ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

IPA

Simon Wittum
23-941-370

# Discrimination of $2\gamma$ and $3\gamma$ Events in Positronium Annihilations Using Machine Learning

**Semester Thesis**

MSc High Energy Physics
Exotic Matter Group
Institute for Particle Physics and Astrophysics
Swiss Federal Institute of Technology (ETH) Zurich

**Supervision**

Benjamin Banto Oberhauser
Prof. Dr. Paolo Crivelli
May 31, 2024

# Abstract

Currently, the experimental uncertainty in determining the o-Ps decay rate is at 150 ppm precision, which is two orders of magnitude lower than the theoretical precision of 1 ppm. A new method proposed in [1] aims to achieve an ultimate precision of 1 ppm, enabling a direct examination of higher-order QED corrections. A deviation between theoretical and experimental results would indicate a potential hidden sector in the standard model, which could be a candidate for dark matter.

To achieve this level of precision, it is essential to differentiate between $2\gamma$ and $3\gamma$ events in the detector, thereby eliminating the time-dependent pick-off annihilation rate that distorts the observed o-Ps decay rate. This thesis presents a discrimination model based on a cut-based pre-selection and a machine learning algorithm for fine-tuning the $2\gamma$ and $3\gamma$ event separation. The model has been trained and tested using data generated with the Geant4 simulation framework. Finally, an outlook on how to apply the model to real data is provided.

# Contents

# Chapter 1

# Introduction

The bound state between an electron $e^-$ and a positron $e^+$ is an exotic atom known as Positronium (Ps). Unlike the hydrogen atom, Ps is a purely leptonic system and therefore does not undergo any influence of the strong force described by Quantum Chromodynamics (QCD). It is therefore an ideal system to perform high precision tests of Quantum Electrodynamics (QED) and to search for new physics beyond the Standard Model (SM).

The Ps atom can exist in two different spin states, the singlet state $^1S_0$ where the spins of the electron and positron are anti-parallel (p-Ps) and the triplet state $^3S_1$ where the spins are parallel (o-Ps). Due to conservation of charge conjugation and parity in QED, the o-Ps state decays into three photons, while the p-Ps state decays into two photons. The decay time of the o-Ps state is therefore suppressed compared to the p-Ps state.

The theoretical prediction for the decay width of the o-Ps state is up to $\mathcal{O}(\alpha^2)$ given by $\lambda_{\text{o-Ps}} = 7.039\,934(10)\,\mu\text{s}^{-1}$ [2]. The most accurate experimental measurements are in accordance with this prediction and deliver a value of $\lambda_{\text{o-Ps}} = 7.0404(10)\,\mu\text{s}^{-1}$ [3]. However, the theoretical prediction for the decay width of the p-Ps state is two orders of magnitude more accurate. In order to test QED to the next order, it would thus be beneficial to have more accurate measurements.

One of the main challenges in the experimental determination of the o-Ps decay width are pick-off annihilations of the produced Ps atoms with electrons of the target material or the vacuum cavity walls. Due to these pick-off annihilations, the observed decay rate of o-Ps is modified to

$$\lambda_{\text{obs}}(t) = \lambda_{\text{oPs}} + \lambda_{\text{pick}}(t), \tag{1.1}$$

with $\lambda_{\text{pick}}(t)$ being proportional to the rate of o-Ps collisions with the target material or

the vacuum cavity

$$\lambda_{\text{pick}}(t) = n\sigma_a v(t), \tag{1.2}$$

where $n$ is the atomic density of the material, $\sigma_a$ the annihilation cross section, and $v(t)$ the time dependent velocity of o-Ps. In order to extract the intrinsic o-Ps decay rate, it is thus crucial to precisely account for the contribution of the pick-off annihilation rate.

Past o-Ps lifetime measurements conducted in Ann Arbor consisted of an $e^+$ beam hitting a target material with a variable density known to have a high Ps production rate. By varying the target density and the beam intensity in a controlled manner, it has been achieved to extract the intrinsic o-Ps decay rate in vacuum by extrapolating the results, which however also forms the main source for the systematic uncertainty of 100 ppm [3].

A different approach to measure the o-Ps decay rate without the need of extrapolation has been employed by a research group in Tokyo. The group uses a $^{22}$Na source undergoing a $\beta^+$ decay to produce positrons surrounded by a silicon-dioxide ($SiO_2$) powder. The time and energy of the o-Ps annihilation photons are measured simultaneously with germanium detectors. By subtracting a $3\gamma$ spectrum obtained with Monte Carlo simulations from the experimental data, they were able to isolate the $2\gamma$ spectrum and determine $\lambda_{\text{pick}}(t)/\lambda_{3\gamma}$ as a function of time. After this step, the cleansed data has been fitted to determine the vacuum decay rate of o-Ps. The normalization procedure of the $3\gamma$ spectrum gives the biggest contribution to the systematic uncertainty of 90 ppm [4].

A new way of measuring the decay rate of o-Ps that could eventually reach a precision of 1 ppm has been proposed by [1] in 2018. The idea is to use a high intensity $e^+$ beam to form o-Ps with a porous $SiO_2$ film and confine it in a vacuum cavity surrounded by a nearly hermetic, granular calorimeter to detect the annihilation photons. The cavity is additionally sealed with a thin membrane transmissive to $e^+$, but not to o-Ps, therefore confining the o-Ps in the cavity. Due to the granularity of the calorimeter, it is possible to discriminate $2\gamma$ and $3\gamma$ events, making it possible to more accurately isolate the intrinsic o-Ps decay rate.

# Chapter 2

# Monte Carlo (MC) Simulations in Geant4

Before the experimental results are being discussed, let us take some time to take a look at a simulation within the Geant4 framework. The procedure will be as follows: First, the geometry and implementation of the "Super Duper Mega Detector" (SDMD) in Geant4 will be introduced. The simulation allows to generate data for $2\gamma$ and $3\gamma$ events independently to an arbitrary amount. For each simulated event, it is possible to store the energy detected in each LYSO crystal of the SDMD. The simulation can be made more realistic by adding additional complexities such as a Gaussian energy smearing to resemble the limited energy resolution of the crystals.

The ultimate goal of the simulation is to discriminate between $2\gamma$ and $3\gamma$ events to later be able to identify and subtract the pick-off annihilations from the o-Ps decay spectrum. This makes it possible to isolate the intrinsic o-Ps decay rate and reduce the systematic uncertainty of the o-Ps lifetime measurement.

## 2.1 Geant4 Simulation of the SDMD

A simplified version of the SDMD being used at ETH Zurich has been implemented in Geant4 by Benjamin Banto Oberhauser. A drawing of the detector is shown in figure 2.1. It consists of two modules A and B, attached to the left and right side of the $SiO_2$ target. Each module consists of two detector units, which form an 8 by 8 matrix of LYSO crystals. The following parameters have been chosen for the geometry of the detector:

- Distance between the two modules: 100 mm

- Distance between the two units: 2.8 mm
- Geometry of LYSO crystals: $x$- and $y$-direction: 2.5 mm, $z$-direction: 30 mm.

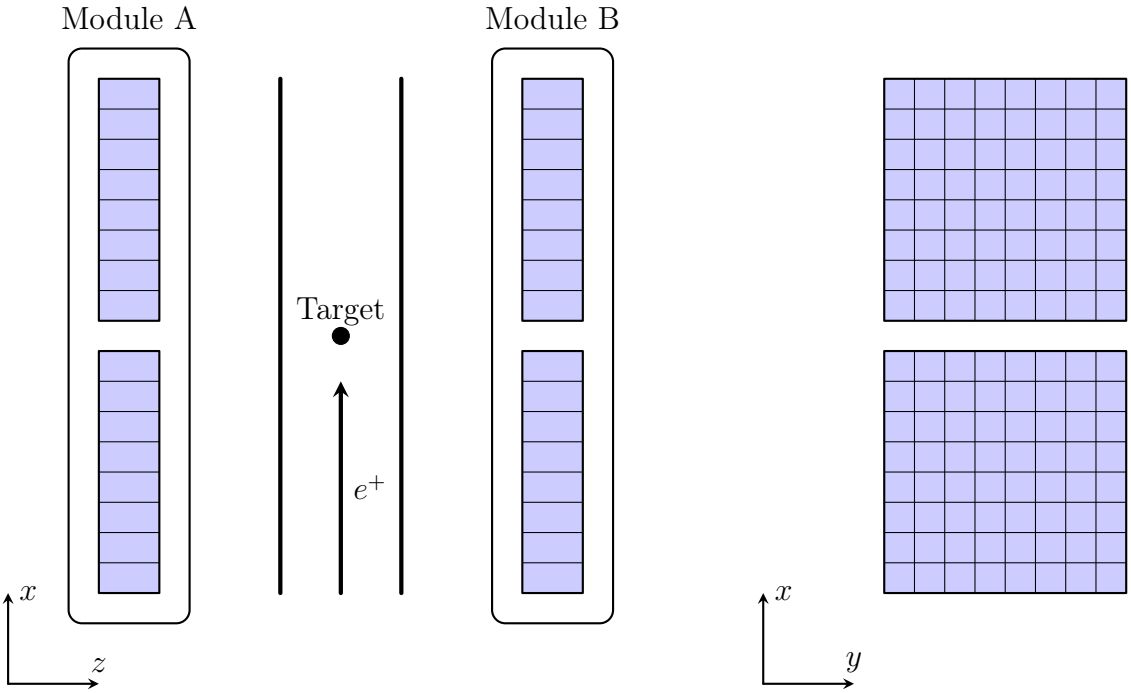


Figure 2.1: Drawing of the SDMD detector. Left: $x-z$ plane view, right: $x-y$ plane view. The blue areas represent the LYSO crystals. The $e^+$ beam emitted by a radioactive $^{22}$Na source undergoing $\beta^+$-decay enters from the $-x$-direction and impinges on an $SiO_2$ target to produce Ps at high rates.

## 2.2 Results of the Geant4 Simulation

The Geant4 simulation has been run for $10^7$ $2\gamma$ and $3\gamma$ events respectively, but most of them are not measured in the detector due to small solid angle coverage. In the end, more than twice as many $3\gamma$ events are captured in the SDMD, implying a signal-to-background ratio of less than one half.

As a first step, the total energy depositions in the two modules are computed and shown in a two dimensional histogram in figure 2.2. The histogram for the $2\gamma$ case is shown on the left and for the $3\gamma$ case on the right. As expected, for most of the $2\gamma$ events, the energy deposition in the modules is either 0 keV or 511 keV, visualized by the three yellow vertices. On the other hand, due to the increased phase space of the $3\gamma$ case, the energy distribution is more spread out. Still, the probability for an energy deposition of 511 keV

in both detectors is not zero as can be seen by the green vertex. This is due to the fact that the angle between the directions of propagation of two out of the three photons might be very small, causing them to be detected in the same module. On the other hand, it can also happen that one of the three photons only has a small energy, while the other two carry an energy around 511 keV.

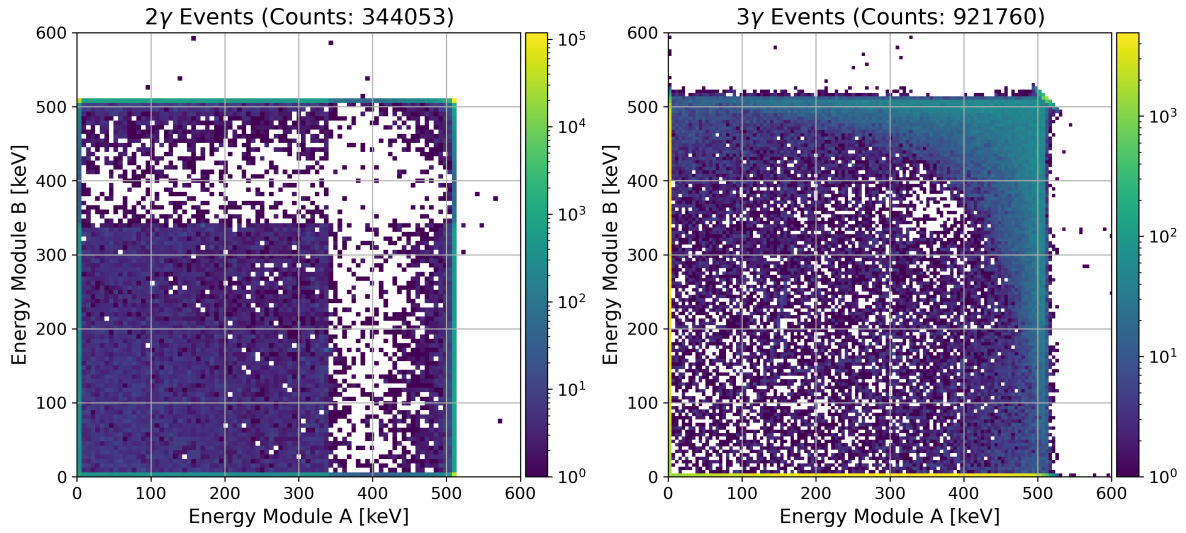


Figure 2.2: 2D histogram of the total energy deposited in the two modules of the SDMD (cf. fig. 2.1) for $2\gamma$ (left) and $3\gamma$ (right) events. Although a total of $10^7$ events were simulated for both types, more than twice as many $3\gamma$ as $2\gamma$ events are detected in the SDMD due to the increased angular coverage of the $3\gamma$ events.

The same procedure has been repeated for the more realistic case of a smeared energy deposition which in the experiment is given due to the limited energy resolution of the LYSO crystals. In this case, a Gaussian energy smearing with an energy dependent standard deviation of[1]

$$\sigma(E) = \sigma(E_0)\sqrt{\frac{E}{E_0}} \tag{2.1}$$

has been chosen. In practice, this means that the results from the simulations run with no energy smearing are taken, and every energy $E$ measured in the LYSO crystals is replaced by a new energy $E'$ drawn from a Gaussian distribution with mean $E$ and standard deviation $\sigma(E)$. According to the datasheet, the LYSO crystals used in the SDMD have a relative energy resolution of $\frac{\mathrm{FWHM}(E_0)}{E_0} = 8\,\%$ for $E_0 = 662\,\mathrm{keV}$, which will be used in the following. FWHM stands for full width at half maximum and it can easily

[1]The proportionality to the square root of the energy is a result from photo statistics and is linked to the central limit theorem.

be verified that it is related to the standard deviation $\sigma$ by

$$\sigma = \frac{\mathrm{FWHM}}{2\sqrt{2\ln(2)}}. \tag{2.2}$$

In addition to the smearing, a coincidence condition has been applied to the data, i.e. only events where the energy deposition in both modules is non-zero are being tracked. The results are shown in figure 2.3. Looking at the amount of entries in the two histograms, it can be seen that the number of $2\gamma$ events is now three times as large as the number of $3\gamma$ events. This means that the signal-to-background ratio can be improved significantly by only considering events satisfying the coincidence condition. Of course, this procedure is not perfect as it also discards some of the $2\gamma$ events, but it is a step in the right direction.

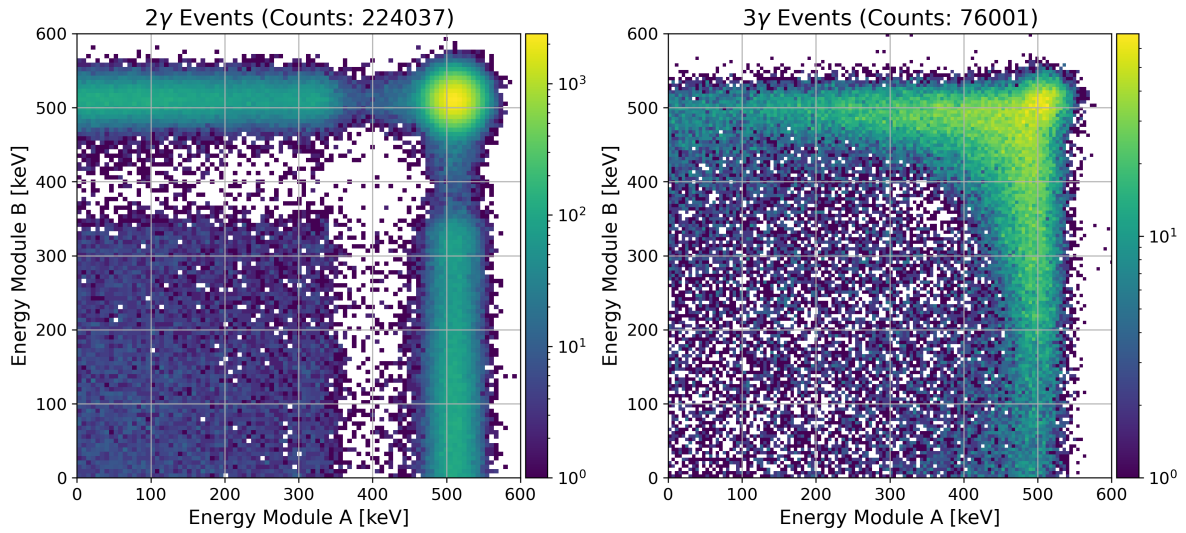


Figure 2.3: 2D histogram of the total energy deposited in the two modules of the SDMD (cf. fig. 2.1) for $2\gamma$ (left) and $3\gamma$ (right) events with Gaussian energy smearing of 8 %. In both cases, a coincidence condition has been imposed, i.e. only events where the energy deposition in both modules is non-zero are being considered.

## 2.3 Tuning of the Signal and Background Efficiency

In the last section, it has been shown that the signal and background efficiencies can be improved by only considering events depositing energies in both modules. This can be pushed further by requiring that the energy deposition in both modules has to not only be non-zero, but lie above a threshold $E_{\mathrm{module}}$. Furthermore, before checking for the

coincidence condition, energy detections of less than $E_{\text{crystal}}$ measured in a single crystal can be discarded. The results obtained with different thresholds are shown in table 2.3. The $2\gamma$ and $3\gamma$ efficiencies are defined as the fraction of $2\gamma$ and $3\gamma$ events that are not filtered out by the cut. It can be seen that the signal and background efficiencies strongly depend on the choice of the thresholds $E_{\text{crystal}}$ and $E_{\text{cut}}$. Which values are chosen in the end depends on the experimental conditions and the desired signal-to-background ratio. In the next section, the data after the cuts have been applied will be used to train different machine learning models to improve the quality of the discrimination even further.

| $E_{\text{crystal}}$ [keV] | $E_{\text{cut}}$ [keV] | $2\gamma$ Eff. [%] | $3\gamma$ Eff. [%] |
|---|---|---|---|
| 1 | 1 | 40.10 | 6.76 |
| 100 | 400 | 22.66 | 1.89 |
| 100 | 450 | 19.87 | 0.86 |
| 100 | 480 | 16.73 | 0.39 |
| 400 | 400 | 9.32 | 0.74 |
| 450 | 450 | 8.01 | 0.26 |
| 480 | 480 | 6.90 | 0.07 |

Table 2.1: Optimizing the signal and background efficiencies using a cut-based approach. All energies below $E_{\text{crystal}}$ detected in a LYSO crystal are discarded. Afterwards, the crystals of the two modules are integrated over separately and only events with a total energy deposition above $E_{\text{cut}}$ in both modules are being considered. The $2\gamma$ and $3\gamma$ efficiencies are defined as the fraction of $2\gamma$ and $3\gamma$ events that survive the cut.

# Chapter 3

# Multivariate Analysis (MVA)

In the last chapter, the response of the SDMD to $2\gamma$ and $3\gamma$ events has been simulated. Furthermore, the influence of different energy cuts on the signal and background efficiencies has been investigated. In this chapter, the efficiencies will be improved even further by using machine learning (ML) algorithms to discriminate between $2\gamma$ and $3\gamma$ events based on the data after the cut. The plan is to use the cut-based approach to do a first pre-selection and then make use of more sophisticated tools to perform the fine-tuning. For this purpose the performance of three different ML models will be compared:

- Gradient Boosted Decision Tree (GBDT)
- Ada Boosted Decision Tree (ABDT)
- Multi Layer Perceptron (MLP).

In the following, the models and evaluation metrics will briefly be introduced. We will then investigate how the combined model of cut and ML performs on the dataset and how different cut parameters influence the signal and background efficiencies.

## 3.1 Models and Metrics

Before beginning with the analysis, the models and evaluation metrics have to be reviewed. For all of the models at use, a demonstration code is provided in this GitLab repository.

### 3.1.1 Decision Tree (DT)

A DT is an ML algorithm commonly used for classification tasks. It works by recursively splitting the dataset into subsets depending on the value of a certain feature $E_i$. If the value of $E_i$ is below a given threshold $E_{\text{cut},i}$, the event is assigned to the left child node, otherwise to the right child node. This process is repeated until a stopping criterion is met, e.g. a maximum depth of the tree. At each node, it is possible to determine the optimal feature $E_i$ and threshold $E_{\text{cut},i}$ by minimizing a loss function. The leaf nodes of the tree are assigned a class label, which is the majority class of the events that have been sorted into the leaf node during the training. A sketch of a DT is shown in figure 3.1.

The biggest problem of DTs is that they tend to overfit the training data. This can easily be seen by imagining a DT with infinite depth and the stopping criterion being the presence of only one data point in each leaf node. In this case, the DT would perfectly classify the training data, but fail to generalize to unseen data. A common way to counteract this is to use ensemble methods, which combine multiple DTs to form a more robust model. One such method is boosting, which trains multiple shallow DTs that have a poor performance on their own in a sequential manner, where each new DT is trained on the mistakes of its predecessors. In our case, we will use two different boosting methods, AdaBoost (Adaptive Boosting) and Gradient Boosting. AdaBoost adjusts the weights of misclassified instances to focus subsequent learners on harder cases, while Gradient Boosting sequentially optimizes a loss function by training new learners on the residuals of previous learners' predictions.

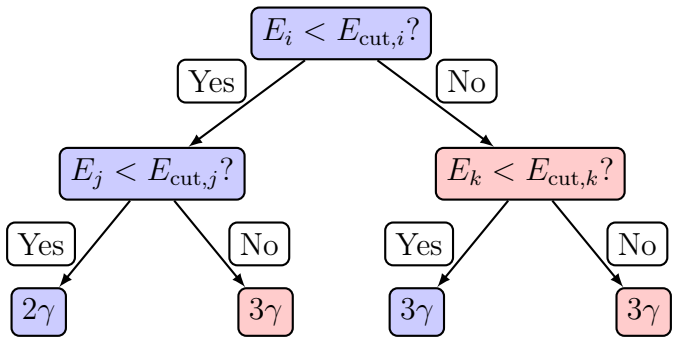


Figure 3.1: Sketch of a simple decision tree (DT). At each node, the data is being split depending on the value of a feature $E_i$. If the value of $E_i$ is below a threshold $E_{\text{cut},i}$, the event is assigned to the left, otherwise to the right child node. At each leaf node, an optimal feature and an optimal threshold can be found by minimizing an appropriate loss function. The leaf nodes are assigned a class label, which is the majority class of events that have been sorted into the leaf node in the training process.

### 3.1.2 Multi Layer Perceptron (MLP)

An MLP is a specific type of an artificial neural network. It takes in an input vector $\mathbf{x}$ consisting of $n$ features, which are being multiplied by a weight matrix $\mathbf{W}^{(1)}$ and added to a bias vector $\mathbf{b}^{(1)}$. The result is then passed through and activation function $\sigma$ to form the output of the first layer $\mathbf{a}^{(1)}$, making the process non-linear. This is repeated for $L$ layers, where the output of the $l$-th layer is the input of the $(l+1)$-th layer. The output of the last layer is then used to make a prediction. In the training process, the weights and biases are adjusted to minimize an appropriate loss function. The backpropagation algorithm is used to calculate the gradients of the loss function with respect to the weights and biases, which are then used to update the weights and biases. A sketch of an MLP is shown in Figure 3.1.2. In the drawing, we have 3 input features, 2 hidden layers with 5 neurons each and 1 output neuron. In our case, we will use 256 input features, 2 hidden layers with 100 neurons each and 1 output neuron. As an activation function, we will use the rectified linear unit (ReLU) function for the hidden layers

$$\mathrm{ReLU}(x) = \max(0, x), \tag{3.1}$$

and the sigmoid function for the output layer given by

$$\sigma(x) = \frac{1}{1 + \exp(-x)}. \tag{3.2}$$

The action of both functions on a given input vector is defined component-wise.

### 3.1.3 Receiver Operator Characteristic (ROC) Curve

A good model should be able to accept many signal events while accepting only few background events. The signal acceptance rate is usually referred to as the true positive rate (TPR), while the background acceptance rate is called the false positive rate (FPR). At the end of the classification, all models predict a value between 0 and 1 for each event. We can then choose a threshold between 0 and 1, where all events with a predicted value above the threshold are classified as signal and all other events as background. The lower the threshold, the higher the probability to correctly classify an actual signal event as a signal event, but also to confuse an actual background event with a signal event. On the other hand, the higher the threshold, the lower the probability to confuse an actual background event with a signal event, but also to correctly classify an actual signal event as a signal event.

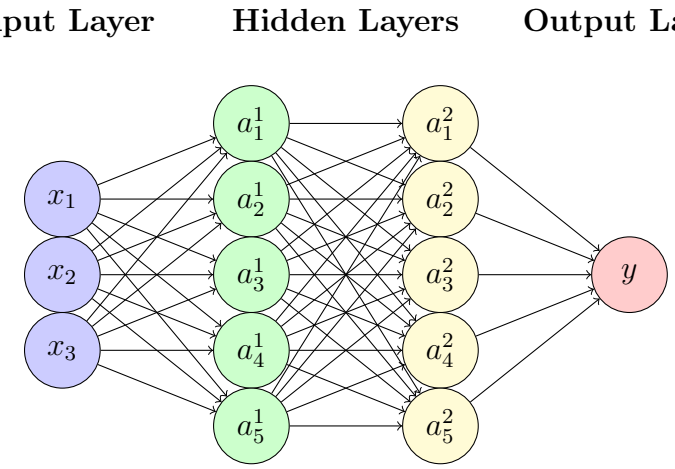


Figure 3.2: Sketch of a multi layer perceptron (MLP). The input vector $\mathbf{x}$ is multiplied by a weight matrix $\mathbf{W}^{(1)}$ (indicated by the arrows) and added to a bias vector $\mathbf{b}^{(1)}$ (not shown). The result is then passed through an activation function $\sigma$ (not shown) to form the output of the first layer $\mathbf{a}^{(1)}$. This process is repeated for $L$ layers ($L = 2$ in the picture), where the output of the $l$-th layer is the input of the $(l+1)$-th layer. The output of the last layer is then used to make a prediction. In the training process, the weights and biases are adjusted to minimize an appropriate loss function.

The ROC curve graphically illustrates this trade-off by plotting the TPR as a function of the FPR. For a random classifier, the TPR and FPR are equal, resulting in a linear ROC curve with a slope of 1. In contrast, a perfect classifier has a strong separation between signal and background events and therefore has a TPR of 1 as soon as the FPR becomes non-zero. The area under the ROC curve (AUC) is a common metric to quantify the performance of a binary classifier. A perfect classifier has an AUC score of 1, while a random classifier has an AUC score of 0.5.

The AUC score will be one of the metrics used to quantify the performance of the models. In the end, we are interested in having a high TPR and a low FPR. The signal-to-background ratio of the model can then be computed by

$$\eta = \frac{\mathrm{TPR}}{\mathrm{FPR}} \frac{n_{2\gamma}}{n_{3\gamma}}, \tag{3.3}$$

where $n_{2\gamma}$ and $n_{3\gamma}$ are the number of $2\gamma$ and $3\gamma$ events within the raw data without any processing.

## 3.2 Data Preprocessing

In order to support the ML models with their task, the data needs to be preprocessed in an appropriate way. The dataset consists of 256 input features which are the energy deposits in the 128 LYSO crystals of the two modules. A very simple preprocessing strategy has been applied in the cut-based approach, where the 256 input features have been reduced to 2 by integrating over the energy deposits in the crystals of each module. Unfortunately, this method erases most of the geometric information of the detector, which contains valuable information due to the different phase space distributions of the two events. The algorithms are expected to extract to work based on the geometric information hidden in the data, which is why a different preprocessing strategy has to be applied. One option is to apply a principal component analysis (PCA). In short, PCA exploits the fact that the covariance matrix of the dataset is symmetric and therefore possesses an orthogonal basis of eigenvectors covering the entire feature space. Switching from the canonical basis (the energy deposits in the crystals) to the eigenbasis of the covariance matrix corresponds to a re-arrangement of the features. The absolute values of the eigenvalues measure how relevant the corresponding eigenvector is for the dataset[1]. To reduce the dimensionality of the data, one could simply discard the eigenvectors with smaller eigenvalues. In this particular case however, all of the eigenvectors will be kept. After the PCA, the data is normalized to have zero mean and unit variance. This is important because the ML algorithms at use are sensitive to the scale of the input features and have been optimized for standardized data.

Moreover, it should be noted that as has been shown in figure 2.2, the raw data output of the simulation contains more than double the number of $3\gamma$ than $2\gamma$ events. This imbalance in the dataset can lead to a bias in the ML algorithms. To counteract this, all of the surplus $3\gamma$ events have been removed from the dataset such that it contains the same amount of $2\gamma$ and $3\gamma$ events. In the future, it would also be interesting to create more data using data augmentation techniques, such as adding noise to the data or rotating the dataset. For the scope of this thesis, however, this has not been done.

## 3.3 Model Training

It is now time to train the three models introduced in section 3.1 on the preprocessed dataset. In order to get optimal results, it is beneficial to feed as much data as possible into

[1] Actually, this is only true if the values of the features have a mean of zero. Therefore, the data is set to be zero-centered in the process of the PCA, and then normalized to have unit variance.

the training process. However, the data generation and preprocessing is computationally expensive, making it necessary to perform a trade-off between the amount of data and the computational resources available. In figure 3.3, it is shown how the AUC score and the accuracy[2] of the models change with an increasing number of training samples. For both the training and the evaluation, a balanced dataset with the same number of $2\gamma$ and $3\gamma$ events and a relative energy smearing of 8 has been used. Each evaluation has been performed with a balanced dataset containing $2 \cdot 10^3$ events. For each model and for each number of the sample size, the evaluation has been repeated 5 times. The mean has been taken as the final result and the error bars correspond to the standard deviation. As can be seen, the GBDT model converges at a sample size of $10^4$, while the ABDT seems to need even less data. The MLP on the other hand does not quite converge, but the rate of improvement decreases around a sample size of $10^4$. Therefore, for future training processes, a balanced dataset containing $10^4$ events will be used. It should however be kept in mind that better results could in principle be achieved for the MLP by increasing the sample size even further.

## 3.4 Model Evaluation

It has been shown that a balanced training dataset containing $10^4$ events is enough for the models to get sufficiently good results. To construct a discriminator, the three ML models have been trained based on the output of the cut-based approach for different cut parameters $E_{\text{crystal}}$ and $E_{\text{module}}$, using the same values as in 2.3. The evaluations have been performed using a balanced dataset with $2 \cdot 10^3$ events. The results are shown in table 3.1. As can be seen, the GBDT model is consistently outperforming both the ABDT and the MLP model as it achieves a higher $2\gamma$ efficiency and a lower $3\gamma$ efficiency for all cut parameters. Considering that the training time of the GBDT model is also significantly shorter than the training times of the ABDT and MLP model, the GBDT is the best choice amongst the tested models for the discrimination of $2\gamma$ and $3\gamma$ events

For each ML model, a discrimination threshold of 0.5 has been chosen. However, the signal efficiency and the background rejection rate can be changed by adjusting this threshold and depending on the experimental conditions, a different threshold might be more suitable. An illustration of this is shown in figure 3.4 using the GBDT model being trained on data filtered with $E_{\text{crystal}} = 480\,\text{keV}$ and $E_{\text{module}} = 480\,\text{keV}$ (see the last entry for the GBDT model in table 3.1). On the left, the $2\gamma$ and $3\gamma$ efficiencies are shown as a

[2]The accuracy is defined as the number of correctly classified events divided by the total amount of events

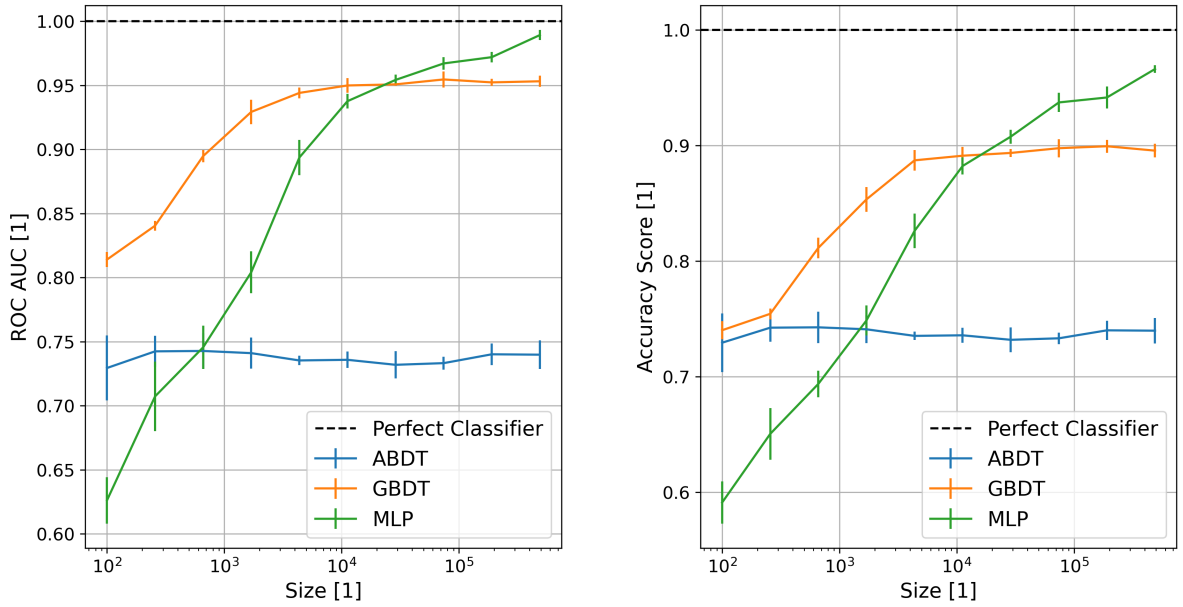


Figure 3.3: Investigating the performance of the models as a function of the number of training samples using the AUC score (left) and the accuracy (right) of the models as metrics. The training has been performed with a balanced dataset containing the same number of $2\gamma$ and $3\gamma$ events and a relative energy smearing of 8 %. Each evaluation has been performed with a balanced dataset containing $2 \cdot 10^3$ events. For each model and for each number of the sample size, the evaluation has been repeated 5 times. The mean has been taken as the final result and the error bars correspond to the standard deviation.

| $E_{\text{crystal}}$ [keV] | $E_{\text{module}}$ [keV] | $2\gamma$ Eff. [%] | $3\gamma$ Eff. | $2\gamma$ Eff. [%] | $3\gamma$ Eff. |
|---|---|---|---|---|---|
| | | **Cut Only** | | **Cut & GBDT** | |
| 1 | 1 | 40.10 | 6.76 | 32.40 | 1.22 |
| 100 | 400 | 22.66 | 1.89 | 18.56 | 0.42 |
| 100 | 450 | 19.87 | 0.86 | 15.88 | 0.25 |
| 100 | 480 | 16.73 | 0.39 | 12.36 | 0.11 |
| 400 | 400 | 9.32 | 0.74 | 7.97 | 0.11 |
| 450 | 450 | 8.01 | 0.26 | 6.85 | 0.05 |
| 480 | 480 | 6.90 | 0.07 | 5.53 | 0.02 |
| | | **Cut & ABDT** | | **Cut & MLP** | |
| 1 | 1 | 30.87 | 1.42 | 31.52 | 1.64 |
| 100 | 400 | 18.45 | 0.44 | 17.63 | 0.46 |
| 100 | 450 | 14.84 | 0.26 | 14.31 | 0.25 |
| 100 | 480 | 11.44 | 0.12 | 11.69 | 0.13 |
| 400 | 400 | 7.90 | 0.14 | 7.76 | 0.16 |
| 450 | 450 | 6.75 | 0.05 | 6.42 | 0.06 |
| 480 | 480 | 5.36 | 0.02 | 5.00 | 0.02 |

Table 3.1: Results of the model evaluation. We first apply the cut-based approach with different cut parameters $E_{\text{crystal}}$ and $E_{\text{module}}$. After the cut, a classification is performed by one of the three models. The signal efficiencies are computed by dividing the number of correctly classified $2\gamma$ and $3\gamma$ events by the total number of $2\gamma$ and $3\gamma$ events before applying a cut (corresponding to TPR and FPR).

function of the threshold. On the right, the ROC curve is shown, which is visualizing the possible combinations of the simultaneously achievable signal acceptance and background rejection rates by plotting the $2\gamma$ efficiency as a function of the $3\gamma$ efficiency. The red line in the left plot and the red cross in the right plot indicate the results for a discrimination threshold of 0.5. The efficiencies only refer to the classification by the GBDT. In order to get the efficiencies for the full discriminator the rates need to be multiplied by the last entries of the "cut only" column in table 3.1.

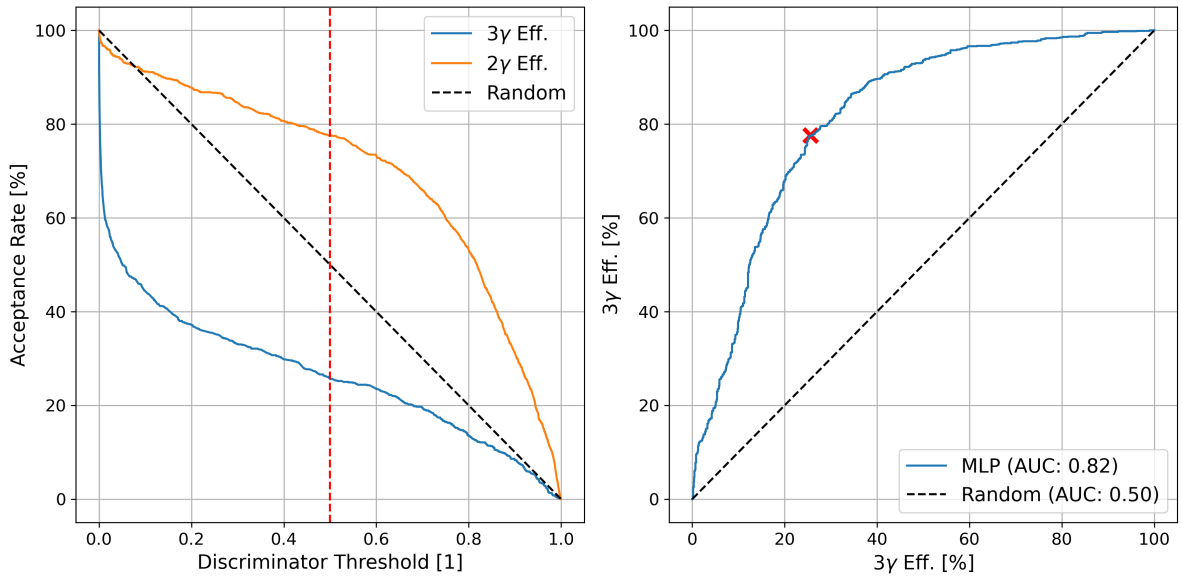


Figure 3.4: Adapting the $2\gamma$ and $3\gamma$ efficiencies by changing the discrimination threshold for the GBDT model. The model is trained on data filtered with $E_{\text{crystal}} = 480\,\text{keV}$ and $E_{\text{module}} = 480\,\text{keV}$. The left plot shows the $2\gamma$ and $3\gamma$ efficiencies as a function of the threshold. The right plot displays the ROC curve, visualizing the possible combinations of simultaneously achievable signal acceptance and background rejection rates. The red line and the red cross indicate the results for a discrimination threshold of 0.5. The efficiencies shown only refer to the classification by the GBDT. To get the efficiencies for the combined cut and GBDT model, the rates need to be multiplied by the last entries of the "cut only" column in table 3.1.

# Chapter 4

# Measurements

So far, the SDMD response to $2\gamma$ and $3\gamma$ events has been simulated and a model to discriminate between the two has been developed. Based on the model parameters ($E_{\text{crystal}}$, $E_{\text{module}}$, and discriminator threshold), it is possible to achieve different $2\gamma$ and $3\gamma$ efficiencies, which are summarized in table 3.1. In this chapter, the experimental setup and measurement principle will be described. Afterwards, the data will be preprocessed and put in a form that can be used for the model training and evaluation. In this context, two main challenges arise: Firstly, the simulated data comes with the information of which energy deposits belong to the same event. This is not the case for the experimental data, so a clustering algorithm for the individual measurements has to be developed. Secondly, the data in the simulation is labeled, meaning that the true decay type of each event is known. This is also not true for the measured data, so a labeling procedure has to be introduced. Both problems need to be solved before any further analysis can be performed.

## 4.1 Measurement Principle

In order to address the two problems mentioned above, the measurement principle has to be understood. A schematic of the most important components of the setup is shown in figure 4.1. A radioactive $^{22}$Na source undergoes a $\beta^+$ decay, emitting a positron which is magnetically guided to an $SiO_2$ target, where it is accelerated with a voltage around 2 keV to 4 keV and either forms Ps, i.e. o-Ps or p-Ps, or annihilates into two photons. In this process, secondary electrons (SEs) are emitted, which are being accelerated in the opposite direction towards the micro-channel plate (MCP) detector due to the sign difference of their charge. The SEs can therefore be used to tag the arrival of the positron at the target. The two modules of the SDMD detector from figure 2.1 are placed at

opposite sites of the target and are used to detect the emitted photons.

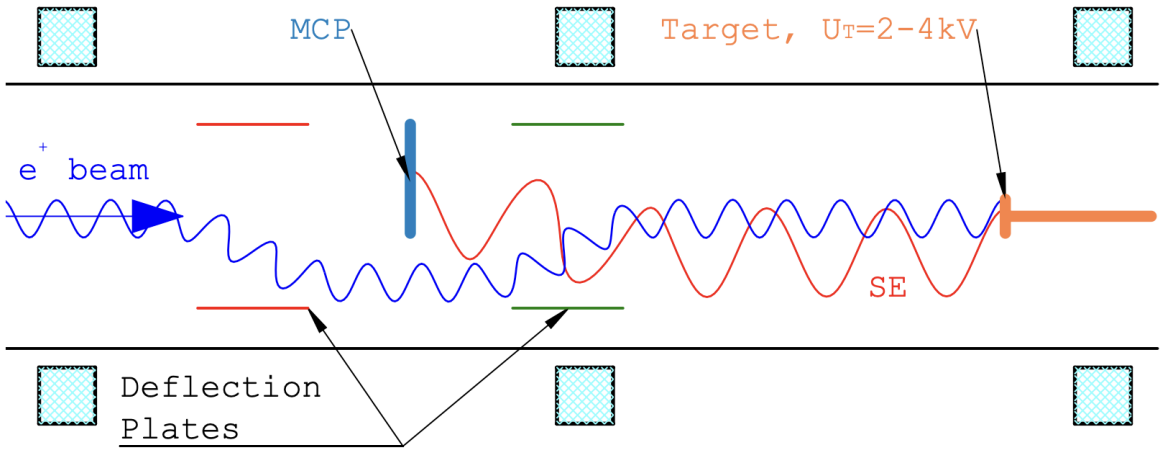


Figure 4.1: Positron tagging scheme with a micro-channel plate (MCP). The positron beam (blue helix, coming from the left) is deflected off axis by the deflection plates (red and green) to bypass the MCP. Secondary Electrons (SE, red helix) are released when the positron impinges the $SiO_2$ target and guided back and detected by the MCP. Note that the positron and electron trajectories are only sketches, the actual deflection is perpendicular to the drawing plan [1].

## 4.2 Data Preparation

It has already been said that for the experimentally obtained data, two additional steps have to be performed before the analysis: The individual measurements have to be clustered to events, which then need to be labeled. In the following, these problems will be tackled.

### 4.2.1 Clustering

The signal of the SDMD is continuously read out by two ASICSs. Every time an SE hits the MCP, all measurements of the SDMD until 1 µs before the SE hit are stored. Each measurement contains the following information:

- Energy deposit exceeding a certain threshold
- ID of the crystal that was hit by the photon
- Trigger time when the SEs hit the MCP

- Time difference between energy deposit and trigger time.

For some trigger times, the ASICs record multiple hits because multiple photons are being emitted. In order for two measurements to belong to the same event, it is therefore a necessary condition that the measurements have the same trigger time. Furthermore, using the fact that the distance between the two detector modules in reality is of the order of 1 m and the speed of light $c$ is roughly given by $3 \cdot 10^8\,\mathrm{m\,s^{-1}}$, we can estimate that the time difference between two hits belonging to the same event can be at most 3.33 ns. Thus, a sufficient condition for two measurement to belong to the same event is that the time difference between two measurements does not exceed this value. A histogram of the photon numbers associated to the same event by the clustering procedure is shown in the left plot of figure 4.2. The graphic shows that the majority of events only contain a single photon. As there are no single photon events, it is clear that the other photon of the event is not recorded by the SDMD, either because it is not hitting the detector or because of the limited efficiency of the crystals.

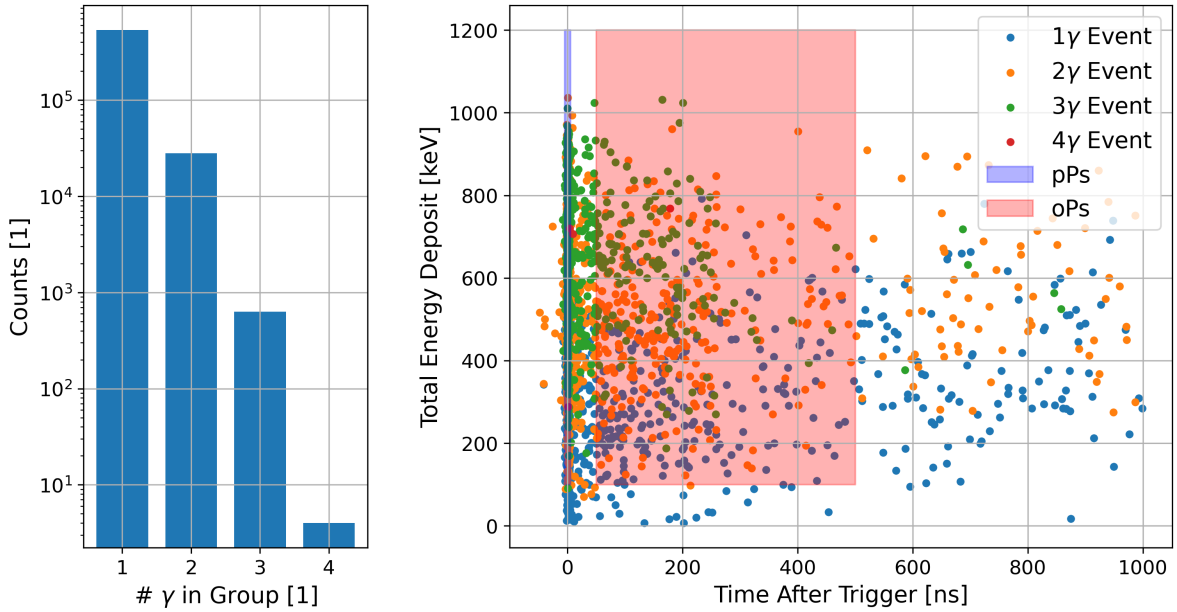


Figure 4.2: The clustering procedure gathers single measurements into events based on their time tags. The left plot shows the histogram of the number of measurements per cluster. After the clustering, the mean time delay and the total energy deposit of each cluster has been calculated. A fraction of the clusters is shown in the right plot. In order to label the clusters, one can use the fact that p-Ps decays into two photons with a lifetime of 0.1244 ns [5]. On the other hand, o-Ps decays into three photons with a much longer lifetime of 142 ns [3, 4]. Clusters with short time delays can therefore be classified as $2\gamma$ events. For the late events, the situation is more complicated, as they get contributions from $2\gamma$ pick-off annihilations.

### 4.2.2 Labeling

Having solved the clustering problem, it is still necessary to find a way of labeling the events. This can be done by exploiting the fact that p-Ps decays into $2\gamma$ with a lifetime of 0.1244 ns [5], whereas o-Ps decays into $3\gamma$ with a much longer lifetime of 142 ns [3, 4]. It is therefore reasonable to label clusters with a short time delay as $2\gamma$ events. Unfortunately, it is not possible to similarly label late events as $3\gamma$ events because they get contributions from $2\gamma$ pick-off annihilations and hence consist of a mixture of $2\gamma$ and $3\gamma$ events. Of course, this complication is the reason why a good discrimination model is needed in the first place. In order to determine the time delay of a cluster, the mean delay within the cluster is calculated. It is possible to additionally compute the total energy deposit of the cluster and plot the two quantities in a scatter plot, as shown in the right of figure 4.2. The blue stripe indicates the region where p-Ps decays are expected, whereas the red stripe tags the time interval of o-Ps decays (including $2\gamma$ pick-off events).

To conclude, it is possible to extract $2\gamma$ events from the data based on the time delay. However, one cannot easily extract $3\gamma$ events in a similar manner because the late events are a mixture of $2\gamma$ and $3\gamma$ events. There are several approaches that could be used to solve this problem. One could for example train an unsupervised clustering algorithm on the $2\gamma$ events and then apply it to the late events to filter out the $2\gamma$ pick-off annihilations. Another approach would be to make the simulation more realistic by comparing simulated and measured $2\gamma$ events. Once the simulation is sufficiently realistic, it could be used to generate $3\gamma$ events and thus create a labeled set out of measured $2\gamma$ and simulated $3\gamma$ events. Both approaches however require more time and resources than are available for this thesis and will therefore not be pursued further.

### 4.2.3 A Look on the Final Data

Although the data is not labeled and we would need more realistic Geant4 simulations to create an appropriate training set, it is still interesting to take a closer look at the data in order to see to which extent the simulation is able to resemble the real data. For this reason, the plots shown in figures 2.2 and 2.3 have been recreated for the experiment. In order to do this, the total energy deposit in the two modules has to be computed for each group. As discussed, a clear separation into $2\gamma$ and $3\gamma$ events is not possible for the data, making it necessary to take into account all events simultaneously. The result is shown in figure 4.3. As can be seen, the data is much more spread out and the back-to-back peak where the energy deposit in module A and B lies at 511 keV is not there. This

is a sign that something is wrong either with the simulation or the experimental data. The sufficient condition of the clustering procedure might be responsible for the missing peak in part, as it removes 95 % of events where the energy deposits in both modules are around 511 keV. However, this is not the full explanation as the amount of back-to-back events in the data after the necessary condition is very low by itself.

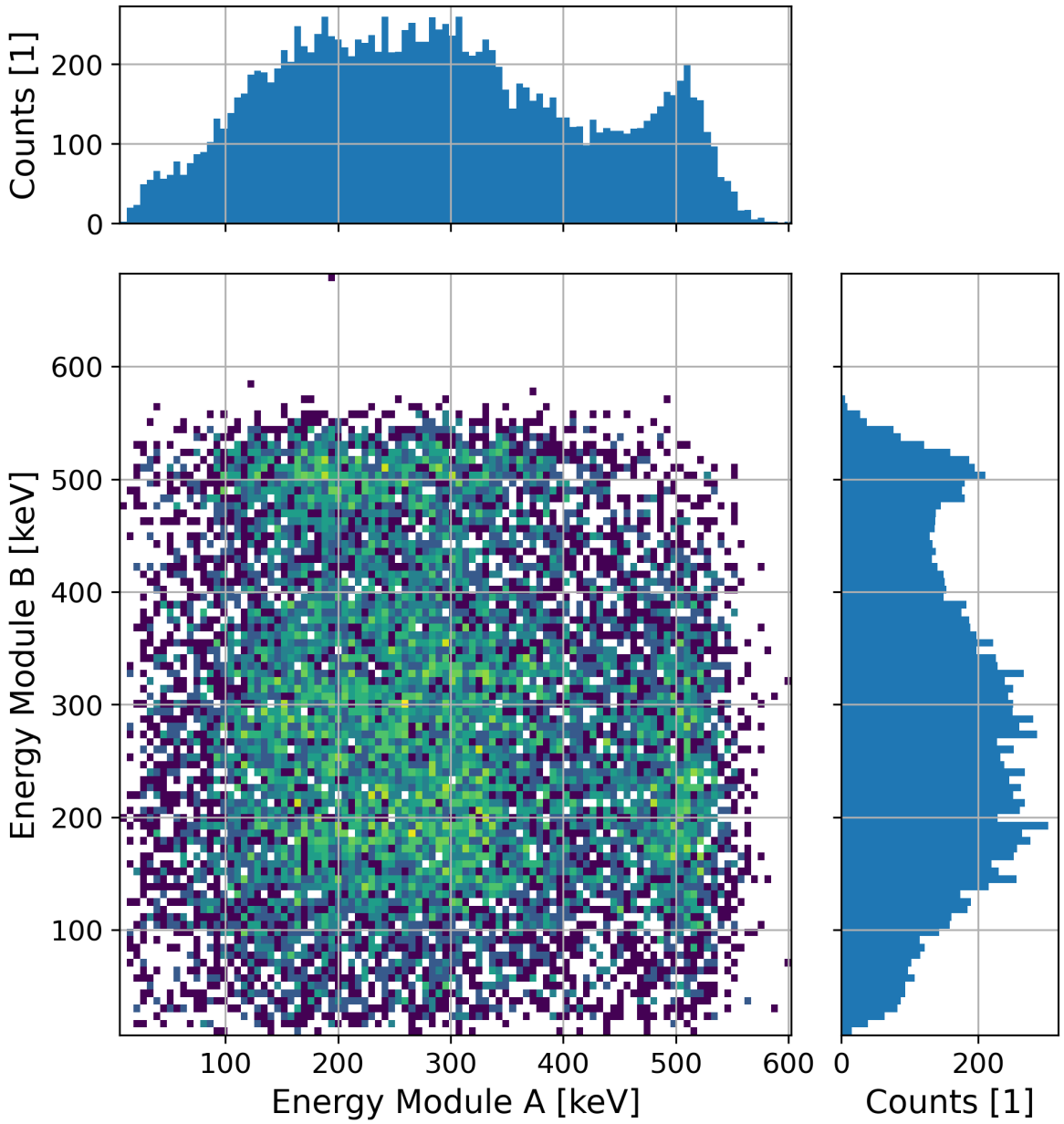


Figure 4.3: After the clustering, the total energy deposit in the two modules of the SDMD can be computed for each group. As it is not possible to separate $2\gamma$ and $3\gamma$ events for the experimental data, all events need to be taken into account simultaneously. The correlation plot shows the total energy deposit in module A and module B for each group. The two one dimensional histograms show the energy deposit in module A and B independently.

# Chapter 5

# Conclusion and Outlook

This thesis has presented an approach to discriminate between $2\gamma$ and $3\gamma$ events based on energy depositions in the SDMD. This is a necessary step to isolate the o-Ps decay width in vacuum from the measured value, which is getting unwanted contributions from pick-off events due to the surrounding material. The approach is based on a cut-based pre-selection of events, followed by an ML algorithm for the fine-tuning. The model has been trained and tested using data generated with the Geant4 simulation framework. An extensive performance analysis of the discriminator has been computed for a wide range of the cut parameters and different ML tools. Depending on the cut values and the discrimination threshold, different $2\gamma$ and $3\gamma$ efficiencies can be achieved, summarized in table 3.1. A specific choice of the model parameters has to be made based on experimental considerations, such as the event rate and the desired signal-to-background ratio.

In order to apply the discriminator to real data, the experimental data has to be clustered into events. This has been done based on the time difference between the energy depositions in the detector. The resulting clusters can however not be used to form a full training set, as it is only possible to isolate $2\gamma$ events from the data using the p-Ps decay peak, but not $3\gamma$ events because the o-Ps decays are modified by pick-off events. In the future, one could therefore try to assemble a training set using actual $2\gamma$ events and simulated $3\gamma$ events, which need to be generated more realistically.

To see the extent of deviation between the simulated and the real data, correlation plots of the energy depositions in the two SDMD modules have been created for both cases (see figures 2.2 and 2.3 for the simulated and 4.3 for the real data). The 511 keV peak does not show up in the real data, which is a problem that has not been solved yet and should be addressed in the future.

All the data and code used in this thesis have been published for reproducibility and further research. The code can be found on this GitLab repository and the data can be accessed on this Polybox.

# Bibliography


[1] Carlos Vigo, Mark Raaijmakers, Lars Gerchow, Balint Radics, André Rubbia, and Paolo Crivelli. A new approach for the ortho-positronium lifetime determination in a vacuum cavity. In *Journal of Physics: Conference Series*, volume 1138, page 012006. IOP Publishing, 2018.

[2] GS Adkins, RN Fell, and Jonathan Sapirstein. Order $\alpha$ 2 corrections to the decay rate of orthopositronium. *Physical Review Letters*, 84(22):5086, 2000.

[3] Richard Sterling Vallery, PW Zitzewitz, and DW Gidley. Resolution of the orthopositronium-lifetime puzzle. *Physical review letters*, 90(20):203402, 2003.

[4] Osamu Jinnouchi, S Asai, and T Kobayashi. Precision measurement of orthopositronium decay rate using sio2 powder. *Physics Letters B*, 572(3-4):117–126, 2003.

[5] Savely G Karshenboim. Precision study of positronium: Testing bound state qed theory. *International Journal of Modern Physics A*, 19(23):3879–3896, 2004.

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

**Title of work:**

# Discrimination of $2\gamma$ and $3\gamma$ Events in Positronium Annihilations Using Machine Learning

**Thesis type and date:**

Semester Thesis, May 31, 2024

**Supervision:**

Benjamin Banto Oberhauser
Prof. Dr. Paolo Crivelli

**Student:**

Name: Simon Wittum
E-mail: swittum@student.ethz.ch
Legi-Nr.: 23-941-370

**Statement regarding plagiarism:**

By signing this statement, I affirm that I have read and signed the Declaration of Originality, independently produced this paper, and adhered to the general practice of source citation in this subject-area.

Declaration of Originality:

http://www.ethz.ch/faculty/exams/plagiarism/confirmation_en.pdf

Zurich, 31.5.2024: