TR
Technical Report

ML4NP: First Quarterly Report

Brief summary of the GARR Scholarship work

Luigi Berducci

Istituto Nazionale di Fisica Nucleare
SEZIONE DI ROMA TRE

Contents

CHAPTER 1

Introduction

This chapter aims to summarize the activities carried out in the first quarterly of the project. They covered the preliminary steps of the definition of the problem and the creation of data in a suitable format for the next phase of feature engineering.

To this aim, we investigated on the methodologies and tools needed to complete this task, establishing collaboration with other experts in particle physics and simulations.

The workflow to produce a Machine Learning dataset is depicted in Figure 1.1 and consists of several activities:

1. Data generation: Simulation of the physics processes that are involved in the LEGEND200 experiments.

2. Data cleaning: Select only the data we are interested in, discarding data that are not observable in the real experimental setup.

3. Data preparation: Convert the simulated data in a format consistent with the experimental data acquisitions.

4. Data analysis: investigate the characterization of the simulated processes to acquire a better understanding of the context.

1.1 Problem definition

In this section, we briefly remark the experimental setting and report a more precise definition of the problem, that we refined in this preliminary period.

The experiment LEGEND-200 aims to study a rare Neutrino-Physics process that occurs at very low energy, using Germanium detectors in liquid Argon (LAr). The experiment results extremely sensitive to background processes such as radioactive decays, cosmogenic background, and electromagnetic noise. These events could distort the readout signals in the low-energy spectrum in which we are interested in, and then drastically affect the outcome of the physical analysis.

For this reason, a lot of work is being done to reduce this background in several ways: improving the electronics, engineering the detector geometry and designing a proper trigger system. From our side, we are interested in proving that cosmogenic

events can be recognized based on the energy detection in LAr and implement a module that automatically discards them, saving time and storage resources.

With this clear objective in mind, we propose to train a Machine Learning classifier based on the readouts of SiPMs sensors. We formalize the problem as binary classification of the following classes of events:

- **Signal**: processes induced by the muon passage in LAr. They includes ^{41}Ar de-excitation from neutron capture, or electrons produced by muon ionization.

- **Background**: intense ^{39}Ar activity in natural LAr, and Ambiental radioactivity produced by surrounding materials, such as Thorium, Uranium, and Potassium decays.

The experiment is not running yet, then we aim to reproduce the events through Monte Carlo simulations as realistic as possible. Once having the event simulations, we aim to make a faithful reproduction of the SiPMs readouts to work on a format consistent with the experimental setup.

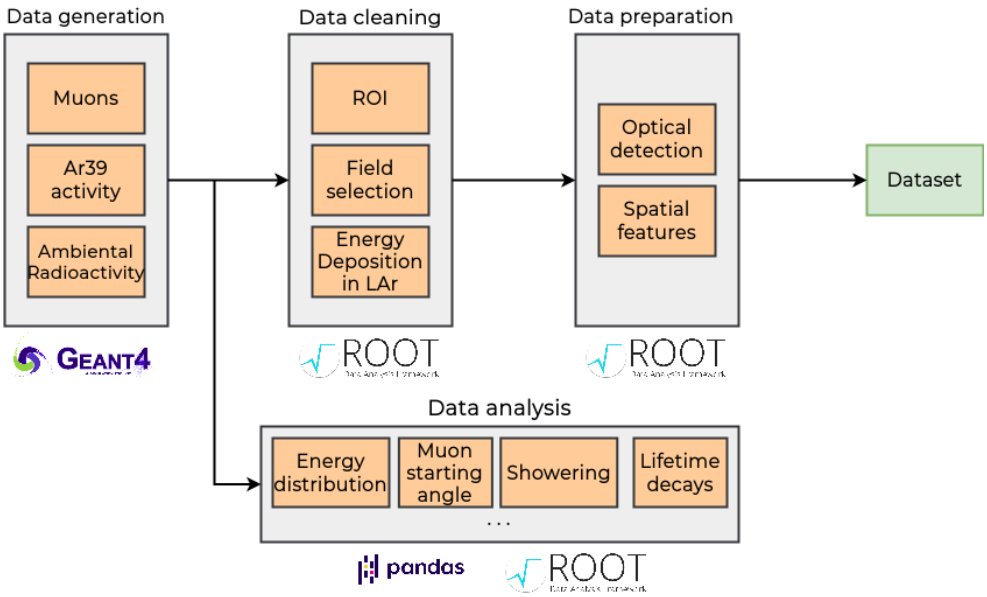

Figure 1.1: Diagram of the activities (main blocks) carried out in the first quarterly of the project, the related tasks (inner blocks) and tools adopted.

1.2 Outline

Each phase is the result of several design choices that we decided according to the conducted data analysis. For this reason, we dedicated a separate chapter to the data analysis. We suggest referring to it to have a better understanding of the work done so far.

The rest of the report is structured as follows:

- Chapter 2 covers the *preprocessing*, describing the details of each phase. Moreover, it will report the tools (e.g. software, libraries, and programming languages) that we adopted in this work.
- Chapter 3 presents the analysis of data from simulations and shows some of the obtained results.

CHAPTER 2

Preprocessing

2.1 Data generation

In this section, we report the work done to produce data with Monte Carlo simulation. We remark that simulation represents only the first step of preprocessing because it produces a large amount of data that we have to clean and process to obtain a suitable format for ML.

First of all, we identified three physical processes to simulate:

- Muon propagation in the LEGEND200 geometry,
- Ar39 radioactivity in liquid argon,
- Ambiental radioactivity typically experienced at the INFN Gran Sasso National Laboratory (LNGS).

The occurrence of these events is not disjoint in the real context, but they all contribute to the observed data. However, we decided to simulate them independently and produce full-scale simulation. Once defined trigger details in collaboration with the Data Acquisition (DAQ) team, we will extract the data in a finite time window using an integration time consistent with the electronic devices. In this way, we guarantee flexibility for the later stages of the study. The procedure to produce the ML dataset will be discussed in Section 2.3.

2.1.1 General information

The primary tool for simulating the interaction of particles through matter is GEANT4 (G4). This toolkit allows the users to define the detector geometry, the materials used, and simulate physical processes, according to a comprehensive library of particle interactions modeled according to the scientific literature. LEGEND200 collaboration uses an additional tool called MaGe. It consists of a G4 wrapper created by the GERDA and Majorana collaborations a few years ago.

We are collaborating with CJ Douglas Barton, an expert on G4 simulation of the University of South Dakota (USA) and member of the LEGEND collaboration, for the simulations of muons and Ar39. Figure 2.1 illustrates the geometry of the detector used for the simulations of muons and Ar39 processes. Besides, we are

currently analyzing Ambiental radioactivity simulations from Matthew Green of NC State University (USA) to understand if they can be used to our purposes or we need to simulate them.

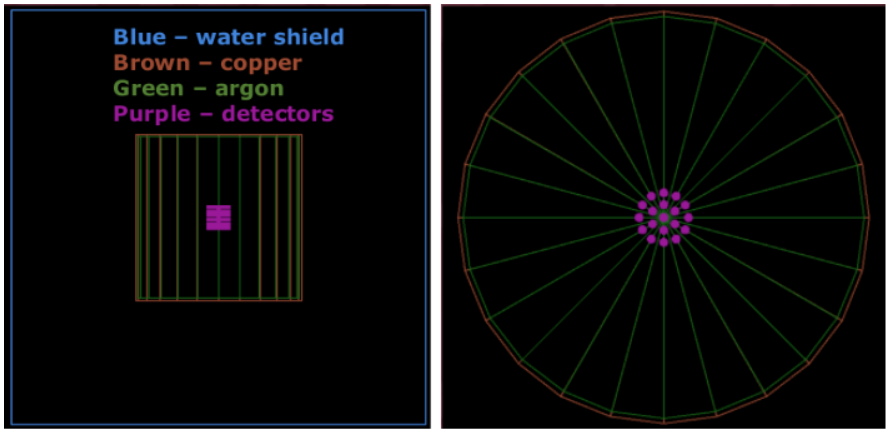

Figure 2.1: Side- and top-view of the detector geometry used in muon simulation. Figure generated by CJ Douglas Barton.

The simulated data are stored in ROOT files. ROOT is a C++ library diffused in the physical field and largely used for data analysis. However, the output format (i.e. data structures) adopted from the two researchers are largely different, and we adapt their format by implementing conversion scripts. We report some terminology to understand the format of the simulated data:

- A simulated event starts with a primary particle (e.g. muon, Ar39) placed at some point in the geometry volume.
- We define the state of the primary particle (i.e. kinetic energy, momentum, ...).
- The simulation proceeds in a discrete fashion and secondary events are generated according to the definition of physical processes involved and the probability that a specific process takes place (i.e. cross-section).
- We also define an output scheme, meaning a set of writing rules that define when and what the simulator has to write in the output file. The entries in the output file correspond to a snapshot of the simulation when it matches the output scheme's rules.

The output scheme that we adopt in this Monte Carlo simulation takes in consideration the energy deposited in liquid argon by the particles. We report a description of the recorded fields.

- Particle ID: particle identifier according to the Monte Carlo numbering scheme [Gar+00].
- Parent Track ID: track number of the parent particle.
- Energy deposition: the amount of energy (in KeV) deposited in liquid argon.
- Kinetic energy: the kinetic energy (in KeV) of the particle that deposited the energy.
- Time: the time of deposition (in nanoseconds) with respect to the event starting.
- x, y, z: the location (in millimeters) in which deposition occurred.
- Event number: unique marker used by G4 for each event in a simulation.
- Track number: unique marker used by G4 for each particle track in one event.
- Creator process: the name of the physical process that created the particle.
- Parent Nucleus PID: the PID of the parent particle, if the parent is a nucleus, otherwise is 0.

As already said, the simulation produces data at the level of energy deposition. However, not all the entries contribute to energy depositions and later scintillation. We will discuss the data cleaning in Section 2.2. Moreover, the current output is far from the SiPM readouts observable in the experiment. We will discuss this further step in Section 2.3.

2.1.2 Definition of the starting muons

Another significant point concerns the definition of primary events for muon simulation. In particular, the conditions of LNGS (e.g. mountain's valleys, rock material) strongly determine the angle and the kinetic energy of the muons that reach the experimental setup. Note that without considering these environmental conditions, the simulated data would result far from the real data because the energy deposited in liquid argon depends on the muon starting energy and the space covered by the propagation. The starting angle mainly determines this second factor.

To address this issue, we use the muon data produced with MUSUN [Kud09], a simulation code that reproduces the muon condition in several underground laboratories, among which the LNGS. The data generated are then adapted to the LEGEND200 geometry and converted in ROOT files to make them compatible with the simulation code in G4. The simulation code is finally changed to sample the starting condition of the muon from the MUSUN data uniformly.

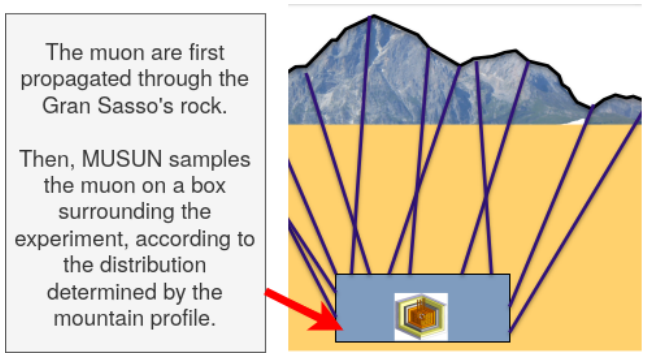

Figure 2.2: Production of starting muons according to the conditions of the LGNS through the MUSUN simulation code. Figure taken from [Kli].

2.1.3 Showering process

The last interesting point to discuss concerns the simulation of a particular process known as *showering* typical of the LNGS. The mountain rock that surrounds the experimental setup is the main shield against cosmogenic events and obviously determines the kind of process experienced in the experiment run (e.g. Ambiental radioactivity, showering). In particular, when a muon is propagated through the rock could produce a shower of high-charged particles (i.e. neutrons) that will propagate in the liquid argon, even if the muon propagates outside the volume. This phenomenum is different from direct muon propagation in liquid argon, but is an important secondary event that characterizes cosmogenic activity close to the experiment.

The simulation of the showering is then important for our purposes and we add 3 meters of rock material around the simulated volume to reproduce it. Figure 2.3 shows the passage of muon in the geometry extended with rock and the production of neutrons that propagates their energy in liquid argon.

2.2 Data cleaning

In this section, we discuss the data selection from the full-scale simulation produced. The two cuts performed are the product of analysis and time, and consider the following information:

1. Energy deposition: the amount of deposited energy is not always greater than zero because most of the simulation steps reports information on the propagation of particles. The first cut discards all the entries that have zero energy deposited.

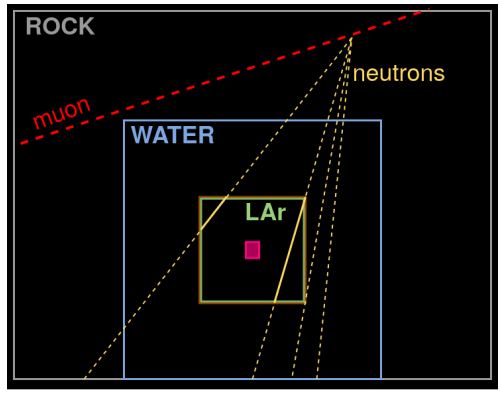

Figure 2.3: An example of showering from the interaction of muon particle with rock. Only the tracks propagated in liquid argon (solid lines) are potentially observable, the other (dotted lines) will be unobservable and hide cosmogenic events..

2. Location of deposition: the energy depositions that are observable through the SiPM sensors must be "close" to the sensors, as will be deeply explained later. The second cut discard all the entries that are out of the observable region.

Despite the first rough cut on zero energy depositions, the second deserves a deeper explaination.

2.2.1 Region of Interest

To understand the cut according to the location of deposition, we have to briefly explain how the detection happens.

The process of scintillation consists of production of optical photons (OP) by the liquid argon, that is a scintillator. The SiPMs are photomultiplier sensors that can detect OP and thanks to their high sensitivity are suitable to this low-energy setting. However, to lead the OPs that travel in liquid argon to the SiPM sensor, the LEGEND200 geometry uses two fibers shroud, illustrated in Figure 2.4. These fibers force the OP to run along them and lead the OPs towards the SiPM sensors.

As a consequence, the OP production is not uniformly observable in the liquid argon volume. Conversely, OP are observable through SiPM sensors if only if they reach the fibers.

Then, we define a Region of Interest (ROI) based on the observability of the OP produced by scintillation, that we remark is the product of energy deposition in liquid argon. The entries that are outside this ROI are discarded because their chance to be detected is zero.

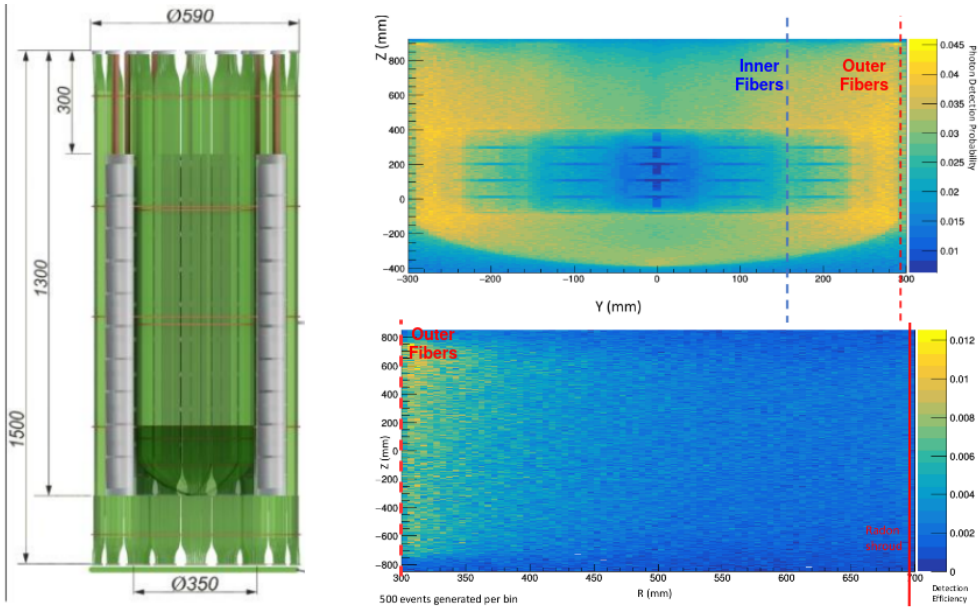

Figure 2.4: Left: The geometry of LEGEND200 with Germanium detectors (grey cylinders) surrounded by inner and external fibers shrouds (green) to detect scintillation. Right: Internal and External Optical Maps that describe the detection efficiency in each coordinate of the LAr volume. Both the images taken from [McF]..

To define the ROI, we based our reasoning on the Optical Map by Neil McFadden, described in Section 2.3. For now, it is enough to know that the Optical Map provides a simulation-based approximation of experimental detection efficiency.

Then, we defined the ROI as

$$
\begin{aligned}
x &\in [-500, +500] \\
y &\in [-500, +500] \\
z &\in [-1\,000, +1\,000]
\end{aligned}
$$

because the detection efficiency out of these box is negligible. As a consequence, any energy deposition that fall out of the ROI is discarded because unobservable with SiPM sensors. Figure 2.4 reports the Optical Maps adopted in the ROI definition.

2.3 Data preparation

In this section, we describe the work done to reproduce the experimental-data format (i.e. SiPM readouts) from the cleaned simulation data (i.e. step-level information). To complete the convertion we need to perform the following operations:

1. Reproduce the number of OP detected.
2. Reproduce where the detection occurs, meaning which is the SiPM activated.
3. Integrate the OP detection in time.

We aim to represent the data in a 2D matrix, as shown in Figure 2.5. To keep the procedure as flexible as possible, we defined the convertion based on the following parameters:

- N: Number of SiPM modules connected to fibers, that covers an equally-spaced area.
- ΔT: Time in which the OP detected are integrated (i.e. the sensor sums up the OPs detected).
- T: Time window, in case we can perform multiple consecutive integrations.

The integration of OP detection over time is performed by iterating over all the steps of the simulated event. However, the convertion from deposited energy to SiPM detection deserves a deeper treatment.

2.3.1 Mapping deposited energy to OP detected

The simulation of scintillation process and propagation of OP photons through fibers is extremely expensive with respect to computational time. For this reason, Neil

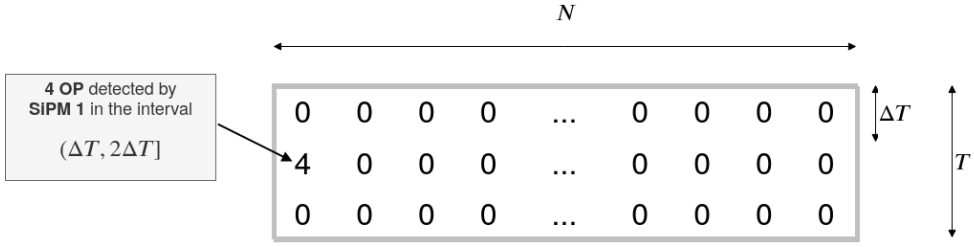

Figure 2.5: Data representation in 2D matrix representing the OP detections over a time window T in consecutive integrations ΔT, for each of the N SiPM modules..

McFadden produced a 3D Optical Map [McF] by running a massive simulation of OPs once and for all.

The Optical Map allows the users to retrieve the optical detection efficiency for each coordinate in the LAr volume. To obtain the detection efficiency, 500 OPs have been simulated for each $5mm \times 5mm \times 5mm$ bin of the LAr volume. Each OP has been propagated in liquid argon and through the fibers. From this simulation, the detection efficiency of the bin is approximated as the ratio between OPs detected and OPs simulated. The resulting Optical Map is illustrated in Figure 2.4.

We used this map to compute the number of OP detected. Given a deposited energy of $EKeV$ at coordinate (x, y, z), we compute the OP detected as

$$E \cdot LY \cdot DE(x, y, z)$$

Where $DE(\cdot)$ is the Optical Map and LY is the light-yield parameter that correspond to the OP production per KeV and depends on liquid argon purity.

2.3.2 Mapping deposited energy to SiPM readouts

The Optical Map is not enough for our purposes because it does not provide any information about the SiPM sensors that perform the OP detection.

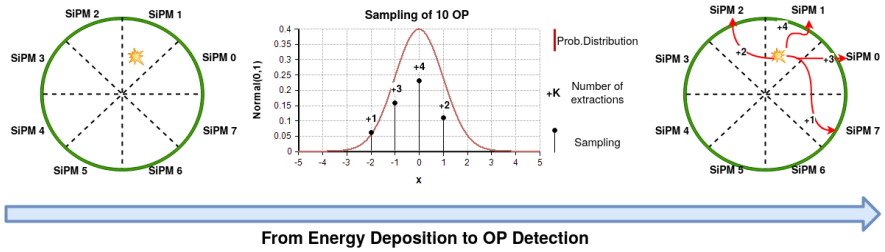

Figure 2.6: Workflow of convertion of energy deposition to OP detection with spatial mapping to SiPM sensors. In the figure, 10 OPs detected, computed according to the amount of energy deposited, are mapped to 8 SiPM sensors according to the deposition location (icon of explosion)..

To solve this issue, we proposed to reproduce the SiPM readouts by Gaussian sampling of the closest sensors from the location where the deposition occurs. The procedure is depicted in Figure 2.6 and can be described as follows:

1. We segment the $[0, 2\pi]$ range in N slices, where N is the number of SiPM sensors.

2. We assume that the depositions within the same slice present the same probability distribution concerning the SiPM activation.

3. We assume that the slice's probability distribution follows a Gaussian distribution centered in 0. We interpret the closest SiPM as 0, the successive SiPMs as positive integers (e.g. $+1$, $+2$) and the preceeding SiPM as negative integers (e.g. -1, -2).

4. Sampling values from a Gaussian distribution and rounding them to the closest integer will correspond to the displacement of activated SiPM with respect to the closest one. Since the distribution is centered in 0, the closest SiPM will have more chance to be activated, whilst the farther SiPM will have decreasing probability. It is consistent with our expectation and provides a first-level approximation of spatial informations.

CHAPTER 3

Data analysis

In this chapter, we briefly report some analyses performed on simulated data. With the following plots, we aim to describe the trend of the interaction of the particles in the liquid argon, in correspondence of the investigated processes. Moreover, we also aim to validate the distributions with the literature measures.

3.1 Ar41 de-excitation and gamma production

The passage of mouns in liquid argon leads to several processes that we simulated in these months. The subsequent deposition of energy is crucial for our purpose because it causes scintillation (i.e., emission of visible light that SiPM sensors can detect).

Among the various physics processes induced by the muon passage, there is a specific process characterized by a high-energy gamma (i.e., photons) emission. While the usual photons (primaries or optical photons) are in the order of a few KeV, the gammas produced in this event have energy in the order of MeV. Informally, what happens is the following:

1. The passage of muon particles in liquid argon releases neutrons (neutron spallation).

2. The free neutron interacts with the Ar40 particle, which becomes Ar41.

3. The Ar41 nucleus is in an excited state, and from its de-excitation, a bunch of gammas is produced.

The following signature formally describes the process:

$$\mu \rightarrow Ar40 + n \rightarrow Ar41^{*} \rightarrow Ar41 + gammas$$

Even if this process is not frequent, its peculiar gamma emission characterizes cosmogenic events and could help us in the classification task. To validate our expectation about the gamma's energy spectrum, we processed the simulation data to extract only the entries of this specific process. Figure 3.1 reports the obtained energy spectrum of gammas.

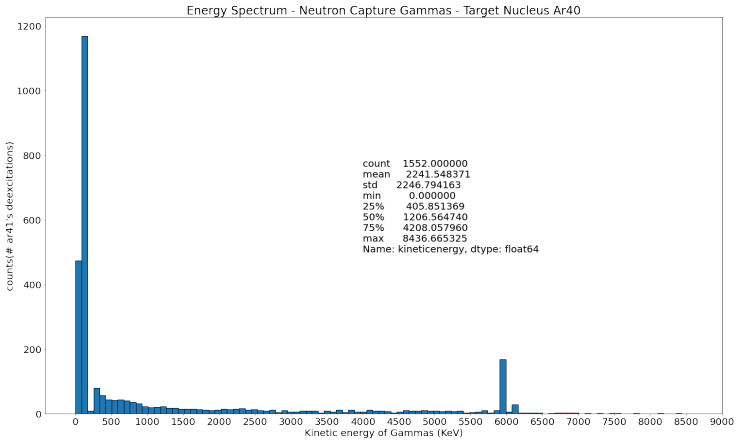

Figure 3.1: Energy spectrum of gammas produced by Ar41 de-excitation.

3.2 Starting muon's conditions with MUSUN

As already said in the previous chapters, we use data produced with MUSUN to reproduce the conditions of muons entering in Gran Sasso Laboratory.

The incoming rate and energy depend on the mountain rock and profile. Despite the natural shield of the mountain, the muons reach the Gran Sasso Laboratory with high energy in the order of hundreds of MeV or even GeV. The upper part of Figure 3.2 confirms this expectation, showing an average moun energy of about 250 MeV.

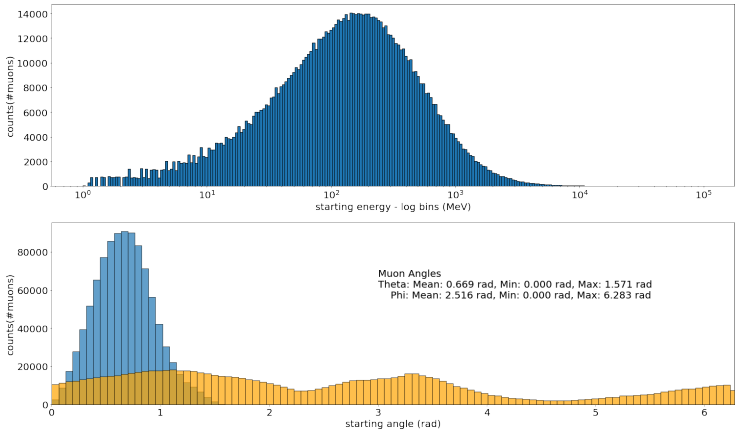

Figure 3.2: Energy spectrum and angles distribution of incoming muons, simulated through MUSUN.

Moreover, the mountain profile affects the angle distribution of incoming muons. In particular, the presence of valleys allows the mouns to pass in certain parts of the mountain more easily. For this reason, the resulting angle spectrum presents a precise shape, as depicted in the bottom part of Figure 3.2. The two angles θ and ϕ characterize the direction of moun entering in the simulated 3d volume. On the one hand, the angle ϕ shows a continuous distribution over the full range in radians (i.e., between 0 and 6.28). On the other hand, the angle θ is distributed only in the interval between 0 and 1.57 radians. This interval corresponds to orientations between 0 and 90 degrees, meaning that the mouns enter in the Laboratory from a specific side and following a precise direction.

3.3 Energy depositions in liquid argon

In this section, we aim to analyze the two main events that constitute the binary classification problem.

Considering cosmogenic events, they produce a large amount of energy deposited in liquid argon within a small time-interval and localized in part of the volume. Other events induced by Ar39 or Ambiental Radioactivity, deposit a small amount of energy but are frequent and isotropically spread in the volume.

Figure 3.3 shows the energy depositions induced by the passage of muons in liquid argon.

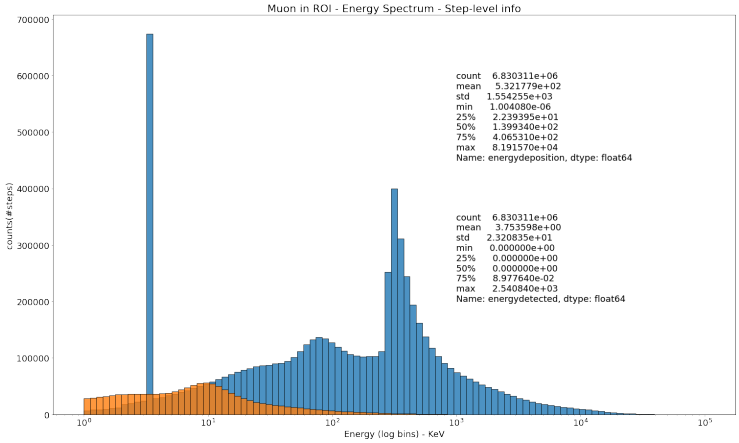

Figure 3.3: Spectrum of step-level depositions in liquid argon induced by the passage of muons. The figure illustrates the energy deposits (in blue) and the detected energy (in orange) using the Optical Map in the inner region.

The average energy deposited in a single simulation step is about 500 KeV. This

value is not small, considering that it is only one deposition, and other events, such as Ar39 decays, produce Beta particles with an energy spectrum up to 560 KeV, as shown in Figure 3.4. Then, summing up all the energy deposited by muons-induced particles, we obtain a large amount of deposited energy in most of the case.

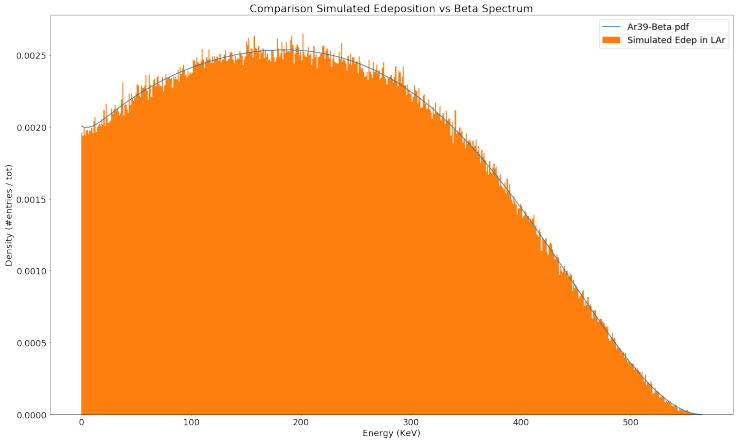

Figure 3.4: Energy spectrum of betas produced by Ar39 radioactive decay. The figure compares the simulated spectrum with the expected one from literature (blue line)..

However, this is not enough for our purposes because we are also interested in the useful classification of pathological cases, where the cut on energy is not enough.

Always in Figure 3.3, we report the distribution of energy detected by SiPM sensors, obtained using the Optical Map in the inner region of the volume. The detected energy depends on the location of deposition and is subject to propagation through fibers and other technical details of the electronics adopted.

Considering the second event of Ar39 decays, they release electrons (i.e., beta particles) with a low-energy spectrum, reported in Figure 3.4. The Ar39 half-life is 249 years, then the complete Ar39 lifetime is long, and its activity decreases over time following an exponential shape, as depicted in Figure 3.5.

However, the Ar39 activity in natural argon is intense, measured at 1.41 Bq/l in the literature. Considering an instrumented liquid-argon volume of $2.6m^3$, equivalent to 2 600 liters, the Ar39 activity in the LEGEND200 geometry is about 3 666 decays per second. Then, considering the limits of integration time due to the electronics, in the order of tens of microseconds, the cumulative energy deposited by thousands of beta particles makes challenging the classification of cosmogenic events.

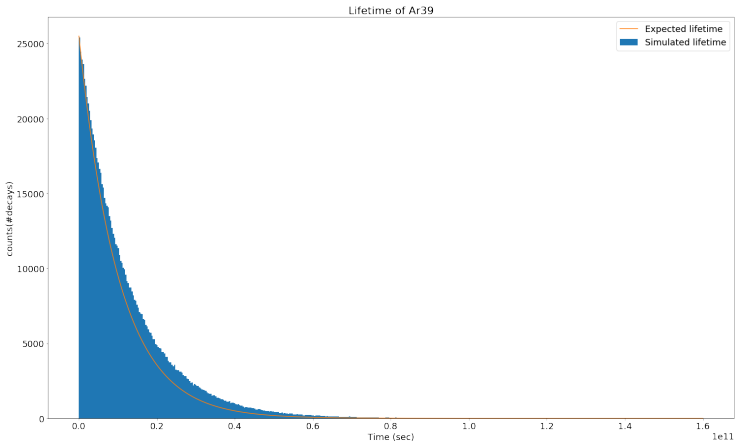

Figure 3.5: The lifetime of Ar39 particles exponentially decreases. The figure compares the simulated lifetime, described by time of beta emission, with the expected lifetime decay from literature (blue line).