concretely, consider a one-dimensional linear regression, where the goal is to find a mapping of the form:

$$f(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 \cdot x \,. \tag{4.1}$$

In general, θ_0 is denoted as the bias, and θ_1 is the weight. Simply, the goal is to find the parameter vector $\boldsymbol{\theta}$ that minimizes the discrepancy between the model predictions $f(x, \boldsymbol{\theta})$ and the true outputs y_i. A common choice is the sum of squared errors, known as least squares loss:

$$L[\theta] = \sum_{i}^{N} \left(f(x_i, \boldsymbol{\theta}) - y_i\right)^2 . \tag{4.2}$$

Minimizing this loss yields the optimal parameters $\hat{\boldsymbol{\theta}}$ for which the model's predictions best approximate the training data:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, L\left[\boldsymbol{\theta}\right] . \tag{4.3}$$

This procedure is referred to as training the model. Once trained, the model can be used to make predictions on new, unseen input data. If the model is too simple to capture the true underlying structure of the data, it is said to be underfitted. Conversely, if the model is excessively complex and learns noise or random fluctuations in the data, it is overfitted and likely to generalize poorly to unseen inputs [57].

4.1.1 Multilayer perceptrons

Finding suitable features for machine learning models is often a challenging and time-consuming task. In fact, until recently, progress in machine learning was largely constrained not by model complexity but by the need for effective, manually crafted feature transformations. Deep learning (DL) addresses this limitation by integrating the feature transformation step directly into the model itself [58]. Instead of requiring handcrafted input features, DL models learn them automatically from raw data. This is accomplished by introducing additional parameters $\boldsymbol{\vartheta}$ that govern the structure of the feature transform:

$$f(x; \boldsymbol{\theta}, \boldsymbol{\vartheta}) = \theta_0 + \theta_1 \cdot \phi(x; \boldsymbol{\vartheta}) . \tag{4.4}$$

The key idea is that a recursive composition of simple functions can model highly complex, non-linear mappings. Non-linearity in the functions is essential, as without it, the stacked layers would collapse to a single affine transformation. This is formalized by expressing the overall model as a composition of layers:

$$f(\mathbf{x}; \boldsymbol{\theta}) = f_L(f_{L-1}(\ldots f_1(\mathbf{x}))\ldots)) . \tag{4.5}$$

where each f_l represents a transformation at layer l, which often involves a linear mapping followed by a non-linear activation. The input vector $\mathbf{x}$ is successively transformed through these layers, enabling the network to

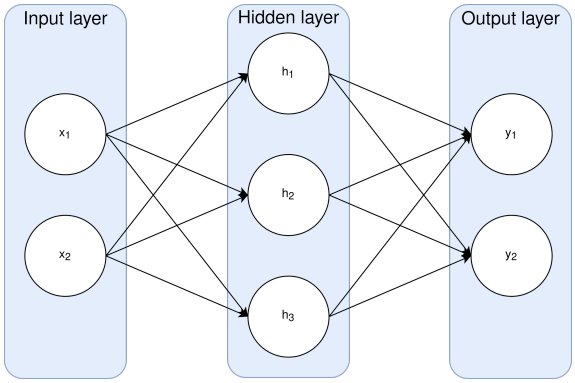

Figure 16: Visualization of a feedforward neural network with a single hidden layer containing three neurons, two inputs, and two outputs. The network is fully connected, meaning every unit in one layer is connected to every unit in the next. This is also referred to as a multilayer perceptron.

extract hierarchical features of increasing complexity. Models that follow this idea are referred to as feedforward neural networks, or more specifically, multilayer perceptrons (MLPs) [58, 59].

At the core of these architectures lies the perceptron, a computational unit that performs a weighted sum of its inputs followed by a nonlinear activation. By arranging many such units in layers, MLPs are able to model a wide range of nonlinear functions through learned hierarchical representations.

Figure 16 shows a shallow feedforward neural network, i.e., one with only a single hidden layer. In a fully connected layer, each neuron is connected to every neuron in the adjacent layers. Although fully connected layers are standard in MLPs, alternative connection patterns exist. The first layer is the input layer, the last is the output layer, and all intermediate layers are referred to as hidden layers [57]. In the figure, the hidden layer consists of three computational units, commonly referred to as neurons.

The computation of a hidden layer is shown in equation (4.6), where $\mathbf{x} \in \mathbb{R}^d$ is the input vector, $\mathbf{W}_\theta \in \mathbb{R}^{m \times d}$ the weight matrix and $\mathbf{b}_\theta \in \mathbb{R}^m$ is the bias vector. The non-linear activation function F is applied element-wise. The pre-activation (linear part) is given by $\mathbf{W}_{\boldsymbol{\theta}} \cdot \mathbf{x} + \mathbf{b}_{\boldsymbol{\theta}}$, and the activation function transforms it into the hidden representation $\mathbf{h}$.

An alternative form uses the augmented weight matrix $\mathbf{W'}_{\boldsymbol{\theta}}$ and augmented input $\mathbf{x}'$ to absorb the bias term:

$$\mathbf{h} = F\left[\mathbf{b}_\theta + \mathbf{W}_\theta \cdot \mathbf{x}\right] = F\left[\mathbf{W'}_\theta \cdot \mathbf{x'}\right] , \tag{4.6}$$

$$\mathbf{y} = \mathbf{b}_\vartheta + \mathbf{W}_\vartheta \cdot \mathbf{h} = \mathbf{W'}_\vartheta \cdot \mathbf{h'} . \tag{4.7}$$

where $\mathbf{h}$ is the hidden layer. The final output $\mathbf{y}$ is typically a linear transformation of the last hidden layer, optionally followed by an output activation depending on the task.

While the transformations used in neural networks are commonly referred to as linear layers, they are affine transformations. The inclusion of a bias term makes them not strictly linear in the mathematical sense. However, this can be reformulated as a purely linear operation by augmenting the input vector and weight matrix. Specifically, by extending the input as

$$\mathbf{x'} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} , \tag{4.8}$$

and defining the corresponding augmented weight matrix as

$$\mathbf{W'} = \begin{pmatrix} \mathbf{W} & \mathbf{b} \\ \mathbf{0}^\intercal & 1 \end{pmatrix} . \tag{4.9}$$

This trick is widely used in machine learning to streamline notation and implementation, and as such, the bias term is often implied rather than explicitly written. The use of this augmented formulation also clarifies why these layers are often referred to as linear, even though they technically are not.

4.1.2 Activation functions

Historically, the original perceptron used the Heaviside step function as its activation function [60]. However, modern neural networks typically employ smooth, non-linear, and almost everywhere differentiable activation functions, which enable both more expressive modeling and efficient optimization via gradient-based methods [59]. The non-linearity introduced by the activation function is essential. Without it, no matter how many layers are stacked, the overall mapping would collapse to a single affine transformation. Differentiability, on the other hand, is crucial for computing gradients during training. It is a remarkable theoretical result that a MLP with just a single hidden layer (given enough hidden units) can approximate any continuous function on a compact domain to arbitrary precision [61]. While

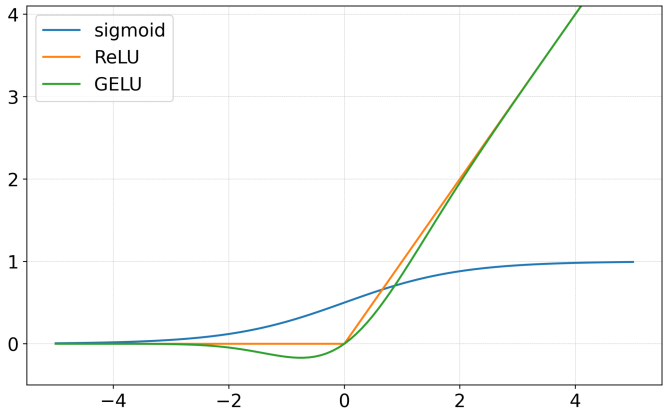

Figure 17: Common activation functions applied to inputs in the range $x \in [-5, 5]$.

this universal approximation property is foundational, in practice, deep networks tend to outperform shallow ones; they can more efficiently capture hierarchical or compositional structure, often requiring exponentially fewer units than shallow counterparts to represent the same function class [59, 62].

One of the most widely used activation functions today is the Rectified Linear Unit (ReLU), defined as:

$$\mathrm{ReLU}(x) = \max(x, 0)\,. \tag{4.10}$$

ReLU introduces sparsity by setting all negative inputs to zero while leaving positive values unchanged [63–65]. A more recent alternative is the Gaussian Error Linear Unit (GELU), a smooth activation function with a non-zero gradient everywhere, which can be advantageous in certain Transformer training setups. Its approximate form is given by:

$$\mathrm{GELU}(x) = \frac{x}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right]. \tag{4.11}$$

Unlike ReLU, GELU is smooth and differentiable everywhere, which can improve convergence in deeper networks. GELU is used as the default activation function in many Transformer-based architectures – deep learning models designed for sequential data processing [66]. Common activation functions are shown in figure 17.

4.2 Optimization

As seen in the case of linear regression, machine learning aims to determine the optimal parameters $\hat{\boldsymbol{\theta}}$ that describe the mapping $f(\mathbf{x};\boldsymbol{\theta})$ given a set of input-output pairs $\{\mathbf{x}_i, \mathbf{y}_i\}$. To do so, one minimizes a loss function, a quantitative measure for the discrepancy between the model's predictions and the true target values [57].

One of the most widely used principles for parameter estimation is maximum likelihood estimation (MLE). In this approach, we seek the parameters $\boldsymbol{\theta} \in \mathcal{T}$ that maximize the likelihood of the data given the parameters. Assuming that the training samples are drawn independently from the same distribution, the likelihood function factorizes, and the MLE takes the form:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, P\left(\mathcal{T} \mid \boldsymbol{\theta}\right) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} P\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}\right) . \tag{4.12}$$

In practice, working with products of probabilities is numerically unstable and analytically inconvenient. Instead, it is customary to use the negative log-likelihood (NLL) as a loss function, which transforms the product into a sum:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathrm{NLL}\left(\boldsymbol{\theta}\right) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[-\sum_{i=1}^{N} \log P\left(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}\right) \right] . \tag{4.13}$$

This formulation is widely used across regression and classification tasks. In fact, the least-squares loss introduced in equation (4.2) is a special case of equation (4.13), where we assume a Gaussian likelihood with a fixed variance [59, 67].

A widely used loss function in supervised classification tasks is the cross-entropy loss, which measures the dissimilarity between the empirical distribution $q(y)$ of the observed data and the predicted distribution $p(y)$ of the model. Minimizing cross-entropy is equivalent to minimizing the Kullback-Leibler divergence between the true and predicted distributions when the true distribution is fixed [57]. The cross-entropy is defined as:

$$\mathbb{H}(q, p) = -\sum_{y} q(y) \log p(y) . \tag{4.14}$$

In classification settings, the true distribution $q(y)$ is often represented as a one-hot vector, where the entry corresponding to the correct class is 1

and all others are 0. The predicted distribution $p(y)$ is typically the output of a softmax function applied to the final layer of the network. However, cross-entropy loss may perform poorly when the dataset is class-imbalanced, as the model is biased toward majority classes. To address this, the focal loss was introduced [68]. It extends the cross-entropy by a modulating factor $(1 - p_t)^\gamma$, where p_t is the model's predicted probability for the true class. This factor down-weights well-classified examples and focuses learning on hard, misclassified instances [68]:

$$L_{\text{focal}}\left(\mathbf{y}, \hat{\mathbf{p}}\right) = -\left(1 - \hat{p_y}\right)^\gamma \log \hat{p_y}\,. \tag{4.15}$$

Here, $\mathbf{y}$ is the true class label (as an integer) and $\hat{\mathbf{p}}$ is a vector representing an estimated probability distribution over the classes [69].

Another loss function commonly used in imbalanced classification problems is the Dice loss, derived from the Dice-Sørensen coefficient. It measures the overlap between predicted and true class distributions and is especially useful in segmentation tasks with extreme class imbalance, such as in medical imaging. The multi-class Dice loss is given by:

$$L_{\text{Dice}} = 1 - \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{2 y_i \hat{p}_i + \epsilon}{y_i + \hat{p}_i + \epsilon}\,, \tag{4.16}$$

where $\mathbf{y}, \hat{\mathbf{p}} \in \mathbb{R}^{N_c}$ represent the true and predicted class probabilities, respectively, and ϵ is a small factor to ensure this function is always well-defined. It is possible to combine different loss functions, for example, by defining a combined loss as the sum of Dice and focal loss:

$$L_{\text{combined}} = L_{\text{focal}} + L_{\text{Dice}}\,. \tag{4.17}$$

Taghanaki et al. showed that a weighted sum of Dice and focal loss outperforms other state-of-the-art methods in medical imaging analysis [70].

4.2.1 Optimization algorithms

Various optimization algorithms are available for minimizing the loss function, but the most fundamental is gradient descent. It works as follows: At each iteration i, the gradient of the loss with respect to the parameters is computed (equation (4.18)), and the parameters are then updated by a step of size α (the learning rate):

$$\boldsymbol{\nabla} L = \frac{\partial L}{\partial \theta}\,, \tag{4.18}$$

$$\theta_{i+1} = \theta_i - \alpha \cdot \boldsymbol{\nabla} L\,. \tag{4.19}$$

When the gradient approaches zero, successive updates become negligible. In practice, most algorithms terminate if the change in parameters falls below a predefined threshold [57]. It can be very challenging to optimize loss functions using gradient descent. If the learning rate is too small, convergence is slow, and if it's too large, the algorithm may overshoot or diverge. In high-dimensional, non-convex problems such as deep learning, plateaus, saddle points, and poor curvature are problematic [59].

Stochastic gradient descent (SGD) reduces the computational cost and helps escape poor regions of the loss function by replacing the full gradient with an expectation over a random subset of the data:

$$\mathbb{E}_{q(\mathbf{z})}\left[L\left(\boldsymbol{\theta}, \mathbf{z}\right)\right]\,. \tag{4.20}$$

Here, $\mathbf{z}$ denotes a mini-batch, and the expectation is taken over the empirical data distribution $q(\mathbf{z})$, which is typically approximated as uniform over the available training samples. The inherent noise introduced by this stochastic sampling not only reduces computation but also improves generalization by preventing the model from settling into sharp or narrow local minima that might overfit the training data [57]. Computing the gradient over a batch $\mathcal{B}$ of size $\|\mathcal{B}\|$ gives:

$$\boldsymbol{\nabla} L_{\mathcal{B}} = \frac{\partial}{\partial \boldsymbol{\theta}} \frac{1}{\|\mathcal{B}\|} \sum_{i \in \mathcal{B}} L(\mathbf{x}_i, \boldsymbol{\theta})\,. \tag{4.21}$$

This notably reduces gradient variance compared to single-sample updates [67]. Momentum incorporates a running average of past gradients. Instead of moving purely in the direction of the current gradient, the update accumulates a velocity term that reflects recent gradient history. This approach dampens oscillations in noisy directions and accelerates along shallow but consistent slopes. Momentum often leads to faster convergence and better traversal of narrow valleys in the loss landscape [59, 67]. Defining:

$$m_{i+1} = \beta \cdot m_i + \boldsymbol{\nabla} L\,. \tag{4.22}$$

The parameters are updated via

$$\theta_{i+1} = \theta_i - \alpha \cdot m_{i+1}\,, \tag{4.23}$$

where β is the momentum coefficient, typically ~ 0.9, that controls the contribution of past gradients.

The Adaptive Moment Estimation (Adam) optimizer combines stochastic optimization and momentum [71]. For this, we define:

$$\tilde{m}_{i+1} = \beta_1 \cdot \tilde{m}_i + (1-\beta_1)\boldsymbol{\nabla} L\,, \tag{4.24}$$

$$\tilde{s}_{i+1} = \beta_2 \cdot \tilde{s}_i + (1-\beta_2)\left(\boldsymbol{\nabla} L\right)^2\,, \tag{4.25}$$

where $\tilde{m}_i$ is the normalized momentum and $\tilde{s}_i$ is the normalized second moment of the gradient. The normalization is to avoid bias at the beginning of the training, where only a few past gradients exist. It is given as $\tilde{m}_i = \frac{m_i}{(1-\beta_1)}$ and analogous for $\tilde{s}_i$. The parameters are updated as follows:

$$\theta_{i+1} = \theta_i - \alpha \frac{\tilde{m}_i}{\sqrt{\tilde{s}_i} + \epsilon}\,, \tag{4.26}$$

where α is the learning rate and ϵ is a small constant to avoid division by zero. The standard values are $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-6}$ [59]. Figure 18 shows the implementation of SGD, momentum, and Adam on an ill-conditioned quadratic loss surface, $L(\boldsymbol{\theta}) = \frac{1}{2}\left(1 \cdot {\theta^{(0)}}^2 + 0.5 \cdot {\theta^{(1)}}^2\right)$.

4.2.2 Backpropagation

As described in Section 4.2.1, optimization requires calculating the gradients of the loss functions with respect to every trainable parameter in the model. In deep neural networks, the number of parameters can be in the millions. It is therefore important to compute these gradients efficiently [57]. This is the role of the well-known backpropagation algorithm.

Let us consider the shallow neural network shown in figure 16, which consists of a single hidden layer. The network's operations during a forward pass can be described by the following sequence of functions:

$$x_2 = f_1(x;\theta_1) \qquad \text{(linear combination)} \tag{4.27}$$

$$x_3 = f_2(x_2) \qquad \text{(activation)} \tag{4.28}$$

$$x_4 = f_3(x_3;\theta_3) \qquad \text{(linear combination)} \tag{4.29}$$

$$L = f_4(x_4, y) \qquad \text{(loss)} \tag{4.30}$$

Here, x and y are the input and true output, respectively. The network as a whole can be written as the composition $f = f_4 \circ f_3 \circ f_2 \circ f_1$. This is

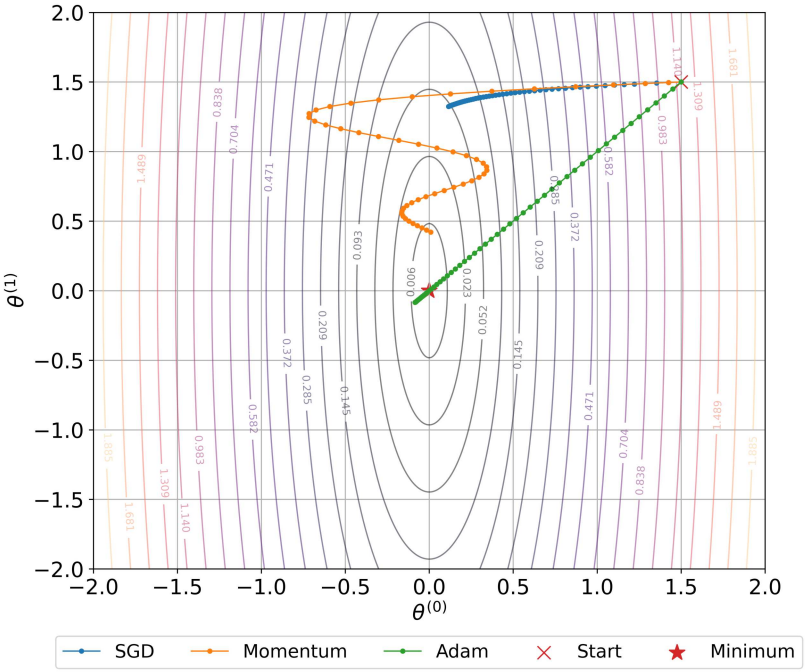

Figure 18: Optimizer trajectories on an ill-conditioned quadratic loss surface, illustrating the behavior of SGD, momentum ($\beta = 0.9$) and Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The horizontal and vertical axes correspond to the two components of the parameter vector; colors show the loss value. SGD follows a slow path down the valley before curving toward the minimum, momentum overshoots and oscillates along the narrow direction before converging, and Adam adapts the step sizes to move almost directly toward the minimum. The number of iterations is set to 40, and the learning rate is set to $\alpha = 0.05$ for all algorithms.

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \xrightarrow[f_1]{\text{Linear}} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} \xrightarrow[f_2]{\text{ReLU}} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} \xrightarrow[f_3]{\text{Linear}} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \xrightarrow[f_4]{\text{Loss}} L$$

Figure 19: Visualization of the MLP in figure 16, where each operation is displayed separately. The hidden layer is split into pre-activation and activation. Plot created with [72].

illustrated in figure 19. Note that the activation function f_2 has no trainable parameters, as it only introduces a non-linearity.

Already small changes to parameters can be amplified as they propagate through the network. To compute how a change in a parameter, such as θ_3 affects the loss, we need to know how intermediate quantities like x_4 respond. For a parameter further away from the output, such as θ_1, the influence must be computed through a chain of dependencies: From x_2 to x_3, from x_3 to x_4, and finally from x_4 to the loss. The backpropagation algorithm exploits this structure: Rather than recomputing gradients from scratch at every layer, it reuses intermediate results. This leads to significant computational savings. The backward pass proceeds from the output layer back toward the input:

$$\frac{\partial L}{\partial \theta_3} = \frac{\partial L}{\partial x_4}\frac{\partial x_4}{\partial \theta_3}\,, \tag{4.31}$$

$$\frac{\partial L}{\partial \theta_1} = \frac{\partial L}{\partial x_4}\frac{\partial x_4}{\partial x_3}\frac{\partial x_3}{\partial x_2}\frac{\partial x_2}{\partial \theta_1}\,. \tag{4.32}$$

Each $\frac{\partial L}{\partial \theta_k}$ is a row vector that can be computed recursively by propagating the upstream gradient through the network. Specifically, it is obtained by multiplying the gradient from the previous layer of the Jacobian $\frac{\partial x_k}{\partial x_{k-1}}$, which captures how changes in the input layer $k-1$ affect the output of layer k. This recursive structure enables efficient computations of gradients. The Jacobian matrix itself is defined as:

$$\mathbf{J_f} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\nabla} f_1(\mathbf{x})^\intercal \\ \vdots \\ \boldsymbol{\nabla} f_m(\mathbf{x})^\intercal \end{pmatrix} = \begin{pmatrix} \frac{\partial \mathbf{f}}{\partial x_1} \cdots \frac{\partial \mathbf{f}}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m\times n}\,. \tag{4.33}$$

Depending on the dimensions of the input n and output m, the Jacobian is calculated differently. If $n < m$, it is more efficient to calculate each row $\frac{\partial \mathbf{f}}{\partial x_j}$. However, in practice, scalar outputs are common ($m = 1$), making it more efficient to compute the Jacobian columns $\boldsymbol{\nabla} f_i(x)^\intercal$.

The backward algorithm then proceeds recursively, starting from $u_{K+1}^\intercal = 1$ and iterating k from K to 1:

$$g_k = u_{k+1}^\intercal \frac{\partial \mathbf{f}_k(x_k, \theta_k)}{\partial \theta_k}\,, \tag{4.34}$$

$$u_k^\intercal = u_{k+1}^\intercal \frac{\partial \mathbf{f}_k(x_k, \theta_k)}{\partial x_k}\,. \tag{4.35}$$

This algorithm is computationally very efficient, as the most expensive operations are matrix multiplications, which can be parallelized and executed rapidly on modern hardware.

4.2.3 Traning stability

Despite the efficiency of the backpropagation algorithm, training large neural networks remains challenging. During backpropagation, gradients are computed using the chain rule and propagated recursively through each layer. However, if these derivatives are small, their repeated multiplication can lead to vanishing gradients, where gradients become too small to update parameters effectively. Similarly, very large derivatives can result in exploding gradients, where parameters become unstable. The latter can be controlled using gradient clipping, where gradients exceeding a certain threshold c are scaled down [59]:

$$\mathbf{g}' = \min(1, \frac{c}{\|\mathbf{g}\|})\,\mathbf{g}\,, \tag{4.36}$$

where $\mathbf{g}'$ is the scaled gradient that goes in the same direction as the gradient $\mathbf{g}$.

The vanishing gradient can be mitigated by using a loss function whose gradient is not too small, which is the case for ReLU and GELU. An effective approach is the use of residual connections, illustrated in figure 20. In a residual network, each perceptron computes:

$$\mathcal{F}'(x) = \mathcal{F}(x) + x\,, \tag{4.37}$$

where $\mathcal{F}(x)$ is the standard non-linear transformation (e.g., a layer or network), as defined earlier in equation (4.6) [73]. While this does not increase the number of trainable parameters, it improves trainability [59]. To see this, consider the gradient in the residual case:

$$\frac{\partial L}{\partial \theta_3} = \frac{\partial L}{\partial x_4}\frac{\partial x_4}{\partial x_3}\frac{\partial x_3}{\partial \theta_2} = \frac{\partial L}{\partial x_4}\left(\frac{\partial f_3}{\partial x_3} + \mathbb{1}\right)\frac{\partial x_3}{\partial \theta_2}\,. \tag{4.38}$$

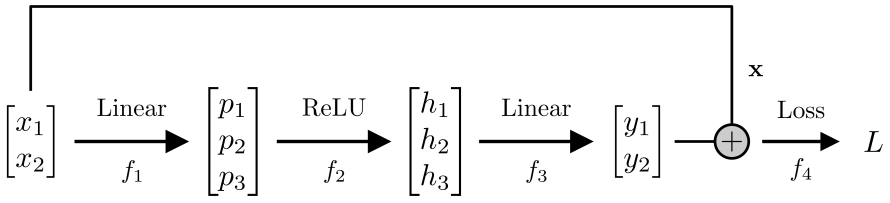

Figure 20: Multilayer perceptron of figure 19, but including a residual connection, which is indicated by the line adding $\mathbf{x}$ to the output $\mathbf{y}$. Plotted with [72].

The identity term $\mathbb{1}$ arises because x_3 is directly added to f_3. Hence even if $\frac{\partial f}{\partial x_3}$ is very small, the gradient will not vanish [59].

Another key technique to improve numerical stability is normalization. Two commonly used methods are batch normalization and layer normalization.

Batch normalization standardizes the pre-activations to zero mean and unit variance. Recall that the equation for a hidden unit in a fully connected NN was $h_d = F\left[\mathbf{b}_\theta + \mathbf{W}_\theta \cdot \mathbf{x}\right]$. We can add a batch normalization as follows:

$$h_d = F\left[\mathrm{BN}(\mathbf{b}_\theta + \mathbf{W}_\theta \cdot \mathbf{x})\right] . \tag{4.39}$$

For a given batch $\mathcal{B}$ the batch normalization is defined as:

$$\mathrm{BN}(\mathbf{x}) = \gamma \circ \frac{\mathbf{x} - \mu_\mathcal{B}}{\sigma_\mathcal{B}} + \beta \,, \tag{4.40}$$

Here $\mu_\mathcal{B}$ and $\sigma_\mathcal{B}$ indicate the mean and the standard deviation of the batch, calculated as in equations (4.41) and (4.42) where ϵ is again a small parameter (we used $\epsilon = 10^{-5}$) to avoid division by zero. This introduces two new parameters γ and β that are learned during training [59, 67].

$$\mu_\mathcal{B} = \frac{1}{\|\mathcal{B}\|} \sum_{x \in \mathcal{B}} \mathbf{x} \,, \tag{4.41}$$

$$\sigma_\mathcal{B} = \sqrt{\frac{1}{\|\mathcal{B}\|} \sum_{x \in \mathcal{B}} (x - \mu_\mathcal{B})^2 + \epsilon} \,. \tag{4.42}$$

Batch normalization works well for large enough batch sizes. Layer normalization is defined similarly. However, the normalization is applied over

all the hidden units in a single layer of a single vector $\mathbf{x}$, making it independent of batch size. It works well for recurrent and Transformer-based models because it avoids dependence on batch-level statistics, which may be unstable or poorly defined when working with variable-length sequences or in small batch sizes, common in NL and sequence modelling. The mean and variance are computed as:

$$\mu_{\mathcal{L}} = \frac{1}{H}\sum_{i=1}^{H} x_i \,, \tag{4.43}$$

$$\sigma^2_{\mathcal{L}} = \frac{1}{H}\sum_{i=1}^{H} (x_i - \mu_{\mathcal{L}})^2 \,, \tag{4.44}$$

where H is the number of hidden units and x_i refers to the activation of the i-th hidden unit in the current layer. Both $\mu_{\mathcal{L}}$ and $\sigma^2_{\mathcal{L}}$ are scalars. As with batch normalization, learnable scale and shift parameters γ and β are applied [67, 74].

4.3 The Transformer network

Although MLPs are very powerful in approximating any function, they are limited in their ability to capture relationships across structured or sequential input. In particular, when the input consists of temporally ordered data such as waveforms, MLPs struggle to model long-range dependencies. This limitation motivates the use of architectures specifically designed for such data structures, such as the Transformer network, which was introduced by Vaswani et al. in 2017 [75]. Unlike previous neural network models for sequential data, the Transformer relies exclusively on attention mechanisms. However, later variants reintroduce these components for efficiency or domain-specific modelling. Initially developed for natural language processing, Transformer models have since been successfully adapted to a wide variety of domains. Prominent examples include AlphaFold for protein folding [76], DALL-E for image generation [77], and large language models such as ChatGPT [78].

Transformers typically follow an encoder-decoder structure. However, depending on the specific task, encoder-only or decoder-only architectures are often sufficient. For instance, ChatGPT uses a decoder-only model, whereas the Transformer used in this work relies only on an encoder.

Figure 21 illustrates an encoder-only Transformer model. The input sequence is first divided into tokens, where a token can represent a word

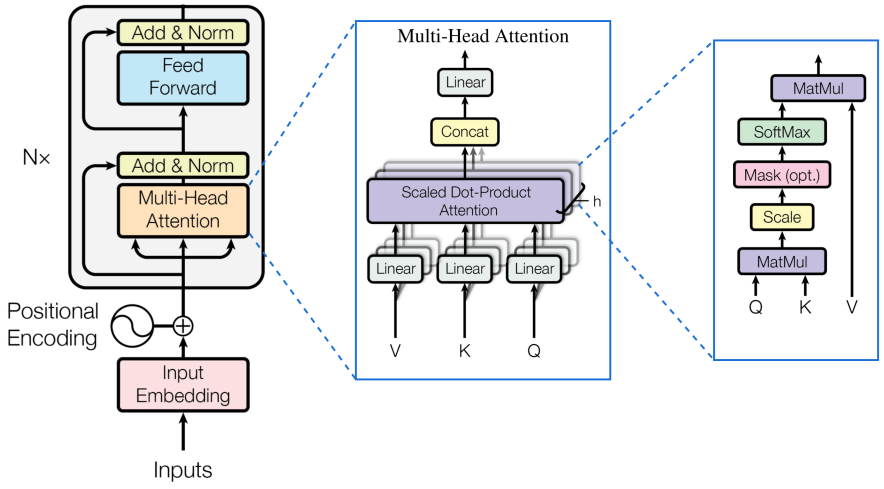

Figure 21: Overview of the Transformer encoder. In the first step, the inputs are embedded and positional encodings are added. Each of the N_x encoder blocks contains a multi-head attention mechanism and a feedforward network. Both components are followed by layer normalization and are wrapped in residual connections, indicated by the arrows pointing to the Add & Norm blocks. The middle panel shows the multi-head attention mechanism in more detail. Queries, keys, and values are computed by linearly transforming the inputs with their respective weight matrices. Scaled dot-product attention is then applied to these projections, and the outputs from all attention heads are concatenated and linearly transformed. The right panel illustrates the computational steps of scaled dot-product attention. This figure is adapted from [75].

fragment, a segment of a protein sequence, or a portion of a waveform. Each token is mapped to an embedding vector. Before entering the encoder blocks, positional encoding is added to these embeddings.

The encoder consists of a series of identical encoder blocks, each composed of a multi-head attention mechanism followed by a feedforward network. The feedforward component closely resembles the multilayer perceptron described in section 4.1.1. Both components are wrapped in residual connections and followed by layer normalization, which together is denoted as Add & Norm. Decoder blocks use similar structures but are more complex and not discussed here [57, 59, 75].

4.3.1 Input representation

The input sequence is split into n_t sub-sequences called tokens. In natural language processing applications, each token is then typically mapped to a learned embedding vector $\mathbf{x} \in \mathbb{R}^{d_{\text{emb}}}$ representing semantic or contextual features in a continuous vector space. All token embeddings are stored in a matrix $\mathbf{T} \in \mathbb{R}^{d_{\text{emb}} \times n_t}$, where each column corresponds to a token embedding [57, 67]. In contrast, our approach operates on fixed-length time-series data. Each waveform is divided into equally sized segments of 10 consecutive ADC samples, resulting in 140 uniformly spaced tokens per waveform. Each token is then projected into a 128-dimensional embedding vector via a learnable linear transformation. Using lower-dimensional embeddings can reduce training complexity but may result in underfitting. Conversely, higher-dimensional embeddings increase computational cost and risk of overfitting.

Because attention mechanisms are inherently permutation-invariant, they lack information about the order of input tokens. Positional encoding addresses this by injecting sequence order into the model.

The positional encoding is implemented as a matrix $\mathbf{PE} \in \mathbb{R}^{n_t \times d_{\text{emb}}}$, where n_t is the number of tokens and d_{emb} the embedding dimension. Each row corresponds to a token and each column to a position in the embedding space [67]. Vaswani et al. proposed a sinusoid positional embedding, where the elements are computed as:

$$\text{PE}_{i,\,2j} = \sin\left(\frac{i}{C^{2j/d}}\right), \tag{4.45}$$

$$\text{PE}_{i,\,2j+1} = \cos\left(\frac{i}{C^{2j/d}}\right), \tag{4.46}$$

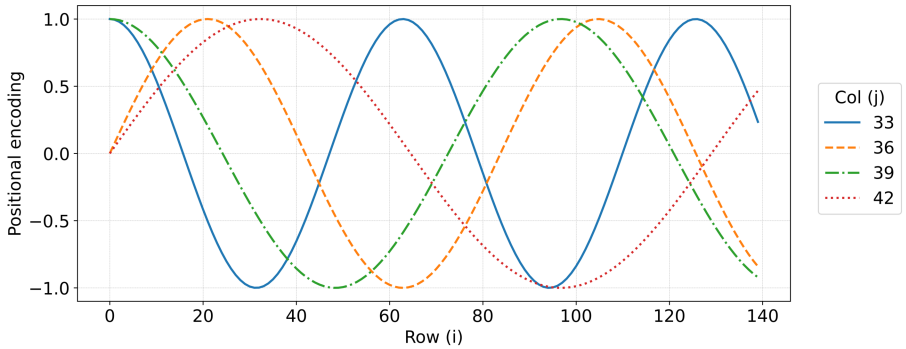

Figure 22: Sinusoidal positional encoding used in this work. Note that higher dimensions (columns) encode lower-frequency information. Even columns start at 0, and odd columns start at 1.

where i is the token index, j is the embedding dimension, and C is a scaling constant. We use the common choice of $C = 10^4$, which is long enough to provide a wide range of frequency scales, allowing the model to capture both short- and long-range dependencies in the sequence. Since even indices are computed with sine and odd ones with cosine, j runs up to $(d/2 - 1)$ [59, 75]. The resulting positional encoding is shown in figure 22.

Although seemingly complex, this encoding offers two major advantages. First, it allows for arbitrary sequence lengths up to C without retraining. Second, the encoding for one position can be linearly computed from another. This linearity allows the model to reason about relative positions and extrapolate to unseen sequence lengths, which supports better generalization in tasks involving variable-length inputs [59, 75]. For example, in a low-dimensional case with $d = 2$ and $C = 1$, the positional encoding k steps away from position p is:

$$\begin{pmatrix} \mathrm{PE}_{p+k,0} \\ \mathrm{PE}_{p+k,1} \end{pmatrix} = \begin{pmatrix} \sin(p+k) \\ \cos(p+k) \end{pmatrix} = \begin{pmatrix} \sin p \cos k + \cos p \sin k \\ \cos p \cos k - \sin p \sin k \end{pmatrix} \tag{4.47}$$

$$= \begin{pmatrix} \cos k & \sin k \\ -\sin k & \cos k \end{pmatrix} \begin{pmatrix} \sin p \\ \cos p \end{pmatrix} . \tag{4.48}$$

Once computed, positional encodings are added to the embeddings $\mathbf{x}$:

$$\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{PE} . \tag{4.49}$$

4.3.2 Attention mechanism

The multilayer perceptron described in section 4.1.1 applies a linear transformation to the input vector $\mathbf{x}$, followed by an activation function F. Each layer has its own set of learnable parameters.

The attention mechanism takes a different approach. Conceptually, it operates like a database: Given a set of N keys $\mathbf{k}_i$ and values $\mathbf{v}_i$, a query vector $\mathbf{q}$ is used to retrieve information. This design allows the model to process variable-length input sequences and dynamically adapt the output based on contextual relevance. When every token attends to every other token, this mechanism allows each token to incorporate information from all other tokens, regardless of distance. This property makes attention particularly powerful for capturing long-range dependencies in sequential data. The general form of attention is:

$$\text{Attn}[\mathbf{q}] = \sum_{i=1}^{N} \alpha\left[\mathbf{q}, \mathbf{k}_i\right] \mathbf{v}_i \tag{4.50}$$

Here, attention is a weighted sum over the values, where each scalar weight $\alpha\left[\mathbf{q}, \mathbf{k}_i\right]$ reflects how much attention is paid to value $\mathbf{v}_i$ given the query. These weights are computed by an attention-scoring function. In transformers, the most common method is the scaled dot product:

$$\alpha[\mathbf{q}, \mathbf{k}_i] = \text{softmax}\left(\frac{\mathbf{q}^\intercal \cdot \mathbf{k}_i}{\sqrt{d_k}}\right) = \frac{\exp\left[\mathbf{q}^\intercal \cdot \mathbf{k}_i / \sqrt{d_k}\right]}{\sum_{j=1} \exp\left[\mathbf{q}^\intercal \cdot \mathbf{k}_j / \sqrt{d_k}\right]} \tag{4.51}$$

Scaling by $\frac{1}{\sqrt{d_k}}$ improves numerical stability. The softmax ensures all attention weights are positive and sum to one, so they act like a probability distribution over the values [57, 67]. The scaled-dot product is illustrated in figure 23.

In Transformer models, keys, queries, and values are computed as linear projections of the same input embeddings $\mathbf{x}_n$ using learned weight matrices:

$$\mathbf{k}_n = \mathbf{W}_\mathbf{K}\, \mathbf{x}_n \qquad \mathbf{W}_\mathbf{K} \in \mathbb{R}^{d_k \times d_\text{emb}} \tag{4.52}$$

$$\mathbf{v}_n = \mathbf{W}_\mathbf{V}\, \mathbf{x}_n \qquad \mathbf{W}_\mathbf{V} \in \mathbb{R}^{d_\text{emb} \times d_\text{emb}} \tag{4.53}$$

$$\mathbf{q}_n = \mathbf{W}_\mathbf{Q}\, \mathbf{x}_n \qquad \mathbf{W}_\mathbf{Q} \in \mathbb{R}^{d_q \times d_\text{emb}} \tag{4.54}$$

Because all three vectors are derived from the same embedding, this is referred to as self-attention:

$$\begin{bmatrix} q_1k_1 & \boxed{q_2k_1} & q_3k_1 \\ q_1k_2 & \boxed{q_2k_2} & q_3k_2 \\ q_1k_3 & \boxed{q_2k_3} & q_3k_3 \end{bmatrix}$$
$$\sum_{i=1}^{3} \frac{q_2k_i}{\sqrt{d_k}} = 1 \implies \text{softmax}$$

Figure 23: Visualization of scaled dot-product attention. Each row corresponds to a query vector $\mathbf{q}_i^\intercal$ and each column to a key vector $\mathbf{k}_j$. The boxed row shows attention weights from query $\mathbf{q}_2$ to all keys. These are normalized via softmax and used to compute a weighted sum over value vectors $\mathbf{v}_j$. Figure created with [72].

$$\text{SelfAttn}[\mathbf{x}_n] = \sum_{m=1}^{N} \alpha\left[\mathbf{x}_m, \mathbf{x}_n\right] \mathbf{v}_m \,. \tag{4.55}$$

In practice, we vectorize the computation by stacking all embeddings into a matrix $\mathbf{X} \in \mathbb{R}^{d_{\text{emb}} \times N}$, with each column $\mathbf{x}_n$ representing a token [57, 67]. The keys, queries, and values are then computed in matrix form:

$$\mathbf{Q} = \mathbf{W_Q}\,\mathbf{X} \qquad \mathbf{Q} \in \mathbb{R}^{d_k \times N} \tag{4.56}$$
$$\mathbf{K} = \mathbf{W_K}\,\mathbf{X} \qquad \mathbf{K} \in \mathbb{R}^{d_k \times N} \tag{4.57}$$

The self-attention can then be calculated compactly as:

$$\text{SelfAttn}[\mathbf{Q}, \mathbf{K}, \mathbf{V}] = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^\intercal}{\sqrt{d_k}}\right) \mathbf{V} \,. \tag{4.58}$$

Instead of using a single attention mechanism, Transformers apply multi-head attention, where h attention heads run in parallel. Each head computes its own self-attention using separate learned projections. The outputs of all heads are concatenated and linearly transformed:

$$\text{MultiHeadSelfAttn}[\mathbf{X}] = \mathbf{W_C} \begin{pmatrix} \text{SelfAttn}^1[\mathbf{X}] \\ \vdots \\ \text{SelfAttn}^h[\mathbf{X}] \end{pmatrix} \qquad \mathbf{W_C} \in \mathbb{R}^{d_k \times d_k} \quad (4.59)$$

The final projection matrix $\mathbf{W_C}$ ensures that the concatenated output has the same shape as the original input $\mathbf{X}$. A typical and efficient choice is to set $d_k = d_{\text{emb}}/h$ so that the total output dimension matches the embedding dimension [57, 59, 67].

4.3.3 Transformer architecture for LEGEND waveforms

Let us now consider the specific Transformer model used in this work. It was implemented by Marta Babicz in Python using the Pytorch framework [79]. The architecture is encoder-only and closely resembles BERT (Bidirectional Encoder Representations from Transformers), originally introduced by Google [80].

The input data consists of physical waveforms recorded by the LEGEND-200 experiment. Each waveform spans 20.4 μs sampled at 1400 time steps. These waveforms are segmented into $n_t = 140$ tokens, each representing 10 time steps, and embedded into a vector of size $d_{\text{emb}} = 128$. After tokenization and embedding, positional encoding is added as described in section 4.3.1. The model contains $n_l = 6$ encoder layers, each using 8 attention heads in parallel. Each attention block and feedforward block is wrapped in a residual connection and followed by layer normalization, following the original Transformer design. Altogether, the network has nearly 1.2 million trainable parameters. Roughly one-third of these lie in the attention layers, while two-thirds are trained in the feedforward network. Table 2 gives a detailed parameter breakdown.

For the task at hand – namely analyzing long-sequence waveform data – Transformers are particularly well suited. Convolutional neural networks and Recurrent neural networks are generally limited in capturing long-range dependencies due to their fixed receptive fields or sequential processing. By contrast, multi-head attention can focus on multiple regions of the waveform simultaneously, enabling the model to learn subtle temporal patterns beyond classical features such as A/E. The Transformer's ability to model global context, apply dynamic attention, and process sequences in parallel is particularly valuable for pulse-shape analysis. The waveforms encode rich spatial and temporal information about energy depositions in the detector, with features that range across the full waveform. In the following chapter,

Table 2: Overview of the trainable parameters of the Transformer model used in this work. The dimension of the embedding is $d_{\text{emb}} = 128$, which is equal to the dimension of the keys and queries. The number of encoder layers is $n_l = 6$, the number of labels $n_{\text{labels}} = 4$, and the number of neurons in the feedforward network $n_n = 512$. The factor 2 in the embedding comes from the fact that we embed not only the waveform but also the gradients.

Architecture	**Weights**	**Bias**	**Parameters**
Tokenization	d_{emb}	-	128
Embedding	$2 \times 10 \times d_{\text{emb}}$	$2 \times d_{\text{emb}}$	2816
Keys	$d_{\text{emb}} \times d_{\text{emb}} \times n_l$	$d_{\text{emb}} \times n_l$	99072
Queries	$d_{\text{emb}} \times d_{\text{emb}} \times n_l$	$d_{\text{emb}} \times n_l$	99072
Values	$d_{\text{emb}} \times d_{\text{emb}} \times n_l$	$d_{\text{emb}} \times n_l$	99072
Output matrix	$d_{\text{emb}} \times d_{\text{emb}} \times n_l$	$d_{\text{emb}} \times n_l$	99072
Linear layer 1	$d_{\text{emb}} \times n_n \times n_l$	$n_n \times n_l$	396288
Linear layer 2	$n_n \times d_{\text{emb}} \times n_l$	$d_{\text{emb}} \times n_l$	393984
Normalization 1	$d_{\text{emb}} \times n_l$	$d_{\text{emb}} \times n_l$	1536
Normalization 2	$d_{\text{emb}} \times n_l$	$d_{\text{emb}} \times n_l$	1536
Decoder	$n_{\text{labels}} \times d_{\text{emb}}$	n_{labels}	516
De-embedding	$2 \times d_{\text{emb}}$	-	256
Total:			1193348

we describe how this Transformer model was trained and evaluated on waveform data, including dataset preparation, model configuration, and training procedures.

5 Pulse shape discrimination with Transformers

Having introduced the theoretical foundations of machine learning and the specific architecture of the Transformer model used in this work, we now turn to their practical implementation in the context of this work. This chapter begins by outlining the structure and organization of waveform data in the LEGEND-200 experiment. We then describe the procedure for constructing training datasets, which includes data selection and data cleaning. The different models developed in this work are presented alongside their classification performance. We continue with a brief overview of the pulse shape simulation (PSS) framework in LEGEND-200, and conclude this chapter by summarizing key results.

5.1 Data in the LEGEND-200 experiment

The DAQ process in LEGEND-200 follows a structured hierarchy to ensure consistency across the dataset. Each DAQ cycle – a continuous block of data-taking, synchronously started and stopped by the DAQ system – is stored as a separate file and marked with a GPS timestamp. A sequence of DAQ cycles recorded under a consistent detector configuration is grouped into a run. Each run begins with a series of calibration cycles, followed by physics cycles. Runs are then grouped into larger units called periods, which are separated by major changes in the detector setup or operating conditions that could impact analysis results. However, not all periods and runs are suitable for physics analysis. Only a subset has passed the data quality checks established by the collaboration [81]. Table 3 summarizes all physics runs that were deemed usable for this work.

The raw waveforms recorded by the LEGEND-200 experiment are stored

Table 3: Summary of all physics runs that passed the LEGEND quality checks [81], including live time and total exposure (ICPC detectors only).

Period	**Runs**	**Start date**	**Live time [yr]**	**Exposure [kg·yr]**
3	0-5	2023-03-12	0.08	4.85
4	0-3	2023-04-15	0.047	2.89
6	0-5	2023-06-11	0.110	6.68
7	2-7	2023-07-31	0.096	5.82
8	0-4, 6-14	2023-10-03	0.237	14.39
9	0-5	2024-01-11	0.093	5.67

at NERSC in HDF5 (Hierarchical Data Format version 5). HDF5 is a binary file format designed to store and organize large, complex datasets. It is well-suited for large datasets because it supports fast I/O and efficient partial data access. Therefore, the entire file need not be loaded into memory.

The collaboration has developed a sophisticated digital signal processing pipeline. Its initial stage operates directly on the raw waveforms, with subsequent stages performing higher-level processing, as illustrated in figure 24.

In the initial stage, the digitized waveforms recorded by the DAQ system are stored in the RAW (raw waveform and ADC information) tier. This tier also includes metadata such as the timestamp of the recording and the measured energy in ADC units. For the DSP (digitally processed features) tier, each waveform is processed independently to extract signal features. This tier contains several parameters directly derived from the waveform, for example, various energy estimates, amplitude, and baseline. The HIT (hit-level calibrated parameters) tier contains parameters derived from a multi-waveform analysis. This includes calibrated energy values, A/E ratios, and LQ cut values. All data tiers up to this point are organized in separate subdirectories and indexed by timestamp, enabling efficient access to specific waveforms. In contrast, the EVT (event cluster level) tier clusters individual waveforms into events. Each event may consist of multiple coincident hits in the HPGe detectors, as well as in other subsystems, such as the SiPMs and the muon veto. Finally, the TCM (trigger and calibration metadata) tier includes information related to event building, along with data from calibration pulses.

5.2 Data preparation for Transformer training

In supervised machine learning, the quality of the training data is critical. The performance of a model is fundamentally limited by the cleanliness and reliability of the dataset it is trained on.

To address this, we developed a code framework that combines waveforms and processing parameters from multiple data tiers into a unified dataset suitable for training machine learning models. Each waveform is then classified (labeled) according to the established LEGEND-200 analysis methods.

The dataset is subsequently cleaned to maximize class purity and minimize contamination from misclassified events. The impact of this data-cleaning procedure is illustrated in figure 25, where we show data from period 3 (only ICPC detectors).

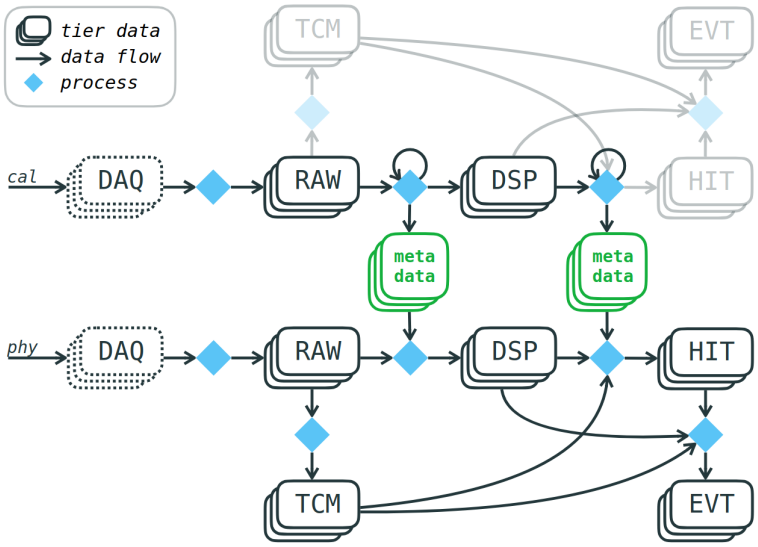

Figure 24: Data flow for the LEGEND-200 experiment. Starting from the DAQ output, the data undergoes multiple processing steps before being clustered into coincident events in the EVT tier. The calibration data, acquired using sources that produce signals at well-known energies, is used to define and validate analysis cuts and to establish a precise energy scale. These calibrations are then applied to the physics data. Image credit to Luigi Pertoldi [82].

5.2.1 Data selection

The data selection begins by loading the calibrated energies from the HIT tier. In this work, we use the energy estimated with the trapezoid filter (explained in section 3.2.3). The HDF5 files, which have a specific structure, are accessed with *legend-pydataobj*, a package that provides a Python implementation of the LEGEND Data Objects to HDF5 [83]. Detector metadata is used to select only detectors that were active and functioning properly during data taking.

For events falling within a specified energy range, the corresponding indices are stored for further processing. For datasets intended for training, the algorithm automatically selects a $\pm 3\,\sigma$ window around the four prominent calibration peaks in the ^{228}Th spectrum. This window is estimated from the FWHM of the energy resolution. Alternatively, a custom energy

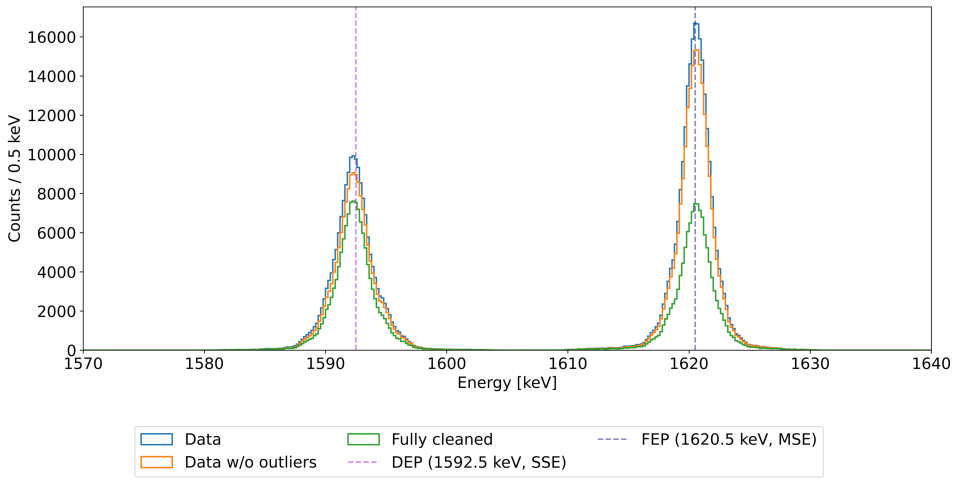

Figure 25: Effect of the data cleaning procedure on two pre-selected peaks ($\pm 3\,\sigma$) from the ^{228}Th calibration in period 3, using only ICPC detectors. The plot shows the impact of the two cleaning steps applied to pre-selected data (blue): removal of statistical outliers (orange) and exclusion of events with non-matching numbers of peaks in the gradient waveforms (green, combined effect). The latter has a strong impact on the full energy peak, as expected, since SSEs are already very pure, while the MSE class tends to be heterogeneous.

window can be manually defined, for example, to load a continuous dataset for benchmarking the Transformer model against conventional PSD methods. Once the energy range is specified, waveforms are loaded from the RAW tier by indexing the file location within the subdirectory structure. Throughout this work, we use waveforms windowed around the rising edge, sampled 1400 times at 16 ns intervals. Pre-selecting events in the HIT tier significantly reduces runtime by up to two orders of magnitude, since only waveforms in the relevant energy range are loaded. These energy windows typically span tens of keV and represent only a small fraction of the total dataset.

We then apply quality cuts to remove waveforms that deviate from expected detector response characteristics. For this purpose, we apply the same set of quality cuts developed and validated for the LEGEND-200 physics analyses, which are defined in [84]. We remove all waveforms that contain:

Table 4: Analytical labels for the waveforms, based on the LEGEND-200 analysis. The most important component is the A/E parameter. For ICPC detectors produced by Mirion Technologies, the LQ cut is not applied because it is not reliable.

Label	**A/E value**	**LQ value**
SSE	intermediate	low
MSE	low	-
p-contact	high	-
n-contact	low	high

- discharges (is_delayed_discharge)
- unstable baselines (is_valid_bl_slope, is_valid_bl_slope_rms)
- noisy tails (is_valid_tail_rms)
- noise bursts in the rising edge (is_not_noise_burst)
- invalid energies (is_valid_cuspEmax, is_valid_cuspEmin, is_low_cuspEmax)
- invalid trap filter (is_valid_trap_tpmin, is_valid_trap_tpmax)
- invalid rise times (is_valid_t0, is_valid_rt, is_valid_dteff)

In the final step of the data selection, the remaining events are classified. As discussed in section 3.4 and shown in figure 14, the amplitude-to-energy ratio is particularly well suited for distinguishing different event topologies. Table 4 summarizes the classification scheme.

The resulting dataset consists of waveforms within the desired energy range that have been classified accordingly. While the class labels are already very pure, the events within each class still exhibit a degree of heterogeneity that will affect model training.

5.2.2 Data cleaning

In the second stage of the data preparation, the dataset is further cleaned to increase the purity of the class labels. Two methods are applied: peak estimation and outlier removal. For peak estimation, the derivative of each waveform is computed. To increase the robustness of the peak identification, the gradients are smoothed using a rolling window with a Gaussian kernel

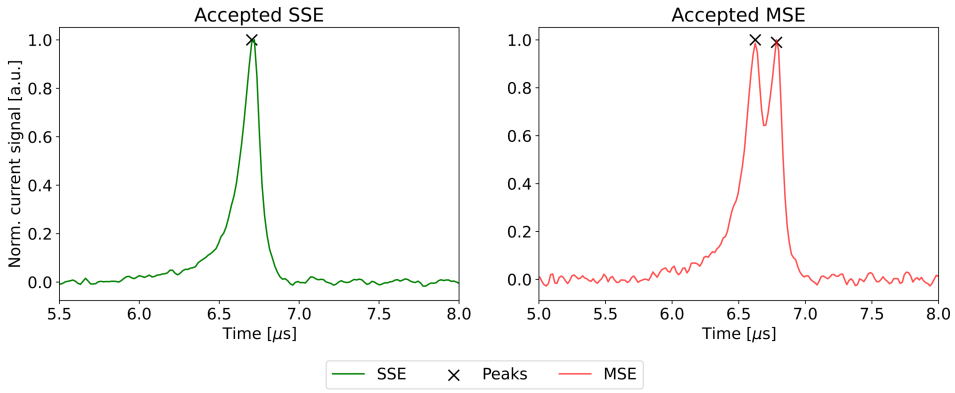

Figure 26: Peak identification for ICPC detector V02160A during period 3. The left panel shows a single-site event with one identified peak; the right panel shows a multi-site event with multiple peaks. Both are accepted.

(applied over 5 timestamps with a width of $1.5\,\sigma$). The resulting derivative signal is then normalized to reduce energy dependence.

To identify distinct signal components in the derivative waveform, we used the *find_peaks* function from the scipy module [85], which allows for peak detection based on height and prominence. An example of single-site and multi-site peak identification is shown in figure 26. For waveforms labeled as MSE, multiple peaks are required, as this corresponds to multiple time-separated charge drifts within a germanium diode. For all other event categories, only a single peak per waveform is permitted. Note that the labels of the waveforms are not changed during this process; instead, waveforms that do not meet the peak criteria are simply rejected. This cleaning cut is very strict, and removes more than half of the events (51.2 %, benchmarked on period 3). In this step, we also extract the FWHM of the largest peak, which is used for outlier rejection.

To identify outliers, we calculate the mean and standard deviation of all parameters listed in table 5, separately for each detector and label. Waveforms with a peak FWHM outside a $\pm 3\,\sigma$ window are rejected – except for MSE events, where we apply only a lower-bound cut. This exception accounts for cases where two nearby peaks merge slightly, causing the FWHM of the dominant peak to broaden. Such events are not rejected to preserve valid MSE samples.

For the other quantities described in table 5, a symmetric $\pm 2\,\sigma$ window is applied for outlier rejection. In total, around 8% of events are rejected in

Table 5: Parameters where outliers are removed in the data cleaning procedure. Generally, only a few events are removed in each cut ($< 3\%$). Efficiency values are determined from data taken during period 3. The total efficiency is obtained by applying all cuts simultaneously and is smaller than the product of individual efficiencies, as some waveforms fail multiple cuts.

Quantity	**Description**	**Range**	**Cut Eff.** [%]
tp-10	Time where A reaches 10 %	$3\,\sigma$	97.7
tp-50	Time where A reaches 50 %	$3\,\sigma$	98.0
tp-80	Time where A reaches 80 %	$3\,\sigma$	97.8
rt1	Time between tp-10 and tp-90	$3\,\sigma$	97.8
rt2	Time between tp-10 and tp-50	$3\,\sigma$	97.6
AoE	Area divided by energy	$3\,\sigma$	99.0
LQ	Measure for the late charge	$3\,\sigma$	99.1
Peak width	FWHM of the largest peak	$2\,\sigma$	98.0
Total	All cleaning cuts applied	-	91.9

this step. Outlier removal is illustrated in figure 27, which shows rise time cleaning cuts for single-site and multi-site events, and in figure 28, which shows cuts on the FWHM of the peak with the highest amplitude. Both figures include example waveforms of accepted and rejected events.

5.3 Model training and evaluation

The Transformer model architecture used in this work is described in section 4.3.3. It was trained on several datasets, each prepared according to the procedures outlined in the previous subsections. An overview of all datasets used for training is provided in table 6. The models are trained using four NVIDIA V100 GPUs on the Perlmutter supercomputer at the National Energy Research Scientific Computing Center (NERSC) in Berkeley, California. The datasets were split into three parts: 60% for training (used to optimize model weights), 10% for validation (used to monitor generalization performance and prevent overfitting), and 30% for testing. The validation set allows for early stopping, terminating training when the validation loss is no longer improving, and performance monitoring without influencing the model's parameters. Since its purpose is monitoring, a relatively small share is sufficient. The test set is strictly held out and used only to report the final model performance; it must be large enough to provide statistically meaningful results without excessively reducing the training data. We used

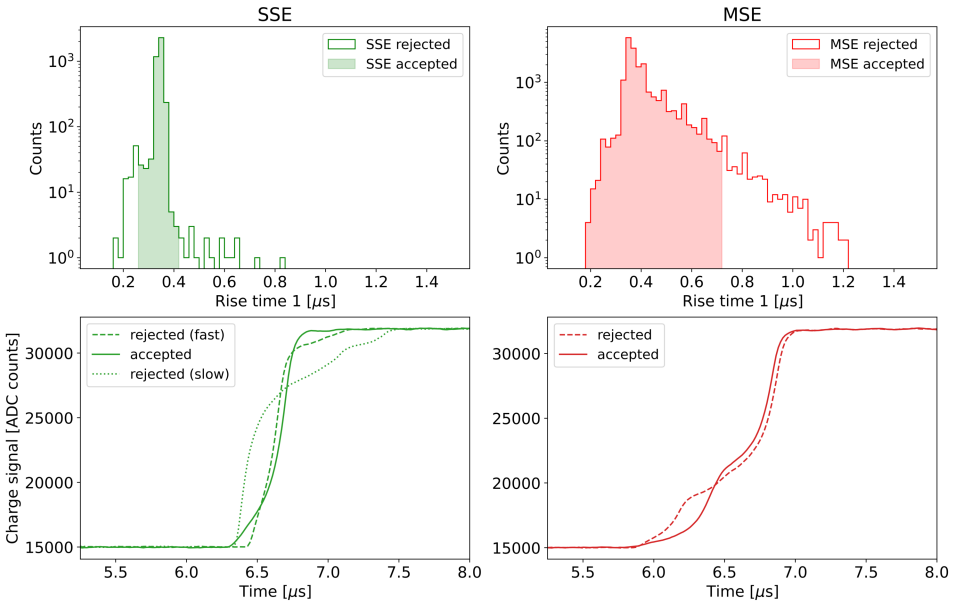

Figure 27: Rise time outlier removal ($3\,\sigma$) for ICPC detector V02160A during period 3. The top panels show histograms of accepted and rejected events; the bottom panels show example waveforms. Single-site events (left) are more homogeneous than multi-site events (right). In the bottom left, an accepted waveform (solid) is compared to outliers with too fast (dashed) and too slow (dotted) rise times. Similar rejection is applied to long-rise-time MSE events (bottom right).

a batch size of 512 per GPU, resulting in an effective batch size of 2048. The loss function used in this work is the combined loss shown in equation (4.17). It is optimized using ADAM, described at the end of section 4.2.1.

The different models are then evaluated and assessed with several metrics. The precision score, defined in equation (5.1), measures the proportion of correctly identified positive instances among all instances classified as positive. In the context of single-site events, it quantifies the fraction of events classified as SSE that are truly single-site. A high precision score indicates a low number of false positives.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}\,. \tag{5.1}$$

Sensitivity addresses the complementary aspect by quantifying the proportion of true SSE events that are correctly identified. A high sensitivity

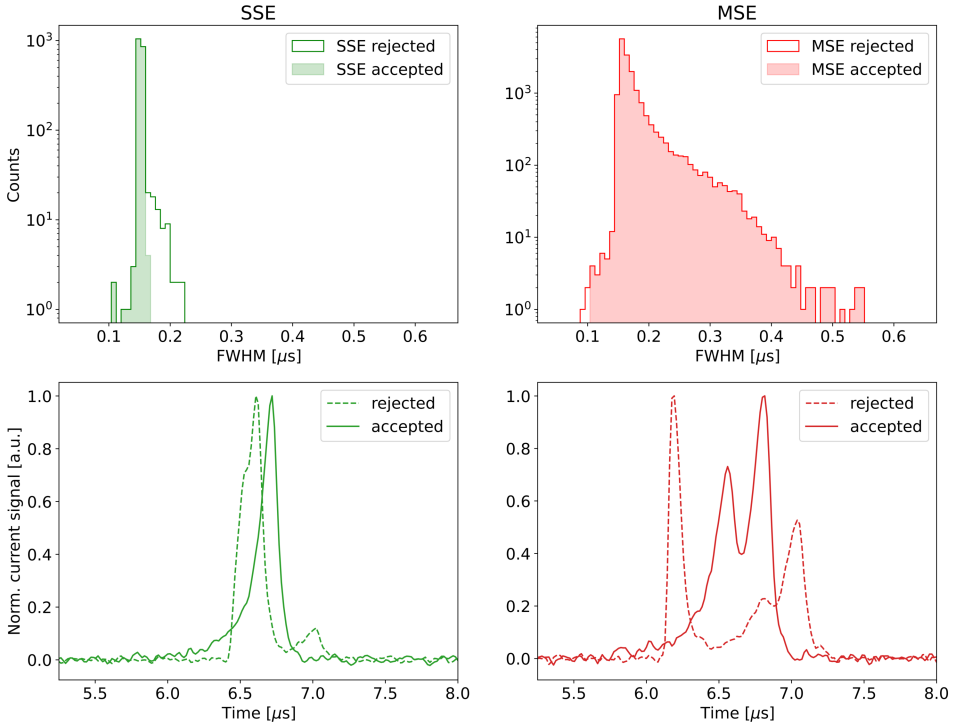

Figure 28: Peak width outlier removal ($2\,\sigma$) for ICPC detector V02160A during period 3. The top panels show histograms of accepted and rejected events; the bottom panels show example differentiated waveforms. For multi-site events, only a lower peak width bound is applied to avoid rejecting merged peaks. Bottom left: accepted SSE waveform (solid) vs. rejected broadened peak (dashed). Bottom right: rejected waveform (dashed) with a sharp rise, indicating interaction near the p^+ electrode.

Table 6: Datasets used for training the Transformer, only including ICPC detector waveforms. The data size indicates the total number of waveforms in the dataset. In the second Transformer model, p^+ and n^+ events are combined into a single surface label.

Transformer	**Periods**	**Data size**	**Isotopes**	**Labels**
Model 1	3 - 10	12.5 $\times 10^6$	^{228}Th	SS, MS, p^+, n^+
Model 2	3 - 11	14.0 $\times 10^6$	^{228}Th, ^{56}Co	SS, MS, surface
Model 3	3 & 11	3.0 $\times 10^6$	^{228}Th, ^{56}Co	SS, MS, p^+, n^+

indicates a low number of false negatives, meaning the model effectively captures most of the true positive cases:

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \,. \tag{5.2}$$

In machine learning, the F1-score is commonly reported alongside precision and sensitivity as a measure of the balance between the two:

$$\text{F}_1 = 2\,\frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Sensitivity} + \text{Precision}} \,. \tag{5.3}$$

A confusion matrix provides a compact graphical representation of classification performance. For n_{label} classes, it is an $n_{\text{label}} \times n_{\text{label}}$ matrix in which each row corresponds to the ground truth and each column to the model's predicted class. The diagonal entries indicate correct classifications, while the off-diagonal entries represent misclassifications, including false positives and false negatives [59].

5.3.1 First Transformer model

The first model configuration was trained on 12.5 million waveforms extracted from the energy ranges of the four dominant ^{228}Th calibration peaks. These include the double-escape peak of ^{208}Tl at 1592.5 keV, the full-energy peak of ^{212}Bi at 1620.5 keV, and the single-escape and full-energy peaks of ^{208}Tl at 2103.5 keV and 2614.5 keV, respectively. We used waveforms from periods 3 to 10 for this dataset.

The true labels were assigned according to the classification scheme outlined in table 4. The classification performance is illustrated by the confusion matrix in figure 29, which shows the precision. The corresponding scores are summarized in table 7. The model achieves very high precision for both single-site and multi-site events, as well as strong precision for surface event

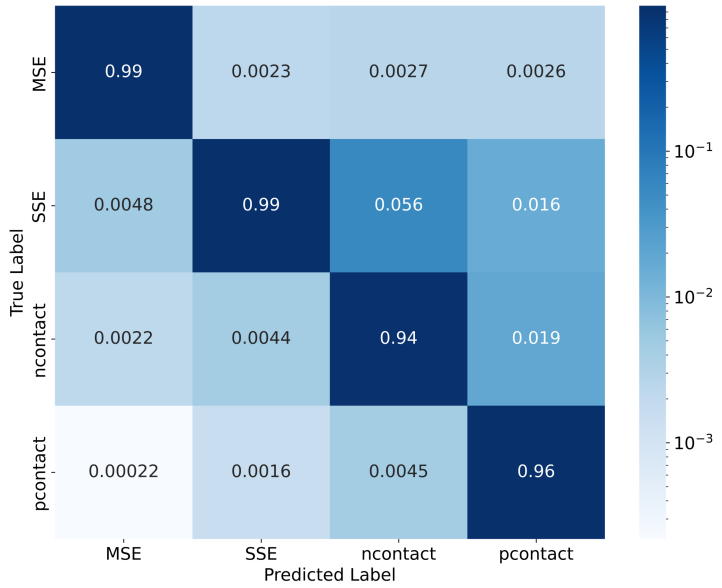

Figure 29: Confusion matrix of model one, with precision scores shown on the diagonal. The model was trained on 12.5 million ^{228}Th calibration waveforms. Multi-site and single-site events are classified with high precision ($> 99\%$), while both surface event types show some misclassification but still achieve precision scores above 90%.

classification. Sensitivity is notably lower for SSE and p^+ events in contrast to MSE and n^+ events, indicating some degree of cross-contamination between these classes. Despite this, all F1-scores exceed 95%, demonstrating an overall excellent classification performance.

5.3.2 Second Transformer model

For the second model, the training data was extended. In addition to the four primary ^{228}Th calibration peaks, we included seven additional energy peaks from a dedicated ^{56}Co calibration run, all of which are DEPs: 1012.8 keV, 1576.5 keV, 1987.6 keV, 2180.0 keV, 2231.5 keV, 2251.1 keV, and 2429.2 keV.

Using additional energy regions should improve the model's flexibility. Furthermore, the two surface event classes (n^+ and p^+) were merged into a single surface label. Apart from these changes, the second model is architecturally identical to the first.

Table 7: Classification scores for the first Transformer model. Precision is the fraction of correct positive predictions, while sensitivity refers to the fraction of true positives correctly identified. The F1-score combines both.

Class	**Precision**	**Sensitivity**	**F1-score**	**Support** $[\times 10^6]$
SSE	0.992	0.925	0.957	1.03
MSE	0.993	0.996	0.994	1.47
n^+	0.937	0.992	0.964	1.15
p^+	0.963	0.937	0.950	0.12
Weighted avg	0.975	0.974	0.974	3.76

Table 8: Classification scores for the second Transformer model: Precision and sensitivity exceed 98% for all classes.

Class	**Precision**	**Sensitivity**	**F1-score**	**Support** $[\times 10^6]$
SSE	0.981	0.989	0.985	0.61
MSE	0.995	0.993	0.994	0.40
surface	0.990	0.986	0.988	0.55
Weighted avg	0.989	0.989	0.989	1.56

This resulted in a total of 11 distinct energy regions, used to extract 14 million waveforms for training. The resulting model demonstrates classification precision across all event types (98%), shown in the confusion matrix in figure 30, and the summary table 8. Furthermore, the model achieves exceptionally high sensitivity across all classes, all with F1-scores exceeding 98%. While such uniformly high scores indicated excellent classification performance, it's important to check that this isn't due to overfitting.

5.3.3 Third Transformer model

A third model was also prepared using a new dataset based on ^{228}Th and ^{56}Co calibration data. While it covered the same energy regions as the second model, we wanted to maintain the balance between the two different calibrations. Therefore, we selected only a single ^{228}Th calibration period (period 3). In addition, the surface-event labels were separated again into distinct p^+ and n^+ categories. However, this model performed poorly: it failed to generalize beyond the training data and exhibited inconsistent classification across detectors. Due to these issues, the model was deemed unsuitable for further evaluation and was excluded from the analysis. The exact cause of the poor performance remains unclear, but it may be related

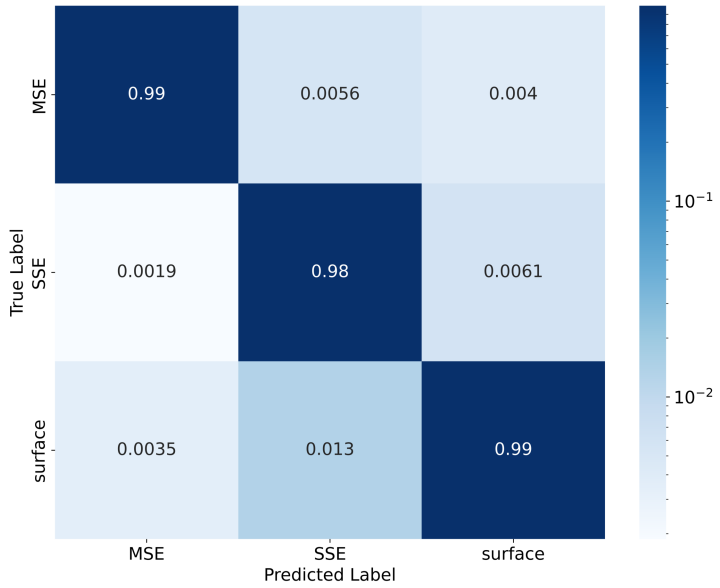

Figure 30: Confusion matrix of model two, where we trained on 14.0 million events. This testing includes 3.3 million events. All event categories are classified with high precision, each exceeding 98%.

to the label noise in the ^{56}Co dataset.

5.3.4 Network comparison and performance on $2\nu\beta\beta$ decay events

The classification scores of the three Transformer models were evaluated on a test dataset consisting of calibration events at well-defined γ-ray energy peaks. For the PSD cut to be applicable in LEGEND-200, the models must generalize beyond these narrow energy regions, in particular to the vicinity of $Q_{\beta\beta} = 2039$ keV. Reliable single-site classification at lower energies is important, as it enables validation of PSD efficiency using the continuous $2\nu\beta\beta$ decay spectrum present in physics data. In particular, the energy range between 1000 and 1300 keV, used in this analysis to extract the PSD efficiency at $Q_{\beta\beta}$, is a key validation region. Figure 31 shows that, for $E < 1$ MeV, only model 1 maintains a stable SSE efficiency across the spectrum, whereas the other models exhibit inconsistent behavior.

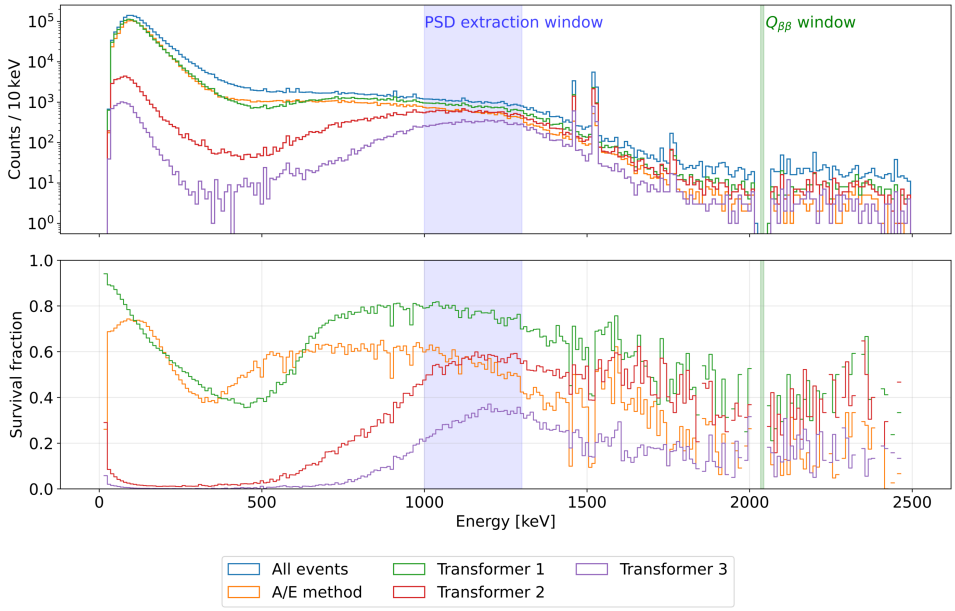

Figure 31: Energy spectra of the $2\nu\beta\beta$ decay population, with different PSD cuts, applied (top), and the corresponding survival fractions (bottom). The spectra include all physics runs listed in section 5.1, corresponding to a total exposure of $40.3\,\mathrm{kg}\cdot\mathrm{yr}$. We see that Transformer models 2 and 3 generalize very poorly for the low-energy part of the spectrum, while the A/E cut generally works better at very low energies.

5.4 Pulse shape discrimination efficiency at $Q_{\beta\beta}$

To evaluate the effect of PSD efficiencies on the expected $0\nu\beta\beta$ decay half-life, the PSD efficiency at $Q_{\beta\beta}$, denoted as ϵ_{PSD}, must be determined. We adopt a similar procedure as developed by the LEGEND-200 collaboration, described in [55]. In this analysis, only Mirion ICPC detectors are considered, which allows for consistent PSD cut definitions throughout the full analysis. Furthermore, the LEGEND-200 collaboration has decided to focus on ICPC detectors. Since we have no pure SSE populations at $Q_{\beta\beta}$ to estimate the efficiencies, we assume a linear energy dependence and extrapolate the efficiency at $Q_{\beta\beta}$.

We therefore calculate the PSD efficiency as:

$$\epsilon_{\mathrm{PSD}} = \tilde{\epsilon}_{\mathrm{PSD}} + \langle \delta_{\mathrm{run}} \rangle + \epsilon_{2\nu\beta\beta}\,, \tag{5.4}$$

where $\tilde{\epsilon}_{\mathrm{PSD}}$ denotes the uncorrected PSD efficiency at $Q_{\beta\beta}$. The correction

term $\langle \delta_{\text{run}} \rangle$ accounts for time-dependent variations in the PSD performance during data taking, and $\epsilon_{2\nu\beta\beta}$, compensates for differences in the single-site event population between $2\nu\beta\beta$ decays and the double escape peak calibration data. The energy dependence of SSE detection efficiency using the A/E parameters arises from three main physical effects [86]:

1. Bremsstrahlung becomes more probable at higher energies, leading to the emission of secondary photons. This can spatially separate the energy deposition and reduce the A/E value, causing SSEs to be misclassified as MSE, thereby lowering the detection efficiency at $Q_{\beta\beta}$ compared to the ^{208}Tl DEP [46].

2. Charge cloud self-repulsion increases with energy due to a higher number of induced charges. This results in longer drift times and broader waveforms, which in turn reduce the signal amplitude and the A/E value. As a result, the detection efficiency at $Q_{\beta\beta}$ is further reduced [87].

3. Electronic noise becomes less relevant at higher energies, which tends to increase the detection efficiency. However, this effect is generally outweighed by the first two.

Together, these effects imply that the PSD efficiency decreases at higher energies. In this work, we estimate the PSD efficiency at three different energy regions. The method for extracting the efficiency in each region is presented in the following subsections:

- In the 1000-1300 keV interval, using the $2\nu\beta\beta$ decay background (PSD extraction window)
- At the 1592.5 keV DEP from ^{208}Tl
- At the 2231.5 keV DEP from ^{56}Co.

5.4.1 PSD efficiency for $2\nu\beta\beta$ decay events

To estimate the PSD efficiency for $2\nu\beta\beta$ decay events, the Transformer models were tested on all periods listed in table 3 for a total exposure of $40.3\,\text{kg}\cdot\text{yr}$. The PSD efficiency can be evaluated at different energy regions within the spectrum. For this analysis, we use the 1000-1300 keV window because this region offers high statistics and is relatively free from prominent gamma lines. An alternative energy range between 1525 and 1750 keV could be used if contributions from gamma lines are subtracted. However,

this region suffers from significantly lower event statistics. Therefore, we rely exclusively on the lower-energy window for the PSD efficiency extraction. It is estimated as follows:

$$\epsilon_{2\nu\beta\beta} = \frac{\frac{T_p}{T_p+T_f} - \lambda_B \lambda_{Bp}}{1-\lambda_B} \,. \tag{5.5}$$

Here, T_p and T_f denote the number of events that pass or fail the PSD cuts, respectively. The parameter λ_B represents the fraction of events in the window that originate from the background, and λ_{Bp} is the fraction of those background events that pass the cuts.

The value of λ_B is derived from the expected number of $2\nu\beta\beta$ decays based on exposure, detection efficiency, and the known half-life. A detailed background analysis for LEGEND-200 by Calgaro et al. determined $\lambda_B = (8.7 \pm 1.7)\%$, which we adopt in this analysis [88].

Assuming the background events are equally likely to pass or fail the PSD cut, we set $\lambda_{Bp} = (50 \pm \frac{1}{\sqrt{12}})$ %, corresponding to a uniform distribution. The uncertainty on T_p and T_f is assumed to follow Poisson statistics. The uncertainty on $\epsilon_{2\nu\beta\beta}$ is given by

$$\begin{aligned}
\sigma_{\epsilon_{2\nu\beta\beta}} = \Bigg[& \left(\frac{\partial \epsilon}{\partial T_p} \cdot \sigma_{T_p} \right)^2 + \left(\frac{\partial \epsilon}{\partial T_f} \cdot \sigma_{T_f} \right)^2 \\
& + \left(\frac{\partial \epsilon}{\partial \lambda_B} \cdot \sigma_{\lambda_B} \right)^2 + \left(\frac{\partial \epsilon}{\partial \lambda_{Bp}} \cdot \sigma_{\lambda_{Bp}} \right)^2 \Bigg]^{1/2} \\
= \Bigg[& \left(\frac{\frac{T_f}{(T_p+T_f)^2}}{1-\lambda_B} \cdot \sigma_{T_p} \right)^2 + \left(\frac{\frac{-T_p}{(T_p+T_f)^2}}{1-\lambda_B} \cdot \sigma_{T_f} \right)^2 \\
& + \left(\frac{\frac{T_p}{T_p+T_f} - \lambda_{Bp}}{(1-\lambda_B)^2} \cdot \sigma_{\lambda_B} \right)^2 + \left(\frac{-\lambda_B}{1-\lambda_B} \cdot \sigma_{\lambda_{Bp}} \right)^2 \Bigg]^{1/2} .
\end{aligned} \tag{5.6}$$

5.4.2 PSD efficiency at double escape peaks

The PSD efficiency at the double escape peaks is determined by fitting the peaks separately for events that pass and events that fail the PSD cut. Example fits are shown in figure 32 and figure 33 for the ^{208}Tl and ^{56}Co DEPs, respectively. The full fit function is given by:

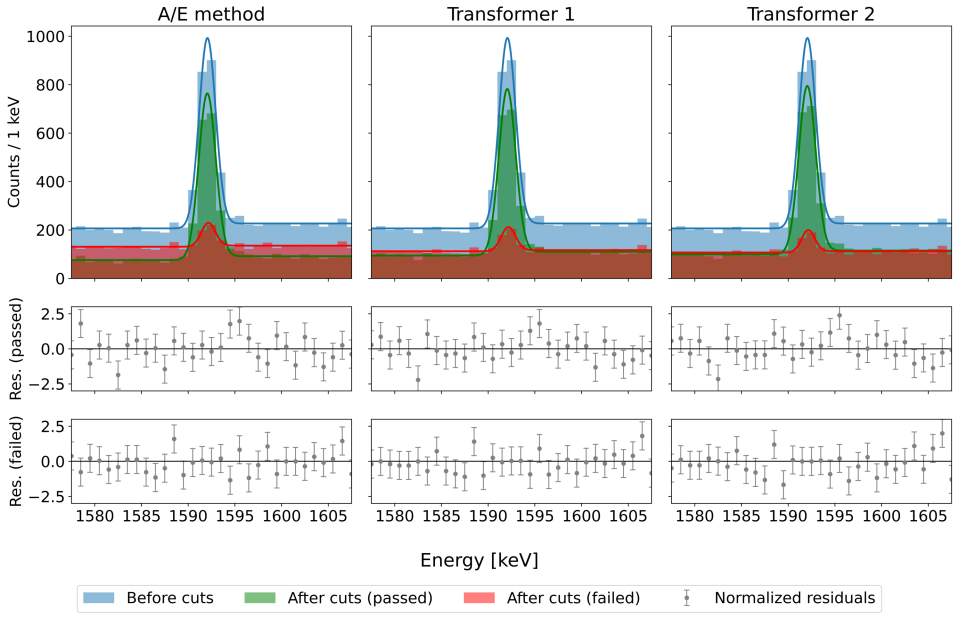

Figure 32: Peak fitting results in the ^{208}Tl DEP region. Each subfigure shows three histograms: all events (blue), events passing the PSD cut (green), and events failing the cut (red). The corresponding fits are shown as line plots. The two lower panels display the normalized residuals for the fit to the passing (middle) and failing (bottom) event populations. The residuals confirm the quality and stability of the fits.

$$f(x, A, \mu, \sigma, a, b, d) = A \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}} + \frac{d}{2} \cdot \operatorname{erfc}\left[\frac{x-\mu}{\sqrt{2}\cdot\sigma}\right] + a \cdot x + b\,, \tag{5.7}$$

where $\operatorname{erf}(x)$ is the complementary Gaussian error function, given by equation (5.8). The parameters of the fit are the amplitude A, mean μ, and standard deviation σ of the Gaussian, the slope a and intercept b of the background, and the size of the step function d.

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \mathrm{d}t\,. \tag{5.8}$$

The linear background is justified since we fit a narrow window of ± 20 keV around the peak. The step function is to account for low- and high-energy tails. The PSD efficiency and its uncertainty are subsequently computed from the background-subtracted peak amplitude determined in the fit:

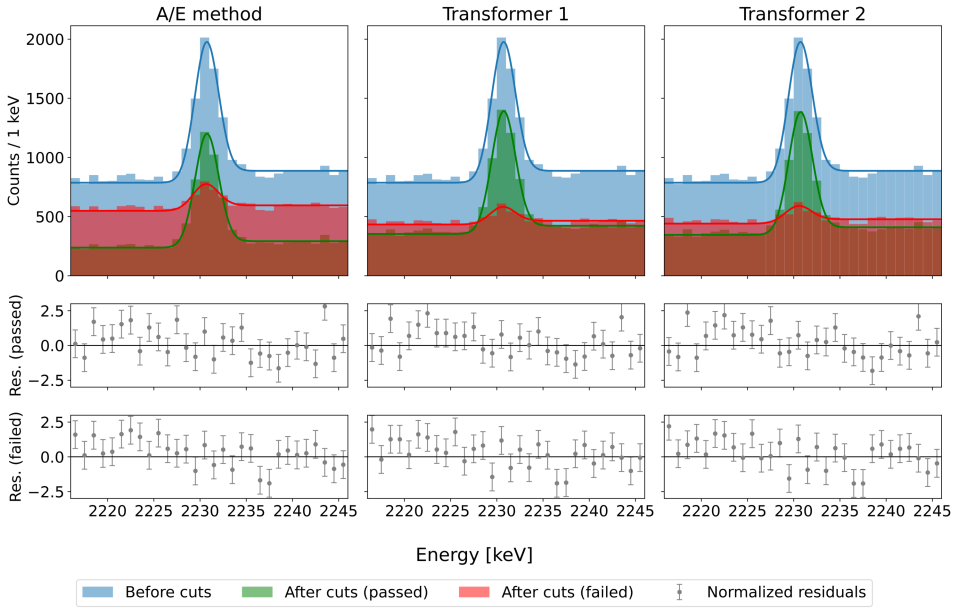

Figure 33: Peak fitting results in the ^{56}Co DEP region. Each subfigure shows three histograms: all events (blue), events passing the PSD cut (green), and events failing the cut (red). The corresponding fits are shown as line plots. The two lower panels display the normalized residuals for the fit to the passing (middle) and failing (bottom) event populations. The residuals confirm the quality and stability of the fits.

$$\epsilon_{\text{DEP}} = \frac{A_p}{A_p + A_f} \tag{5.9}$$

$$\sigma_{\epsilon_{\text{DEP}}} = \epsilon_{\text{DEP}} \cdot (1 - \epsilon_{\text{DEP}}) \cdot \sqrt{\left(\frac{\sigma_p}{A_p}\right)^2 + \left(\frac{\sigma_f}{A_f}\right)^2}, \tag{5.10}$$

where A_p and A_f are the amplitudes of the Gaussian fits for events that pass and fail the cut, respectively, σ_p and σ_f are their corresponding uncertainties. The upper plots in figures 32 and 33 show these fits. The A/E parameter is calibrated such that approximately 90% of events in the DEP region are classified as single-site. Consequently, only a small number of events are expected to fail the cut, which can lead to large uncertainties in A_f due to low statistics.

We determine the PSD efficiency individually for each detector. To ob-

tain an overall value, we initially computed a weighted average of the individual efficiencies. To account for potential inconsistencies between detectors, we applied Particle Data Group-style uncertainty scaling: if the spread of the efficiencies is larger than expected from their statistical uncertainties, the combined uncertainty is scaled by a factor $S = \sqrt{\chi^2/\mathrm{dof}}$. This avoids underestimating the total uncertainty if there are unaccounted systematic uncertainties between detectors [3]. In principle, this method is justified since the event sets that pass and fail the cut are disjoint.

However, a poor goodness-of-fit ($\chi^2 = 9$) indicated that the uncertainty reported by the initial fit underestimates the true dispersion in the measurements. This discrepancy likely stems from detector-specific effects that are not captured in the statistical errors alone. To account for variations in readout electronics, differences in supply voltages, or other subtle hardware-related factors, we introduce an additional detector-specific uncertainty term, denoted by $\langle\delta_{\mathrm{det}}\rangle$. We model the total variance by augmenting each statistical uncertainty with this extra contribution and fit the data using a least-χ^2 method[1]:

$$\chi^2\left(\epsilon_i, \sigma_{\epsilon,i} \mid \langle\varepsilon\rangle, \langle\delta_{\mathrm{det}}\rangle\right) = \sum_i \left[\frac{(\epsilon_i - \langle\varepsilon\rangle)^2}{\sigma_{\epsilon,i}^2 + \langle\delta_{\mathrm{det}}\rangle^2} + \log\left(\sigma_{\epsilon,i}^2 + \langle\delta_{\mathrm{det}}\rangle^2\right)\right], \quad (5.11)$$

where ϵ_i is the PSD efficiency measured for detector i, $\sigma_{\epsilon,i}$ the corresponding statistical uncertainty, $\langle\varepsilon\rangle$ the global average PSD efficiency, and $\langle\delta_{\mathrm{det}}\rangle$ is the additional uncertainty representing unaccounted detector-to-detector variations. The logarithmic term ensures proper normalization of the likelihood. The resulting PSD efficiencies for the double escape peaks of ^{208}Tl and ^{56}Co are shown in figures 34 and 35, respectively. The additional detector-to-detector variances for both DEPs are shown in table 9.

The global PSD efficiency at the DEP peak is obtained as the inverse-variance weighted mean:

$$\langle\hat{\epsilon}\rangle = \frac{\sum_i \frac{\epsilon_i}{\left(\sigma_{\epsilon,i}^2 + \langle\delta_{\mathrm{det}}\rangle^2\right)}}{\sum_i \frac{1}{\left(\sigma_{\epsilon,i}^2 + \langle\delta_{\mathrm{det}}\rangle^2\right)}}. \quad (5.12)$$

The corresponding $1\,\sigma$ uncertainty on the mean efficiency is given by:

[1] To make the connection to section 4.2: The least χ^2 is equivalent to the negative log-likelihood if we assume the likelihood to be Gaussian.

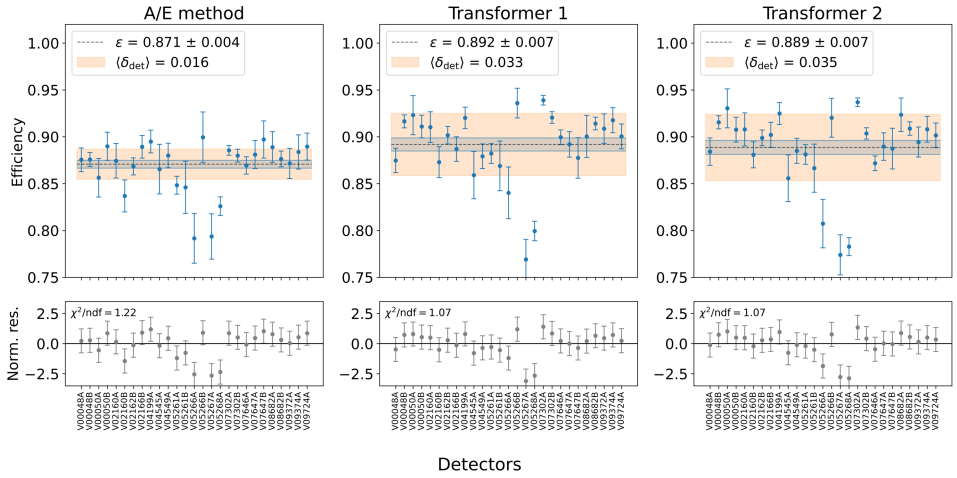

Figure 34: PSD efficiencies in the ^{208}Tl DEP (1592.5 keV) for Mirion detectors used in this analysis. Most detectors exhibit efficiencies (blue band) between 85% and 90%, although a few show noticeably lower performance across all PSD methods. The additional detector-to-detector uncertainty is indicated by the orange band. The A/E method shows a small spread (1.6%), while the Transformer classifications exhibit larger systematic variations ($> 3\%$). The lower panel shows the normalized residuals, demonstrating the quality of the fit.

$$\sigma_{\langle\hat{\epsilon}\rangle} = \left(\sum_i \frac{1}{\sigma_{\epsilon,i}^2 + \langle\delta_{\text{det}}\rangle^2} \right)^{-1/2}, \tag{5.13}$$

which incorporates both statistical fluctuations and the additional variance due to detector-to-detector differences.

5.4.3 Time dependence

To estimate how the PSD efficiency varies over time across different data-taking periods, we repeat the procedure explained in the previous section. Instead of grouping the data by detector, we now group it by run. Each subset thus contains events from all detectors used in the analysis during that run. We then fit the efficiency data per run using the same statistical model, but we reinterpret the additional uncertainty term. In equation (5.11), the detector-specific term $\langle\delta_{\text{det}}\rangle$ is replaced by a run-specific term $\langle\delta_{\text{run}}\rangle$, which quantifies the spread in the efficiencies due to potential

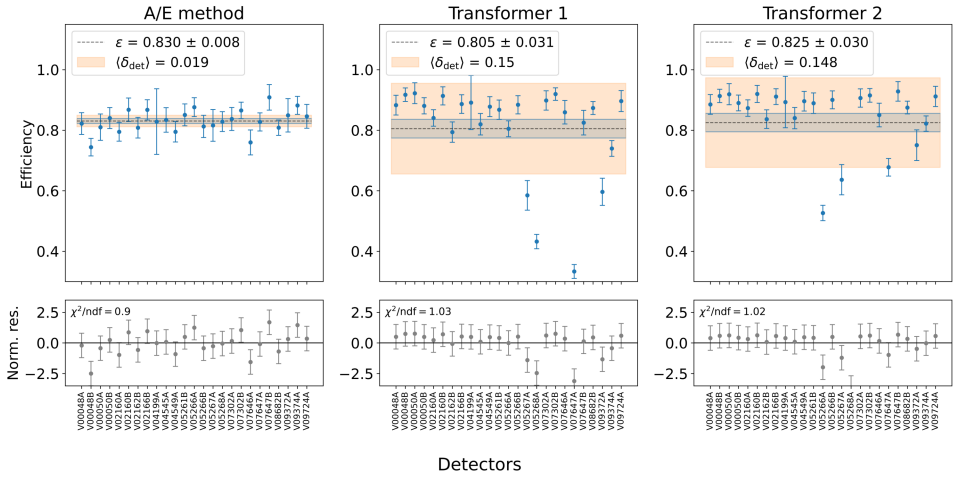

Figure 35: PSD efficiencies in the ^{56}Co DEP (2231.5 keV) for Mirion detectors used in this analysis. The efficiency is indicated by the blue band and the detector-to-detector uncertainty by the orange band. The Transformer models show less robustness compared to the A/E method when applied to this DEP. In several detectors, the Transformer classification fails completely. Possible causes include detector-specific waveform variations not captured in training or sensitivity to isotope-dependent event topologies, leading to significantly increased uncertainty. The lower panel shows the normalized residuals, demonstrating the quality of the fit.

time-dependent effects, such as drifts in detector response, electronic noise variations, or changes in environmental conditions.

$$\chi^2(\epsilon_i, \sigma_{\epsilon,i} \mid \langle\epsilon\rangle, \langle\delta_{\text{run}}\rangle) = \sum_i \left[\frac{(\epsilon_i - \langle\epsilon\rangle)^2}{\sigma_{\epsilon,i}^2 + \langle\delta_{\text{run}}\rangle^2} + \log\left(\sigma_{\epsilon,i}^2 + \langle\delta_{\text{run}}\rangle^2\right) \right] . \quad (5.14)$$

This approach allows us to quantify any systematic run-to-run variations in the PSD efficiency. The PSD efficiencies are illustrated in figure 36, the spread due to time-dependent effects is shown in table 9.

5.4.4 Combined efficiency at $Q_{\beta\beta}$

Although the ^{56}Co source produces several higher-energy DEPs, most exhibit limited statistics and are therefore not suitable for a reliable estimation of the PSD efficiency estimates. To estimate the PSD efficiency at the region

Table 9: Additional uncertainty representing unaccounted detector-to-detector variations at the ^{208}Tl and ^{56}Co DEPs, and the systematic run-to-run variation. The PSD performance at the ^{208}Tl DEP exhibits some spread ($< 4\%$). While the efficiency is very stable over time ($< 1\%$ for all models), the Transformer models show a very large detector-to-detector variation.

Quantity	**A/E [%]**	**Model 1 [%]**	**Model 2 [%]**
$\langle\delta_{\text{det}}\rangle$ (^{228}Th)	1.6	3.3	3.5
$\langle\delta_{\text{det}}\rangle$ (^{56}Co)	1.9	15.0	14.8
$\langle\delta_{\text{run}}\rangle$	0.8	0.6	0.9

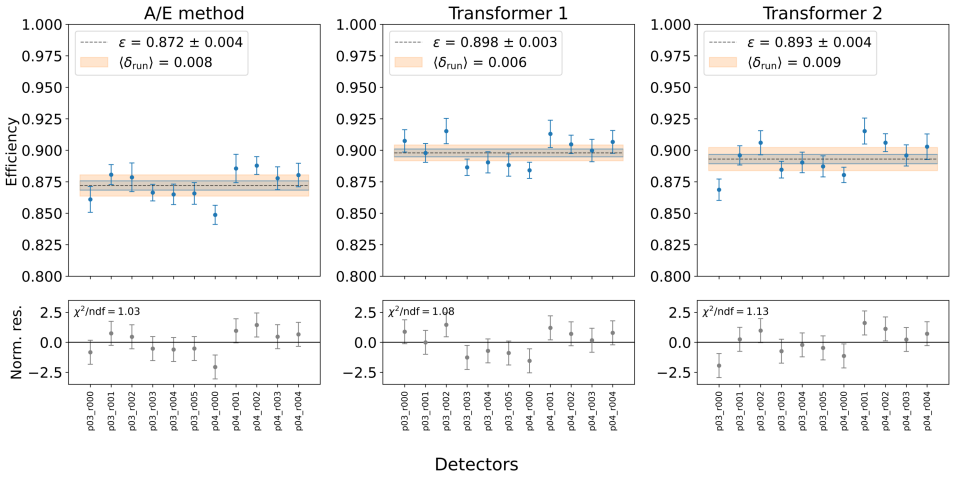

Figure 36: Time stability of the PSD efficiency (blue band) in the ^{208}Tl DEP, shown per run for all detectors used. The efficiency remains very stable over time, and there is very little time-dependent uncertainty ($< 1\%$ throughout all models). The bottom panel shows the normalized residuals.

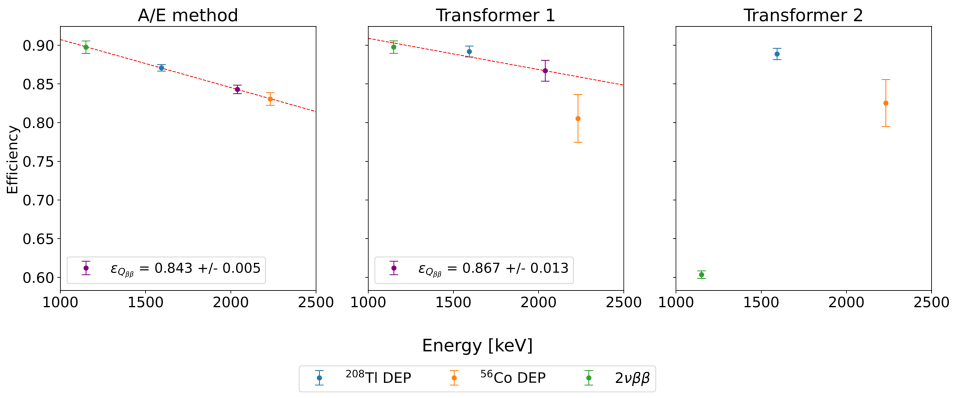

Figure 37: PSD efficiency as a function of energy for the three PSD methods. We evaluate the PSD efficiency at three different energies and extrapolate to $Q_{\beta\beta}$ using a linear fit. The A/E method is the most precise. The PSD efficiency of Transformer 1 is higher but exhibits a larger uncertainty. The Transformer model 2 failed for the $2\nu\beta\beta$ decay events, which is why no fit was performed.

of interest, $\epsilon_{Q_{\beta\beta}}$, we perform a linear fit to the PSD efficiencies measured at the three selected energy regions. This is shown in figure 37. The uncertainty is obtained by propagating the uncertainties from the fit parameters using the covariance matrix:

$$\sigma_{\epsilon_{Q_{\beta\beta}}} = J \cdot \mathrm{Cov} \cdot J^{\mathsf{T}} \,, \tag{5.15}$$

where $J = \left[\frac{\partial y}{\partial a} \; \frac{\partial y}{\partial b}\right]$ is the Jacobian and Cov the covariance matrix of the fitted parameters.

The combined experimental efficiency of the experiment ϵ, which enters the Bayesian fit as a nuisance parameter, includes also the liquid argon veto efficiency ϵ_{LAr}, the quality cut efficiency $\epsilon_{\mathrm{quality}}$, the fraction of active detector mass $\epsilon_{\mathrm{active}}$ and the ^{76}Ge enrichment fraction ϵ_{Ge}. Therefore, the total experimental efficiency is given by:

$$\epsilon = \epsilon_{\mathrm{PSD}} \cdot \epsilon_{\mathrm{LAr}} \cdot \epsilon_{\mathrm{quality}} \cdot \epsilon_{\mathrm{active}} \cdot \epsilon_{\mathrm{Ge}} \,. \tag{5.16}$$

The associated uncertainty is estimated using Gaussian error propagation. Factoring out the total efficiency ϵ, we find:

Table 10: Summary of non-PSD experimental efficiencies in the range around $Q_{\beta\beta}$ used in this analysis. Values are taken from the LEGEND-200 internal metadata database for periods 3 and 4 only. For each efficiency type, the per-detector values were averaged, and the quoted uncertainty is the standard error of the mean across detectors.

Efficiency	**Value** [%]
Liquid Argon veto	93 ± 1
Quality cuts	97.48 ± 0.01
Active volume	92.5 ± 0.3
^{76}Ge enrichment	92.6 ± 0.1
Total (excluding PSD)	77.7 ± 0.9

$$\sigma_\epsilon = \epsilon \cdot \sqrt{\sum_i \left(\frac{\sigma_{\epsilon_i}}{\epsilon_i} \right)}, \tag{5.17}$$

where the sum runs over all individual efficiency contributions listed above. The experimental efficiencies, not including PSD efficiencies, are summarized in table 10. These values were obtained from the LEGEND-200 internal metadata database, which is maintained by the collaboration and not publicly released. Efficiencies are provided for each detector and data-taking period; in this analysis, only periods 3 and 4 are used. For each efficiency type, the per-detector values were combined into a single number by taking the arithmetic mean, with the uncertainty given by the standard error of the mean.

5.5 Pulse shape simulation

The LEGEND-200 collaboration began developing a software package called LegendGeSim [89] with the goal of simulating detector waveforms for comparison studies with real data. This pulse shape simulation (PSS) framework is intended to support the validation of signal processing, and the goal was to use it to increase event classification across different energies. Written in Julia, LegendGeSim generates both idealized and partly realistic waveforms starting with a Geant4 simulation output. These results are stored in so-called PET files, which contain event position, energy and time information needed to simulate charge collection and signal formation in germanium detectors. The detector geometries required for these simulations are provided through metadata.

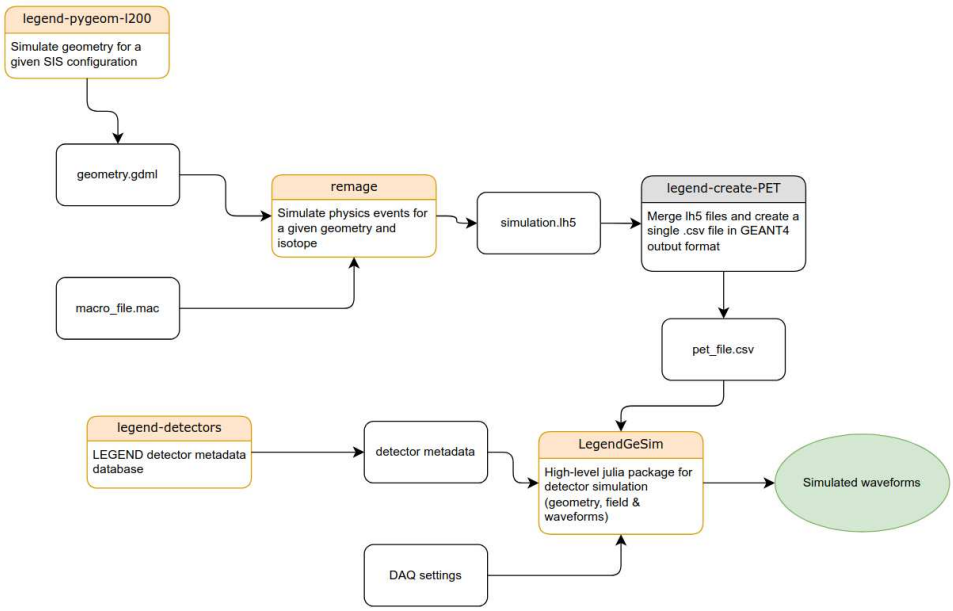

Figure 38: Connection between different LEGEND packages for the use of LegendGeSim.

The PET files are produced using ReMaGe, a modern `C++` Geant4-based simulation framework for germanium experiments [90]. The geometries themselves are defined in the legend-pygeom-l200 package [91], developed within the LEGEND-200 collaboration. Finally, a dedicated interface script was developed to convert the Geant4 output into a format compatible with LegendGeSim. Additionally, a coordinate transformation is required, since ReMaGe and LegendGeSim use different coordinate systems[2].

Figure 38 illustrates the interaction between the various software components used in the waveform simulation pipeline.

Despite the promising concept, the LegendGeSim framework remains under development and is not yet fully functional. In attempting to use it, we encountered several limitations that hindered its use for this project:

- **Incomplete physics modeling**: Essential mechanisms such as diffusion, charge cloud drift, charge trapping, and temperature dependence are either not fully implemented or not comparable with data.
- **Systematic uncertainties**: Systematics associated with different

[2]Credit to Giovanna Saleh for identifying and resolving this inconsistency.

PSD parameters and the simulated electronics chain are not yet characterized.

- **Inaccurate energy reconstruction**: The trapezoidal filter used in the framework systematically underestimates the true energy of events.

The inaccurate energy reconstruction is evident in figure 39, which compares simulated data to calibration measurements. Example waveforms are shown in figure 40, where a systematic slowly rising current signal is visible.

The current pulse shape simulation workflow suffers from significant inefficiency, with runtimes so long that large-scale waveform generation is impractical. A major bottleneck is that, in its present implementation, ReMaGe must fully simulate each event before deciding whether to retain it. Consequently, complete decay chains (e.g., the ^{228}Th chain) are simulated, and the full energy spectrum is generated, including events outside the peaks of interest. There is the option to restrict the decay chain (e.g., to only ^{208}Tl $\rightarrow^{208}$ Pb); this has to be handled with care, as it might introduce bias. While it is possible to apply generator-level energy cuts, this approach is non-trivial. These inefficiencies occur before the actual pulse-shape simulation in LegendGeSim. For the dataset shown in figure 39, the simulation required 27.5 hours of wall time on NERSC (using 128 CPU cores in parallel for ReMaGe only) for a single detector (V09372A), yielding approximately 5000 events in the ^{208}Tl DEP.

The Pulse Shape Simulation (PSS) framework was expected to serve as a practical tool for this study. Although some development work was anticipated, the framework turned out to be less mature and more complex to integrate than expected. At the time of writing, further development is ongoing within the LEGEND-200 collaboration, with significant contributions from the University of Zurich. Nonetheless, this work established a functioning processing pipeline for handling simulated PSS data from a single detector. Due to time constraints, a full-scale simulation was not achieved, but the essential tools required for generating, processing, and analyzing simulated waveforms are now in place for use in future studies.

5.6 Summary and combined results at $Q_{\beta\beta}$

In this work, a Transformer-based PSD method for the LEGEND-200 experiment was developed. Of the three transformer models tested, only Transformer model 1 demonstrated stable and reliable performance across the full energy range. The PSD efficiency at $Q_{\beta\beta}$ was obtained by interpolating between three reference points: $2\nu\beta\beta$ decay events in the 1000-1300 keV

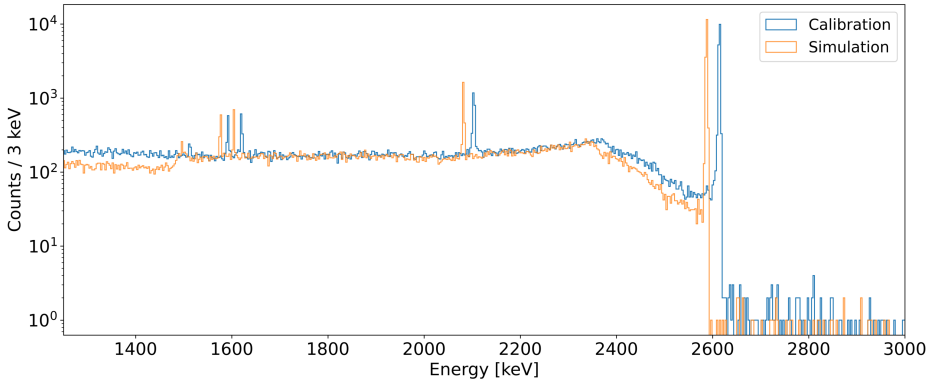

Figure 39: Energy spectrum for the IC detector V09372A, comparing calibration data (blue) from period 3 with simulated data (orange). In the simulation, the decay chain is restricted to nuclei with $208 \leq A \leq 212$ and $81 \leq Z \leq 83$, and only events with energies above 1200 keV are retained. The simulated spectrum shows a systematic shift of the characteristic γ-lines toward lower energies, indicating inaccuracies in the energy reconstruction.

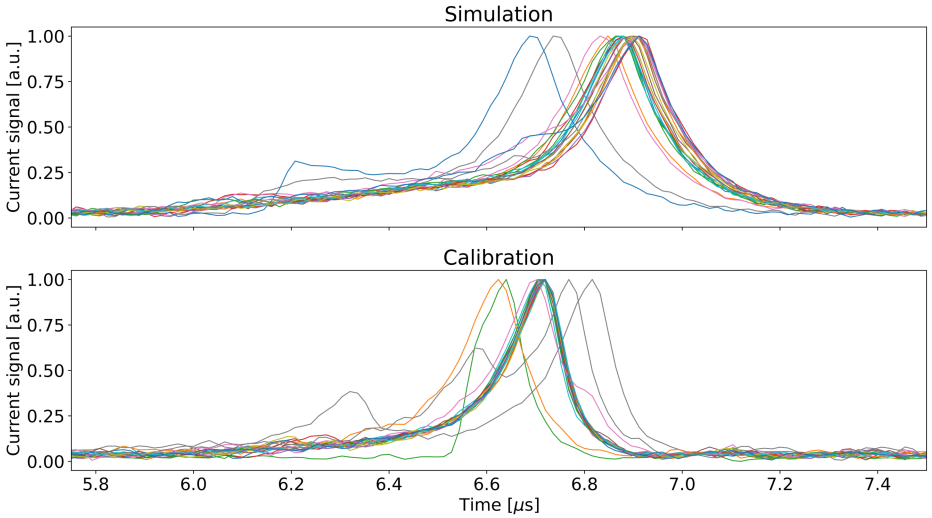

Figure 40: Example waveforms for the IC detector V09372A in the ^{208}Tl DEP at 1592.5 keV, comparing simulated waveforms (top panel) with calibration data from period 3 (bottom panel). The simulated waveforms exhibit a systematically slower, early rise compared to the measured data.

interval, the 1592.5 keV DEP of ^{208}Tl, and the 2231.5 keV DEP of ^{56}Co. Additional systematic uncertainty was included to account for time-dependent stability, detector-to-detector variations, and differences in PSD acceptance between $2\nu\beta\beta$ and DEP events. The final PSD efficiency yields $(86.7 \pm 1.3)\%$ at $Q_{\beta\beta}$ for the Transformer. This is consistent with the efficiency obtained by the conventional A/E method, $(84.3 \pm 0.5)\%$, with a two-sided p-value of $p = 0.08$ indicating no significant difference between the two approaches at the 5% level. Taking into account the experimental efficiencies listed in table 10, the overall detection efficiency achieved with the Transformer-based approach is $(66.8 \pm 2.1)\%$, compared to $(65.0 \pm 2.0)\%$ for the A/E method.

6 Sensitivity study for the $0\nu\beta\beta$ decay

Bayesian statistics offers a powerful framework for modeling uncertainty and performing probabilistic inference, making it particularly well-suited for rare-event searches such as $0\nu\beta\beta$ decay. In this chapter, we explain the essential principles of Bayesian inference and present the statistical model used to estimate the experimental sensitivity for $0\nu\beta\beta$ decay discovery under a background-only hypothesis and varying pulse shape discrimination efficiencies. We do not apply the likelihood model (equation (6.7)) directly to real data. Although the LEGEND-200 data in the $(Q_{\beta\beta} \pm 25)$ keV window has been unblinded, waveform-level information is not accessible to the full collaboration. Instead, we estimate the expected 90% credible upper limit on the signal half-rate, defined as the inverse of the signal half-life (see equation (2.22)), using toy Monte Carlo datasets. The usage of half-rate $\mathcal{S} = 1/T^{0\nu}_{1/2}$ instead of half-life improves numerical stability. To this end, we generate toy spectra and fit them using a signal-plus-background model within a Markov Chain Monte Carlo framework. From the resulting posterior distributions, we derive upper limits on the signal half-rate and define the experimental sensitivity as the median of the 90% credible interval (CI) distributions. We also detail the fit procedure, including the treatment of nuisance parameters.

6.1 Introduction to Bayesian inference

Bayes' theorem forms the foundation of Bayesian inference, providing a formal mechanism for updating probabilistic beliefs based on observed data:

$$p(\boldsymbol{\theta} \mid \text{data}) = \frac{p(\text{data} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\text{data})} \,. \tag{6.1}$$

Here, $p(\boldsymbol{\theta})$ denotes the prior distribution, which encodes our prior knowledge or assumptions about the parameters before observing any data. The likelihood $p(\text{data} \mid \boldsymbol{\theta})$ quantifies how probable the observed data is under a given choice of parameters. The resulting posterior distribution $p(\boldsymbol{\theta} \mid \text{data})$ reflects the updated knowledge after accounting for the observed data. If the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ are independent, the posterior becomes:

$$p(\boldsymbol{\theta} \mid \text{data}) = \frac{p(\text{data} \mid \boldsymbol{\theta}) \cdot \prod_{i=1}^{n} p(\theta_i)}{p(\text{data})} \,. \tag{6.2}$$

This factorization is an approximation that holds only if the parameters are uncorrelated under the prior. In practice, some parameters may exhibit

correlations, and modeling their joint prior distribution may be necessary to accurately capture posterior dependencies. The normalization factor, $p(\text{data})$, known as marginal likelihood, ensures the posterior is normalized:

$$p(\text{data}) = \int_{\theta} p(\theta) p(\text{data} \mid \theta)\, d\theta\,. \tag{6.3}$$

This integral marginalizes over all possible values of θ, weighted by their prior probabilities. Marginalization is particularly useful to extract the distribution for a given parameter by integrating out the joint distribution over the unknowns that are not of interest. We later apply it to remove nuisance parameters – quantities that affect the model but are not of direct interest [92].

Once the posterior distribution is obtained, it can be used to make predictions about new or unseen data via the posterior predictive distribution. It describes the distribution of potential future observations from a repeated experiment conducted under identical conditions:

$$p(y \mid \text{data}) = \int p(y, \theta \mid \text{data})\, d\theta = \int p(y \mid \theta) p(\theta \mid \text{data})\, d\theta\,. \tag{6.4}$$

Here, y represents a hypothetical data, θ are the model parameters, and $p(y \mid \theta)$ is the likelihood of observing y given those parameters. The posterior distribution $p(\theta \mid \text{data})$ captures what we have learned about θ from the observed data.

Posterior predictive studies are particularly useful for model checking and sensitivity analysis, as they naturally incorporate uncertainty in the model parameters. They are often used to compare the posterior predictive distribution to newly observed data, because this provides insight into how the model and prior assumptions align with reality. If the original data were modeled appropriately, then the generated values of y under the model should exhibit a distribution similar to that of the observed data.

6.2 Statistical model for sensitivity estimation

To evaluate the effect of the different PSD efficiencies on the half-life sensitivity of the $0\nu\beta\beta$ decay, we perform a Bayesian sensitivity study using toy Monte Carlo datasets generated under the null hypothesis (background-only scenario). The $0\nu\beta\beta$ decay analysis is performed over the energy range from 1930 to 2190 keV. Two γ-lines at (2104 ± 5) keV and (2119 ± 5) keV are excluded, resulting in a net analysis window of 240 keV [28]. The expected

number of $0\nu\beta\beta$ decay events in toy MC datasets, as a function of the signal half-rate $\mathcal{S} = 1/T_{1/2}^{0\nu}$, is given by:

$$s\left(\mathcal{S}\right) = \frac{\log 2 \cdot \mathcal{E} \cdot N_A \cdot \epsilon}{m_{\text{mol}}} \cdot \mathcal{S} = \frac{\log 2 \cdot \mathcal{E} \cdot N_A \cdot \epsilon}{m_{\text{mol}}} \cdot \frac{1}{T_{1/2}}\,. \tag{6.5}$$

Here, we have the following parameters [93]:

- $\mathcal{E} = 1000$ kg $\cdot$ yr is the exposure
- $N_A = 6.022 \times 10^{23}$ mol^{-1} is Avogadro's number
- ϵ is the detection efficiency, calculated as in equation (5.16)
- $m_{\text{mol}} = 75.92$ g/mol is the molar mass of ^{76}Ge
- $T_{1/2}^{0\nu}$ is the $0\nu\beta\beta$ decay half-life

The number of background events b is given by:

$$b\left(\mathcal{B}\right) = \mathcal{E} \cdot \Delta E \cdot \mathcal{B}\,, \tag{6.6}$$

where $\Delta E = 240$ keV is the net width of the energy window used for the fit and $\mathcal{B}$ the background index in counts/(keV·kg·yr). Assuming a counting experiment with a total of n observed events surviving the analysis cuts in the analysis window, the full unbinned extended likelihood takes the form:

$$L\left(\text{data} \mid \mathcal{S},\ \mathcal{B},\ \vartheta\right) = \frac{(s+b)^n \cdot e^{-(s+b)}}{n!} \cdot \prod_i^n \left[\frac{1}{s+b}\Big(s \cdot p_s(E_i) + b \cdot p_b\Big)\right] \tag{6.7}$$

The first factor is the Poisson probability of observing n events given the total expected rate $s + b$, while the product accounts for the probability density functions of each event's energy E_i, normalized over signal and background contributions. Further, ϑ denotes the nuisance parameters. In this analysis, the signal distribution $p_s(E)$ is modeled as a Gaussian centered at the decay Q-value, which is valid if we assume negligible detector-specific non-Gaussian tails or energy leakage. The background distribution $p_b(E)$ is modeled as flat across the analysis window:

$$p_s(E) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(E-x)^2}{2\sigma^2}}\,, \tag{6.8}$$

$$p_b = \frac{1}{E_{\text{max}} - E_{\text{min}}}\,, \tag{6.9}$$

where $x = Q_{\beta\beta} - \Delta$ is the energy bias, with $\Delta = E_{\text{true}} - E_{\text{cal}}$. In addition, we consider two more nuisance parameters: the energy resolution σ and the signal efficiency ϵ. The prior distributions are modeled as Gaussian centered around their nominal values ($\hat{\sigma}$, $\hat{\epsilon}$, and $\hat{\Delta}$). Assuming the nuisance parameters to be independent, the joint prior distribution is:

$$\begin{aligned}\mathcal{P}(\sigma, \epsilon, \Delta) &= \mathcal{P}(\sigma) \cdot \mathcal{P}(\epsilon) \cdot \mathcal{P}(\Delta) \\ &= \frac{1}{\sqrt{2\pi}\sigma_\sigma} \cdot e^{-\frac{(\sigma-\hat{\sigma})^2}{2\sigma_{\hat{\sigma}}^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \cdot e^{-\frac{(\epsilon-\hat{\epsilon})^2}{2\sigma_{\hat{\epsilon}}^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_\Delta} \cdot e^{-\frac{(\Delta-\hat{\Delta})^2}{2\sigma_{\hat{\Delta}}^2}} .\end{aligned} \tag{6.10}$$

Here, σ_ϑ is the standard deviation of the respective prior, which quantifies the uncertainty in the measurements. For the signal half-rate $\mathcal{S}$ and the background index $\mathcal{B}$, we use uniform priors over physically allowed ranges, which are $[0, 1000] \times 10^{27}$ yr^{-1} for $\mathcal{S}$ and $[0, 0.1]$ counts/(keV·kg·yr) for the background index. Applying Bayes' theorem, the posterior probability density becomes:

$$\begin{aligned}p\left(\mathcal{S},\ \mathcal{B},\ \sigma,\ \epsilon,\ \Delta \mid \text{data}\right) \propto\ & L\left(\text{data} \mid \mathcal{S},\ \mathcal{B},\ \sigma,\ \epsilon,\ \Delta\right) \\ & \times \mathcal{P}(\sigma, \epsilon, \Delta) \cdot \mathcal{P}(\mathcal{S}) \cdot \mathcal{P}(\mathcal{B}) .\end{aligned} \tag{6.11}$$

This formulation enables the systematic propagation of uncertainties in resolution and efficiency to the final inference on the signal half-rate and background index. In general, the marginalization over nuisance parameters cannot be solved analytically. Therefore, we perform the integration numerically using a Markov chain Monte Carlo (MCMC) approach. Specifically, we employ the Metropolis-Hastings sampling algorithm to generate samples from the posterior distribution. For this analysis, we use 5 independent MCMC chains, each with 10^5 sampling steps. Multiple MCMC chains are employed to mitigate the dependence of finite-length chains on their initial values, thereby enhancing the reliability of the resulting estimates.

6.3 Toy Monte Carlo simulations

The procedure for the generation and analysis of toy datasets is as follows. First, we sample nuisance parameters. For each toy dataset, the signal detection efficiency ϵ and the energy resolution σ are drawn from their respective priors. Since only Mirion ICPC detectors are considered in this analysis, we achieve an excellent energy resolution of $\sigma = (0.914 \pm 0.062)$ keV. Furthermore, Δ was set to zero, neglecting a potential energy bias. This is in

agreement with the recently published first results from the LEGEND-200 experiment, which show an energy bias of (0.3 ± 0.3) keV for ICPC detectors [94]. In that analysis, an energy-bias correction was applied; however, the fitted bias was consistent with zero within uncertainties.

In the second step, we generate the number of events. The total number is drawn from a Poisson distribution, with an expected mean determined by the assumed background index, as shown in equation (6.6). To determine the final experimental sensitivity of LEGEND-200, we use the background goal of $\mathcal{B} = 2 \cdot 10^{-4}$ counts/(keV·kg·yr). Finally, we construct the toy spectrum: assuming background-only, event energies are sampled from the background probability density function $p_b(E)$.

Each toy dataset is then analyzed using the full signal-plus-background likelihood model defined in section 6.2. To extract the 90% CI upper limit $\mathcal{S}_{90}$ on the signal half-rate, we first marginalize the joint posterior over all other parameters:

$$p(\mathcal{S} \mid \text{data}) = \int p(\mathcal{S}, \mathcal{B}, \sigma, \epsilon, \Delta \mid \text{data})\, d\mathcal{B}\, d\sigma\, d\epsilon\,. \tag{6.12}$$

The 90% CI upper limit is then given by:

$$\int_0^{\mathcal{S}_{90}} p\left(\mathcal{S} \mid \text{data}\right)\, d\mathcal{S} = 0.9\,. \tag{6.13}$$

This procedure is repeated for 10^4 independent toy datasets. This number ensures statistical robustness of the $\mathcal{S}_{90}$ distribution and smooth convergence of CI estimates. The median of $\mathcal{S}_{90}$, denoted $\tilde{\mathcal{S}}_{90}$, serves as the expected upper limit on the signal half-rate for repeated experiments under the same conditions. This can be translated into a lower bound on the $0\nu\beta\beta$ decay half-life.

Figure 41 shows the distribution of best-fit $\mathcal{S}$ obtained from each toy fit for both A/E and Transformer-based PSD methods. As expected under the background-only hypothesis, most fits yield $\mathcal{S} = 0$, though discrete populations at higher $\mathcal{S}$ appear due to statistical fluctuations from events near $Q_{\beta\beta}$. These effects are shown in figure 43. Figure 42 shows the corresponding distributions of best-fit background indices $\mathcal{B}$. Both PSD methods recover consistent values with the input background index, indicating that the MCMC fits are unbiased and stable.

This analysis is performed using ZeroNuFit.jl [95], a specialized Bayesian analysis tool built on the Bayesian Analysis Toolkit (BAT.jl) [96]. ZeroNuFit.jl implements extended unbinned maximum likelihood fits tailored for

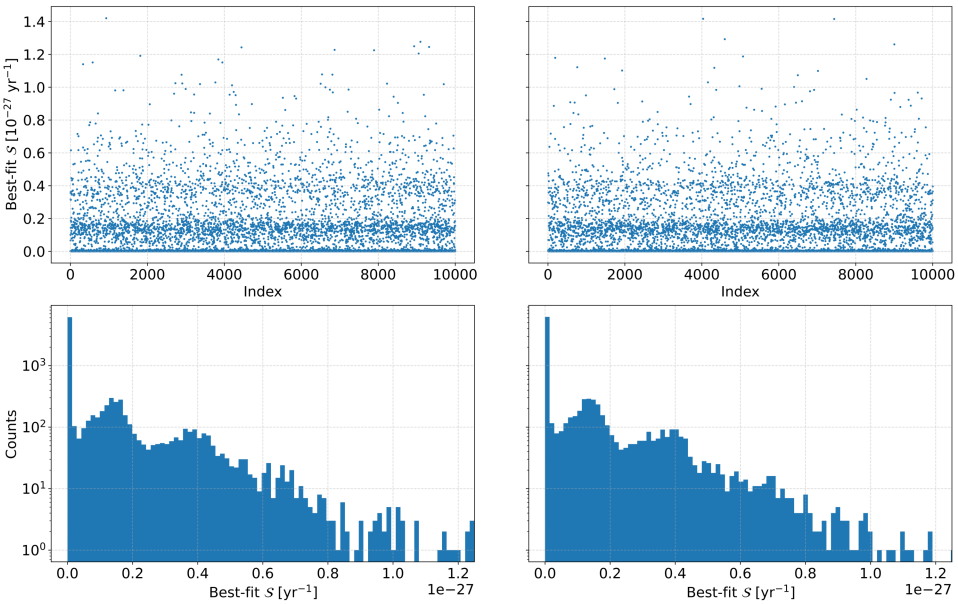

Figure 41: Scatter plots (top) and distributions (bottom) for the best-fit signal half-rate $\mathcal{S}$ obtained from MCMC fits to 10^4 toy Monte Carlo datasets generated under the background-only hypothesis. The A/E method (left) and the Transformer-based PSD method (right) correspond to different assumed PSD efficiencies in the toy generation. Both yield similar results, with the majority of toys returning $\mathcal{S} = 0$, because no events fall within the signal region around $Q_{\beta\beta}$. Non-zero $\mathcal{S}$ values arise when background fluctuations place one or more events close to $Q_{\beta\beta}$, producing discrete populations at characteristic $\mathcal{S}$ levels. The correlation between region of interest occupancy and fitted $\mathcal{S}$ is shown in figure 43.

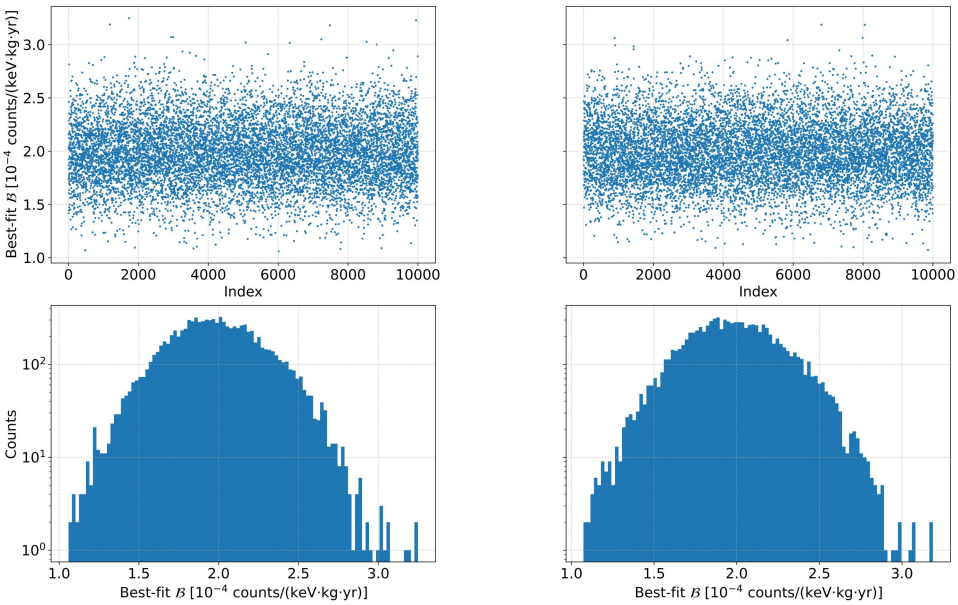

Figure 42: Scatter plots (top) and distributions (bottom) of best-fit background indices $\mathcal{B}$ from 10'000 toy fits. Both PSD methods (A/E on the left, Transformer-based on the right) correctly recover the assumed background level of 2×10^{-4} counts/keV/kg/yr. The distributions are approximately Gaussian, validating the fit model and confirming the stability of the inference procedure.

$0\nu\beta\beta$ decay searches, allowing flexible modeling of both signal peaks and background components. The statistical framework we used is further described in [18].

6.4 Results of the Bayesian analysis

To assess the physics reach of the Transformer PSD strategy, in comparison with the conventional A/E method, a Bayesian half-life sensitivity analysis for the $0\nu\beta\beta$ decay was performed separately for each PSD method. Using the LEGEND-200 goal exposure and background index, the final $0\nu\beta\beta$ decay half-life sensitivity of LEGEND-200 was estimated, with the PSD efficiencies treated as Gaussian-distributed nuisance parameters in the fit.

For each toy dataset, we extracted the 90% one-sided upper CI limit on the signal half-rate, $\mathcal{S}_{90}$. Its distribution over all toys reflects the expected sensitivity of the experiment: the median, $\tilde{\mathcal{S}}_{90}$ defines the central value, and

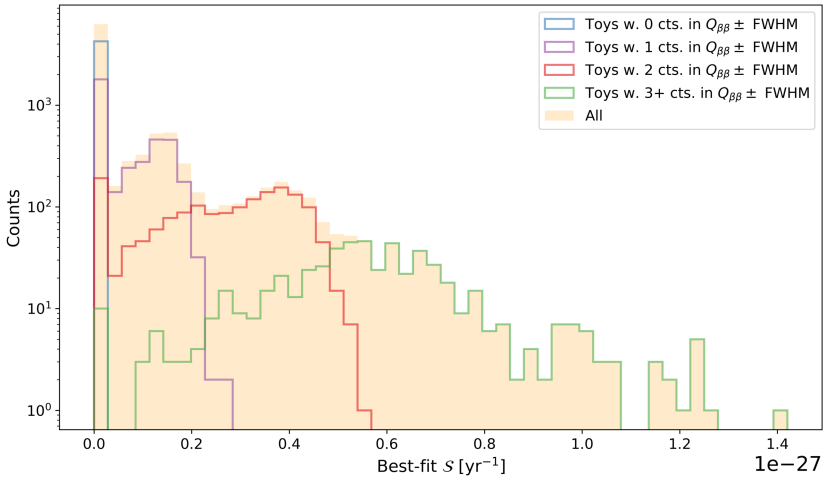

Figure 43: Distribution of best-fit $\mathcal{S}$ from A/E-based MCMC fits, grouped by the number of events observed within $Q_{\beta\beta}\pm$ FWHM. Modes at non-zero $\mathcal{S}$ are associated with datasets containing one or more background events near the region of interest.

Table 11: Comparison of PSD efficiency, resulting signal half-rate limits and half-life limits for the Transformer and A/E methods.

Method	**PSD eff. [%]**	$\tilde{\mathcal{S}}_{90}$ **[10^{-28} yr^{-1}]**	$\mathbf{T}^{0\nu}_{1/2}$**[10^{27} yr]**
A/E	84.3 ± 0.5	$8.23^{+3.65}_{-1.76}$	$1.22^{+0.33}_{-0.37}$
Transformer	86.7 ± 1.3	$8.00^{+3.42}_{-1.71}$	$1.25^{+0.34}_{-0.37}$

the 68% credible interval quantifies the associated uncertainty. This procedure incorporates both statistical fluctuations and systematic uncertainties, as the signal detection efficiency and energy resolution are sampled from their priors and marginalized over in each fit. Since the signal half-rate is inversely proportional to the half-life, its median and 68% CI range can be directly translated into a median and 68% CI range for the $0\nu\beta\beta$ decay half-life. By comparing the half-lives derived from each PSD method, we can evaluate their respective impacts on the exclusion sensitivity.

All results are summarized in table 11, and the signal half-rate 90% CI upper limits distributions are shown in figure 44. For the A/E cut, we obtain:

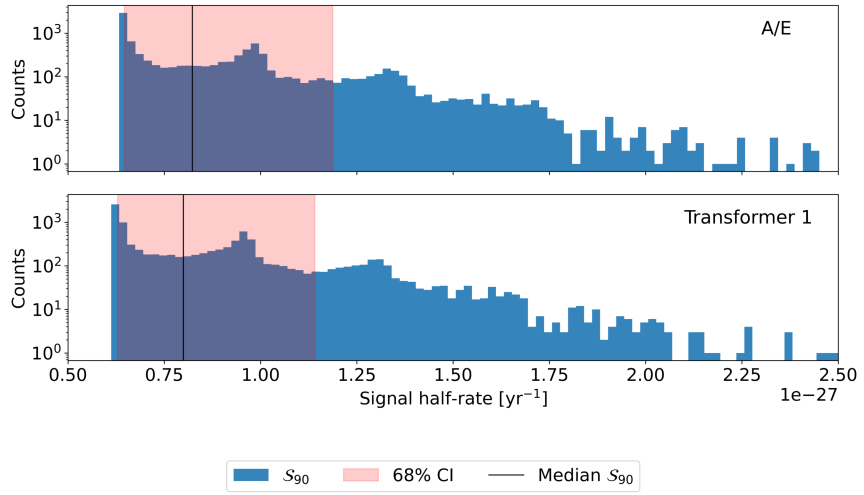

Figure 44: Distribution of 90% upper credible limits on the signal half-rate $\mathcal{S}$ obtained from 10'000 toy Monte Carlo simulations under the background-only hypothesis. The upper plot shows results obtained from the A/E cut, the lower from the Transformer-based PSD cut. The Transformer method achieved a slightly improved sensitivity, reflected by the lower median.

$$\mathcal{S} = 8.23^{+3.65}_{-1.76} \times 10^{-28} \tag{6.14}$$

With the definition $T^{0\nu}_{1/2} = 1/\mathcal{S}$, this translates to a Bayesian 90% credible lower bound on the $0\nu\beta\beta$ decay half-life of:

$$T_{1/2} > 1.22^{+0.33}_{-0.37} \times 10^{27}\ \text{yr} \tag{6.15}$$

Similarly, using the Transformer-based PSD efficiency, the Bayesian analysis yields:

$$\mathcal{S} < 8.00^{+3.42}_{-1.71} \times 10^{-28}\ \text{yr}^{-1} \tag{6.16}$$

and

$$T_{1/2} > 1.25^{+0.34}_{-0.37} \times 10^{27}\ \text{yr} \tag{6.17}$$

7 Conclusion and outlook

In this work, a Transformer-based approach to pulse shape discrimination was developed for the LEGEND-200 experiment, aiming to improve the separation of signal-like events from background noise in the search for neutrinoless double beta decay. The model developed achieved a PSD efficiency of $(86.7 \pm 1.3)\%$ at $Q_{\beta\beta}$, which agrees with the conventional A/E method $(84.3 \pm 0.5)\%$. A two-sided test of the difference between Transformer-based PSD efficiency and A/E yields $p = 0.08$, indicating that the observed 2.4% difference is not statistically significant at the 5% level. While the Transformer-based method demonstrates a higher central efficiency, it also exhibits a larger uncertainty. This broader uncertainty reflects not just statistical variation, but also contributions from model variance and possible sensitivity to detector-specific features. In contrast to the A/E method, which is calibrated per detector and period, the Transformer is trained globally across all detectors and periods. This choice enables cross-detector generalization but can also introduce performance fluctuations when the training data does not sufficiently capture detector-to-detector variation. The elevated uncertainty may therefore indicate a trade-off between generalization and robustness, and suggests the need for better-balanced training data.

A Bayesian $0\nu\beta\beta$ decay sensitivity analysis yields signal half-rate limits of $\mathcal{S} = 8.23^{+3.65}_{-1.76} \times 10^{-28}$ (A/E) and $\mathcal{S} < 8.00^{+3.42}_{-1.71} \times 10^{-28}$ yr^{-1} (Transformer), if we assume the LEGEND-200 design goals are reached (background index 2×10^{-4} counts/(keV·kg·yr), exposure 1000 kg $\cdot$ yr). Both Transformer-based method ($T^{0\nu}_{1/2} = 1.25^{+0.34}_{-0.37} \times 10^{27}$yr) and conventional A/E ($T^{0\nu}_{1/2} = 1.22^{+0.33}_{-0.37} \times 10^{27}$yr) cuts agree within uncertainty, validating the Transformer-based approach as a viable alternative for pulse shape discrimination.

Although a significant sensitivity improvement was not yet achieved, these results demonstrate that machine-learning techniques, particularly Transformer architectures, can achieve performance comparable to well-established PSD methods refined and optimized extensively over many years. However, the Transformer introduces additional computational overhead, since training requires labeled datasets generated using conventional methods. This cost is largely a one-time expense, as the training would be repeated or updated only when detector conditions change. Once trained, classification of new waveforms is relatively fast, so the computational cost is amortized over large amounts of data. With further development, the

Transformer-based PSD method shows promise for enhancing sensitivity in rare-event searches, though several aspects still require improvement.

First, the quality and consistency of training data should be improved by applying stricter data cleaning procedures, more rigorous detector selection, and better data balance (across energies and detectors) to increase the model's robustness and reduce sensitivity to unwanted features.

Second, further investigation is needed into the ^{56}Co waveforms, as their inclusion in the training set led to poor generalization of events below 1.5 MeV, resulting in reduced PSD performance in the $2\nu\beta\beta$ decay energy region. This again highlights the need for improved training data, particularly in terms of data quality and data balance.

Third, incorporating pulse shape simulations and domain adaptation techniques could help mitigate detector-specific effects and improve the model performance across a larger energy region.

If Transformer-based (or other ML-based) PSD methods continue to demonstrate competitive performance, future work should also address their integration into the experiment's digital signal processing pipeline. As LEGEND progresses towards its ton-scale goals, deep learning-based PSD methods may become a critical component in maximizing the experiment's sensitivity to rare physics beyond the Standard Model. Their scalability could not only enhance event classification but also enable advanced techniques such as unsupervised anomaly detection, offering new avenues for discovery.

Acknowledgements

I would like to thank the LEGEND group at the University of Zurich for their support throughout this project. I am especially grateful to Aravind Remesan Sreekala and Gloria Senatore for their help with detector simulations, to Giovanna Saleh for her support on pulse shape simulations, to Francesco Piastra for his guidance on detector physics, and to Dr. Sofia Calgaro for her invaluable support with the Bayesian sensitivity analysis.
I am grateful to Prof. Laura Baudis for welcoming me into her group and for her guidance and insightful feedback at key stages of the project, as well as to Dr. Marta Babicz for her guidance and supervision. I also thank Prof. Laura Baudis, Dr. Marta Babicz, and Dr. Sofia Calgaro for their feedback and suggestions, which helped improve this thesis. Finally, I would like to thank my partner and my parents for their support and encouragement throughout my studies.

References

[1] M. E. Peskin and D. V. Schroeder, *An introduction to quantum field theory*, The advanced book program (CRC Press, Taylor & Francis Group, Boca Raton London New York, 2019).

[2] S. Weinberg, *The quantum theory of fields* (Cambridge University Press, Cambridge ; New York, 1995).

[3] S. Navas, C. Amsler, et al., "Review of particle physics", Physical Review D **110**, 030001 (2024).

[4] S. M. Bilen'kij, *Introduction to the physics of massive and mixed neutrinos*, Second edition, Lecture notes in physics volume 947 (Springer, Cham, 2018).

[5] B. Pontecorvo, "Neutrino experiments and the problem of conservation of leptonic charge", Soviet Journal of Experimental and Theoretical Physics **26**, 984 (1968).

[6] N. Agafonova, A. Alexandrov, et al. (OPERA Collaboration), "Final results on neutrino oscillation parameters from the OPERA experiment in the CNGS beam", Phys. Rev. D **100**, 051301 (2019).

[7] M. A. Acero, P. Adamson, et al. (The NOvA Collaboration), "Improved measurement of neutrino oscillation parameters by the NOvA experiment", Phys. Rev. D **106**, 032004 (2022).

[8] Super-Kamiokande Collaboration, Y. Fukuda, et al., "Evidence for oscillation of atmospheric neutrinos", Physical Review Letters **81**, 1562–1567 (1998).

[9] Super-Kamiokande Collaboration, S. Fukuda, et al., "Solar ^{8}B and hep neutrino measurements from 1258 days of Super-Kamiokande data", Physical Review Letters **86**, 5651–5655 (2001).

[10] Q. R. Ahmad, R. C. Allen, et al. (SNO Collaboration), "Direct evidence for neutrino flavor transformation from neutral-current interactions in the sudbury neutrino observatory", Phys. Rev. Lett. **89**, 011301 (2002).

[11] M. Thomson, *Modern particle physics* (Cambridge University Press, Cambridge, United Kingdom ; New York, 2013).

[12] K. Zuber, *Neutrino physics*, 3rd ed. (CRC Press, Boca Raton, May 11, 2020).

[13] C. Giunti and C.-W. Kim, *Fundamentals of neutrino physics and astrophysics* (Oxford university press, Oxford, 2007).

[14] T. Yanagida, "Horizontal symmetry and masses of neutrinos", Progress of Theoretical Physics **64**, 1103–1105 (1980).

[15] KamLAND Collaboration, A. Gando, et al., "Reactor on-off antineutrino measurement with KamLAND", Physical Review D **88**, 033001 (2013).

[16] J. Engel and J. Menéndez, "Status and future of nuclear matrix elements for neutrinoless double-beta decay: a review", Reports on Progress in Physics **80**, 046301 (2017).

[17] S. M. Bilenky and C. Giunti, "Neutrinoless double-beta decay: a probe of physics beyond the standard model", International Journal of Modern Physics A **30**, 1530001 (2015).

[18] S. Borden, S. Calgaro, et al., *L-Note 25-001: $0\nu\beta\beta$ statistical analysis of the first year of LEGEND-200 data*, LEGEND Collaboration Internal Note, 2025.

[19] W. Rodejohann, "Neutrino-less double beta decay and particle physics", International Journal of Modern Physics E **20**, 1833–1930 (2011).

[20] M. J. Dolinski, A. W. P. Poon, and W. Rodejohann, "Neutrinoless double-beta decay: status and prospects", Annual Review of Nuclear and Particle Science **69**, 219–251 (2019).

[21] W. Huang, M. Wang, F. Kondev, G. Audi, and S. Naimi, "The AME 2020 atomic mass evaluation (i). evaluation of input data, and adjustment procedures*", Chinese Physics C **45**, 030002 (2021).

[22] J. Kotila and F. Iachello, "Phase-space factors for double-β decay", Phys. Rev. C **85**, 034316 (2012).

[23] V. D'Andrea, N. Di Marco, et al., "Neutrinoless double beta decay with germanium detectors: 10^{26} yr and beyond", Universe **7**, 341 (2021).

[24] S. Dell'Oro, S. Marcocci, M. Viel, and F. Vissani, "Neutrinoless double beta decay: 2015 review", Advances in High Energy Physics **2016**, 2162659 (2016).

[25] KamLAND-Zen Collaboration, S. Abe, et al., "Search for the Majorana nature of neutrinos in the inverted mass ordering region with KamLAND-Zen", Physical Review Letters **130**, 051801 (2023).

[26] D. Q. Adams, C. Alduino, et al., "Search for Majorana neutrinos exploiting millikelvin cryogenics with CUORE", Nature **604**, 53–58 (2022).

[27] C. Augier et al., "Final results on the $0\nu\beta\beta$ decay half-life limit of ^{100}Mo from the CUPID-Mo experiment", Eur. Phys. J. C **82**, 1033 (2022).

[28] GERDA Collaboration, M. Agostini, et al., "Final results of GERDA on the search for neutrinoless double-β decay", Physical Review Letters **125**, 252502 (2020).

[29] Majorana Collaboration, I. J. Arnquist, et al., "Final result of the Majorana demonstrator's search for neutrinoless double-β decay in ^{76}Ge", Physical Review Letters **130**, 062501 (2023).

[30] M. M. Ivanov, M. Simonović, and M. Zaldarriaga, "Cosmological parameters and neutrino masses from the final Planck and full-shape BOSS data", Physical Review D **101**, 083504 (2020).

[31] M. Aker, A. Beglarian, et al., "Direct neutrino-mass measurement with sub-electronvolt sensitivity", Nature Physics **18**, 160–166 (2022).

[32] I. Esteban, M. C. Gonzalez-Garcia, et al., "NuFit-6.0: updated global analysis of three-flavor neutrino oscillations", Journal of High Energy Physics **2024**, 216 (2024).

[33] R. Abbasi, M. Ackermann, et al., "Measurement of atmospheric neutrino mixing with improved IceCube DeepCore calibration and data processing", Physical Review D **108**, 012014 (2023).

[34] M. Agostini, G. Benato, and J. A. Detwiler, "Discovery probability of next-generation neutrinoless double-β decay experiments", Physical Review D **96**, 053001 (2017).

[35] J. Torres, *Toej93/LobsterPlot*, Oct. 8, 2024.

[36] M. Agostini, A. M. Bakalyarov, et al., "Upgrade for Phase II of the GERDA experiment", The European Physical Journal C **78**, 388 (2018).

[37] N. Abgrall, A. Abramov, et al., "The large enriched germanium experiment for neutrinoless double beta decay (LEGEND)", AIP Conference Proceedings **1894**, 020027 (2017).

[38] L. Collaboration, N. Abgrall, et al., *LEGEND-1000 preconceptual design report*, arXiv.org, (July 23, 2021) `https://arxiv.org/abs/2107.11462v1` (visited on 01/10/2025).

[39] P. Hofmann, *Solid state physics: an introduction*, 1st ed, New York Academy of Sciences Series (John Wiley & Sons, Incorporated, Newark, 2015).

[40] S. H. Simon, *The Oxford solid state basics*, Reprinted (with corrections, twice) (Oxford University Press, Oxford, 2017).

[41] S. M. Sze, M.-K. Lee, and M. K. Lee, *Semiconductor devices, physics and technology*, 3. ed (Wiley, Hoboken, N.J, 2012).

[42] G. F. Knoll, *Radiation detection and measurement*, 3rd ed (Wiley, New York, 2000).

[43] J. Walker, D. Halliday, and R. Resnick, *Fundamentals of physics*, Tenth edition (Online-Ausg.) (Wiley, Hoboken, New Jersey, 2014).

[44] M. Agostini, G. Araujo, et al., "Pulse shape analysis in gerda phase II", The European Physical Journal C **82**, 284 (2022).

[45] M. Berger, J. Hubbell, S. Seltzer, J. Coursey, and D. Zucker, *XCOM: photon cross section database (version 1.2)*, en, Jan. 1999.

[46] T. Comellato, M. Agostini, and S. Schönert, "Charge-carrier collective motion in germanium detectors for β-decay searches", The European Physical Journal C **81**, 76 (2021).

[47] Z. He, "Review of the shockley–ramo theorem and its application in semiconductor gamma-ray detectors", Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **463**, 250–267 (2001).

[48] S. Ramo, "Currents induced by electron motion", Proceedings of the IRE **27**, 584–585 (1939).

[49] W. Shockley, "Currents to conductors induced by a moving point charge", Journal of Applied Physics **9**, 635–636 (1938).

[50] I. Abt, C. Gooch, et al., "Temperature dependence of the electron-drift anisotropy and implications for the electron-drift model", Journal of Instrumentation **18**, P10030 (2023).

[51] M. Willers and for the LEGEND collaboration, "Signal readout electronics for LEGEND-200", Journal of Physics: Conference Series **1468**, 012113 (2020).

[52] M. Salathe and T. Kihm, "Optimized digital filtering techniques for radiation detection with hpge detectors", Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **808**, 150–155 (2016).

[53] Y. Müller, "Calibration of the LEGEND-200 experiment to search for neutrinoless double beta decay and searches for signatures of new physics with the GERDA experiment", PhD thesis (University of Zurich, 2023).

[54] L. Baudis, G. Benato, et al., "Calibration sources for the LEGEND-200 experiment", Journal of Instrumentation **18**, P02001 (2023).

[55] I. Guinn, V. Biancacci, et al., *L-Note 24-013: Pulse Shape Discrimination for the 2024 Unblinded Analysis of LEGEND-200*, LEGEND Collaboration Internal Note, 2024.

[56] S. Badillo, B. Banfai, et al., "An introduction to machine learning", Clinical Pharmacology and Therapeutics **107**, 871–885 (2020).

[57] S. J. D. Prince, *Understanding deep learning* (The MIT Press, Cambridge, Massachusetts, 2023).

[58] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", Nature **521**, 436–444 (2015).

[59] K. P. Murphy, *Probabilistic machine learning: an introduction*, Adaptive computation and machine learning (The MIT Press, Cambridge, Massachusetts London, England, 2022).

[60] T. M. Mitchell, *Machine learning*, McGraw-Hill series in computer science (McGraw-Hill, New York, 1997).

[61] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators", Neural Networks **2**, 359–366 (1989).

[62] M. Telgarsky, *Benefits of depth in neural networks*, 2016.

[63] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient BackProp", in *Neural networks: tricks of the trade: second edition*, edited by G. Montavon, G. B. Orr, and K.-R. Müller (Springer, Berlin, Heidelberg, 2012), pp. 9–48.

[64] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines vinod nair", Proceedings of ICML **27**, 807–814 (2010).

[65] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks", Neural Information Processing Systems **25**, 10.1145/3065386 (2012).

[66] D. Hendrycks and K. Gimpel, *Gaussian error linear units (GELUs)*, June 6, 2023.

[67] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into deep learning*, Aug. 22, 2023.

[68] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal loss for dense object detection*, Feb. 7, 2018.

[69] A. Mavrin, *Artemmavrin/focal-loss*, Oct. 30, 2024.

[70] S. A. Taghanaki, Y. Zheng, et al., "Combo loss: handling input and output imbalance in multi-organ segmentation", Computerized Medical Imaging and Graphics **75**, 24–33 (2019).

[71] D. P. Kingma and J. Ba, *Adam: a method for stochastic optimization*, Jan. 30, 2017.

[72] The Manim Community Developers, *Manim – mathematical animation framework*, version v0.19.0, Jan. 2025.

[73] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, Dec. 10, 2015.

[74] J. L. Ba, J. R. Kiros, and G. E. Hinton, *Layer normalization*, July 21, 2016.

[75] A. Vaswani, N. Shazeer, et al., *Attention is all you need*, Aug. 2, 2023.

[76] A. Madani, B. Krause, et al., "Large language models generate functional protein sequences across diverse families", Nature Biotechnology **41**, 1099–1106 (2023).

[77] N. Parmar, A. Vaswani, et al., *Image transformer*, June 15, 2018.

[78] OpenAI, J. Achiam, et al., *GPT-4 technical report*, Mar. 4, 2024.

[79] J. Ansel, E. Yang, et al., *PyTorch 2: faster machine learning through dynamic python bytecode transformation and graph compilation*, Apr. 2024.

[80] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: pre-training of deep bidirectional transformers for language understanding*, May 24, 2019.

[81] S. Calgaro, V. D'Andrea, et al., *L-Note 25-010: Data Partitioning and Exposures for the First LEGEND-200 Unblinding*, LEGEND Collaboration Internal Note, 2024.

[82] *PyHEP 2023 - Python in HEP users workshop (online)*, Indico, (Oct. 9, 2023) `https://indico.cern.ch/event/1252095/` (visited on 04/01/2025).

[83] J. Detwiler, L. Pertoldi, et al., *Legend-pydataobj*, version v1.11.13, May 2025.

[84] W. Quinn, S. Calgaro, et al., *L-Note 24-011: Data Cleaning and Quality: HPGe Quality Cuts and Efficiencies*, LEGEND Collaboration Internal Note, 2024.

[85] P. Virtanen, R. Gommers, et al., "SciPy 1.0: fundamental algorithms for scientific computing in python", Nature Methods **17**, 261–272 (2020).

[86] T. Comellato, M. Agostini, and S. Schönert, "Topologies of ge double-beta decay events and calibration procedure biases", The European Physical Journal C **83**, 236 (2023).

[87] P. Mullowney, M.-C. Lin, et al., "Computational models of germanium point contact detectors", Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **662**, 33–44 (2012).

[88] S. Calgaro, T. Dixon, et al., *L-Note 24-007: Background analysis for LEGEND-200 data release at Neutrino 2024*, LEGEND Collaboration Internal Note, 2024.

[89] LEGEND Collab., *Legend-exp/LegendGeSim.jl*, May 10, 2025.

[90] L. Pertoldi, M. Huber, et al., *Remage*, version v0.11.0, May 2025.

[91] LEGEND Collab., *Legend-exp/legend-pygeom-l200*, May 21, 2025.

[92] A. Gelman, J. B. Carlin, et al., *Bayesian data analysis*, Third edition, Texts in statistical science series (CRC Press, Taylor and Francis Group, Boca Raton London New York, 2014).

[93] M. Agostini, M. Allardt, et al., "Background-free search for neutrinoless double-β decay of ^{76}Ge with GERDA", Nature **544**, 47–52 (2017).

[94] H. Acharya, N. Ackermann, et al., "First results on the search for lepton number violating neutrinoless double beta decay with the LEGEND-200 experiment", (2025).

[95] T. Dixon and S. Calgaro, *Legend-exp/ZeroNuFit.jl: v2.3.1*, version v2.3.1, Apr. 2025.

[96] O. Schulz, F. Beaujean, et al., "BAT.jl: a julia-based tool for bayesian inference", SN Computer Science **2**, 210 (2021).