Merged $\pi^0 \to 2\gamma$ identification in LHCb for Run 4 using machine learning

Summer Student Project Report

Author: Jan Gerrit Schulz*

Supervisors: Philipp Roloff†

Sascha Stahl†

September 26, 2023

*RWTH Aachen University, Aachen, Germany

†CERN, Geneva, Switzerland

Abstract

The identification of high-energy neutral pions whose decay products (two photons) merge into a single signal cluster in the electromagnetic calorimeter is a problem in event reconstruction at the LHCb experiment, since such clusters can easily be misinterpreted as other types of clusters, e.g., those of single photons. This summer student project developed and investigated a possible classification of these two cluster types (merged pions and single photons) using machine learning. This, in turn, may help improve current studies of CP-violating decays as well as open up new possibilities in physics analysis at LHCb.

Contents

1. Introduction

The three pions $\pi^{\pm}$ and π^0 are the lightest of all hadrons [1]. Because of this property, they are among the most frequently produced particles in proton-proton collisions at high-energy particle accelerators [2]. This is also the case inside the LHCb detector at the Large Hadron Collider (LHC) at CERN.

While the charged pions themselves leave signals in various subdetectors of LHCb, e.g., tracker or calorimeter [3], the neutral pion decays into two photons in the immediate vicinity of the production point [1]. The two photons are detected in the electromagnetic calorimeter (ECAL) unless they convert beforehand by interacting with other detector material [3].

Because of their high production rate, one would like to detect neutral pions as efficiently as possible among the ECAL data. In addition, there are some interesting CP-violating processes in which the neutral pion appears as a decay product, for example $D^0 \to \pi^+\pi^-\pi^0$ or $B^+ \to K^+\pi^0$ [4], [5]. Going further, improved π^0 detection may open new doors for physics analyses that have not been thought of before. Following this motivation, this summer project will attempt to improve the identification of (merged) pions using machine learning (ML).

2. Identifying π^0 with current algorithms

In 98.8 % of the cases, the neutral pion decays after a few micrometers into two photons [1]:

$$\pi^0 \to 2\,\gamma$$

Accordingly, neutral pions are detected in the LHCb experiment by their decay products, the two photons. If they do not convert beforehand, the photons are intercepted in the ECAL. There, they produce electromagnetic showers whose energy is then measured [1], [3]. The ECAL resembles a large wall perpendicular to the beam axis of the LHC at a distance of about 12.6 m from the collision point. It is composed of individual cells, each of which separately measures the energy deposited in it per event [3]. This results in a two-dimensional image of the energy deposited in the ECAL for each event (see Figure 1).

The electromagnetic showers of the photons can be seen in the image as local maxima. Cells that represent a local maximum in the image are called seed cells. The grouping of such a seed cell and its surrounding cells is called a cluster.

2.1. Resolved neutral pions

In the usual reconstruction procedure of LHCb data, clusters of 3×3 cells are formed. The position and energy information of the cells in the cluster is then used to reconstruct the four-vector of the initial photon. If the photons are resolved in the detector (forming separate clusters), the invariant mass for that combination can be determined from the sum of the four-vectors and can be compared to the pion mass [6].

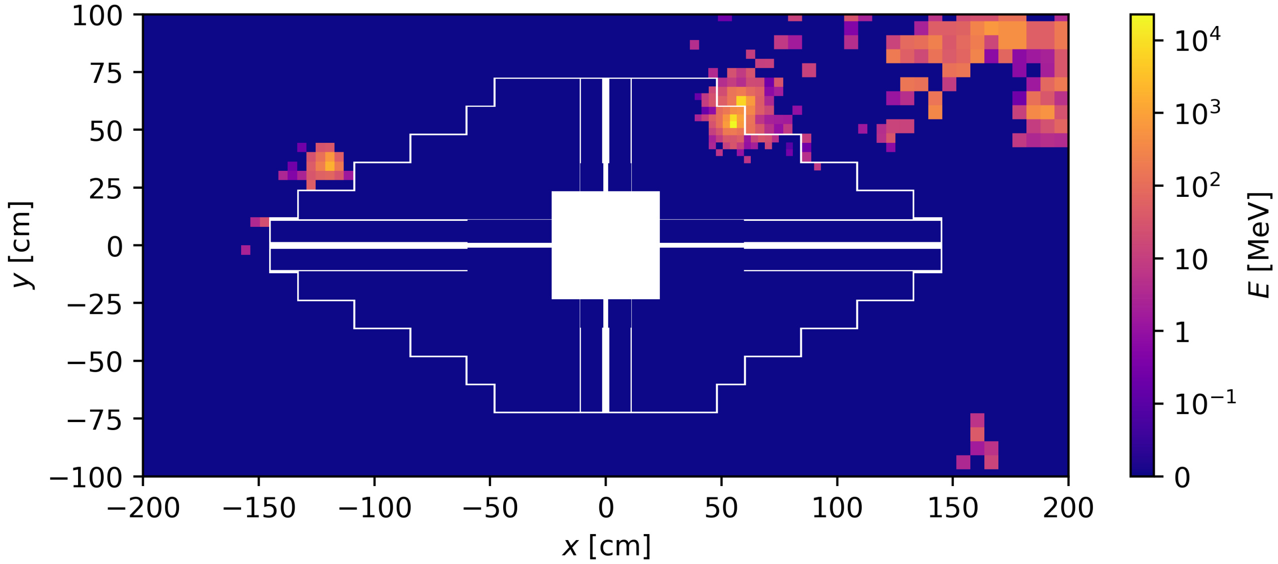


Figure 1: Example event measurement of the LHCb ECAL in Run 4 of LHC (simulated).

2.2. Merged neutral pions

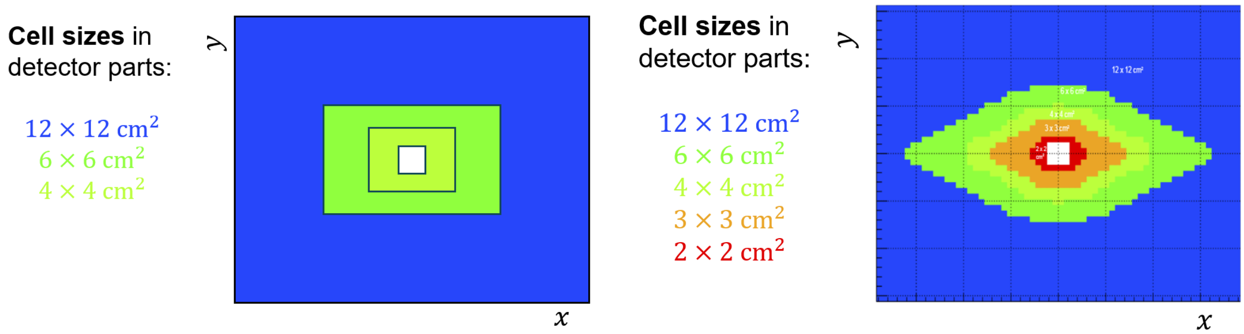
However, if the kinetic energy of the pion is too high, the photons are so close together when they enter the ECAL that only a single local maximum is produced. The two photons merge into a single cluster in the image of the event, a so-called merged π^0 cluster. For such clusters, the usual pion reconstruction procedure explained earlier can no longer be used because the single cluster is reconstructed as a single photon. There is a method to reconstruct the invariant mass from a single cluster (sub-clustering algorithm [6]). However, this method of pion identification is always associated with a relatively large false positive rate. Therefore, in this summer project, a different approach to classification between merged pions and single photons is investigated. This approach is based on machine learning.

2.3. First machine learning approach

A first attempt to identify neutral pions using machine learning for the ECAL of Run 1-3 of LHC already showed promising results [7]. However, the structure of the network applied therein as well as the representation of the problem is not suitable for the future ECAL in Run 4.

The ECAL in Run 1-3 consisted of three different types of cells with different sizes. These were arranged in a very simple rectangular structure (see Figure 2a). One neural network was trained for each cell region and all possible transition regions in between, resulting in nine individual networks in total.

The ECAL in Run 4, on the other hand, will have two additional cell types arranged in a more complicated structure (see Figure 2b). The approach of training a separate network for each part of the detector becomes infeasible and complicated. Instead, a new machine learning approach is chosen in this project, with which it is possible to describe the entire calorimeter with a single network.



(a) ECAL in Run 1-3. **(b)** ECAL in Run 4.

Figure 2: Comparison of the LHCb ECAL structures in different runs of LHC.

3. Simulated data sets

Before the machine learning approach is explained in more detail, the data used for it shall be discussed here.

3.1. Toy model data

In the beginning, an initial study was conducted using simulated data from a simplified toy model. This can be found in Appendix A, but will not be elaborated here, since the central results were generated with detailed ECAL Run 4 simulations.

3.2. Detailed ECAL Run 4 simulations

The main study was performed with simulated data for the LHCb ECAL in Run 4. For this, the simulations were created using LHCb software and a detailed stand-alone simulation package for the ECAL based on Geant4 with a parameterized transport of scintillation photons. The simulated events are clean signal and background events, that is, either a single photon in the ECAL or two photons from a π^0 decay[1].

During the data generation, it cannot be said a priori whether the generated event is a merged or resolved event, because for this the clustering must be carried out first. Therefore, to maximize the fraction of merged events in the simulation, a cut was applied to the energy, $E > 15\,\mathrm{GeV}$, and to the transverse energy, $E_\mathrm{T} > 2\,\mathrm{GeV}$, of the π^0 respectively, when generating the pion data.

Selection of merged pion events Subsequently, clustering was performed for all events. The merged events were then selected from all the simulated events according to the following procedure:

[1]The π^0 here comes from the simulation of a $D^0 \to \pi^+\pi^-\pi^0$ decay.

1. Match photons from the Monte Carlo (MC) simulation to reconstructed clusters in the ECAL: Choose the closest cluster to the photon, that fulfills the condition:

$$\|\vec{r}_\gamma - \vec{r}_{\text{cluster}}\| < 2 \cdot \Delta x_{\text{seed}}$$

   Hereby, $\vec{r}_\gamma$ is the intersection of the photon trajectory and the ECAL plane shifted by the penetration depth of the given cluster, $\vec{r}_{\text{cluster}}$ is the reconstructed coordinate of the cluster and Δx_{seed} is the width/size of the seed cell. If no cluster matches this condition, do not assign any cluster to the photon.

2. The three possibilities that arise are labeled as follows: the event is *resolved* (the photons are attributed to different clusters), *merged* (the photons are attributed to the same cluster), or *not reconstructable* (at least one of the photons is not attributed to any cluster).

Figure 3 shows the number of clusters per event for the differently labeled events. As expected, most merged events have exactly one cluster.

One could naively take as criterion for merged events all events with exactly one cluster. However, this also accepts events in which one of the photons has too little energy to create a cluster (there is a transverse energy threshold for seed cells) or events in which one of the photons does not hit the detector at all.

The resulting energy distribution of the initial particles (single photon or π^0) after applying the described selection rule for merged events is shown in Figure 4. As can be seen, the simulated energies for signal and background are very similar. However, the pion energy distribution has fewer low-energy events, since no merged events occur at low energies.

Preparation of the data for training For the machine learning application, larger clusters than in the usual reconstruction were created to include more information from the shower structure. Usually, clusters are of the order of 3×3 cells and are assembled by including all neighbors of the seed cell. Here, all neighbors of neighbors were additionally added to the cluster, resulting in a scale of 5×5 cells (see Figure 5). The information gathered from each cell in the cluster for the ML input is summarized in Table 1.

[2] In a small ancillary study, it was found that omitting the absolute positions of the cells in the ECAL did not significantly degrade the classification.

Table 1: Inputs for the artificial neural networks used in this project. Note that variables marked with an asterisk are passed to the ParticleNet lite network twice: once as normal input and once as coordinate input.

Inputs for the machine learning network	
Energies of individual cells	E
Relative energies of individual cells w.r.t. seed cell	E/E_{seed}
*Positions of individual cells[2]	(x, y)
*Relative positions of individual cells w.r.t. seed cell	$(x - x_{\text{seed}}, y - y_{\text{seed}})$
Widths/sizes of individual cells	$(\Delta x, \Delta y)$

Figure 3: Number of clusters per π^0 event in the detailed LHCb Run 4 simulations.

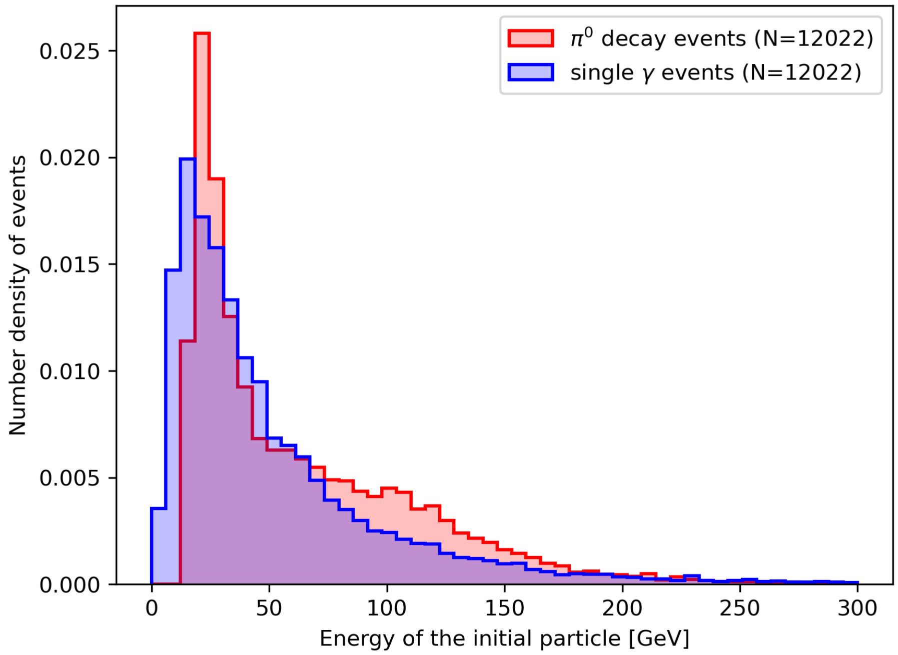
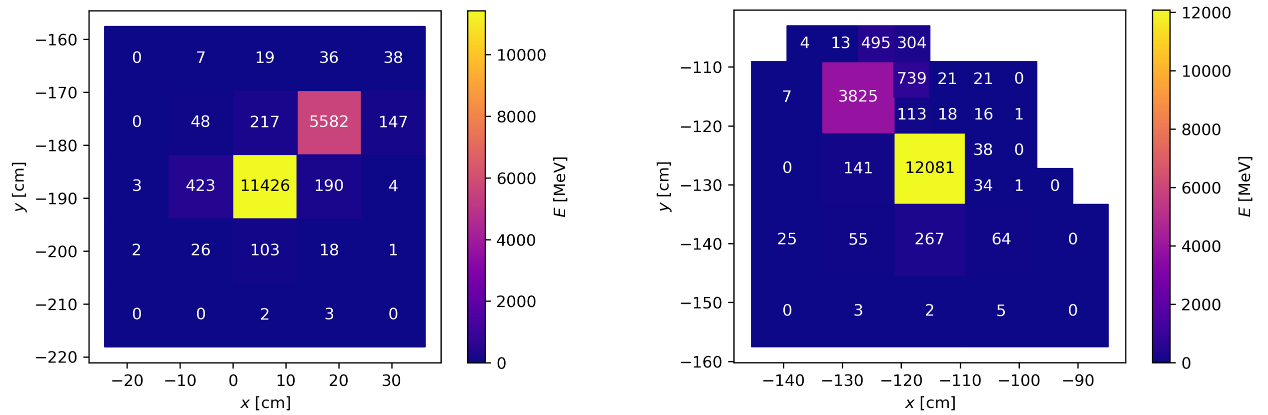


Figure 4: Energy distribution of the initial particles generated (single photon or π^0) in the detailed ECAL Run 4 simulations for the training and validation sets.



(a) Example of a cluster consisting of only one type of cells.

(b) Example of a cluster consisting of several types of cells.

Figure 5: Example clusters for merged pion events in the LHCb ECAL for Run 4 (simulated).

Training, validation and test sets A total of 29 748 events were generated (14 874 single photons and 14 874 merged π^0). This data set was divided into a training set, a validation set and a test set (see Table 2 for the exact numbers).

Table 2: Data set sizes for the training of the artificial neural networks.

Total	Training		Validation		Test	
merged π^0 (signal) 14 874	9618	65 %	2404	16 %	2852	19 %
single photon (background) 14 874	9618	65 %	2404	16 %	2852	19 %

4. The machine learning approach

The code for the ML approach as well as for the toy data generation can be found in the GitLab project linked here[3].

4.1. The used network architectures

Two different network architectures were used for this project, the *ParticleFlowNetwork* [8] and the *ParticleNet lite* [9]. More information about these networks can be found in the respective papers. The major advantage that these networks offer, and thus the reason for their use, is that they accept particle clouds as input. This means that the input consists of an unordered list of identical objects. Identical in this context means that these objects all have the same attributes. In this case, the objects are the cells of the cluster and the attributes are the properties of each cell summarized in Table 1.

The number of cells passed to the network can vary from cluster to cluster. Together with the fact that the positions and sizes of the cells are passed directly to the network, this allows the network to process different cluster shapes without having to handle them explicitly in the reconstruction. This makes the approach both elegant and easy to apply to the complex structure of the ECAL in Run 4.

4.2. Training and validation

Training with the detailed MC simulations was performed using the *weaver-core*[4] framework based on *PyTorch*. The optimized hyperparameters used can be found in Table 3. The training was performed with the training data set up to the specified number of epochs. Then, the network that achieves the best score with respect to the validation data set was selected as the result. The metric used in this process has as its most important component the accuracy in the classification. The number of epochs was chosen sufficiently large so that overfitting already starts in the later course of the training.

Table 3: Optimized hyperparameters for the training of the networks on the detailed LHCb ECAL Run 4 simulations.

Network	**ParticleFlowNetwork**	**ParticleNet lite**
Optimizer	ranger	ranger
Batch size	256	181
Learning rate (at start)	4.6×10^{-3}	4.6×10^{-2}
Learning scheduler	flat+decay	flat+decay
Number of epochs	1500	1500

[3]GitLab project: https://gitlab.cern.ch/spacal-rd/summerstudentprojects/jan-schulz
[4]See weaver-core here: https://pypi.org/project/weaver-core/.

5. Results

The results were produced using the trained networks and the test data set and are shown in Figure 6 in the form of receiver operating characteristic (ROC) curves. Those show the true positive rate (acceptance rate ϵ_s for signal events, in this case π^0) as a function of the false positive rate (acceptance rate ϵ_b for background events, in this case photons).

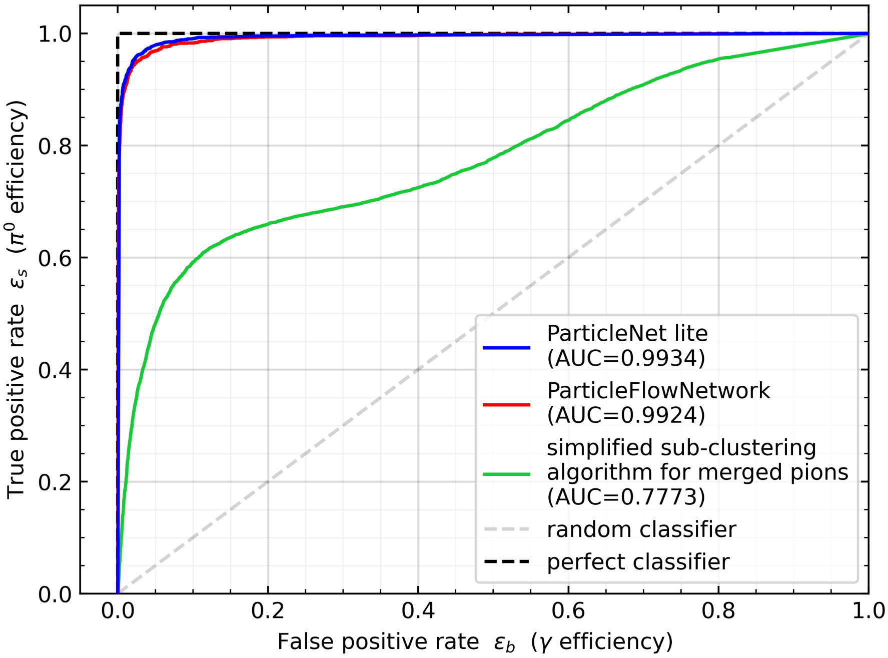


Figure 6: Results for the training of the two network architectures on detailed LHCb ECAL Run 4 simulations shown in the form of a receiver operating characteristic (ROC) curve.

In addition to the ROC curves for the two networks, the figure also shows those of a perfect classifier, a random classifier, and a simplified sub-clustering algorithm for comparison. The latter is similar to the merged reconstruction procedure mentioned in subsection 2.2, which is used for the current data analysis in Run 3. However, its implementation is not yet fully adapted to Run 4. For this reason, it should be assumed that the final reconstruction algorithm may perform somewhat better than the simplified one shown here.

Nevertheless, it can be concluded from the comparison of the curves that the ML approach with both networks clearly provides better results in the classification. Their curves are very close to the perfect curve and achieve very promising results with an area under the curve (AUC) score of more than 0.99. Accordingly, the new approach to describe the problem for the entire ECAL using a single network works well.

Moreover, it can be seen that the ParticleNet lite performs slightly better than the Particle-FlowNetwork, which confirms the results of the corresponding paper [9]. It is also worth mentioning that the ParticleNet lite used here has in its structure less than one third of free parameters compared to the ParticleFlowNetwork used.

6. Conclusion and outlook

In this project, a new, more elegant ML approach for identifying high-energy neutral pions was introduced and shown to work.

It can be concluded that the ML approach yields promising results in the classification of merged neutral pions and single photons, clearly outperforming those of the currently used algorithm. Thus, the approach can potentially help improve physics analysis in LHCb in the future.

Since the events used here are pure and thus pile-up is not present in any form, it is likely that results for pile-up events and event types not yet learned will be somewhat worse. For this reason, more extensive studies and training should be done to obtain networks for actual data analysis.

In addition, other future features of the LHCb ECAL could be included, such as timing information, which will be introduced in Run 5.

Acknowledgments

I would like to thank my awesome supervisors Philipp and Sascha for the great project and the amazing opportunity to work on it. I would also like to thank the lunch group for the many interesting lunches followed by coffee in front of building 38. And finally, I would like to say a special thank you to all my friends (especially those who will be having dinner tonight) who have made this summer better than I could have ever dreamed.

References

[1] Particle Data Group, R. L. Workman *et al.*, "Review of Particle Physics," *PTEP*, vol. 2022, p. 083C01, 2022. DOI: 10.1093/ptep/ptac097.

[2] ALICE collaboration, K. Aamodt *et al.*, "Production of pions, kaons and protons in pp collisions at $\sqrt{s} = 900\,\text{GeV}$ with ALICE at the LHC," *The European Physical Journal C*, vol. 71, no. 6, Jun. 2011. DOI: 10.1140/epjc/s10052-011-1655-9. [Online]. Available: https://doi.org/10.1140%2Fepjc%2Fs10052-011-1655-9.

[3] LHCb collaboration, R. Aaij *et al.*, *The LHCb upgrade I*, 2023. arXiv: 2305.10515 [hep-ex].

[4] LHCb collaboration, R. Aaij *et al.*, "Search for CP violation in the phase space of $D^0 \to \pi^-\pi^+\pi^0$ decays with the energy test," Jun. 2023. arXiv: 2306.12746 [hep-ex].

[5] LHCb collaboration, R. Aaij *et al.*, "Measurement of CP Violation in the decay $B^+ \to K^+\pi^0$," *Phys. Rev. Lett.*, vol. 126, no. 9, p. 091 802, 2021. DOI: 10.1103/PhysRevLett.126.091802. arXiv: 2012.12789 [hep-ex].

[6] O. Deschamps *et al.*, "Photon and neutral pion reconstruction," CERN, Geneva, Tech. Rep., 2003. [Online]. Available: https://cds.cern.ch/record/691634.

[7] V. Chekalina and F. Ratnikov, "Machine Learning approach to γ/π^0 separation in the LHCb calorimeter," *Journal of Physics: Conference Series*, vol. 1085, no. 4, p. 042 036, Sep. 2018. DOI: 10.1088/1742-6596/1085/4/042036. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1085/4/042036.

[8] P. T. Komiske *et al.*, “Energy flow networks: Deep sets for particle jets,” *Journal of High Energy Physics*, vol. 2019, no. 1, Jan. 2019. DOI: 10.1007/jhep01(2019)121. [Online]. Available: https://doi.org/10.1007%2Fjhep01%282019%29121.
[9] H. Qu and L. Gouskos, “Jet tagging via particle clouds,” *Physical Review D*, vol. 101, no. 5, Mar. 2020. DOI: 10.1103/physrevd.101.056019. [Online]. Available: https://doi.org/10.1103%2Fphysrevd.101.056019.
[10] G. Grindhammer and S. Peters, *The parameterized simulation of electromagnetic showers in homogeneous and sampling calorimeters*, 2000. arXiv: hep-ex/0001020 [hep-ex].
[11] LHCb collaboration, C. A. Beteta *et al.*, *Calibration and performance of the LHCb calorimeters in Run 1 and 2 at the LHC*, 2020. arXiv: 2008.11556 [physics.ins-det].

A. First study with a toy model

To test the possible application of machine learning to the merged π^0 classification in a computationally efficient manner, a toy model of an electromagnetic calorimeter was first developed and used for initial approaches. In the following, first the toy model itself with its limitations and then the data simulation procedure will be explained. Finally, the results of the study are shown.

A.1. The toy calorimeter and its limitations

The simulated example calorimeter is, similar to the real LHCb ECAL, a two-dimensional wall with an extension of about $3 \times 4\,\mathrm{m}^2$ (about 1/4-th of the size of the real one). Its orientation is perpendicular to the beam axis and its distance to the collision point is 12.6 m. The calorimeter is composed of individual square cells of two different types (A and B). See Figure 7 for a visualization of the ECAL and for the specifications of the two cell types. As with the real LHCb ECAL, the smaller cells are used in the inner region of the ECAL near the beam axis.

It is important to note that the data generated with this toy model is by no means designed to simulate real LHCb data one-to-one, and an artificial neural network trained with it cannot be applied directly to real data. It is rather an experimental toy example for studying the network performance, the needed amount of Monte Carlo (MC) simulations for the proper network and the general use of machine learning for the given task without having to generate computational more expensive MC simulations for the true LHCb detector. Therefore, this toy data comes with some limitations:

- Only two different cell types are considered in the ECAL while the real one has up to five.
- Statistical fluctuations in the detector resolution are included in an approximate way (see Equation 1 further down).
- The energy calculation for the cells treats all photons as if they hit the detector perpendicularly, i.e., not under an angle. This is a strong simplification especially for photons hitting the cells at the edge of the detector.

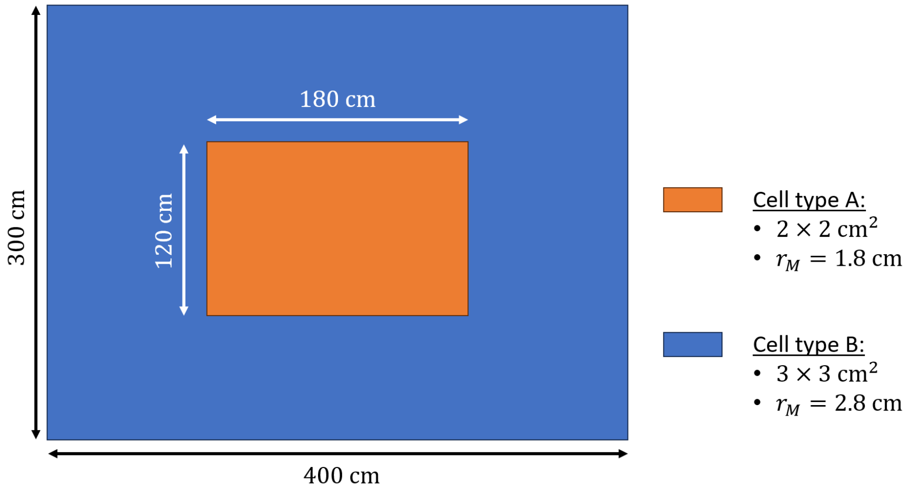


Figure 7: Illustration of the simulated toy calorimeter in the x-y plane including the different cell specifications.

A.2. Generation of the toy data

The data generation for the two event cases (merged π^0 or single photon) differ slightly from one another.

Single photons For a single photon event of given energy, a random position (x_γ, y_γ) is generated inside a random cell of the ECAL. Then, a cluster (usually 5×5 cells) is constructed around this so-called seed cell. Based on the position, the mean energy density distribution of the shower in the x-y plane is calculated using the following formula (based on Equation (23) in [10]):

$$\epsilon(r) = \frac{f(r)}{2\pi r} \quad \text{with} \quad f(r) = \left[p \cdot \frac{2rR_1^2}{(r^2+R_1^2)^2} + q \cdot \frac{2rR_2^2}{(r^2+R_2^2)^2} + (1-p-q) \cdot \frac{2rR_3^2}{(r^2+R_3^2)^2}\right] \cdot \left(\frac{1.8\,\text{cm}}{r_\text{M}}\right)^2$$

Hereby, r_M is the Molière radius of the cell material and r the scaled distance to the shower center (x_γ, y_γ):

$$r = \sqrt{(x-x_\gamma)^2 + (y-y_\gamma)^2} \cdot \frac{1.8\,\text{cm}}{r_\text{M}}$$

This distribution is then integrated over each cell individually to determine the energy deposited in the cell. The values for the parameters p, q, R_1, R_2, R_3 were determined via a fit of MC simulated showers in a material with Molière radius $r_\text{M} = 1.8\,\text{cm}$ (see Figure 8). Although the fit is not perfect, it is of sufficient accuracy for the given application in the toy model.

This approach works fine for clusters consisting only of one cell type and will always yield perfectly symmetric energy showers. However, as soon as multiple cell types are involved, the energy determination gets more complicated.

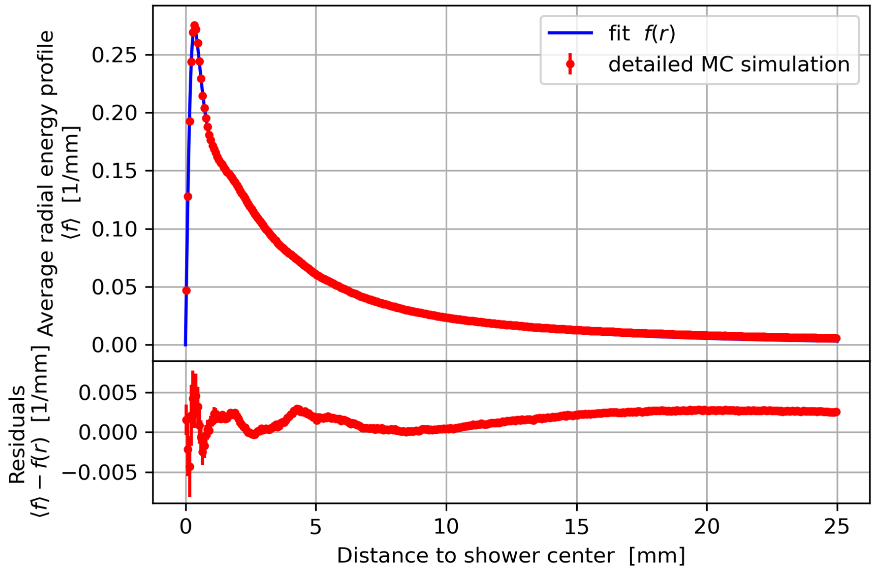


Figure 8: Fit of the mean transversal energy density distribution $f(r)$ of an electromagnetic shower.

Since a shower can now extend into two regions of different cell types, the shower is in general not symmetric anymore as it is broader in cells with larger Molière radius. In order to keep it simple, almost all specific effects that might occur at the boundaries are ignored. The different regions are simply integrated according to the normal energy distribution associated to the Molière radius. The only thing that is taken into account is that in order to keep the total energy fixed, one needs to scale one of the two regions down by a factor (see following calculation).

Integrating over the energy distribution ϵ_A of the cells of type A (determined by their Molière radius) yields 1,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \epsilon_A(r)\,\mathrm{d}x\,\mathrm{d}y = 1,$$

just like integrating over the energy distribution ϵ_B of cells B:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \epsilon_B(r)\,\mathrm{d}x\,\mathrm{d}y = 1$$

However, integrating over one part of the ECAL, e.g., the region $\mathcal{A}$, does **not** yield the same for both distributions:

$$\iint_{\mathcal{A}} \epsilon_A(r)\,\mathrm{d}x\,\mathrm{d}y = P_A \neq P_B = \iint_{\mathcal{A}} \epsilon_B(r)\,\mathrm{d}x\,\mathrm{d}y$$

Therefore, partially integrating over two regions alone is not a physical option:

$$\iint_{\mathcal{A}} \epsilon_A(r)\,\mathrm{d}x\,\mathrm{d}y + \iint_{\mathbb{R}^2\setminus\{\mathcal{A}\}} \epsilon_B(r)\,\mathrm{d}x\,\mathrm{d}y = P_A + (1 - P_B) \neq 1$$

Thus, at least one of the two distributions must be scaled accordingly, e.g.:

$$\epsilon'_B \equiv \frac{1-P_A}{1-P_B} \cdot \epsilon_B$$

$$\Rightarrow \quad \iint_{\mathcal{A}} \epsilon_A(r)\,\mathrm{d}x\,\mathrm{d}y + \iint_{\mathbb{R}^2\setminus\{\mathcal{A}\}} \epsilon'_B(r)\,\mathrm{d}x\,\mathrm{d}y = P_A + (1-P_B)\cdot\frac{1-P_A}{1-P_B} = 1$$

It was choosen to always scale the distribution of the cell region that does not contain the photon. An example for a shower distribution in such a border region is shown in Figure 9.

Additionally, for 50 % of the single photon events, a low-energy photon (0 to 10 GeV) was added to the event to mimic pile-up.

As a final step of data distortion, the energy resolution is taken into account by varying the energy of each cell using the following formula:

$$\frac{\sigma_E}{E} = \frac{10\,\%}{\sqrt{E/(1\,\mathrm{GeV})}} \oplus 1.4\,\% \tag{1}$$

Note on the energy resolution formula The target energy resolution for an entire cluster is given by

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E/(1\ \mathrm{GeV})}} \oplus b \tag{2}$$

where $a = 10\%$ and $b = 1\%$ (approximation of Equation (1) in [11]). Since this formula describes the general behavior of energy resolutions due to their measurement method, it is assumed here that the energy resolution of individual cells is also described by a formula of the given structure.

However, it is by no means obvious whether the parameters a, b are the same as for the cluster or change. In order to study this, different parameters a, b are used to disturb the generated clusters on the level of individual cells. Then, all cell energies of a cluster are summed up to get the *reconstructed energy* of the cluster. From the distribution of those reconstructed energies for clusters of the same *true energy* E, one can estimate a standard deviation σ_E for the given true energy E. Plotting σ_E/E and comparing it to the expectation given by the formula above, one can choose parameters a, b that reproduce the expectation best (see Figure 10). From the plot, one gets:

$$a = 10\% \quad \wedge \quad b = 1.4\%.$$

So, the energy resolution for individual cells is slightly worse than for the cluster.

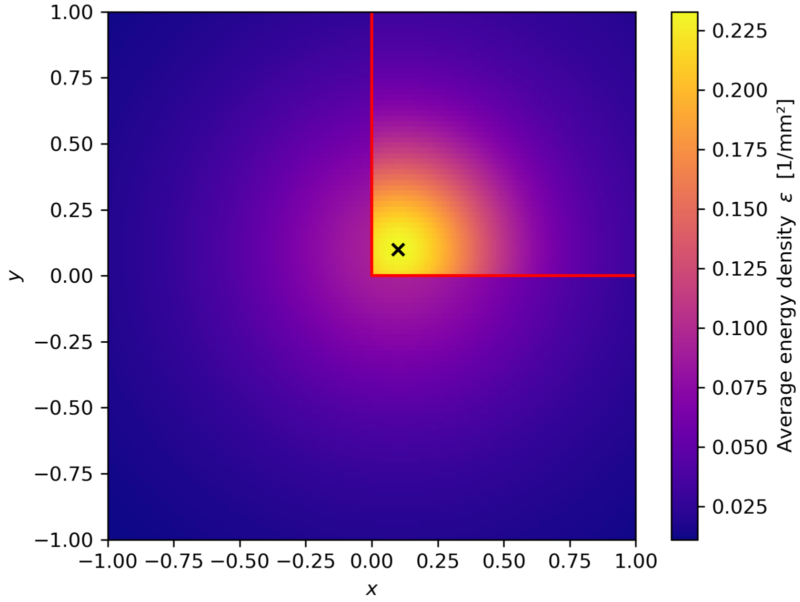


Figure 9: Example shower in border region between two cell types. The upper right region has a Molière radius of 1.8 cm, while the rest has one of 2.8 cm. The × symbol marks the photon hit position.

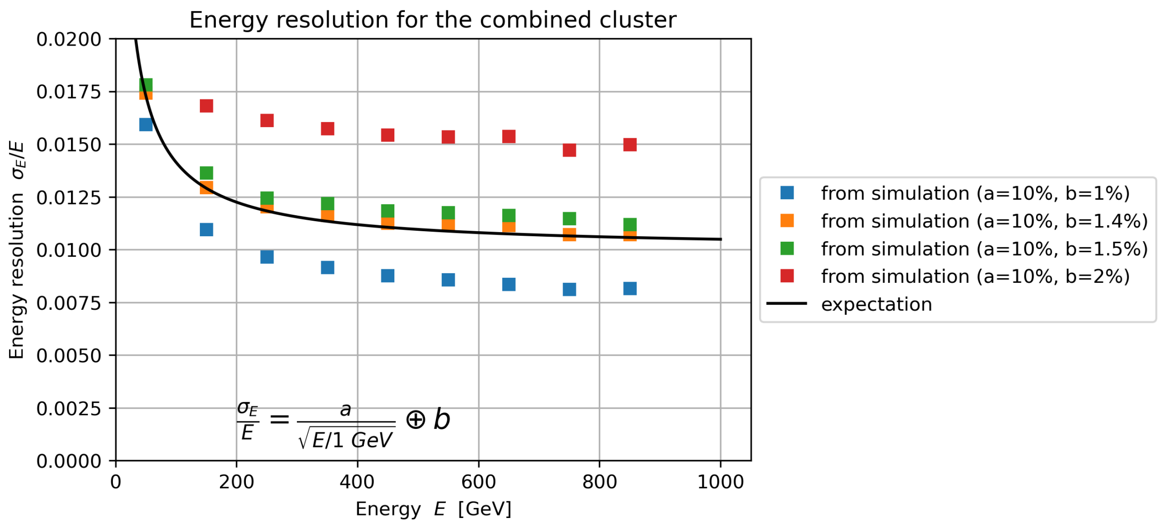


Figure 10: Overall energy resolution of an entire cluster when varying the energy resolution of each individual cell according to Equation 2. The expected curve shows the known energy resolution for clusters in the LHCb experiment.

Pion decays For the neutral pion decay, the simulation procedure is a little bit more complex. First, the decay into two photons is simulated in the pion rest frame. There, the two photons have the four-momenta p_1' and p_2'. Those four-vectors are then boosted in a random direction towards the ECAL according to a given energy E_π of the pion. The $\vec{\beta}$ vector (with $|\vec{\beta}| = \sqrt{1 - m_\pi^2/E_\pi^2}$) is randomly generated so that it hits the ECAL (see Figure 11):

The first element of the resulting boosted four-momentum is the energy and the rest gives the momentum vector which can be propagated to the ECAL plane at a distance of 12.6 m. Here, the energies in the individual cells can be calculated analogously to the single photon case except that one has to sum up the distributions of the two photons.

A.3. Training and results

Using the procedure described in the previous section, a total of 140 000 events were generated, which are divided as follows:

- 50 000 π^0 events, which are randomly divided into a training set (80 %) and a validation set (20 %) for training.
- 50 000 single photon events, which are randomly divided into a training set (80 %) and a validation set (20 %) for training.
- 20 000 π^0 events, which are used as test set for evaluation.
- 20 000 single photon events, which are used as test set for evaluation.

The energy of the generated neutral pions or single photons was uniformly distributed between 80 and 1000 GeV (see energy distribution for the test sets in Figure 12). The background for the choice of the energy is the fact that for the given cell size of the smaller cell type A, the minimum energy for merged π^0 is about 60 GeV.

Training The choice of the variables used for machine learning is described in section 3. The two network architectures trained are the *ParticleFlowNetwork* [8] and the *ParticleNet lite* [9] as later used for the detailed LHCb ECAL simulations (more information about them in section 4).

Training was performed with the open source module *weaver-core*[5] for Python which is based on *PyTorch*[6]. The optimized hyperparameters can be found in Table 4 and are overall the same as for the training on detailed LHCb simulations, except for the number of epochs.

For each of the following results, the networks from the epoch in which the evaluation metric (accuracy in categorizing the validation set) was best were used.

Results All resulting plots were created by applying the networks to the test data, which the networks did not yet know.

[5]See weaver-core here: https://pypi.org/project/weaver-core/.
[6]See PyTorch here: https://pytorch.org/.

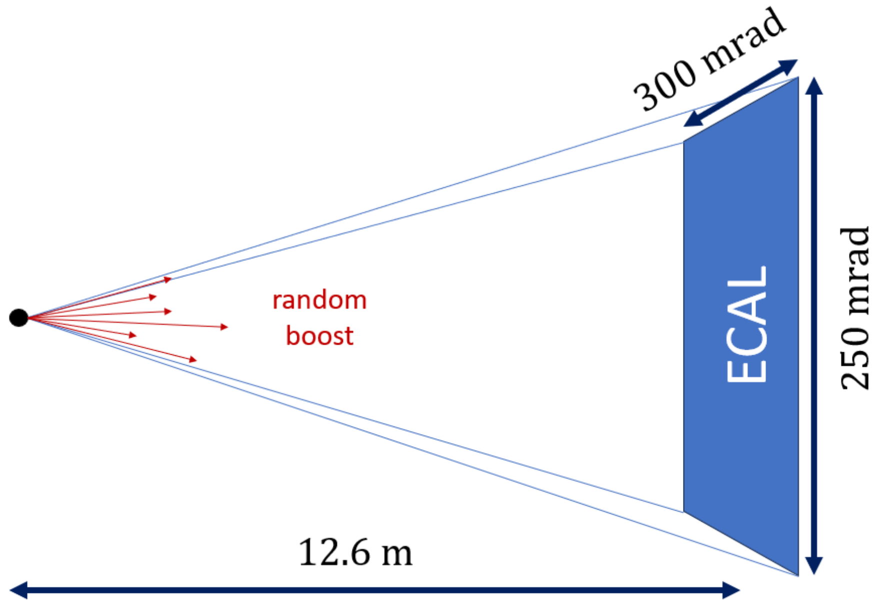


Figure 11: Random boost of the pion onto the ECAL.

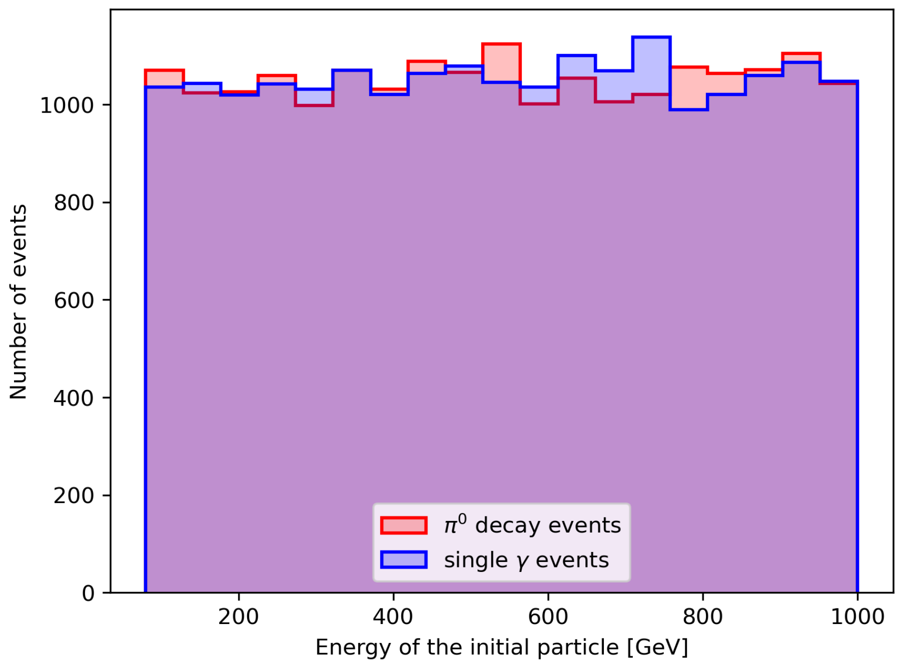


Figure 12: Energy distribution of the initial particles generated (single photon or π^0) for the test sets.

Table 4: Optimized hyperparameters for the training of the networks on the toy data.

Network	**ParticleFlowNetwork**	**ParticleNet lite**
Optimizer	ranger	ranger
Batch size	256	181
Learning rate (at start)	4.6×10^{-3}	4.6×10^{-2}
Learning scheduler	flat+decay	flat+decay
Number of epochs	8500	8500

Figure 13 shows the receiver operating characteristic (ROC) curve with the true positive rate (acceptance rate ϵ_s for signal events, in this case π^0) as a function of the false positive rate (acceptance rate ϵ_b for background events, in this case photons).

It can be seen that both networks learn to distinguish between signal and background. The ParticleNet lite performs slightly better than the ParticleFlowNetwork. This is in agreement with the results of the ParticleNet paper [9]. This is a promising first result suggesting that efficient classification for merged neutral pions is possible with machine learning.

Further, the toy model can be used to check the energy dependence of the performance. This is shown in Figure 14a. As expected, the accuracy of both networks decreases towards higher energies. The reason is that the photons from the pion decay have on average a smaller distance in the ECAL at higher pion energy and thus the signal resembles more and more that of a single photon.

In addition, the performance in different cell regions can be compared (see Figure 14b). Here, as one would naively expect, the accuracy is highest in the region with the smallest cells (type A). The border region between A and B performs worst. This is probably due to the more complicated structure of the clusters as well as the smaller amount of data (relative amount of events in different regions: $N_A : N_B : N_{A+B} \approx 15 : 80 : 5$). This is a first sign that large samples are needed for the training with a complex ECAL geometry.

Another sign can be drawn directly from the comparison of networks trained with data sets of different sizes[7] (see Figure 14c). The increasing accuracy with increasing data set size is clearly visible. Thus, the detailed LHCb ECAL simulation should be as extensive as possible.

Summary Overall, all results with the toy model confirm expectations, suggesting that the problem and the machine learning solution approach were well understood. Furthermore, all plots without exception show the performance superiority of the ParticleNet lite compared to the ParticleFlowNetwork.

[7] It should be noted here that these trainings were performed with photon data without the additional low-energy photons to mimic pile-up (0 to 10 GeV). For this reason, the curves are better than the main results in Figure 13.

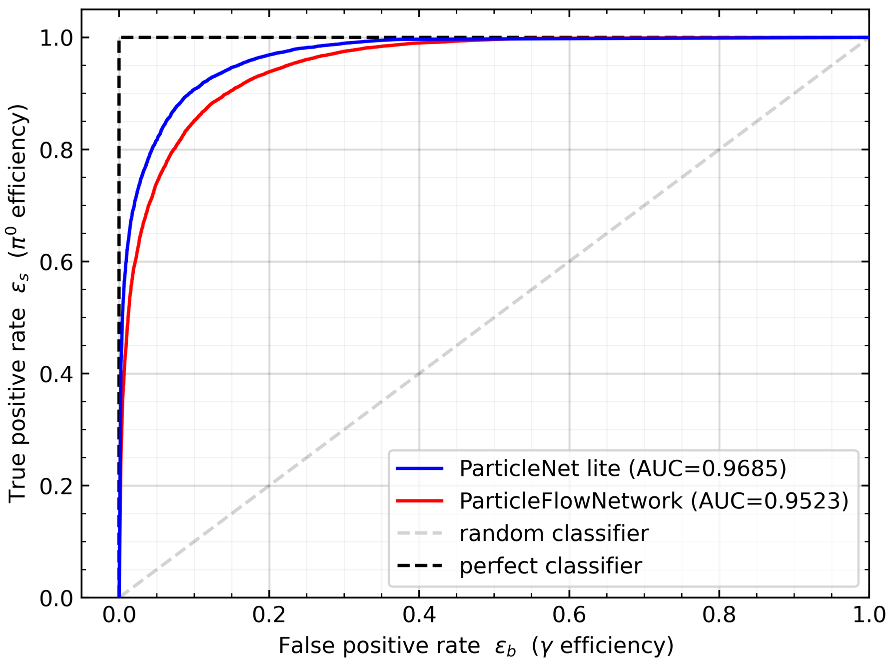
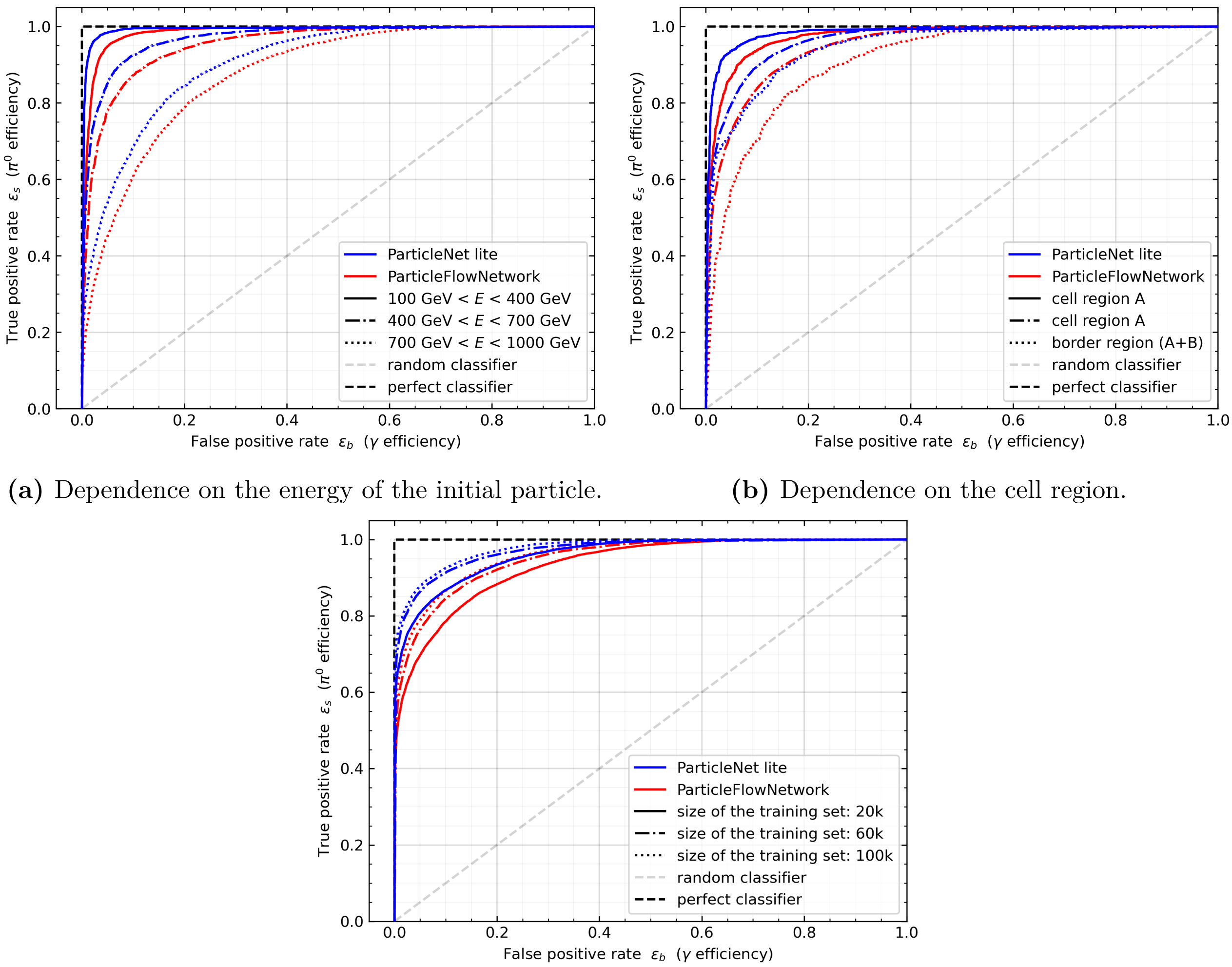


Figure 13: Results for the training of the two network architectures on toy data shown in the form of a receiver operating characteristic (ROC) curve.



(a) Dependence on the energy of the initial particle.

(b) Dependence on the cell region.

(c) Dependence on the size of the training set[7].

Figure 14: Results in form of ROC curves for the training of the two network architectures on toy data compared for different characteristics.