Journal of
Instrumentation

PAPER • OPEN ACCESS

Shower separation in five dimensions for highly granular calorimeters using machine learning

To cite this article: S. Lai *et al* 2024 *JINST* **19** P10027

Published by IOP Publishing for Sissa Medialab

Received: *June 28, 2024*
Revised: *August 28, 2024*
Accepted: *September 25, 2024*
Published: *October 24, 2024*

Shower separation in five dimensions for highly granular calorimeters using machine learning

2024 JINST 19 P10027

The CALICE collaboration

S. Lai,[a] J. Utehs,[a] A. Wilhahn,[a] M.C. Fouz,[b] O. Bach,[c] E. Brianne,[c] A. Ebrahimi,[c] K. Gadow,[c] P. Göttlicher,[c] O. Hartbrich,[c,1] D. Heuchel,[c] A. Irles,[c,2] K. Krüger,[c] J. Kvasnicka,[c,3] S. Lu,[c] C. Neubüser,[c] A. Provenza,[c] M. Reinecke,[c] F. Sefkow,[c] S. Schuwalow,[c,4] M. De Silva,[c] Y. Sudo,[c] H.L. Tran,[c] L. Liu,[d] R. Masuda,[d] T. Murata,[d] W. Ootani,[d] T. Seino,[d] T. Takatsu,[d] N. Tsuji,[d] R. Pöschl,[e] F. Richard,[e] D. Zerwas,[e] F. Hummer,[f] F. Simon,[f] V. Boudry,[g] J-C. Brient,[g] J. Nanni,[g] H. Videau,[g] E. Buhmann,[h] E. Garutti,[h,*] S. Huck,[h] G. Kasieczka,[h] S. Martens,[h] J. Rolph,[h] J. Wellhausen,[h] B. Bilki,[i] D. Northacker,[i] Y. Onel,[i] L. Emberger[j] and C. Graf[j]

[a] *II. Physikalisches Institut, Georg-August-Universität Göttingen, Friedrich-Hund-Platz 1, D-37077 Göttingen, Germany*

[b] *CIEMAT, Centro de Investigaciones Energeticas, Medioambientales y Tecnologicas, Madrid, Spain*

[c] *DESY, Notkestrasse 85, D-22603 Hamburg, Germany*

[d] *ICEPP, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

[e] *Université Paris-Saclay, CNRS/IN2P3, IJCLab, 91405 Orsay, France*

[f] *Institute for Data Processing and Electronics, Karlsruhe Institute of Technology, Kaiserstr. 12, D-76131 Karlsruhe, Germany*

[g] *Laboratoire Leprince-Ringuet (LLR), CNRS, École polytechnique, Institut Polytechnique de Paris, F-91120 Palaiseau, France*

[h] *Institut für Experimentalphysik, Physics Department, University of Hamburg, Luruper Chaussee 149, 22761 Hamburg, Germany*

[i] *Department of Physics and Astronomy, University of Iowa, 203 Van Allen Hall, Iowa City, IA 52242-1479, U.S.A.*

[j] *Max-Planck-Institut für Physik, Föhringer Ring 6, D-80805 Munich, Germany*

E-mail: erika.garutti@uni-hamburg.de

[*] Corresponding author.

[1] Now at Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37830, U.S.A.

[2] Now at Instituto de Física Corpuscular, Parque Científico, Catedrático José Beltrán, 2 | E-46980 Paterna, España.

[3] Also at Institute of Physics, The Czech Academy of Sciences.

[4] Deceased.

https://doi.org/10.1088/1748-0221/19/10/P10027

Abstract: To achieve state-of-the-art jet energy resolution for Particle Flow, sophisticated energy clustering algorithms must be developed that can fully exploit available information to separate energy deposits from charged and neutral particles. Three published neural network-based shower separation models were applied to simulation and experimental data to measure the performance of the highly granular CALICE Analogue Hadronic Calorimeter (AHCAL) technological prototype in distinguishing the energy deposited by a single charged and single neutral hadron for Particle Flow. The performance of models trained using only standard spatial, energy and charged track position information from an event was compared to models trained using timing information available from AHCAL, which is expected to improve sensitivity to shower development and, therefore, aid in clustering. Both simulation and experimental data were used to train and test the models and their performances were compared. The best-performing neural network achieved significantly superior event reconstruction when timing information was utilised in training for the case where the charged hadron had more energy than the neutral one, motivating temporally sensitive calorimeters. All models under test were observed to tend to allocate energy deposited by the more energetic of the two showers to the less energetic one. Similar shower reconstruction performance was observed for a model trained on simulation and applied to data and a model trained and applied to data.

Keywords: Large detector systems for particle and astroparticle physics; Pattern recognition, cluster finding, calibration and fitting methods; Performance of High Energy Physics Detectors

2024 JINST 19 P10027

Contents

1 Introduction **1**

2 Methods and tools **2**
- 2.1 CALICE AHCAL 2
- 2.2 Neural network models 4
- 2.3 Datasets and training 4

3 Results **9**
- 3.1 Reconstruction quality and confusion distribution 9
- 3.2 Fraction of energy reconstructed within calorimeter resolution 15

4 Conclusion **20**

A Summaries of neural networks **21**

B Validation of synthetic neutral hadron showers **24**

C Producing events with a charged and synthetic neutral shower **25**

1 Introduction

A challenging final state jet-energy resolution must be achieved to fulfil the requirements for BSM physics searches and Higgs precision measurements at future linear colliders. For example, for ILC operating at centre-of-mass $\sqrt{s}$= 0.5−1 TeV where typical di-jet energies for interesting physics processes will be in the range 150–350 GeV, a relative jet energy resolution of 2.7 % is crucial [1]. Particle Flow (PF) is a method expected to provide this resolution, which relies upon accurate tracking of charged particles in a jet, sophisticated event reconstruction techniques, and highly granular sampling calorimeters. A prototype of such a detector is the CALICE Analogue Hadronic Calorimeter (AHCAL) [2], a highly-granular steel-scintillator sampling calorimeter designed for PF, with $24 \times 24 \times 38$ individual readout cells. The AHCAL is notable for its capacity to measure a timestamp for each readout channel.

Accurate energy clustering algorithms must exploit highly granular calorimeters as part of PF for future linear colliders to achieve this challenging jet energy resolution. Pandora Particle Flow Algorithm (PFA) [1] is an example of a clustering algorithm for resolving the energy deposits of particles. Explicitly, one of the fundamental tasks of a PFA is the clustering of charged and neutral energy deposits using event information. Furthermore, it has been demonstrated in ref. [1] that the main contributing factor to jet energy resolution for jet energies greater than 50 GeV using Pandora PFA is the confusion between the energy deposits of particles. Thus, it is scientifically important to minimise confusion by improving techniques for clustering energy deposits from hadron showers. Superior pattern recognition of energy deposits in highly granular calorimeters can achieve this.

2024 JINST 19 P10027

Machine learning models provide a powerful tool for developing a bespoke shower separation algorithm using event information. Graph neural networks have been found to provide superior shower separation performance than traditional convolutional neural networks for PF, owing to learning a representation of detector geometry, as demonstrated in ref. [3]. However, in ref. [3], there are several limitations concerning the AHCAL. Explicitly, the sensor density used is more than an order of magnitude smaller than for the AHCAL and the influence of timing information has not been assessed for shower separation.

More recent studies on machine learning-based PF, such as in ref. [4] and ref. [5], also do not include timing information as part of the reconstruction. This further motivates complementary research to assess the possible improvements in PF clustering using this additional observable.

The present study evaluated the performance of three published neural network models for hadron shower separation in the AHCAL between a simultaneous charged and synthetic neutral hadron shower produced using data synthesis techniques. These models use information from the event, consisting of hits in the AHCAL sensor array, the track position of the charged particle and its momentum, which are standard inputs to state-of-the-art PFAs. The models aim to predict the fraction of its energy that belongs to either the charged or neutral shower for each hit. The algorithms' performance was then tested using simulation and experimental data, using models trained with and without timing information.

As an important caveat, this study focuses on the effectiveness of clustering energy deposits from a charged and neutral hadron shower with the AHCAL detector. It neither assesses the effectiveness of the track-cluster association required to 'label' individual energy deposits as either charged or neutral nor the effectiveness of determining the number of simultaneous hadron showers in an event. Results should be interpreted with these caveats in mind.

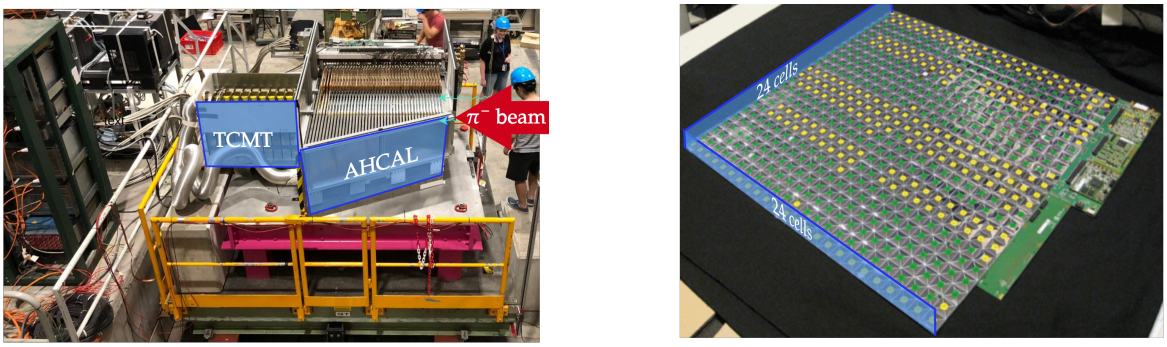
This paper's superscripts Q and N indicate variables associated with charged and neutral hadrons or showers. The variables E and $\widehat{E}$ indicate energy measured by the calorimeter and the value reconstructed by the neural networks. For instance, E^N_{sum} and $\widehat{E}^N_{\text{sum}}$ denote the total neutral energy measured by the calorimeter and the total value reconstructed by the neural networks. Additionally, the predicted and true fractions of energy belonging to a shower in a particular hit of the calorimeter are denoted $\widehat{f}_{\text{hit}}$ and f_{hit}, respectively. Lateral distances between showers are presented in Moliere radii for AHCAL, $\rho_{\text{M}} = 24.9\,\text{mm}$ [6]. Unless otherwise specified, 'data' and 'simulation' are taken to mean experimentally obtained and simulated hadron shower events, respectively.

2 Methods and tools

2.1 CALICE AHCAL

The CALICE AHCAL is a non-compensating steel-scintillator calorimeter prototype designed for future precision e^+-e^- collider experiments. It has a highly granular structure, consisting of $24\times24\times38$ plastic scintillator cells of $30\times30\times3\,\text{mm}^3$ volume each, read out by individual silicon photomultipliers (SiPMs). These cells indicate the spatial position, magnitude and timestamp of energy deposition with an optimal time resolution of 100 ps allowed by the hardware. The detector has a depth of approximately 4.2 nuclear interaction lengths (λ_{I}). The AHCAL has 38 layers, each with a total depth of 26.1 mm, of which 17.2 mm is steel absorber. The hadronic calorimeter is complemented by a steel-scintillator Tail Catcher/Muon Tracker (TCMT) detector, composed of 320 extruded scintillator strips of $1000\times50\times5\,\text{mm}^3$ volume packaged in $16\times1\,\text{m}^2$ planes interleaved between steel plates corresponding to an additional depth of 1.1 λ_{I} [7].

2024 JINST 19 P10027



(a) (b)

Figure 1. Pictures showing the CALICE AHCAL at testbeam. figure 1(a) shows the detector setup for a testbeam performed in June 2018 at the Super Proton Synchrotron (SPS) at CERN, Geneva. Figure 1(b) shows one of the 38 layers, with the individual cells of the calorimeter wrapped in foil to improve photon sensitivity. Reproduced from [10]. The Author(s). CC BY 4.0.

The TCMT is not used in this analysis. Pictures of the AHCAL calorimeter are shown for reference in figure 1. The AHCAL is intended to be part of a complete calorimetric system, including an electromagnetic and hadronic section. The electromagnetic section is not included in this study. This limitation does not prevent relevant studies into algorithms and developments relevant to the hadronic section.

Event information from AHCAL consists of a list of hits (i.e. active cells for which the energy is detected above a noise threshold). The position of a hit indicates the location of an energy deposit in the AHCAL cell matrix (I_{hit}, J_{hit}, K_{hit}). I_{hit} and J_{hit} indicate the lateral spatial position of a hit relative to the longitudinal axis of the calorimeter (I_{hit}, $J_{\text{hit}} \in [1, 24]$ in units of cell index). The longitudinal spatial position (depth in layers) is denoted K_{hit} ($K_{\text{hit}} \in [1, 38]$ in units of layer index). The energy (E_{hit}) is measured in units of MIP, the energy deposited by a minimum-ionising particle in a single layer. E_{hit} is first recorded in Analogue-to-Digital counts and then later converted to the scale of the energy deposited by a minimum ionising particle (MIP) in one cell [8]. E_{hit} takes a value between a noise threshold of 0.5 MIP and the energy corresponding to the SiPM saturation value. The timestamp (t_{hit}) is defined by the first time at which the electronics signal, proportional to the energy, crosses a given threshold relative to an external trigger. It is then converted to nanoseconds based on a TDC voltage ramp. The ramp's pedestal, maximum value, and time between them are calibrated for each SPIROC2E readout chip of AHCAL [9] and are used to reconstruct the time value in nanoseconds relative to the trigger. A second deposit later than a few nanoseconds from the first yields no output. Smearing due to electronic noise can result in timestamps less than 0 ns. This study considers the ultimate 100 ps timing resolution for AHCAL. No charge integration gate length is considered in this study. The calorimeter response is measured as the sum of the individual hits in an event, $E_{\text{sum}} = \sum^{\text{event}} E_{\text{hit}}$. Additionally, the incident position of a charged particle in lateral coordinates is reconstructed using four delay wire chambers (DWC) of $10 \times 10\,\text{cm}^2$ size, which is denoted as a vector $[I_{\text{track}}, J_{\text{track}}]$ [10]. The track position of the charged particle entering the calorimeter is a critical input to the shower separation model. Finally, the energy-weighted mean spatial position of the hadron shower in spatial coordinates is defined as a vector called 'centre-of-gravity' ($\text{CoG} = [\text{CoG}_I, \text{CoG}_J, \text{CoG}_K]$). The shower starting depth is K_S. K_S is calculated using a algorithm described in ref. [11]. Finally, a hit radius is defined, where $R_{\text{hit}} = \sqrt{(I_{\text{hit}} - \text{CoG}_I)^2 + (J_{\text{hit}} - \text{CoG}_J)^2}$, measured in cell units, which describes the distance of an hit to the shower axis.

2024 JINST 19 P10027

2.2 Neural network models

Three neural network models were implemented to assess shower separation for the AHCAL: `PointNet` [12], Dynamic Graph Convolutional Neural Network (`DGCNN`) [13], and `GravNet` [3]. Only `GravNet` is designed explicitly for PF shower separation of these networks. These neural networks were chosen because they support 'point clouds', a set of sampled points in Cartesian space, a natural representation of a hadron shower in the AHCAL. Explicitly, each active sensor is defined as a point in Cartesian space, $[I_{\text{hit}}, J_{\text{hit}}, K_{\text{hit}}, \log E_{\text{hit}}, \operatorname{arcsinh} t_{\text{hit}}]$, where $\operatorname{arcsinh} t_{\text{hit}}$ is optional. The transformations of the active hit energy and hit time are used because the distributions of these variables are highly skewed, which makes them poorly suited to machine learning. In particular, the inverse hyperbolic sine transformation is used to perform a log-like transformation for time information with the possibility of handling smearing for negative values.

Details of the models can be found in appendix section A and the provided references. The fundamental differences between the models are as follows. `PointNet` uses a 'global' approach to clustering, exchanging information without considering the local relationships between the individual hits. By contrast, using a dynamically updated graph, `DGCNN` and `GravNet` exploit local energy distributions and the relationships between hits. `DGCNN` directly constructs the graph from the hits, while `GravNet` projects the hits to a subspace, using these auxiliary hits to cluster. The advantage of `PointNet` is that it is computationally faster than `DGCNN` or `GravNet`, which require sequential k-NN clustering as part of the model design. The advantage of `DGCNN` and `GravNet`, which are broadly similar in overall design, is that the model can more readily learn local distributions of energy and the relationships between hits, which is expected to result in superior clustering performance compared to `PointNet`.

The overall designs of the neural networks were modified from the references to reflect the structure of a Pandora PFA algorithm. Explicitly, the first stage of clustering is performed using the positions of the hits only, $[I_{\text{hit}}, J_{\text{hit}}, K_{\text{hit}}]$, to build a spatial representation of the shower. Track clustering is then encouraged in the next stage by supplying the charged track position and transformed hit energy and hit time $\left[\log E_{\text{hit}}, \operatorname{arcsinh} t_{\text{hit}}, I^{Q}_{\text{track}}, J^{Q}_{\text{track}}\right]$. In the final stages of each model, track energy (E^{Q}_{track}) and unmodified hit energy information, $\left[E_{\text{hit}}, E^{Q}_{\text{track}}\right]$, is added to encourage a 'statistical re-clustering', involving the aggregation of the energy of a cluster and associating it to a charged track energy. This step is used in Pandora PFA to improve performance for jet energies $E_j > 50\,\text{GeV}$, where it is expected that there will be significant confusion between hadron showers. The maximum, mean, and variance were used as aggregation functions in the networks. In certain cases, additional fully connected layers were added to condense the output where necessary. The final output of each network is two sets of fractions, f^{Q}_{hit} and f^{N}_{hit}, predicting the fraction of energy belonging to Q and N respectively. The sum of f^{Q}_{hit} and f^{N}_{hit} is one (i.e. either the energy belongs to Q or N). Each network was designed to have around 2 million weights.

2.3 Datasets and training

2.3.1 Raw datasets

Both simulation and experimental data were used for training and evaluation, respectively. Single π^- hadron shower events observed with the AHCAL detector were studied. The simulation of the particle showers was achieved using `Geant4` [14], with a full detector simulation developed using `DD4hep` [15]. Additional effects, such as digitisation of the analogue signal and reconstruction of the detector

2024 JINST 19 P10027

variables, were achieved for both simulation and data using `CALICESoft` [16]. Timing information from experimental data is not studied due to comparatively poor timing resolution arising from chip occupancy effects [17]. Useful insights may nonetheless be obtained using timing information in simulation. A MIP-to-GeV calibration factor of 37.3 MIP/GeV was used [6]. The statistics of the training, validation and test datasets are shown in table 1.

Table 1. Number of events used for training shower separation models and performing analysis after all cuts. An additional sample of 40 GeV K_L^0 hadrons simulated under the same conditions as the π^- hadrons is incuded for analysis. Hyphens indicate 0 events.

Hadron Type	K_L^0	π^-					
	Simulation	June 2018 SPS Testbeam Data			Simulation		
Purpose	Analysis	Testing	Training	Validation	Testing	Training	Validation
Particle Energy [GeV]							
5	-	-	-	-	40685	36966	4108
10	-	21396	171166	21396	68812	62333	6926
15	-	-	-	-	75224	67379	7487
20	-	32221	257762	32220	77759	70866	7874
25	-	-	-	-	81379	73023	8114
30	-	-	-	-	80971	74180	8243
35	-	-	-	-	78646	75619	8403
40	78146	34428	275424	34428	77055	76142	8461
45	-	-	-	-	73620	76994	8555
50	-	-	-	-	86014	77430	8604
55	-	-	-	-	63218	77881	8654
60	-	44600	356799	44600	72306	78550	8728
65	-	-	-	-	82256	78779	8754
70	-	-	-	-	59806	79042	8783
75	-	-	-	-	49390	79417	8825
80	-	37790	302315	37789	69106	79713	8858
85	-	-	-	-	70091	79848	8872
90	-	-	-	-	80773	79918	8880
95	-	-	-	-	62647	79614	8846
100	-	-	-	-	54433	79292	8811
105	-	-	-	-	53377	77300	8589
110	-	-	-	-	52632	77853	8651
115	-	-	-	-	58750	75091	8344
120	-	34881	279044	34880	56274	74835	8316
Total Events	78146	205316	1642510	205313	1625224	1788065	198686

Data and simulation were subject to the following cuts:

- events were required to be identified using the standard CALICE particle identification algorithm [18] as being a single particle and having less than a 0.5 % probability of being a muon to exclude non-showering, 'punch-through' pions;
- the 38$^{\text{th}}$ layer of the AHCAL is ganged and requires special treatment beyond the scope of this paper. Therefore, energy deposits were considered up to the 38$^{\text{th}}$ layer of the calorimeter;

2024 JINST 19 P10027

- events were required to have a correctly reconstructed track position (i.e. a track position with a corresponding position inside the 24×24 cell AHCAL front-face).

- events were required to have at least 50 hits after the MIP-track cut discussed in the following section. This criterion reduces the influence of partially showering punch-through pions, which may initiate a small cascade and continue through the calorimeter.

2.3.2 Data synthesis and datasets

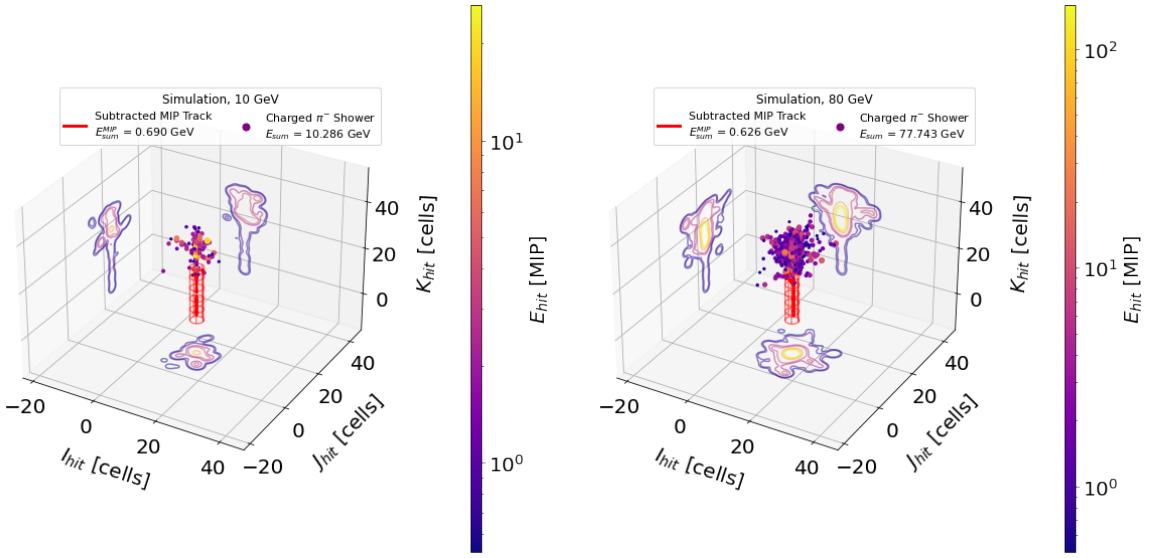
Only single π^- hadron showers from experimental data are available for AHCAL. Therefore, a method is required to produce 'synthetic' showers with two showers, one charged and one neutral. The method must be consistent between the simulation and the data to be compared.

One of the main differences between charged and neutral hadron showers is that charged particles ionise the detector medium before a shower initiates. A set of highly localised, rectilinear, Landau-distributed energy deposits distributed along the axis of motion of the particle before showering is expected for charged particles. This is called a 'MIP-track'. Therefore, synthetic neutral hadron showers can be produced by applying criteria developed in ref. [19] to select and remove the MIP-track, referred to as the 'MIP-track cut'. The cut selects hits with $R_{\text{hit}} < 60\,\text{mm}$, $E_{\text{hit}} < 3\,\text{MIP}$ and $K \leq K_S - 2\,\text{layers}$. It is noted that the requirement on the hit radius biases the shower shape to symmetric hadron showers. However, only around 3 % of showers have a track that is further than 60 mm from the centre-of-gravity, which typically indicates poor track reconstruction. Therefore, the effect is minor and necessary. Removing the energy deposits satisfying the cut from the shower results in a synthetic neutral hadron shower. The MIP-track cut is performed for both simulation and experimental data in the study. The effect of this cut is shown in figure 2.

This method has clear limitations. For example, it subtracts energy from the shower that would have been part of the shower if initiated by a neutral particle of the same energy. Another limitation is partially showering hadrons, which deposit some energy but continue as MIPs through the calorimeter. Additionally, the likelihood of producing high-energy π^0s is lower for K_L^0 than for $\pi^\pm$ due to strangeness conservation, resulting in weaker calorimeter response to K_L^0 compared to $\pi^\pm$ of the same energy [20]. However, a study comparing simulated 40 GeV π^- with the MIP-cut applied and simulated 40 GeV K_L^0 hadron showers found them similar at the event level, making them acceptable replacements for neutral showers. Further study details are provided in appendix section B.

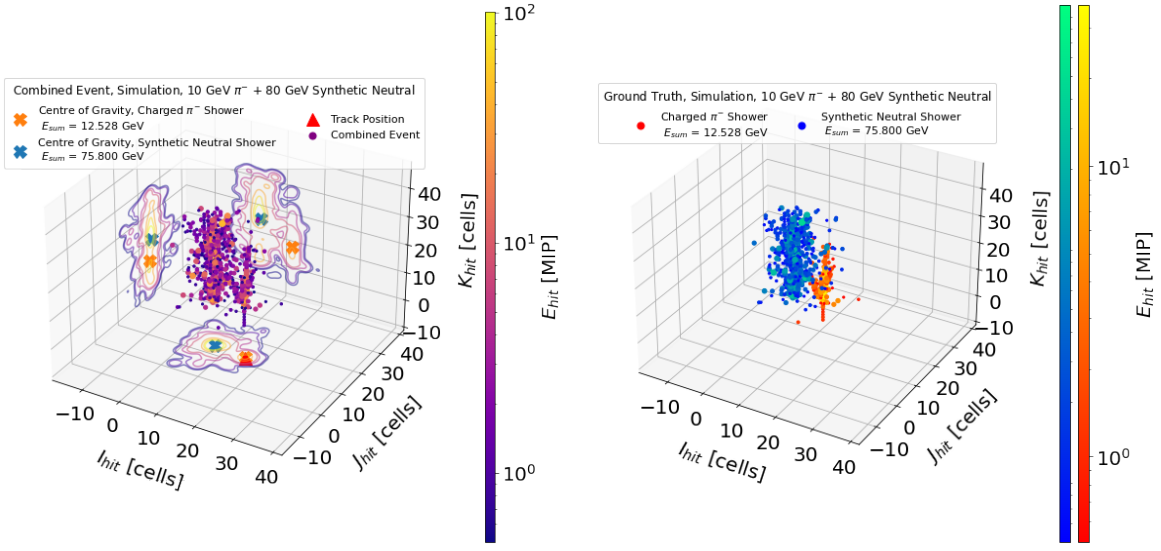
A pair of showers are then overlaid by shifting their lateral positions within a circle to create synthetic events. The circle's radius is based on the radial distance from the centre-of-gravity within which 80% of each shower's energy is contained and depends on the particle energy. Overall, Q and N have a most-probable distance of $5.5\,\rho_{\text{M}}$ apart. For each cell which has deposited energy from both Q and N, E_{hit} and t_{hit} are recalculated as the sum of the energies of the original showers and the earliest timestamp of the two hits. f_{hit}^Q and f_{hit}^N are calculated as the fraction of energy in the cell belonging to either Q or N. The specific details of the method, including how the average inter-shower distance for the data was selected, are discussed in appendix section C. Appendix figure 20 provides a flow-chart of the algorithm for reference. The result of the method is a combined shower with fractions of hit energy belonging to each shower, f_{hit}^Q and f_{hit}^N, that are to be reconstructed by the neural networks. Example event displays of the combined and ground-truth separated hadron showers produced using this method are shown in figure 3.

2024 JINST 19 P10027



(a) (b)

Figure 2. Event displays of simulated π^- hadron showers demonstrating the MIP cut applied to two examples from the training sample. Each axis represents the spatial coordinates of the calorimeter, and the purple disks indicate E_{hit} in a logarithmic scale, both in colour and size. The purple contours on each calorimeter face indicate a smoothed 'energy shadow' of the hadron shower to indicate its profile. Selected MIP-track hits are shown as red circles. The red cylinder indicates the cut region in space. The energy criterion cannot be shown. Figures 2(a) and 2(b) show a 10 GeV and 80 GeV π^- hadron shower, respectively.



(a) (b)

Figure 3. Figure 3(a) shows an event display of two overlayed simulated showers, one 10 GeV charged and 80 GeV synthetic neutral, where track position and centres of gravity of the charged and neutral shower indicated by a red triangle, and an orange and blue cross, respectively. Figure 3(b) shows the same event as figure 3(a), with the individual showers identified by colour. The red and blue points indicate the simulated charged π^- hadron shower and a synthetic neutral hadron shower, respectively. Else, as in figure 2.

2024 JINST 19 P10027

For simulation, 7.2×10^5 and 8×10^4 synthetic charged-neutral hadron showers with an average of 1250 ± 37 events and 139 ± 12 events per particle energy combination were produced for training and validating the neural networks during the training phase, respectively, while 8×10^5 events were produced to test the models, with an average of 1389 ± 39 events per particle energy combination. Each combined sample contained showers purely from the corresponding source samples outlined in table 1. The same number of events for training and validation samples were chosen for data. However, a smaller sample of 2×10^5 events were used for testing than for simulation, owing to a smaller sample of available test events. Also, the average number of events per particle energy for training, validation, and testing in data was 20000 ± 140 events, 2222 ± 40 events and 5556 ± 81 events, respectively. The number of events in data is higher than for simulation due to fewer available energy combinations.

2.3.3 Training

For simulation, two independent neural networks based on each model defined in section 2.2 were trained with and without timing information. For data, a single neural network with the best performance in the simulation was trained without timing information and tested on data. Another instance of the same model was then trained and tested on data. The networks were developed in `PyTorch` [21] and trained using the `PyTorch Lightning` research framework [22] on an NVidia V100 GPU. The ADAM optimiser was used to improve the convergence rate for ten epochs. The hyperparameters used for training are shown in table 2, selected based on the results of a parameter scan using Optuna hyperparameter optimisation framework [23]. It is noted that the γ parameter of `GravNet` was also varied as a hyperparameter, which was not done in ref. [3]. The ADAM hyperparameters β_1 and β_2 were held at nominal values of 0.9 and 0.999 respectively, as tuning these parameters resulted in instabilities during optimisation. The batch size was fixed at 32 samples per batch and was not optimised due to memory limitations.

The loss was chosen to be the same as in the study of ref. [3]. This study applied a square-root energy-weighted mean square loss during training to encourage the models to cluster the most energy-dense parts of the shower correctly. However, to reduce the influence of 'shower-swapping' (i.e. when the neural network correctly separates the two showers but compares them to incorrect permutations during training and evaluation, see ref. [3] for more details), a permutation-invariant training approach was adopted.

Table 2. Hyperparameters used to train the neural network. In this table, β_1 and β_2 are the ADAM momentum parameters, p_{dropout} is the dropout probability, and k is the number of nearest-neighbours per cluster. Hyphens indicate hyperparameters that do not apply to the model. The parameters, excluding β_1 and β were informed by a hyperparameter scan using Optuna [23].

Parameter	PointNet, no Time	PointNet, + Time	GravNet, no Time	GravNet, + Time	DGCNN, no Time	DGCNN, + Time
Learning Rate	2.567×10^{-4}	5.681×10^{-5}	2.012×10^{-4}	5.169×10^{-4}	1.252×10^{-5}	1.660×10^{-4}
p_{dropout}	0.332	0.259	0.268	0.469	0.167	0.164
γ	-	-	8.137	12.815	-	
k	-	-	16	24	15	18
β_1	0.9	0.9	0.9	0.9	0.9	0.9
β_2	0.999	0.999	0.999	0.999	0.999	0.999

2024 JINST 19 P10027

The loss is shown in eq. (2.1):

$$\mathcal{L}(\widehat{f}_{\text{hit}}; f_{\text{hit}}, E_{\text{hit}}) = \sum_{i=0}^{\{Q,N\}} \frac{\sum_{\text{event}} \sqrt{E_{\text{hit}} \cdot f_{\text{hit}}^{i}} \cdot (\widehat{f}_{\text{hit}}^{i} - f_{\text{hit}}^{i})^2}{\sum_{\text{event}} \sqrt{E_{\text{hit}} \cdot f_{\text{hit}}^{i}}} \tag{2.1}$$

where $\{Q, N\}$ is the set of possible hadron showers, Q and N are charged and neutral hadrons inducing showers in the AHCAL, and i is the index of each set element.

The loss is then found for each possible permutation of outputs from the network by re-ordering the output from each network, which could be either $\{f_{\text{hit}}^{Q}, f_{\text{hit}}^{N}\}$ or $\{f_{\text{hit}}^{N}, f_{\text{hit}}^{Q}\}$. The minimum loss of the of permutations is then taken. This consideration helps reduce the influence of 'shower-swapping' on the network's performance during training.

The results must be interpreted with the following biases in mind:

- the study focuses solely on scenarios involving one charged and one neutral hadron shower, without considering an unknown number of showers;
- there is ambiguity in the order of the output of each model to overcome the confusion of 'shower-swapping' (i.e. the model does not label each shower as Q or N, but instead focuses entirely on clustering the energy deposits);
- the distance distribution between incident hadrons is ad-hoc and chosen to train a robust and functional shower separation algorithm. In the ideal case, the inter-particle distance of a jet, which could be obtained from simulation and was not available for this study, would be used instead.

Finally, for the `PointNet` network only, the A matrix as defined in appendix section A is constrained to be close to an orthogonal matrix by adding a regularisation condition to the loss of eq. (2.1) to produce a modified loss function. It is shown in eq. (2.2) [12]:

$$\mathcal{L}_{\text{reg}}(A) = \left\| I - AA^{T} \right\|^2 \tag{2.2}$$

where I is the identity matrix, superscript T indicates the matrix transpose operation.

3 Results

In this section, the shower separation capabilities of each neutral network presented in section 2.2 are presented and analysed.

The neutral hadron shower, N, is considered the reference shower henceforth. Confusion energy is therefore defined as the predicted minus the true reconstructed calorimeter response:

$$E_{\text{confusion}}^{N}(\widehat{E}_{\text{sum}}^{N}; E_{\text{sum}}^{N}) = \widehat{E}_{\text{sum}}^{N} - E_{\text{sum}}^{N} \tag{3.1}$$

3.1 Reconstruction quality and confusion distribution

The energy distributions of reconstructed hadron showers are first evaluated for the original shower energy for the three separation models. The reconstruction quality for reconstructed neutral showers is evaluated using the mean and the most probable value (MPV) obtained from the confusion energy

2024 JINST 19 P10027

distribution, defined in Equation (3.1). The MPV is estimated using a kernel density estimate using `KDEpy` [24], with Silverman's binning rule utilised to determine the bandwidth [25]. A spline is then fitted to the estimate, and the MPV is determined by locating the root of the spline with the maximum probability density.

The confusion energy distributions of the implemented shower separation models are assessed, presenting the RMS_{90} and median absolute deviation (MAD) of each distribution. These statistics are used to compare their resilience to extreme outliers to standard deviation measures. RMS_{90} is defined as the minimum standard deviation within all possible central 90 % percentile ranges permitted by the data, and is commonly used in the Particle Flow community to measure calorimeter response spread. The MAD is defined as the median distance of the absolute deviations of the data to the median [26] and serves as an additional robust statistic commonly employed in statistical analysis.

Simulation. The reconstruction quality for each model applied to the testing dataset is depicted in figure 4. This figure illustrates that, in general, the models tend to reconstruct the neutral shower with energy levels close to the original. The MPVs of confusion energy, as shown in the subplots, exhibit differences of no more than 1.5 GeV. However, a bias is evident from the mean values in the subplot and the green regions in the main figures, indicating a tendency to overestimate the shower energy of neutrals below 60 GeV and underestimate it above that value. This bias is observed across all models under the test, evident from the steeper slope of the mean compared to the MPV of each distribution. The discrepancy between the MPV and mean, along with the asymmetry of the green and blue regions around shower energies of 60 GeV, indicates skewness in confusion distributions. Comparing the left column with the right column of figures reveals significant improvements in neutral hadron shower reconstruction for `DGCNN` and `GravNet` when incorporating 100 ps time resolution, as evidenced by the narrowing of the distributions around the purple line. Conversely, `PointNet` shows no such improvement.

The confusion energy distributions for each model under test are depicted in figure 5. Comparing figure 5(a) with figures 5(b)–5(c) reveals that the `PointNet` model performs similarly in reconstructing hadron showers, regardless of whether timing information is utilised. In contrast, both the `DGCNN` and `GravNet` models exhibit significantly less confusion compared to `PointNet` when timing information is provided to these models. This improvement in performance could be attributed to the ability of graph neural networks (`DGCNN` and `GravNet`) to exploit local geometric structures (i.e. patterns of energy density in space and time), unlike `PointNet`, which treats energy deposits independently [13]. Notably, a slight positive skewness in the distribution is observed for `DGCNN` and `GravNet` models (i.e. where the MPV of the distribution is displaced from 0), without timing information. The reasons for this are unknown.

Overall, the `GravNet` model performed the best of the models under test, both with and without timing information. Without timing information, `GravNet` demonstrated less confusion and comparable performance to `DGCNN` with timing. Introducing 100 ps timing resolution led to a notable 23 % reduction in the median absolute deviation (MAD) of the confusion distribution for `GravNet`. The optimal 'potential scaling' parameter, γ, increased with timing information inclusion (see table. 2), indicating a more localised influence of individual cells in the shower and suggesting a more sophisticated clustering approach. It is noted that no strong correlations were observed between typical shower observables and the improvement due to timing information. Therefore, further research is needed to grasp the full influence of timing information on clustering and how

2024 JINST 19 P10027

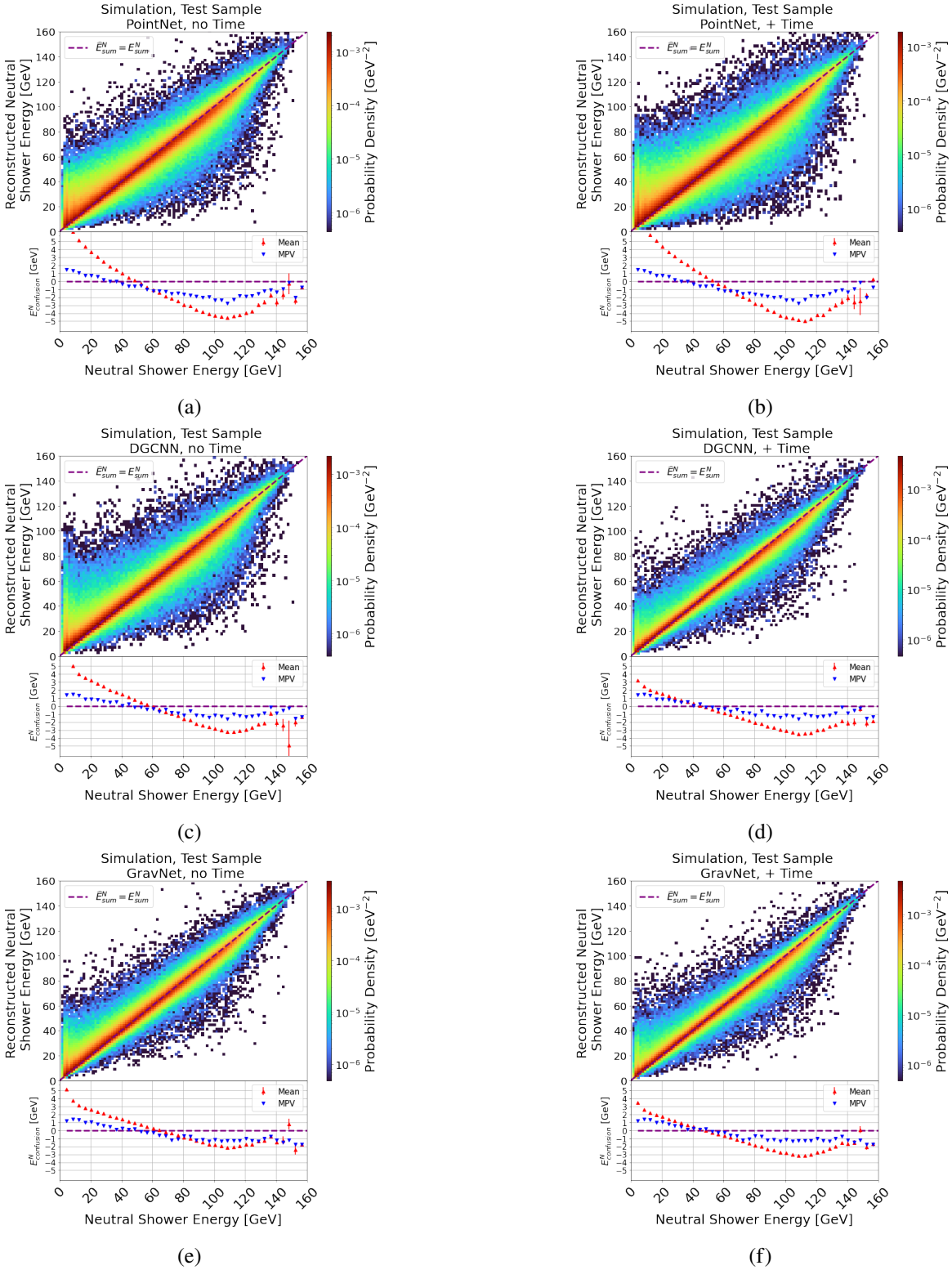


Figure 4. Figures 4(a), 4(c), 4(e) (left) and figures 4(b), 4(d), 4(f) (right) show the joint distributions of the predicted and true reconstructed neutral shower response for `PointNet`, `DGCNN` and `GravNet`, without and with timing information, respectively. The colour scale indicates probability density. The purple dashed line indicates perfect reconstruction. The bottom subplot shows the mean and MPV on the y-axis at each bin along the x-axis.

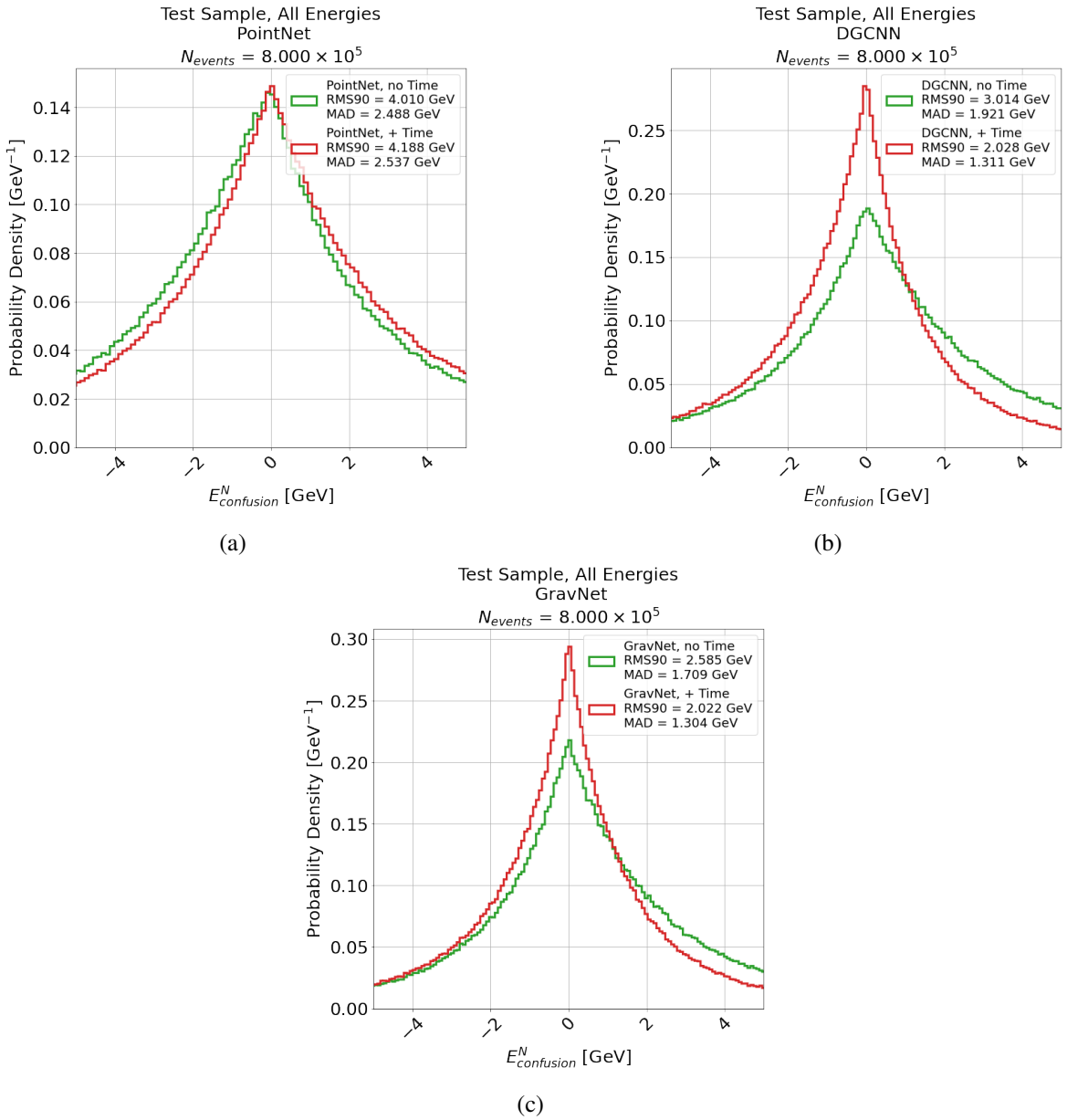


Figure 5. Distributions of neutral confusion energy for each shower separation model under test. The green and red lines indicate the same models, without and with timing information, respectively. RMS_{90} and MAD are shown in the legend for each model.

this is achieved. It is also noted that in a real jet, this improvement due to timing information will likely improve even further due to the delay of charged particles, which bend in a magnetic field and thus travel further to get to the calorimeter.

The skewness of confusion energy distributions for each tested model is displayed using the adjusted Fisher-Pearson standardised moment coefficient (G_1) [27], as depicted in figure 6.

All models exhibit positive skewness when the charged particle energy exceeds the neutral one, implying an overestimation of neutral shower energy. Similarly, underestimation is observed in the opposite case. This pattern suggests that all methods of shower separation lean towards an

2024 JINST 19 P10027

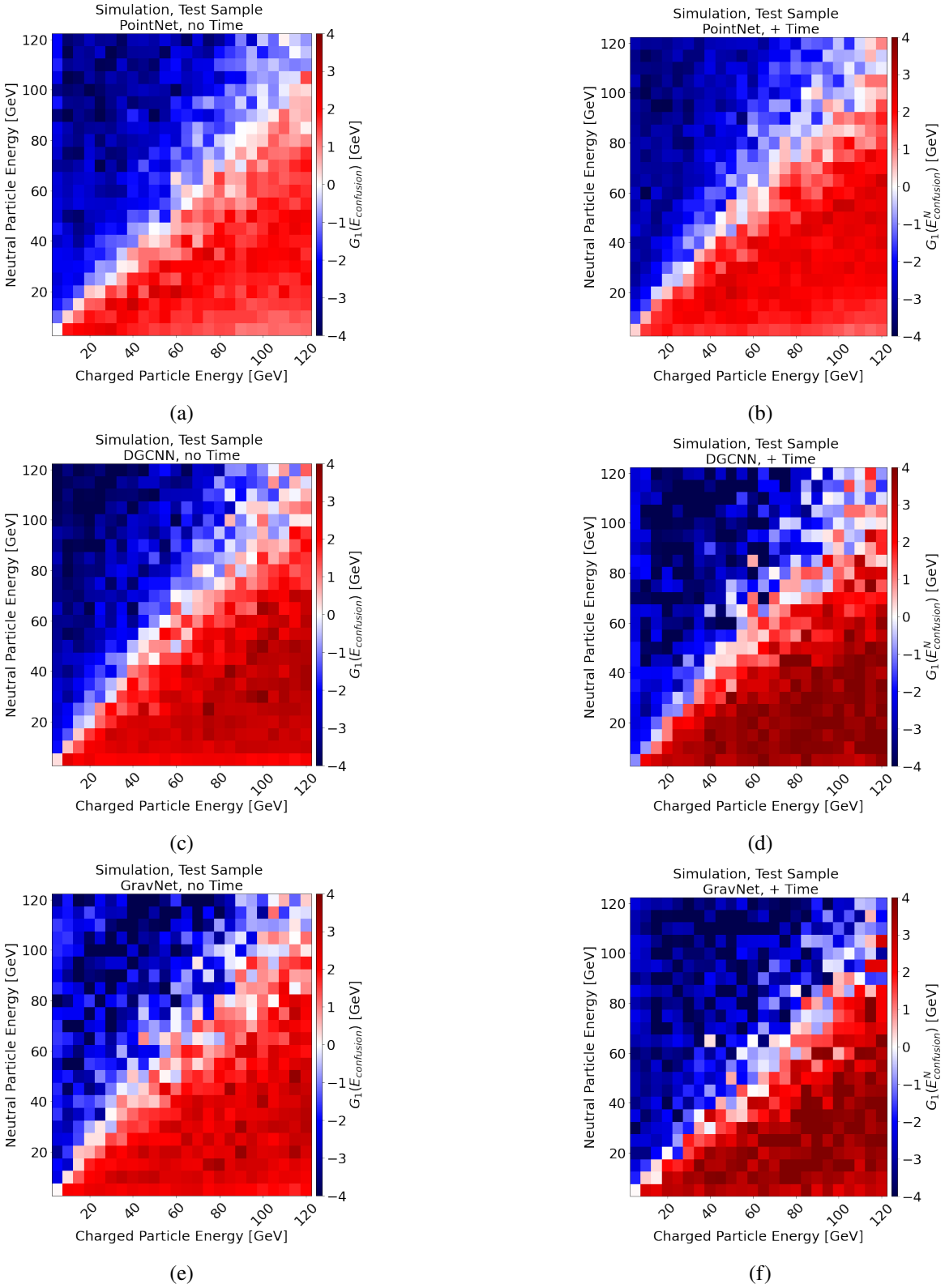
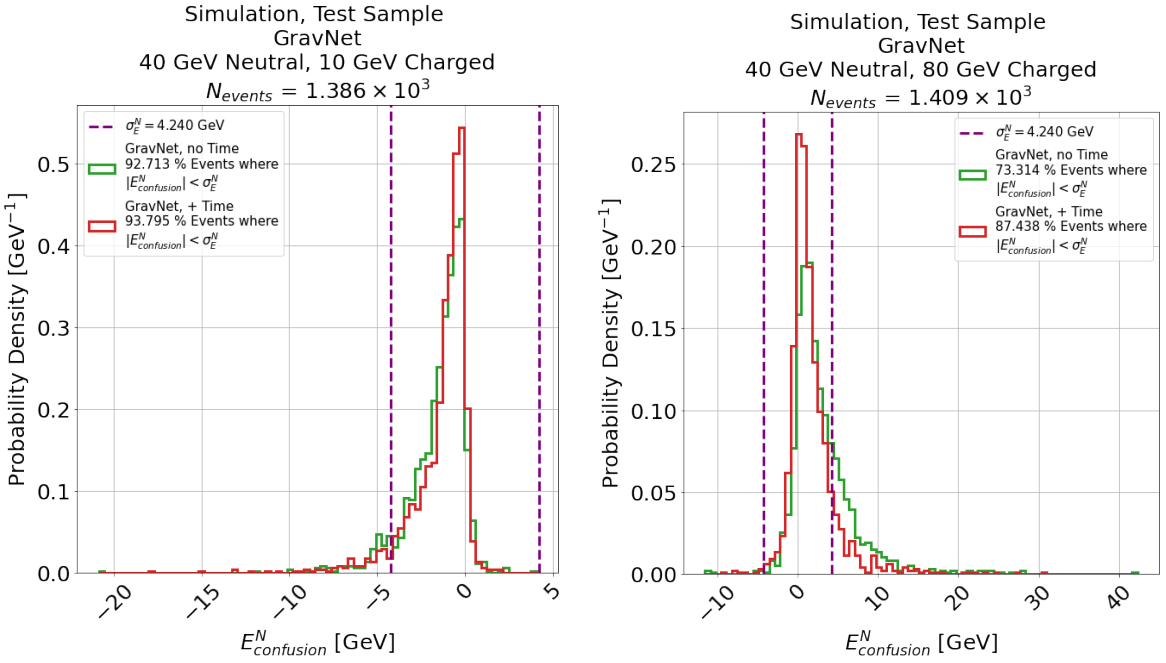


Figure 6. Skewness as a function of charged and neutral shower energy. The x and y-axes indicate the charged and neutral shower energy, respectively. The colour scale indicates skewness, G_1, where red and blue indicate positive and negative skewness, respectively.

2024 JINST 19 P10027



(a) (b)

Figure 7. Figures 7(a) and 7(b) shows a 40 GeV neutral shower separated from a 10 GeV and 80 GeV charged shower. The green and red lines indicate the models trained without and with timing information, respectively. The purple dashed lines indicate the resolution of the AHCAL calorimeter in simulation.

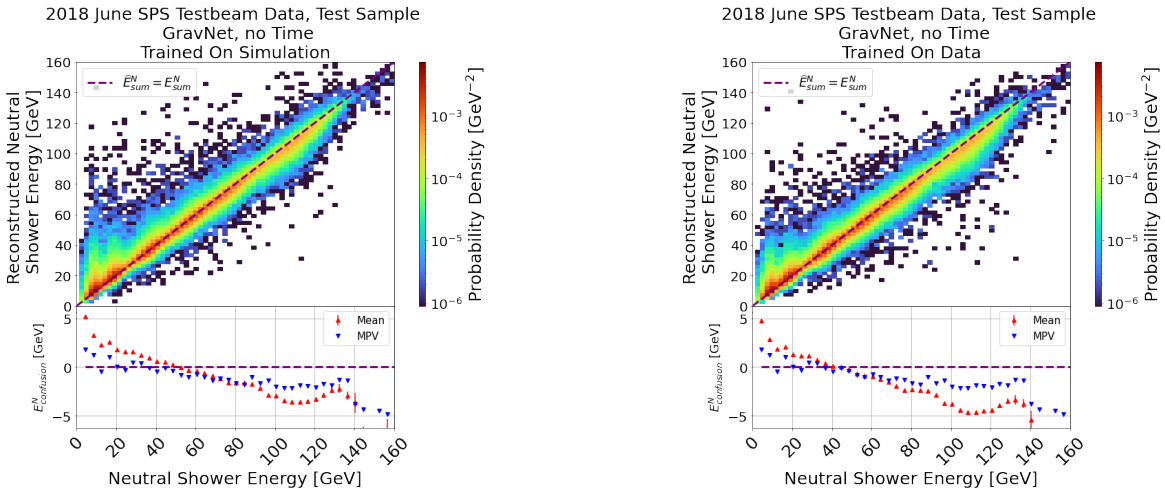
'altruistic' approach, where higher-energy showers transfer energy to lower-energy ones. A plausible explanation for this phenomenon is that redistributing energy from the highest-energy hadron shower to lower-energy ones during clustering is an optimal strategy. Notably, Pandora PFA, a clustering algorithm, tends to split true clusters during its initial stage rather than merging energy deposits from multiple particles into a single cluster [1]. Similar skewness in the confusion distribution has been independently observed in other studies involving Pandora PFA [19]. Further analysis is required to validate these hypotheses beyond the presented studies.

Two example distributions from `GravNet` are presented in figure 7 for further illustration.

2018 June testbeam data. The reconstruction quality, overall confusion distributions and skewness shown in figures 4, 5 and 6 are presented using the `GravNet` model evaluated on data in figures 8, 9 and 10. Two models are assessed: the model trained on simulation and a separate model trained on data.

The similarity of figures 8(a) and 8(b) and the green and brown lines in figure 9 indicate that the performance of the model trained on simulation and data achieve similar levels of confusion. This means models can be trained on simulation and applied to testbeam data with minimal difference in performance. Figure 10 shows the same behaviour of the skewness as previously discussed in figure 6, indicating the behaviour is consistent in simulation and data. The apparent difference in the magnitude of the values between these figures results from the skewness being sensitive to outliers, and no appreciable difference in the shape of the confusion energy distributions of simulation and data is observed.

2024 JINST 19 P10027



(a) (b)

Figure 8. Figures 8(a) and 8(b) show the joint distributions of the predicted and true reconstructed neutral shower response for `GravNet` trained on simulation and data and applied to the test sample of data, respectively. Else, as in figure 4.

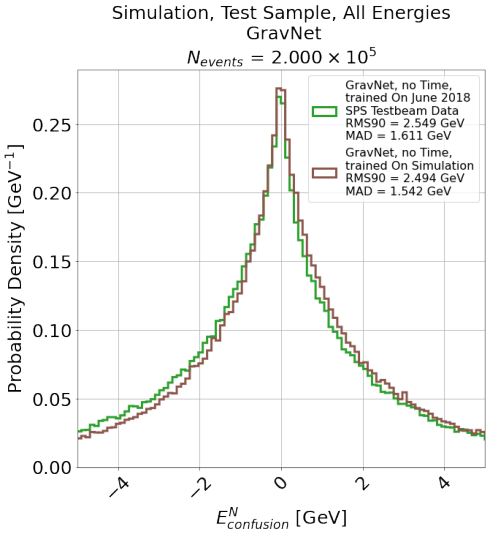


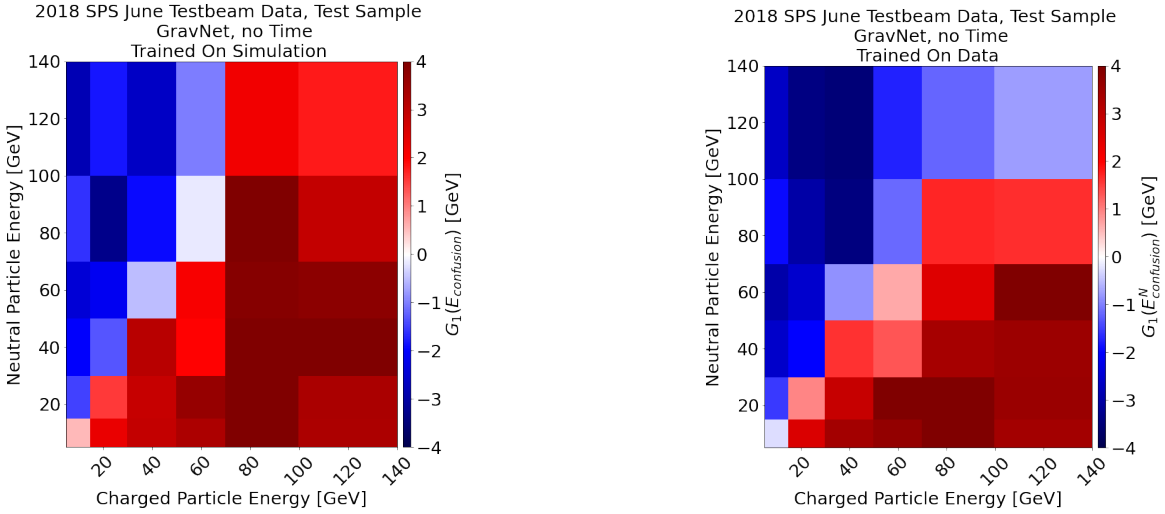
Figure 9. The distributions of the neutral confusion energy for `GravNet` applied to the test sample of data. The green and brown lines indicate the models trained on simulation and data, respectively. Else, as in figure 5.

3.2 Fraction of energy reconstructed within calorimeter resolution

For perfect separation of Q and N, the energy resolution gives the uncertainty on the hadron shower energy. The performance of shower separation can, therefore, be quantified by the fraction of showers with reconstructed energy within one standard deviation from the true one, where the standard deviation is the calorimeter resolution at a given energy, defined f_{rec} and given by eq. (3.2):

$$f_{\text{rec}} = \frac{N_{|E^N_{\text{confusion}}|<\sigma_E}}{N_{\text{events}}} \tag{3.2}$$

2024 JINST 19 P10027



(a) (b)

Figure 10. Figures 10(a) and 10(b) show the skewness of the neutral confusion energy resolution gives the uncertainty on the hadron shower energy. Else, as in figure 6.

where N_{events} are the total number of showers in a studied subsample of the test dataset, $|E^N_{\text{confusion}}| < \sigma_E$ is the condition to be satisfied, $N_{|E^N_{\text{confusion}}|<\sigma_E}$ are the number of showers satisfying that condition and σ_E is the calorimeter resolution, given by eq. (3.3):

$$\sigma_E = a \cdot \sqrt{E} \oplus b \cdot E \tag{3.3}$$

where E is particle energy, a describes the combined sampling and stochastic fluctuations experienced by the calorimeter and b the quality of detector calibration, non-uniformities in the signal collection, imperfections in calorimeter construction, etc., and $\oplus$ is an addition in quadrature.

The resolution terms a and b has been measured for simulation and data in ref. [28] with $a_{\text{sim}} = (49.5 \pm 0.4)\,\%/\sqrt{\text{GeV}}$, $b_{\text{sim}} = (7.1 \pm 0.1)\,\%$ and $a_{\text{data}} = (56.1 \pm 0.7)\,\%/\sqrt{\text{GeV}}$, $b_{\text{data}} = (6.1 \pm 0.1)\,\%$.

Simulation. f_{rec} was calculated for the test sample of simulation for all models under test. The results are shown as a function of the charged and neutral particle energy in figure 11.

Figures 11(a), 11(d), and 11(g), along with figures 11(b), 11(e), and 11(h), reveal an asymmetry in shower reconstruction performance based on whether the charged particle energy surpasses that of the neutral particle. When the neutral hadron possesses more energy than the charged one, from 80 % to well over 90 % of showers are reconstructed within the calorimeter resolution for `DGCNN` and `GravNet`. However, performance deteriorates in the opposite scenario. This result can be attributed to the relative magnitude of σ_E compared to $E_{\text{confusion}}$. To clarify, confusion was observed to be about the same for a particular set of particle energies, regardless of whether it was Q or N that deposited more or less energy. However, the σ_E of the calorimeter is smaller when N has less energy, which means f_{rec} increases in that case. The opposite is true when N has more energy.

Figures 11(f) and 11(i) demonstrate that `DGCNN` and `GravNet` achieve improvements of up to an additional 15–20 % of showers with the inclusion of timing information where the energy of the charged shower surpasses the neutral. The red region below the cyan equality line indicates this. Conversely,

2024 JINST 19 P10027

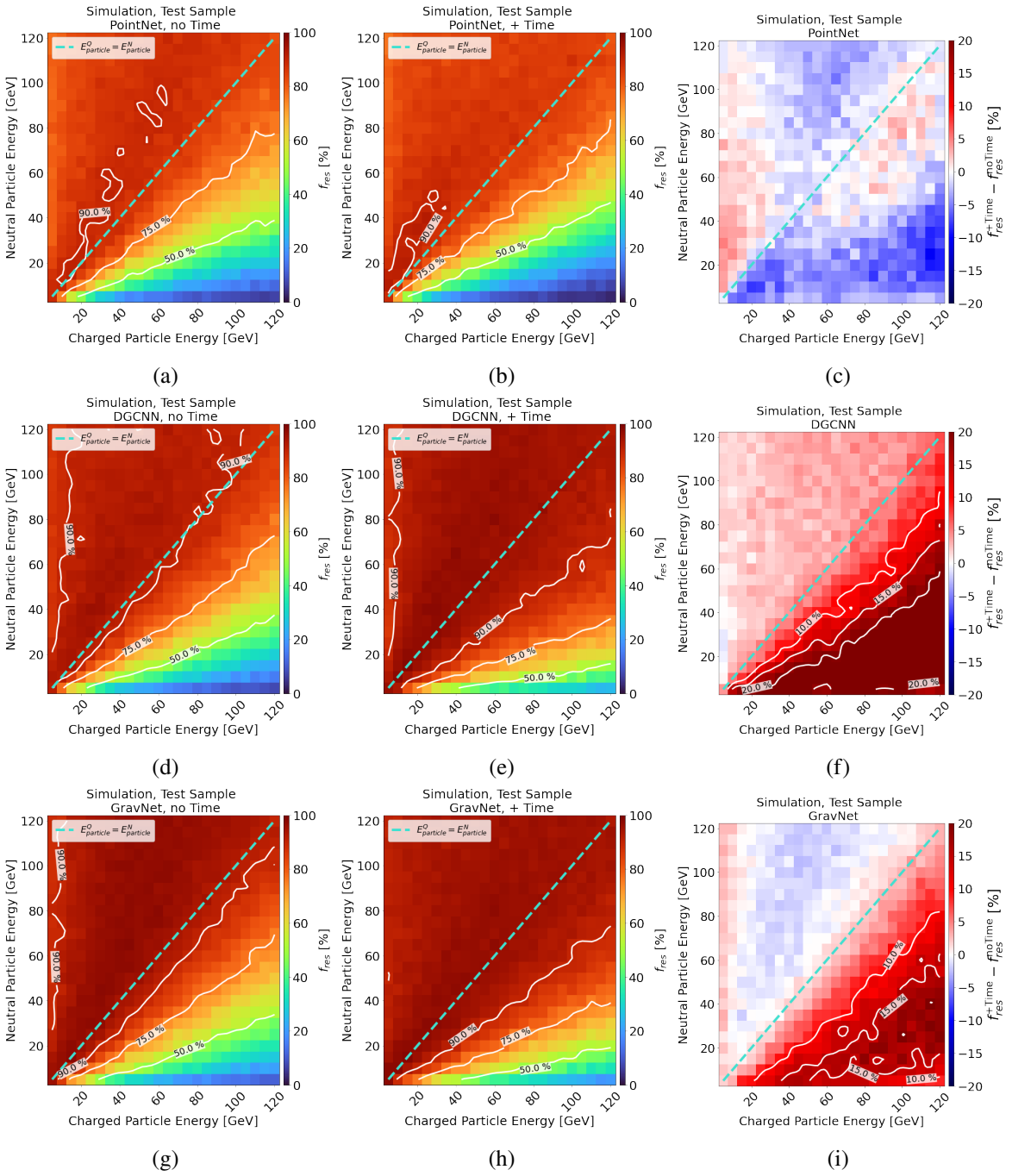
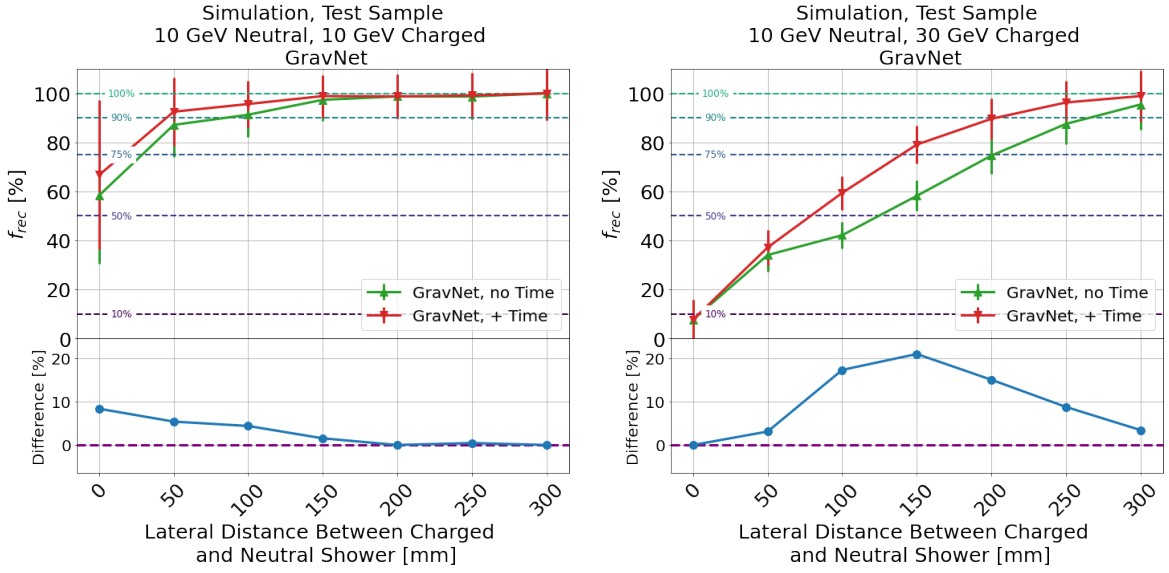


Figure 11. Figures 11(a), 11(d), 11(g) and figures 11(b), 11(e), 11(h) show the matrices of the fraction of showers with energy reconstructed within the calorimeter resolution as a function of the charged and neutral particle energy, for `PointNet`, `DGCNN` and `GravNet`, respectively, where red to blue indicates a higher to lower percentage of events. The turquoise dashed line indicates $E^{Q}_{\text{particle}} = E^{N}_{\text{particle}}$, while white lines indicate contours. Figures 11(c), 11(f) and 11(i) indicate the ratios of the fractions for models trained with and without timing information.

2024 JINST 19 P10027



(a) (b)

Figure 12. Figures 12(a) and 12(b) show f_{rec} (see eq. (3.2)) as a function of the distance between showers Q and N in mm for the test sample of simulation with the `GravNet` model, with $E_Q = E_N = 10\,\text{GeV}$ and $E_Q = 30\,\text{GeV}$, $E_N = 10\,\text{GeV}$, respectively. The red and green lines indicate the models trained with and without time, respectively. The subplots with the blue lines indicate the additional fraction of showers, in percent, reconstructed by the model trained with time than without.

no notable improvement is observed in the opposite scenario. `PointNet`, as depicted in figure 11(c), does not exhibit such enhancements. This result indicates that a more sophisticated clustering method is obtained using timing information than without, particularly in the absence of track information.

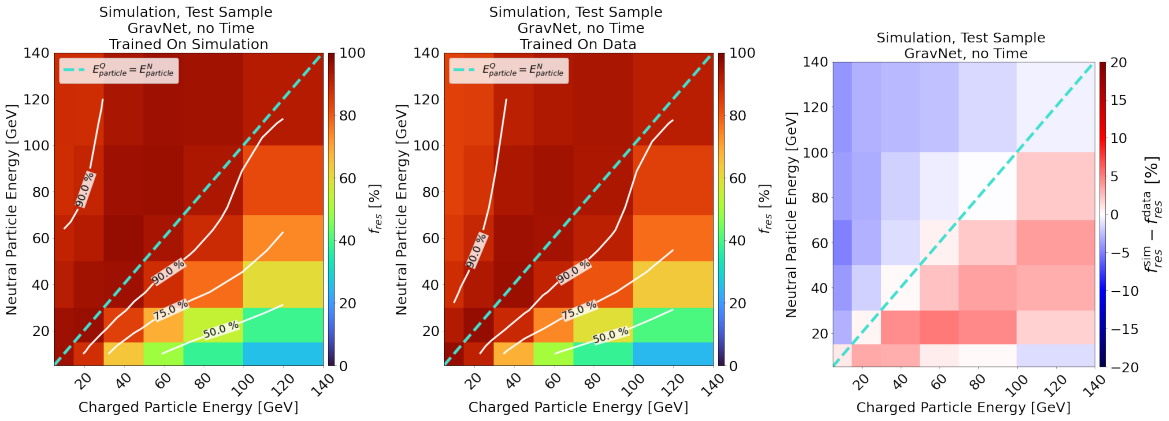
Figure 12 shows the correlation between f_{rec} and the lateral distance between Q and N, highlighting the enhancement achieved by incorporating timing data into the `GravNet` network for two benchmark PF separation scenarios. Specifically, figure 12(a) shows cases where the particle energies are $E_Q = E_N = 10\,\text{GeV}$, while figure 12(b) represents $E_Q = 30\,\text{GeV}$ and $E_N = 10\,\text{GeV}$. Figure 12(a) shows marginal and statistically insignificant improvement by including timing information when the energies of Q and N are comparable. In contrast, figure 12(b) demonstrates a notable and statistically significant improvement, particularly at distances ranging from 50–250 mm, with an increase of up to 20 % in the number of showers reconstructed within the resolution. Once again, these results highlight the relevance of the temporal aspect of the AHCAL to PF shower separation.

2018 June testbeam data. As in figure 11, the `GravNet` model trained on the training sample of simulation and data is evaluated on the test sample of data. The results are shown in figure 13.

Figures 13(a) and 13(b) show the same asymmetry as presented in figure 11. This result indicates the consistent effect between training on simulation or data showers.

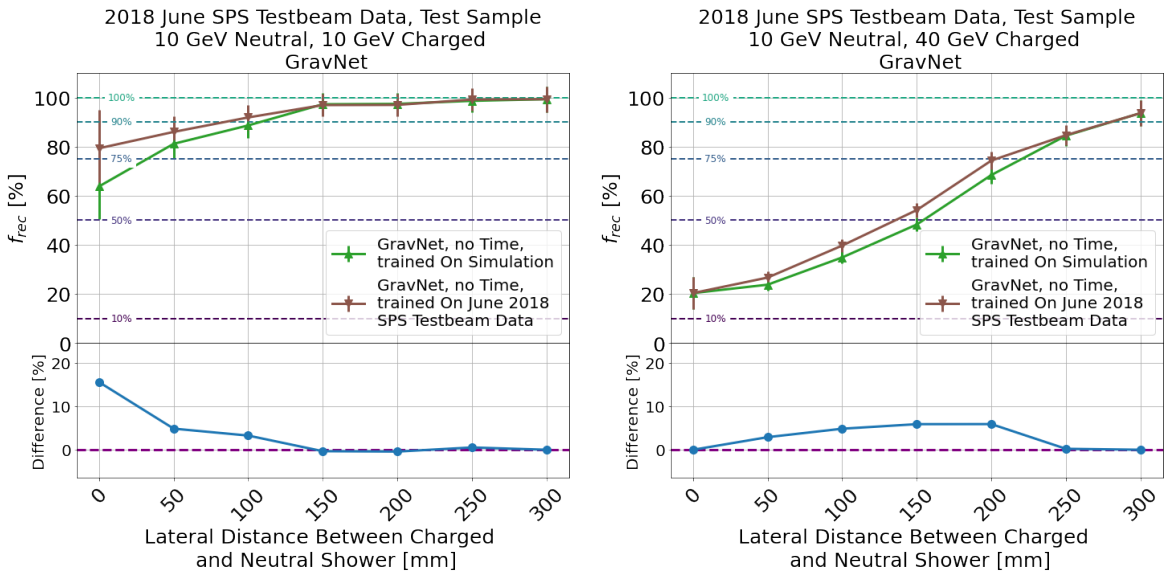
Figure 13(c) indicates that the fraction of showers reconstructed within the resolution of the AHCAL by the `GravNet` network trained on simulation and data varies by no more than around 5 %. This result means that the performance of the neural networks is not strongly related to the choice to use simulation or data for training the models.

2024 JINST 19 P10027



(a) (b) (c)

Figure 13. Figures 13(a) and 13(b) show the matrices of the fraction of showers with energy reconstructed within the calorimeter resolution as a function of the charged and neutral particle energy for `GravNet`, trained on simulation and data, respectively. Figure 13(c) indicates the ratios of the fractions. Else, as in figure 11.



(a) (b)

Figure 14. Figures 14(a) and 14(b) show f_{rec} (see eq. (3.2)) as a function of the distance between showers Q and N in mm for the test sample of data with the `GravNet` model, with $E_Q = E_N = 10\,\text{GeV}$ and $E_Q = 40\,\text{GeV}$, $E_N = 10\,\text{GeV}$, respectively. The green and brown lines indicate the models trained on simulation and data, respectively. Else, as in figure 12.

Figure 14 the same information as figure 12 but for data and with $E_Q = 40\,\text{GeV}$, $E_N = 10\,\text{GeV}$ in figure 14(b), as a 30 GeV sample was not used in the study. The similarity of the green and brown lines indicate agreement with the previous statement. The difference in the trend between figures 12(b) and 14(b) can be attributed to the difference in energy between Q and N being larger in the latter case, resulting in more overall confusion.

2024 JINST 19 P10027

4 Conclusion

Three published neural networks (`PointNet`, `DGCNN` and `GravNet`) were trained to separate a charged and neutral hadron shower with the AHCAL technological prototype to evaluate the shower separation performance of the calorimeter. The neural networks were trained to separate synthetic showers with two showers produced using a method to overlay two hadron showers from single showers. The position of a single shower was uniformly distributed with a fixed most-probable distance between showers and had a uniform energy distribution between 5 and 120 GeV. Simulation and data were used, as well as timing information in simulation with 100 ps resolution. The networks were evaluated, and the results were studied.

Firstly, in simulation, it was observed that `PointNet` did not improve resolution using timing information. By contrast, `DGCNN` and `GravNet` observed a significant reduction in confusion using timing information. For the best-performing neural network (`GravNet`) this corresponded to a reduction of the MAD by around 23 %. This result was speculatively attributed to the improved sensitivity of `GravNet` and `DGCNN` to 'local energy density' compared to `PointNet`, which does not exploit this information by design.

Secondly, all models exhibited asymmetry in the confusion energy distribution, with a tendency to allocate more energy to the less energetic of the two showers rather than the more energetic one. This result was attributed to an 'altruistic' clustering method that produces similar distributions to those observed in similar studies using Pandora PFA.

Thirdly, all models were found to reconstruct 80–90 percent of showers within the calorimeter resolution where the neutral particle had more energy than the charged one and rapidly degraded in the opposite case. This behaviour was attributed to the better performance achievable when there is a significant disparity between track position and the centre-of-gravity of the most energetic shower, which is rarely the case when the charged shower has more energy than the neutral. Timing information was found to explicitly increase the number of showers reconstructed correctly of the latter case by 15–20 % using `DGCNN` and `GravNet`, and motivates the temporal component of the AHCAL calorimeter exploited by graph neural networks.

Finally, the studies made on simulation were repeated for the 2018 June SPS Testbeam data, and the simulation-trained and data-trained models were compared. Regarding performance and properties, almost no difference was observed between the model trained on simulation and the model trained on data. Note that no timing information was available in the study of the test beam data. Therefore, timing was only applied in the simulation studies.

This study suggests that the AHCAL is a highly effective PF calorimeter whose performance can be enhanced using timing information with a resolution of 100 ps. It also concludes that shower separation algorithms can be trained on simulation and applied directly to experimental data with similar performance.

Acknowledgments

We would like to thank the technicians and the engineers who contributed to the design and construction of the CALICE AHCAL prototype detector. We also gratefully acknowledge the CERN management for its support and hospitality and its accelerator staff for the reliable and efficient operation of the test beam. The authors acknowledge the support from the BMBF via the High-D consortium. This

2024 JINST 19 P10027

work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC 2121, Quantum Universe (390833306).

A Summaries of neural networks

`PointNet`. The paper's implementation is based on ref. [30]. Firstly, the hit indices of the shower (I_{hit}, J_{hit}, K_{hit}) pass through a 'transformation network' (T-Net, see ref. [12]), which produces an affine transformation matrix. The T-Net includes an upsampling module with four sequential 1D fully connected layers (64, 128, 256, and 512 channels), using mean, variance, and maximum pooling and a downsampling module with three sequential 1D fully connected layers (256, 128, and 9 channels), with batch normalization and leaky ReLU activation. The output matrix is multiplied by the input. Additional features of the hadron shower relevant to clustering ($\log E_{\text{hit}}$, $\operatorname{arcsinh} t_{\text{hit}}$, I_{track}, J_{track}) are concatenated and upscaled by a 1D fully-connected layer with 64 channels. Then, the output passes through a second T-Net of the same structure as the 3D Transform with 64×64 layers. The output, denoted as A, is initialised as the identity matrix and is used with mean, variance and maximum pooling. The remaining hit energy and track energy information are concatenated to the output. The final module includes four sequential 1D fully connected layers (512, 256, 128, and 64 channels), with batch normalisation, leaky ReLU activation, and dropout in the last layer. The final layer outputs reconstructed energy fractions for each hit in the shower, using Softmax activation. The matrix A is regularized during training to remain symmetric to ensure affine transformation. A diagram indicating the model's design is shown in appendix figure 15.

`DGCNN`. The paper's implementation, based on ref. [31], utilizes four `EdgeConv` operators (see ref. [13]). The input is concatenated with its mean in each operation and passed through a k-NN clustering, followed by two fully connected 2D layers with 64 channels. The mean, variance, and maximum are calculated over the clusters and passed through a 1D fully-connected layer with 64 channels, along with batch normalisation and leaky ReLU activation for all layers. The first `EdgeConv` operator uses only hit indices of the shower (I_{hit}, J_{hit}, K_{hit}). Additional 'features' of the hadron shower relevant to clustering ($\log E_{\text{hit}}$, $\operatorname{arcsinh} t_{\text{hit}}$, I_{track}, J_{track}) are later concatenated to the output of the first `EdgeConv` operator before applying the second operator. At each stage, the output from each `EdgeConv` operator is recorded and later concatenated into a single tensor. Then, one shared 1D fully connected layer with 1024 channels, batch normalization, and leaky ReLU activation condenses the features learned at the clustering stage. Mean, variance and maximum pooling over the points activate the points, with remaining hit energy and track energy information (E_{hit}, E_{tr}^{Q}) concatenated to the output. The final stage of the network includes four sequential 1D fully connected layers (512, 256, 128, and 64 channels), with batch normalization, leaky ReLU activation, and dropout in the last layer. The final layer produces two outputs, one for $\widehat{f}_{\text{hit}}^{Q}$ and one for $\widehat{f}_{\text{hit}}^{N}$, using Softmax activation, resulting in reconstructed energy fractions for each hit in the shower. A diagram indicating the model's design is shown in appendix figure 16.

`GravNet`. The paper's implementation, based on refs. [31] and [29], replaces `EdgeConv` layers with `GravNet` layers in the model. Before each `GravNet` layer, the mean is concatenated to the input and passed through two fully connected 1D layers with batch normalization and leaky ReLU activation, each with 64 channels.

2024 JINST 19 P10027

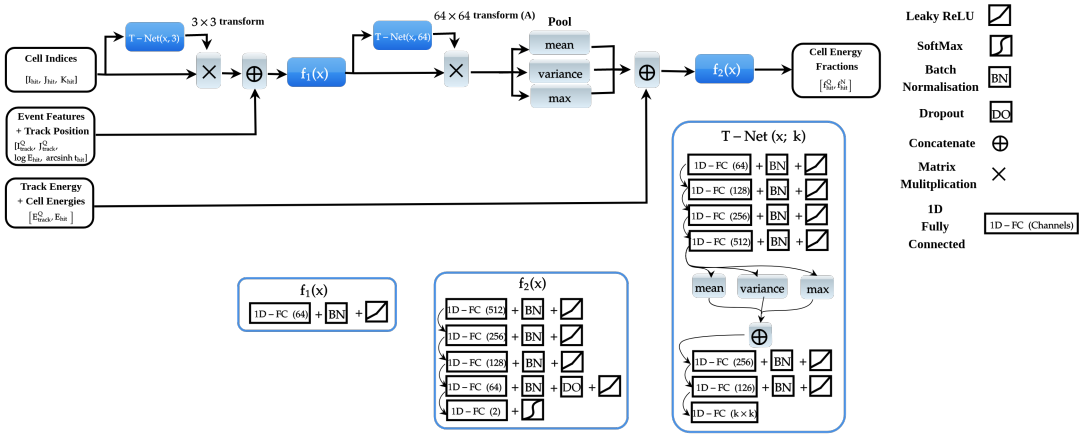
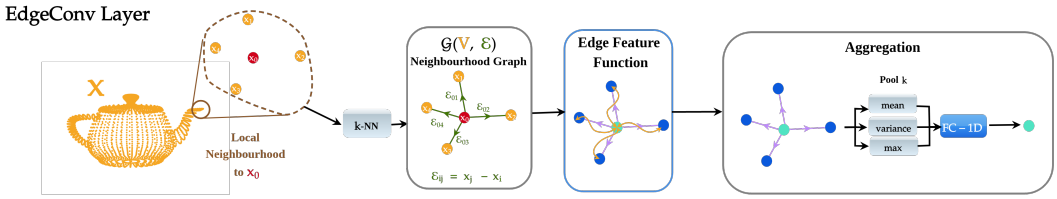
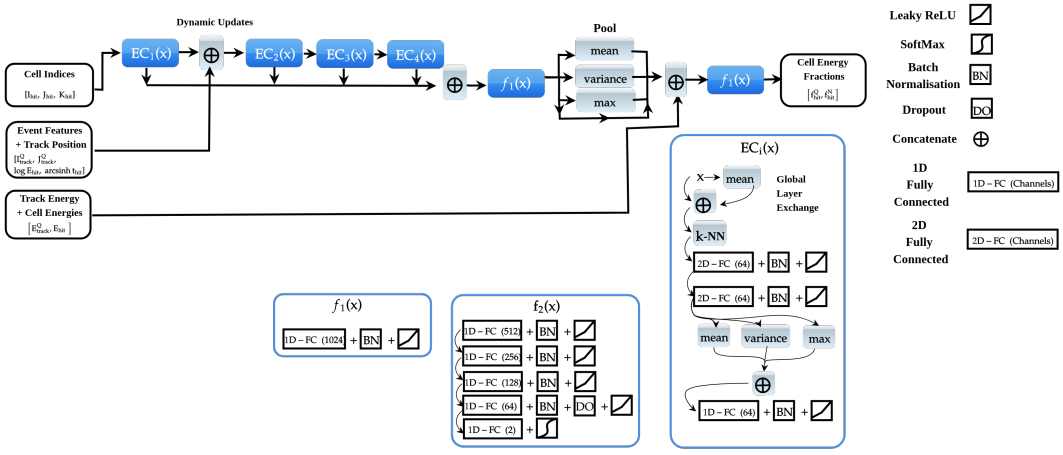


Figure 15. Diagram illustrating the `PointNet` implementation in this study. The black, blue and grey boxes indicate inputs and outputs, convolutional operations and general operations, respectively. Additional operations are specified on the right of the figure. Additionally, matrix multiplication is indicated as a × symbol, batch normalisation is denoted 'BN' and the Softmax activation is indicated in the legend.



(a)

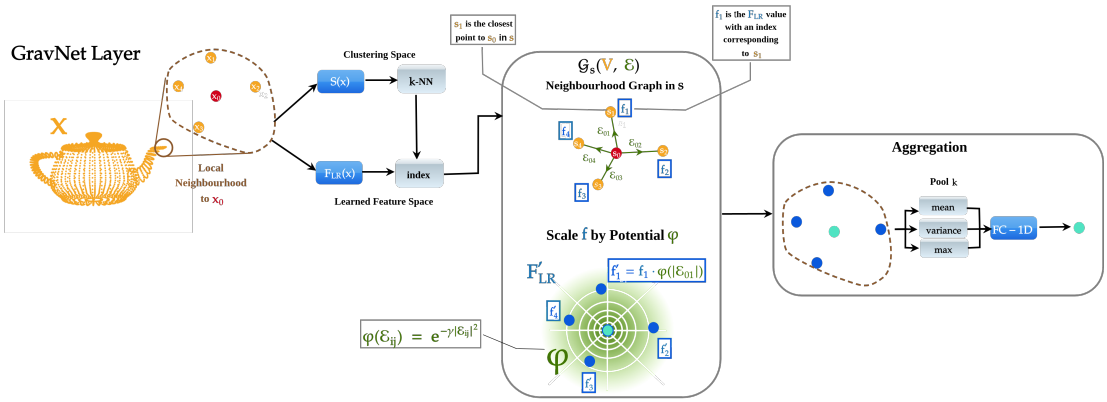


(b)

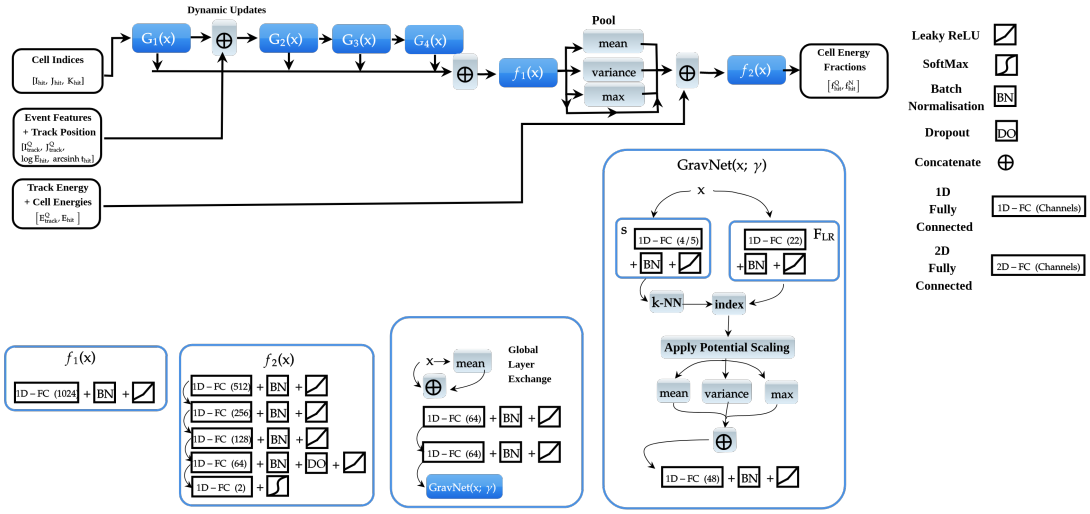
Figure 16. Figure 16(a) illustrates the `EdgeConv` operator. The orange Utah teapot indicates some 'point cloud', or distribution of points, x, with some underlying distribution, to be operated upon. The orange dots indicate vertices, V, of a local neighbourhood k-NN graph around the central red dot. The green arrows indicate the vectors (edges) between the central red dot and its neighbours, $\mathcal{E}$. The orange arrows indicate 'message-passing' between the vertices and the edges, which modifies the graph as indicated by the colour inversion. Figure 16(b) shows a diagram illustrating the `DGCNN` implementation in this study. Else, as in figure 15.

2024 JINST 19 P10027

In this paper, the `GravNet` layer operation starts by projecting points to a low-dimensional clustering space, s, and then to a high-dimensional learned feature space, F_{LR}. This process involves two individual 1D fully connected layers with 4 (5 if time is included) and 22 channels, respectively. Then, a k-NN cluster is found in the s-space, similar to `DGCNN`. Euclidean distances of the neighbours to the graph's origin are calculated for each cluster. These distances are scaled by a Gaussian potential with a hyperparameter 'potential strength' parameter γ, resulting in aggregated values across the cluster using maximum, mean, and variance. Finally, a 1D fully connected layer with 48 channels, batch normalization, and leaky ReLU activation condenses the aggregates to a single value. A diagram indicating the design of the model is shown in appendix figure 17.



(a)



(b)

Figure 17. Figure 17(a) illustrates the `GravNet` operator. The orange Utah teapot indicates some 'point cloud', or distribution of points, x, with some underlying distribution, to be operated upon. The S and F_{LR} operations indicate the clustering and feature space representations learned by the neural network. The green arrows and $\mathcal{E}$ correspond to vectors between the red point and its orange neighbours, its norm indicating a distance in a potential well given by φ, indicated by the shaded green region. f and f' indicate unscaled and potential-scaled features. Else as in figure 16(a).

2024 JINST 19 P10027

B Validation of synthetic neutral hadron showers

A classifier can be used to assess the similarity between two event categories from AHCAL with many correlated observable properties. The classifier's performance can be measured using the receiver-operating curve (ROC) and its area-under-curve (AUC). The ROC measures the true positive rate against the false positive rate, while the AUC indicates classifier performance. An AUC of 0.5 signifies random guessing, and a classifier with an AUC greater than 0.5 is more effective. The AUC can, therefore, be used to quantify the overall difference between a 'real' and a 'synthetic' neutral.

A study was conducted using a training sample consisting of 40 GeV π^- and the entire K_L^0 sample from simulation, as shown in table 1. This dataset was divided into training, validation, and test samples, with event details listed in table 3(a). The study utilised the standard CALICE AHCAL Particle Identification (PID) classifier, which employs a gradient-boosted decision tree implemented in the `LightGBM` framework [16, 18, 32]. This classifier accurately categorises hadrons, electrons, and muon-like events observed with AHCAL using thirteen event-level variables: E_{sum}; the total number of hits in the event; the average hit radius; CoG_K; the fraction of E_{sum} deposited in the first 22 AHCAL layers; K_S; the fraction of E_{sum} deposited after K_S; the fraction of E_{sum} in the 'core' of a hadron shower ($R_{\text{hit}} < 1$ cell, ≥ 2 adjacent hits, and > 0 cells active in the same layer); the 'track-like' fraction of E_{sum} (≥ 2 adjacent hits and 0 cells active in the same layer); the 'detached' fraction of E_{sum}: (0 adjacent hits); the number of hits in the event occurring after K_S; the number of hits contributing to the 'track-like fraction' of E_{sum} and number of hits contributing to the event in the last 4 AHCAL layers.

The gradient-boosted decision tree classifier was re-optimised for classifying simulated K_L^0 from π^- hadron showers with the MIP cut applied in AHCAL using the hyperparameters and loss detailed

Table 3. Table 3(a) shows a subsample of table 1 of 40 GeV particles used for training the gradient-boosted decision tree to assess the performance of the synthetic neutral hadron showers produced via the MIP-track cut. Table 3(b) shows the hyperparameters used for the CALICE PID gradient-boosted decision tree, used in this paper to analyse the effectiveness of the MIP-track cut. Further information about these parameters can be found in [32].

(a)

	Simulation		
Hadron	Testing	Training	Validation
K_L^0	11732	54800	11614
π^-	11411	53201	11530

(b)

Hyperparameter	Value
Objective	SoftMax
Metric	Multi-Log Loss
# Classes	2
Metric Frequency	1
# Leaves	10
Max Depth	10
Min Child Samples	10
Learning Rate	0.1
Feature Fraction	0.9
Bagging Fraction	0.8
Bagging Frequency	5

2024 JINST 19 P10027

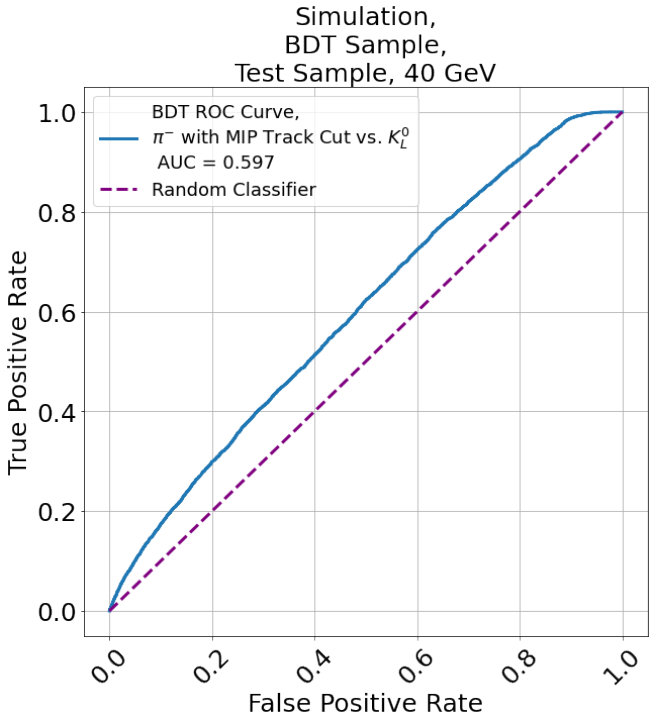


Figure 18. ROCs from the trained classifier applied to the test sample from table 3(a). The blue line indicates the performance of the model applied to the same testing sample of K_L^0 and π^- with the MIP-track cut applied to the event variables specified. The purple dashed line indicates the expected curve for a random classifier.

in table 3(b). The re-optimisation was performed until the loss no longer improved after 5 training steps. Then, the newly optimised classifier was evaluated using the test dataset of π^- hadron showers with the MIP cut applied. The classifier's performance was evaluated using the AUC.

The results are shown in figure 18. They illustrate that the classifier achieved an AUC of 0.597 for classifying π^- with the MIP cut applied and K_L^0 showers in the test sample of table 3(a). It may be concluded that while these showers have some differences, they are challenging to distinguish at the event level. This result indicates the cut can produce 'convincing' synthetic neutrals.

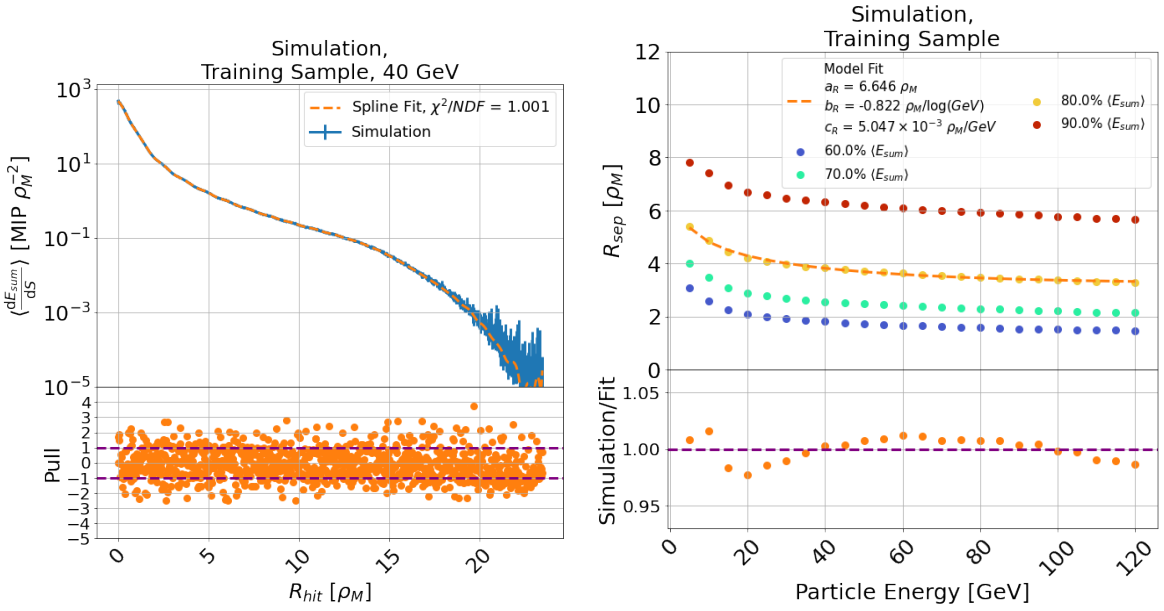
C Producing events with a charged and synthetic neutral shower

A method to combine a sample of single π^- hadron showers into a shower with a charged hadron and a synthetic neutral hadron shower is presented.

Firstly, two π^- hadron showers are selected by a weighted random subsample of either the training, validation or testing sample of table 1 to produce the corresponding sample for training, validating and testing the neural networks. One is designated the charged candidate, Q; the other the synthetic neutral candidate, N. The weight for each particle energy is selected such that approximately equal numbers of each possible combination of particle energies appear in the final sample. Next, the MIP-track cut is applied to N only.

Four integers, ΔI^Q, ΔI^N, ΔJ^Q and ΔJ^N, are then defined as distances in cells by which to displace both Q and N in the I and J directions in calorimeter space to mimic the particles entering at a different position than the beam spot which is centred around $I = J = 12.5$ cells for both simulation

2024 JINST 19 P10027



(a) (b)

Figure 19. Figure 19(a) shows the differential energy loss per unit surface area as a function of R_{hit} for the 40 GeV training sample of π^- simulation in table 1. The blue points show the simulation, and the orange dashed line indicates a spline fit to the points. The residues of the fit are shown in the bottom subplot. Figure 19(b) shows the value of R_{sep} as a function of particle energy for the simulation sample of table 1. The blue, teal, orange, and red circle markers indicate the 60 %, 70 %, 80 %, and 90 % percentiles of $\langle E_{\text{sum}} \rangle$. The dashed orange line indicates an ad-hoc fit to the 80 % percentile.

and data. These displacements are first uniformly sampled within a circle with a radius R_{circ} centred at $I_{\text{hit}} = 0$ cell, $J_{\text{hit}} = 0$ cell, within the circumference of a circle of radius R_{circ}. The radius was chosen so that the showers have a moderate amount of confusion during training on average but not too much so that the training becomes overly challenging.

The average distance between two points uniformly sampled within a circle of radius R_{circ} is denoted as R_{sep} for each shower. The method of estimating R_{sep} for each particle energy in the simulation training sample of table 1 is presented as follows. First, the differential energy loss by a hadron shower per unit area of a circle (i.e., a thin ring) around the centre-of-gravity is calculated, defined as $\langle \mathrm{d}E_{\text{sum}}/\mathrm{d}S \rangle$, where $\mathrm{d}S = 2\pi R_{\text{hit}}\,\mathrm{d}R_{\text{hit}}$ with $\mathrm{d}R_{\text{hit}}$ as a bin width determined using the Freedman-Diaconis bin rule [33]. An example is shown in figure 19(a). A spline is fitted to the distribution, and then cumulatively integrated as a function of radial distance from the shower's centre-of-gravity, $\langle E_{\text{sum}}(R_{\text{hit}}) \rangle = \int_0^{R_{\text{hit}}} \langle \mathrm{d}E_{\text{sum}}/\mathrm{d}S \rangle \cdot 2\pi R_{\text{hit}}\,\mathrm{d}R_{\text{hit}}$. The R_{sep} value for a particular E_{particle} is determined where this cumulative integral reaches 80 % of $\langle E_{\text{sum}} \rangle$.

The relationship between the particle energy and R_{sep} is expected to decrease with particle energy as the electromagnetic fraction of a hadron shower increases, making the shower more energy-dense. This relationship can be approximated using an ad-hoc function, $R_{\text{sep}}(E_{\text{particle}}) = a_R + b_R \cdot \log E_{\text{particle}} + c_R \cdot E_{\text{particle}}$, where a_R, b_R, and c_R are free parameters obtained from a fit. The relationship for simulation is shown in figure 19(b).

2024 JINST 19 P10027

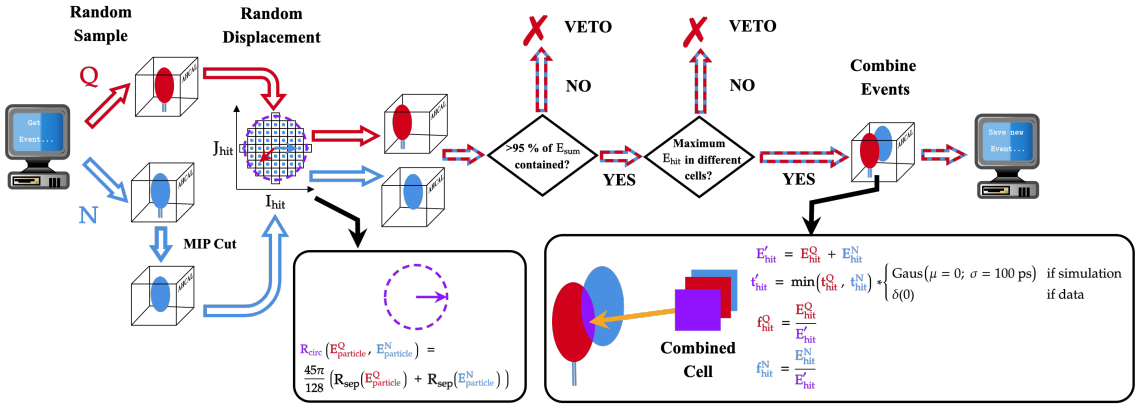


Figure 20. Diagram illustrating the shower-combination algorithm used to produce synthetic charged-neutral hadron shower events from a sample of single π^- hadron showers observed with AHCAL. The red and blue arrows indicate the paths the charged and synthetic neutral hadron shower took. The squared circle indicates the distribution of available cells by which the event can be displaced from its entry position.

Finally, R_{circ} is calculated using the formula $R_{\text{circ}}(E_Q, E_N) = 45\pi/128 \cdot (R_{\text{sep}}(E_Q) + R_{\text{sep}}(E_N))$, where the factor relates a circle's radius to the average distance between two uniformly sampled points within its circumference [34]. The vectors $\{\Delta I^Q, \Delta J^Q\}$ and $\{\Delta I^N, \Delta J^N\}$ obtained by rejection sampling using $\Delta I^2 + \Delta J^2 \leq R^2_{\text{circ}}$.

Next, all of the hits and track positions of Q and N are then shifted by the two sampled integers. For example, all the hits in Q are displaced by ΔI^Q and ΔJ^Q, (i.e. $I^Q_{\text{hit}} \rightarrow I^Q_{\text{hit}} + \Delta I^Q$, $J^Q_{\text{hit}} \rightarrow J^Q_{\text{hit}} + \Delta J^Q$), and the track position of the charged particle is also shifted by the same integers (i.e. $I^Q_{\text{track}} \rightarrow I^Q_{\text{track}} + \Delta I^Q$, $J^Q_{\text{track}} \rightarrow J^Q_{\text{track}} + \Delta J^Q$). At this stage, any hits outside the calorimeter are removed from both showers. If over 95% of each shower's energy remains in the calorimeter and if the highest energy cells are not shared between Q and N, the algorithm continues. Event-level properties like the centre-of-gravity are recalculated based on the MIP-track cut and the removed hits. If the criteria are unmet, the event is rejected, and displacement integers are resampled until they meet the criteria.

Upon satisfying the criteria, Q and N are combined. The energies from Q and N are added for cells with shared energy. The minimum hit time of Q and N is taken for the cell, with a random Gaussian smearing of 100 ps applied after selection in the simulation. As the hit time in data already includes the detector's time resolution, no additional smearing is applied. Finally, energy fractions for each hadron shower, f^Q_{hit} and f^N_{hit}, are calculated from the combined energy. The result is an event containing a charged and synthetic neutral hadron shower from a dataset of single π^- hadron showers, useful for training and validating machine learning algorithms for shower separation. A diagram of the algorithm is shown in figure 20.

Figure 21(a) shows that a wide range of inter-shower distances are available in the training dataset. The most probable distance between Q and N is 5.5 ρ_{M} for both simulation (left) and data (right), determined by kernel density estimate. Figures 21(b)–21(c) show that the distribution of the initial centres-of-gravity of the hadron showers in the calorimeter is convolved with a circle function, indicating a wide variety of shower configurations in the final training sample.

2024 JINST 19 P10027

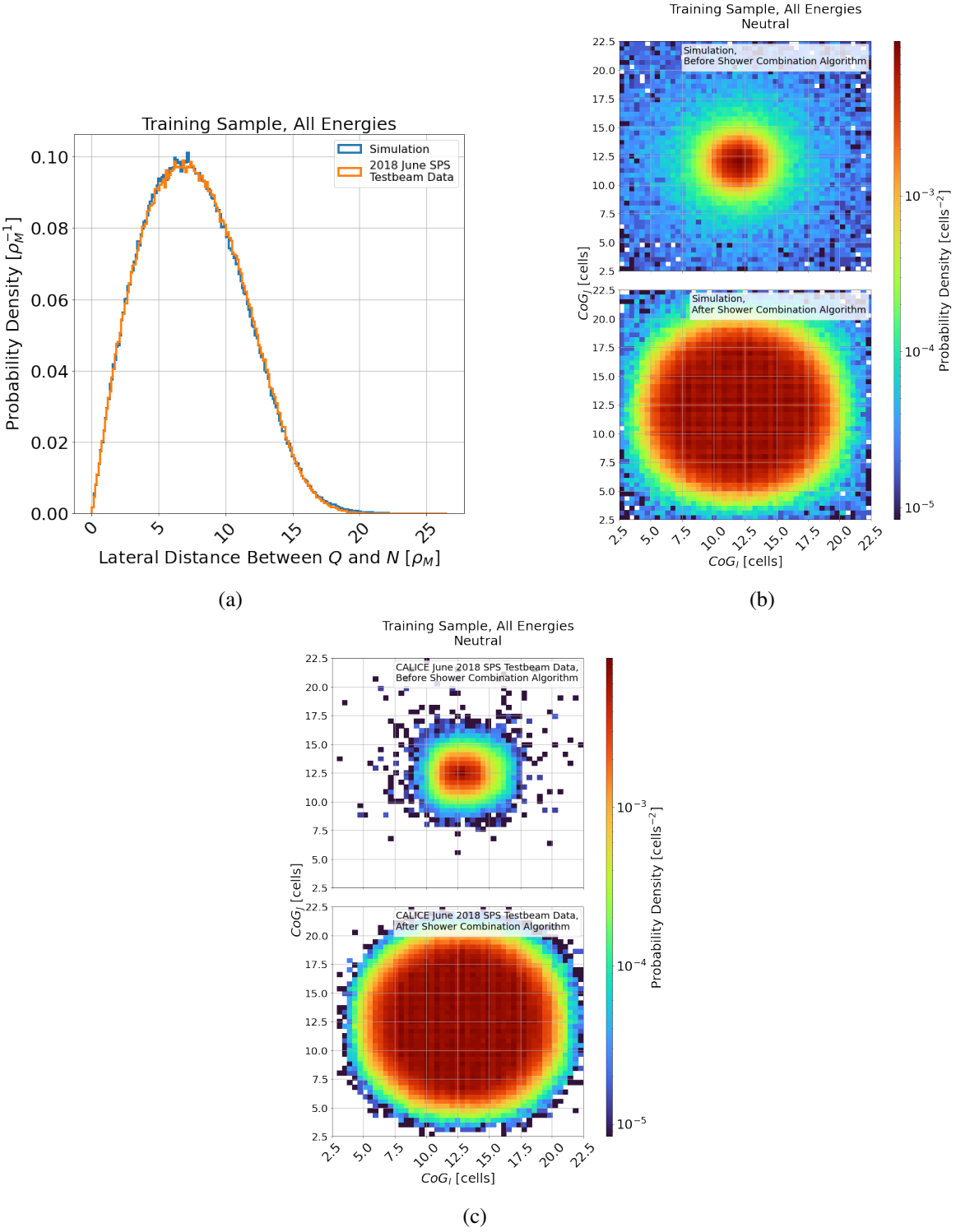


Figure 21. Figure 21(a) shows the distribution of distances between Q and N in Moliere radii (ρ_{M}). The blue and orange lines indicate simulation and data, respectively. Figures 21(b) and 21(c) show the distribution of the centres-of-gravity of the synthetic neutral shower before (top plot) and after (bottom plot) the shower overlay algorithm is applied, for simulation and data respectively. The colour axis indicates probability density.

References

[1] M.A. Thomson, *Particle Flow Calorimetry and the PandoraPFA Algorithm*, *Nucl. Instrum. Meth. A* **611** (2009) 25 [arXiv:0907.3577].

[2] CALICE collaboration, *A highly granular SiPM-on-tile calorimeter prototype*, *J. Phys. Conf. Ser.* **1162** (2019) 012012 [arXiv:1808.09281].

[3] S.R. Qasim, J. Kieseler, Y. Iiyama and M. Pierini, *Learning representations of irregular particle-detector geometry with distance-weighted graph networks*, *Eur. Phys. J. C* **79** (2019) 608 [arXiv:1902.07987].

[4] J. Pata et al., *Improved particle-flow event reconstruction with scalable neural networks for current and future particle detectors*, *Commun. Phys.* **7** (2024) 124 [arXiv:2309.06782].

[5] F.A. Di Bello et al., *Towards a Computer Vision Particle Flow*, *Eur. Phys. J. C* **81** (2021) 107 [arXiv:2003.08863].

[6] O.L. Pinto, *Shower Shapes in a Highly Granular SiPM-on-Tile Analog Hadron Calorimeter*, Ph.D. thesis, University of Hamburg, Hamburg, Germany (2022).

[7] CALICE collaboration, *Design, construction and commissioning of a technological prototype of a highly granular SiPM-on-tile scintillator-steel hadronic calorimeter*, 2023 *JINST* **18** P11018 [arXiv:2209.15327].

[8] CALICE collaboration and ILD Concept Group, *Calibration of the Scintillator Hadron Calorimeter of ILD*, CALICE Analysis Note CAN-018 (2009).

[9] Omega Group, *SPIROC2 User Guide*, (2009).

[10] CALICE collaboration, *Design, construction and commissioning of a technological prototype of a highly granular SiPM-on-tile scintillator-steel hadronic calorimeter*, 2023 *JINST* **18** P11018 [arXiv:2209.15327].

[11] CMS and CALICE collaborations, *Performance of the CMS High Granularity Calorimeter prototype to charged pion beams of 20–300 GeV/c*, 2023 *JINST* **18** P08014 [arXiv:2211.04740].

[12] C.R. Qi, H. Su, K. Mo and L.J. Guibas, *PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation*, arXiv:1612.00593.

[13] Y. Wang et al., *Dynamic Graph CNN for Learning on Point Clouds*, arXiv:1801.07829.

[14] GEANT4 collaboration, *GEANT4 — a simulation toolkit*, *Nucl. Instrum. Meth. A* **506** (2003) 250.

[15] M. Petrič et al., *Detector simulations with DD4hep*, *J. Phys. Conf. Ser.* **898** (2017) 042015.

[16] CALICE collaboration, *CALICESoft*, https://stash.desy.de/projects/CALICE.

[17] L.K. Emberger, *Precision Timing in Highly Granular Calorimeters and Applications in Long Baseline Neutrino and Lepton Collider Experiments*, Ph.D. thesis, Technische Universität München, München, Germany (2022).

[18] V. Bocharnikov, *Particle identification methods for the CALICE highly granular SiPM-on tile calorimeter*, presented at *Verhandlungen der Deutschen Physikalischen Gesellschaft*, Aachen, Germany, March 25–29 (2019).

[19] D. Heuchel, *Particle Flow Studies with Highly Granular Calorimeter Data*, Ph.D. thesis, University of Heidelberg, Heielberg, Germany (2022).

[20] R. Wigmans, *Calorimetry: Energy measurement in particle physics*, Oxford University Press (2000) [DOI:10.1093/oso/9780198786351.001.0001].

2024 JINST 19 P10027

[21] A. Paszke et al., *Automatic differentiation in PyTorch*, in the proceedings of the *31st Conferenece on Neural Information Processing Systems*, Long Beach, CA, U.S.A., December 4–9 (2017).

[22] W. Falcon et al., *PyTorch Lightning*, https://www.pytorchlightning.ai/.

[23] T. Akiba et al., *Optuna: A Next-generation Hyperparameter Optimization Framework*, arXiv:1907.10902.

[24] T. Odland, *KDEpy*, https://github.com/tommyod/KDEpy.

[25] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Routledge (2018) [DOI:10.1201/9781315140919].

[26] P.J. Rousseeuw and C. Croux, *Alternatives to the Median Absolute Deviation*, *J. Am. Statist. Assoc.* **88** (1993) 1273.

[27] S. Kokoska and D. Zwillinger, *CRC Standard Probability and Statistics Tables and Formulae*, Chapman & Hall, New York, NY, U.S.A. (2000) [DOI:10.1201/b16923].

[28] CALICE collaboration, *Software Compensation for Highly Granular Calorimeters using Machine Learning*, 2024 *JINST* **19** P04037 [arXiv:2403.04632].

[29] J. Kieseler, *caloGraphNN*, https://github.com/jkiesele/caloGraphNN.

[30] X.Yan, *Pytorch Implementation of PointNet and PointNet++*, https://github.com/yanx27/Pointnet_Pointnet2_pytorch.

[31] Y. Wang, *Dynamic Graph CNN for Learning on Point Clouds*, https://github.com/WangYueFt/dgcnn.

[32] M. Hajihosseinlou, A. Maghsoudi and R. Ghezelbash, *A Novel Scheme for Mapping of MVT-Type Pb-Zn Prospectivity: LightGBM, a Highly Efficient Gradient Boosting Decision Tree Machine Learning Algorithm*, *Nat. Resources Res.* **32** (2023) 2417.

[33] D. Freedman and P. Diaconis, *On the histogram as a density estimator:L 2 theory*, *Zeit. Wahrschein. Verw. Gebiet.* **57** (1981) 453.

[34] B. Burgstaller and F. Pillichshammer, *The Average Distance Between Two Points*, *Bull. Austral. Math. Soc.* **80** (2009) 353.

2024 JINST 19 P10027