Nuclear Instruments and Methods in Physics Research A 1083 (2026) 171156

Contents lists available at ScienceDirect

Nuclear Inst. and Methods in Physics Research, A

journal homepage: www.elsevier.com/locate/nima

Full Length Article

Study of hadronic shower starting point reconstruction in a 3D imaging calorimeter

X.F. Tang [a,b,c], Z. Quan [a,b,*], Y.W. Dong [a,b,**], T.W. Bao [a,b], C.L. Liao [a,b,c], X. Liu [a,b], J.Y. Sun [a,b,c], J.J. Wang [a,b], R.J. Wang [a,b], Z.G. Wang [a,b], Q. Wu [b,c], M. Xu [a,b], X.G. Yang [a,b,c]

[a] State Key Laboratory of Particle Astrophysics, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, 100049, China
[b] Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, 100049, China
[c] School of Physical Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China

ARTICLE INFO

Keywords:
HERD
High granularity calorimeter
Hadronic shower starting point
Machine learning

ABSTRACT

The High Energy cosmic-Radiation Detection (HERD) facility will be installed as a space astronomy payload on the China Space Station in 2028. The three-dimensional imaging calorimeter (CALO) of HERD comprises about 7500 lutetium yttrium oxy-orthosilicate (LYSO:Ce) cubes, where the topological development of hadronic showers can be measured to determine the shower starting point (SSP). This paper presents two machine learning architectures – Inception Convolutional neural network (IncepCNN) and Transformer – to reconstruct SSP in CALO. These architectures are developed using isotropic proton simulations, and demonstrate superior performance over a wide energy range from 30 GeV up to 1 TeV. Comparison with traditional layer-wise and energy-ratio methods shows that both machine learning methods achieve ∼30 % higher accuracy and ∼1 cm improved spatial resolution. Based on the correlation between SSP and shower leakage, a correction to the visible energy is implemented for protons. The results show that this correction effectively restores the mean value of the measured energy distribution to the expected primary energy, with a bias of about 1 % for isotropic simulations and 0.3 % for test beam data.

1. Introduction

Unlike photons and electrons that initiate showers almost immediately in a calorimeter, hadrons typically travel a significant distance before developing a shower [1]. This distance, known as the nuclear interaction length (λ_I), characterizes the distribution of the first inelastic interaction position for incident hadrons, which is identified as the shower starting point (SSP). Before this interaction, the primary hadron behaves like a minimum ionizing particle (MIP), depositing very small amounts of its energy. The first interaction generates numerous secondary particles, triggering the hadronic cascade and a steep increase in energy deposition. Therefore, SSP marks the transition from a single particle track to a developing shower. The determination of SSP is of importance for data analysis based on a segmented calorimeter:

Particle identification. The difference in interaction scales (radiation length $X_0 \ll \lambda_I$) leads to distinctly different SSP distributions. Electromagnetic showers produce narrowly distributed SSPs concentrated within the first few calorimeter layers. Hadrons (predominantly protons, ∼90 % of cosmic-ray particles) either traverse as MIPs or interact deeply, resulting in a broader SSP distribution. This difference enables effective particle identification techniques [2–4].

MIP calibration. Cosmic-ray protons behave like minimum ionizing particles before shower starts, providing efficient MIP counts for on-orbit calibration. A shower-enabled method has been proposed that uses the MIP track within shower events for calorimeter calibration [5]. A key advantage of this method is that energy deposition information helps to preferentially select high-momentum (e.g., above 15 GeV/c) protons, thereby avoiding the peak shift in the MIP signal caused by low-momentum non-relativistic protons.

Energy leakage correction. Energy leakage is a significant effect for late-starting showers and increases with energy. Such late-starting showers lack sufficient material for full containment, leading to energy leakage that degrades the energy measurement. An SSP-based algorithm effectively corrects for this leakage by using a multi-dimensional lookup table for event-by-event correction [6]. This method restores the mean of the energy distribution to within 2 % of the beam energy and improves the energy resolution by approximately 25 % at 80 GeV.

* Corresponding author at: Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, 100049, China.
** Corresponding author at: Institute of High Energy Physics, Chinese Academy of Sciences, Beijing, 100049, China.
E-mail addresses: quanzheng@ihep.ac.cn (Z. Quan), dongyw@ihep.ac.cn (Y.W. Dong).

https://doi.org/10.1016/j.nima.2025.171156
Received 18 July 2025; Received in revised form 5 November 2025; Accepted 5 November 2025
Available online 6 November 2025

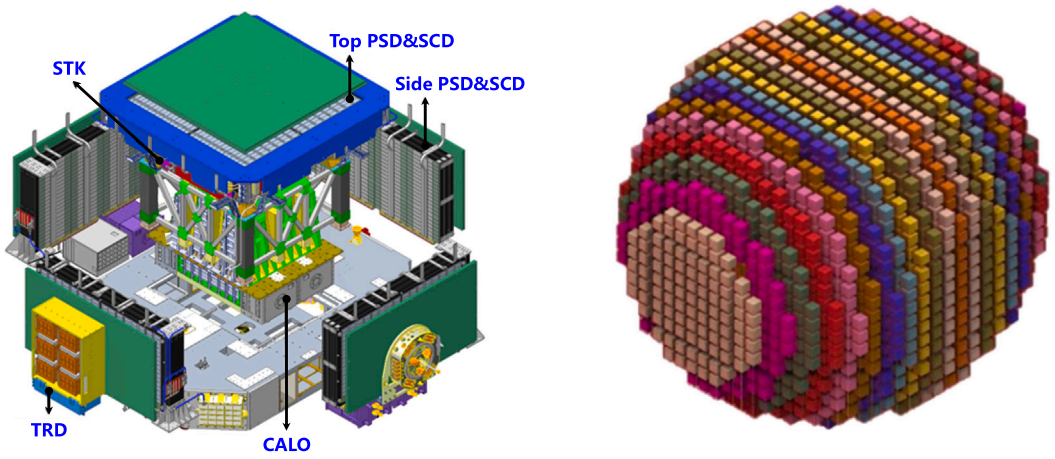


Fig. 1. (a) Schematic diagram of the HERD detector design. The detector components are: 3D cubic calorimeter (CALO), Silicon Tracker (STK), Plastic Scintillator Detector (PSD), Silicon Charge Detector (SCD) and Transition Radiation Detector (TRD). (b) The CALO detector's crystal array comprising 7489 LYSO cubes, each with 3 cm side length.

As an astronomy and particle astrophysics experiment, the High Energy cosmic-Radiation Detection (HERD) facility is scheduled for launch in 2028 and will operate aboard the China Space Station for more than 10 years. Its main scientific objectives are the search for dark matter signals, precise measurements of cosmic-ray energy spectrum and composition, and high energy gamma-ray monitoring and survey [7–9]. Fig. 1(a) gives a schematic view of the HERD detectors. The core instrument in the innermost is a three-dimensional (3D) cubic imaging calorimeter (CALO) [10], surrounded on its top side by the Silicon Tracker (STK). CALO and STK are covered by the Plastic Scintillator Detector (PSD) [11] and the Silicon Charge Detector (SCD) [12] from five sides. Additionally, a Transition Radiation Detector (TRD) [13] is placed on one of the lateral sides.

The CALO detector comprises 7489 cerium-doped lutetium yttrium orthosilicate (LYSO:Ce) cubic crystals, arranged in a spheroidal $23 \times 23 \times 21$ array (see Fig. 1(b)). The total depth of CALO is about 55 radiation lengths and 3 nuclear interaction lengths for centrally incident particles from any direction. LYSO is selected for its high light yield, high density and low temperature coefficient [14]. These properties allow for a compact detector design that increases the acceptance, which is a key advantage for space experiments under strict weight constraints. Moreover, LYSO's large radiation and nuclear interaction lengths enable the total absorption of electromagnetic showers up to tens of TeV and provide a high interaction probability for protons and nuclei. The scintillation light from each crystal is read out by a "double read-out system": (1) Wavelength Shifting Fibers (WLSF) coupled to custom image Intensified scientific CMOS (IsCMOS) cameras, and (2) photo-diodes (PDs) sensors connected to custom front-end electronics chips. This 3D cubic array allows for imaging hadronic showers with high granularity, and enables the reconstruction of the shower starting point using advanced methods — the subject of this paper.

In this paper, we propose two machine learning architectures, the Inception Convolutional Neural Network (IncepCNN) and the Transformer, to reconstruct SSP of proton-induced showers in HERD CALO. This paper is structured as follows: Section 2 outlines the traditional methods and presents the motivation for adopting machine learning techniques. Section 3 describes the simulation setup, the Monte Carlo truth definition of SSP and event selection. Section 4 introduces the implementation of SSP reconstruction, including two traditional methods (layer-wise and energy-ratio methods) and two machine learning methods (IncepCNN and Transformer). Their performance is compared in Section 5. Section 6 presents an application of SSP to energy leakage correction using both isotropic simulation and test beam data. Conclusions are given in Section 7.

2. Methods

Several algorithms for SSP reconstruction have been developed in high energy physics experiments. The CALICE collaboration has developed a layer-wise method based on the increase in energy deposition and hit counts in consecutive calorimeter layers [15]. The CMS collaboration has identified the shower starting layer by requiring both the total measured energy and transverse energy spread to exceed thresholds simultaneously [16]. The HERD collaboration has proposed an energy-ratio method using the longitudinal energy density profile along the shower axis [5].

However, when applied to HERD CALO, these traditional methods face some limitations: (1) They characterize shower development by predefined layer geometry or preselected crystals traversed by the shower axis, but neither approach fully exploits CALO's high granularity information; (2) The unit size (e.g. layer thickness or crystal size) constrains the achievable spatial resolution; (3) The trade-off between accuracy and efficiency demands careful consideration: lower thresholds degrade accuracy due to too early reconstructed SSP from noise or delta electrons, while higher thresholds may miss the true SSP entirely. Furthermore, the correlation between SSP and CALO response is disrupted by event-to-event fluctuations typical of hadronic showers. This intrinsically limits a reconstruction algorithm's potential for accuracy improvement.

Over the years, advancements in machine learning (ML), particularly Convolutional Neural Networks (CNNs) [17–19] and Transformers [20–22], have allowed us to overcome these limitations. The spatial sensitivity of CNN makes it highly suitable for processing 3D images, whereas Transformer is an excellent tool to capture long-range dependencies in signals via self-attention mechanism. These architectures have been widely used for data analysis in high energy physics, demonstrating state-of-the-art performance in particle identification [23,24], energy reconstruction [25,26], and trajectory reconstruction [27,28]. Building on these advances, we develop two end-to-end ML models (IncepCNN and Transformer), which extract features directly from CALO hit information for SSP reconstruction.

3. Simulation setup

3.1. Dataset

The IncepCNN and Transformer models are trained and evaluated using Monte Carlo (MC) simulated events generated by the HERD

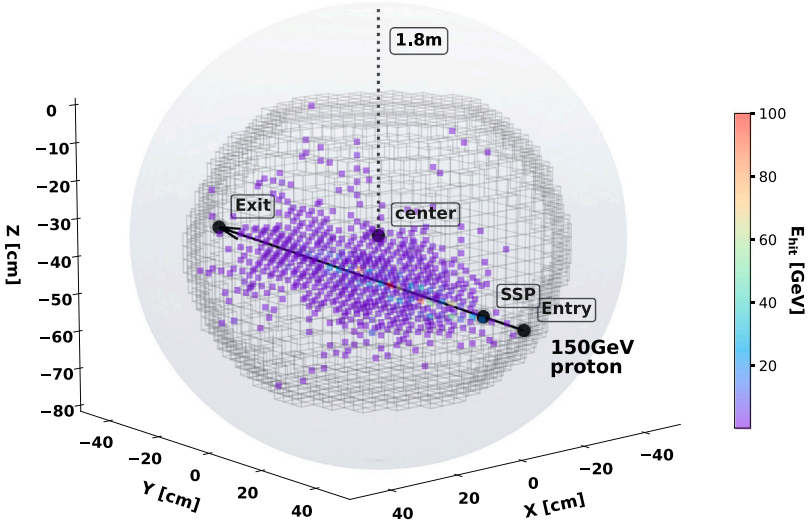


Fig. 2. A simulated 150 GeV proton-induced shower in the CALO detector. The proton is emitted from a spherical surface (gray, radius: 1.8 m). LYSO crystals are depicted as gray cubes, with energy deposits shown as colored cubes. The shower axis (black arrow) intersects the CALO surface at the entry and exit points. SSP marks the shower starting point.

Offline Software (HERDOS), a Geant4-based framework for full detector simulation. Our analysis only focuses on CALO data. Protons are emitted isotropically from a spherical surface of 1.8 m radius[1] – sufficient to cover the entire payload envelope. An example of simulated event is shown in Fig. 2. These simulated events are split into two datasets:

- **Training dataset** A clear energy dependence is observed in CALO response to hadronic showers (e.g. the number of hits increases with the particle's energy). Therefore, the training dataset combines 3×10^5 events at 150 GeV and 3×10^5 events at 330 GeV, allowing models to achieve balanced contributions from low and high energies.
- **Test dataset** To avoid bias toward the training energy points, the ML models are evaluated on a test dataset covering 13 energy points (30, 50, 150, 200, 250, 330, 400, 500, 600, 700, 800, 900 and 1000 GeV), with 1×10^4 events per energy point. All events used in training are excluded from the test dataset.

The MC truth information including SSP and shower axis, and the simulated energy deposition of each cell, are used for model optimization.

3.2. Definition of the shower starting point

For simulated events, the true SSP can be obtained by following the incident particle till its endpoint. The endpoint of the primary track is often the vertex where most first-generation secondary particles (produced directly in the primary particle's inelastic collision with the nucleus) are generated, and this vertex is defined as the MC truth of SSP. However, simply using the vertex that generates most secondaries within CALO geometry as the MC truth is not satisfactory. Some of these interactions are elastic or too soft, which do not initiate a hadronic shower. Such events are excluded from the analysis, as discussed in detail in Section 3.3.

[1] Particles are emitted uniformly in all directions from each point on the sphere surface, with the fluence for each direction being proportional to the cosine of the angle between the source direction and the local normal to the sphere surface.

3.3. Event selection

To achieve better alignment between the simulation and the experimental data, Poisson smearing and noise filtration are applied. The hit energy is first converted to an equivalent photon number using a ratio of 200 photons per MIP. This photon number is subjected to Poisson fluctuation and then reconverted to energy to simulate statistical variability. Subsequently, a noise threshold of 1/3 MIP (equivalent to 10 MeV) is applied. Hit energies below this threshold are set to zero to emulate noise filtration. The following selection criteria are also applied:

- The visible energy, defined as the sum of the hit energies in an event, $E_{\text{vis}} = \sum^{\text{event}} E_{\text{hit}}$, is required to be at least 20 % of the primary particle's energy E_{prim}:

$$E_{\text{vis}}/E_{\text{prim}} \geq 0.2 \tag{1}$$

  This cut selects well-contained showers and is applied to both training and test datasets.
- The total path length L is obtained by summing up the traversal lengths of the incident particle through each crystal along its trajectory. $L \geq 20$ cm is required to ensure full shower development.
- The total kinetic energy of generated secondary particles $E_{\text{kin}}^{\text{gen}}$ is calculated as:

$$E_{\text{kin}}^{\text{gen}} = \sum E_{\text{kin}}^{\text{sec}} - E_{\text{kin}}^{\text{lead}} \tag{2}$$

  where $\sum E_{\text{kin}}^{\text{sec}}$ is the sum of the kinetic energies of all first-generation secondaries, and $E_{\text{kin}}^{\text{lead}}$ is the kinetic energy of the leading secondary proton (i.e. the most energetic proton emerging from the interaction). Fig. 3 illustrates a correlation plot of the number of first-generation secondaries N_{sec} versus $E_{\text{kin}}^{\text{gen}}$. The plot suggests the following cut to reject elastic and too soft interactions and provide a well-defined sample:

$$E_{\text{kin}}^{\text{gen}}/E_{\text{prim}} \geq 1/3 \tag{3}$$

  This cut ensures that a large fraction of the primary particle's energy is transferred to new particle production. At 30 GeV, approximately 7 % of events are rejected. This rejection fraction decreases to 6 % at 50 GeV and stabilizes at 5 % for energies above 150 GeV. Note that this kinetic energy cut is applied only during the ML model training and is removed during inference to reflect a realistic application where $E_{\text{kin}}^{\text{gen}}$ is unknown.

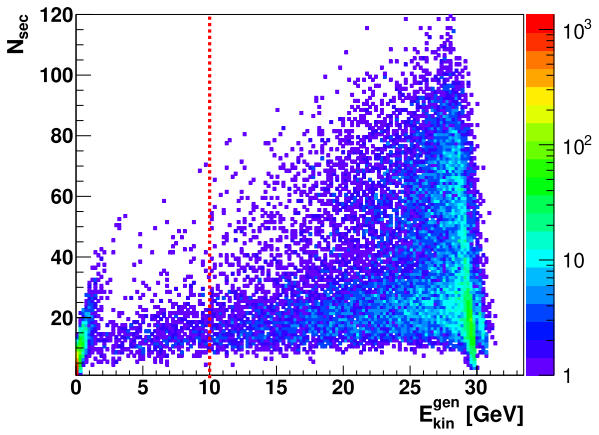


Fig. 3. Correlation plot of the number of secondaries N_{sec} versus the total kinetic energy of generated secondaries $E_{\text{kin}}^{\text{gen}}$ for 30 GeV protons. The frequency of occurrence is represented by a logarithmic color scale. The dashed red line is the threshold to reject elastic and too soft interactions. The small cluster of events around $E_{\text{kin}}^{\text{gen}} = 0\,\text{GeV}$ and $N_{\text{sec}} < 20$ corresponds to elastic interactions. Other events that do not pass the cut are termed soft inelastic interactions.

4. Implementation of SSP reconstruction

4.1. Layer-wise method

The layer-wise method is based on detecting a sudden increase in shower development, as detailed in Ref. [15]. We segment the crystal array along the shower axis into M layers of equal thickness (see Fig. 4(a)), where $M = L/3.458\,\text{cm}$ and L represents the distance between the entry and exit points where the shower axis intersects the CALO boundary. For layer i ($1 \le i \le M$), the moving average of deposited energy A_i is calculated as:

$$\begin{aligned} A_i &= \frac{1}{i}\sum_{k=1}^{i} E_k^{\text{layer}}, \quad i < 4 \\ A_i &= \frac{1}{4}\sum_{k=i-3}^{i} E_k^{\text{layer}}, \quad i \ge 4 \end{aligned} \tag{4}$$

where E_k^{layer} is the energy deposited in the kth layer. SSP is reconstructed at the intersection of the ith layer with the shower axis if both A_i and hit counts N_i in two consecutive layers fulfill:

$$\begin{aligned} &A_i + A_{i+1} > E_{\text{thr}}\ [\text{MIP}], \\ &N_i + N_{i+1} > N_{\text{thr}}. \end{aligned} \tag{5}$$

The thresholds E_{thr} and N_{thr} are dependent on the primary particle's energy E_{prim} (in GeV):

$$\begin{aligned} E_{\text{thr}} &= 0.02 \times E_{\text{prim}} + 9.71, \\ N_{\text{thr}} &= \lfloor 0.68 \times \log E_{\text{prim}} - 1.41 \rfloor. \end{aligned} \tag{6}$$

4.2. Energy-ratio method

The primary proton typically produces a MIP track before initiating a hadronic shower. By analyzing the longitudinal profile of energy deposition density, we can determine the position at which the energy loss transits from minimum ionization to shower cascade [5]. In Fig. 4(b), let N be the number of crystals traversed by the shower axis (marked in red). Their energy deposition densities ($\mathrm{d}E/\mathrm{d}x$) are defined as the energy deposition per unit path length:

$$\frac{\mathrm{d}E}{\mathrm{d}x}[n] = \frac{E_{\text{hit},n}}{l_n}, \quad n = 1,\ 2,\ \ldots,\ N. \tag{7}$$

Here, $E_{\text{hit},n}$ and l_n are the energy deposition and traversal length in the nth crystal respectively. Starting from the first crystal ($n = 1$), the algorithm iterates through each crystal as shown in Fig. 5:

If $1 < n < N - 1$, the energy ratios are defined as follows:

$$\text{ratio}_b[n+i] = \frac{\mathrm{d}E/\mathrm{d}x[n+i]}{\mathrm{d}E/\mathrm{d}x[n-1]} > R_{\text{thr},b}[i], \quad i = 0,\ 1,\ 2. \tag{8}$$

The index n denotes a target crystal, and $n + i$ represent the ith crystal before (-) or after (+) the target crystal. If $\text{ratio}_b[n+i]$ for both the target crystal (n) and the next two crystals ($n+1$, $n+2$) exceed their respective thresholds, SSP is reconstructed at the geometric center of the target crystal.

If $n = 1$ (the first crystal), the most probable value (MPV) of the $\mathrm{d}E/\mathrm{d}x$ distribution for LYSO crystals is used instead of $\mathrm{d}E/\mathrm{d}x[n-1]$, as there is no crystal before the first crystal. In this case, the energy ratios are defined as follows:

$$\text{ratio}_a[n+i] = \frac{\mathrm{d}E/\mathrm{d}x[n+i]}{K_{\text{LYSO}}} > R_{\text{thr},a}[i], \quad i = 0,\ 1,\ 2. \tag{9}$$

where $K_{\text{LYSO}} = 9.6\,\text{MeV/cm}$ is the most probable value of $\mathrm{d}E/\mathrm{d}x$ distribution for MIPs in LYSO crystals.

If $n \ge N - 1$ (the last two crystals), SSP is directly identified as the geometric center of the second to last crystal.

To achieve the best trade-off between accuracy and efficiency, the thresholds $R_{\text{thr},a}[i]$ and $R_{\text{thr},b}[i]$ are set to:

$$\begin{aligned} R_{\text{thr},a}[0] &= 0.0045 \times E_{\text{prim}} + 6.2136, \\ R_{\text{thr},a}[1] &= 0.0356 \times E_{\text{prim}} + 9.0385, \\ R_{\text{thr},a}[2] &= 0.0759 \times E_{\text{prim}} + 8.3348, \\ R_{\text{thr},b}[0] &= 0.0023 \times E_{\text{prim}} + 2.4380, \\ R_{\text{thr},b}[1] &= 0.0090 \times E_{\text{prim}} + 3.5906, \\ R_{\text{thr},b}[2] &= 0.0215 \times E_{\text{prim}} + 3.4964, \end{aligned} \tag{10}$$

where E_{prim} is the primary particle's energy in GeV.

4.3. IncepCNN method

The IncepCNN model is inspired by GoogLeNet [29], which has demonstrated excellent performance in image classification and regression tasks. As shown in Fig. 6, our IncepCNN architecture consists of three main stages: a channel expansion block, a feature extraction block, and a prediction network.

The hit energies are structured into a 4D matrix of dimensions 1×23×23×21 (channel × height × width × depth), with empty positions padded with zeros. This matrix is normalized to the range [0, 1] to aid model convergence. After two convolutional layers that expand the number of channels, the input is sent to the Inception module. The core of IncepCNN's power comes from the use of Inception modules, which enhance feature learning capacity and allow for efficient dimensionality reduction. As shown at the bottom of Fig. 6, the Inception module receives feature maps from the previous layer and processes them through four parallel branches. These branches apply convolutional layers at multiple scales ($1 \times 1 \times 1$, $3 \times 3 \times 3$, $5 \times 5 \times 5$) to capture features at different spatial scales. The $1 \times 1 \times 1$ kernel is used before the $3 \times 3 \times 3$ and $5 \times 5 \times 5$ convolutions to reduce the channel dimensionality, reducing the model's complexity. The outputs from all branches are then concatenated along the channel dimension, forming the module's output.

Before and after the Inception module, a downsampling layer (Max Pooling or Adaptive Average Pooling) is applied to reduce the spatial dimension of feature maps. The extracted features are flattened and then concatenated with the shower axis, thereby constraining the IncepCNN to work along the particle incidence direction and improving accuracy. Finally, these features are fed into the fully connected layers for SSP prediction.

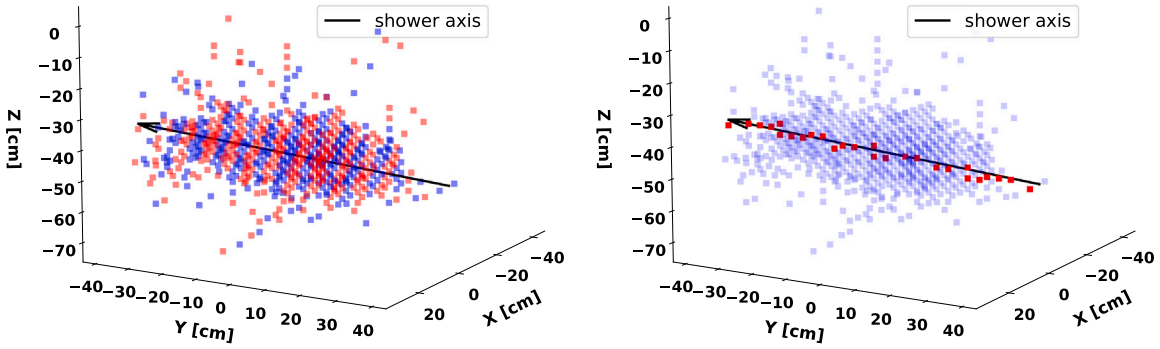


Fig. 4. Simulation of a 150 GeV proton-induced hadronic shower: (a) The layer segmentation in the layer-wise method. Each layer is represented by alternating colors (blue and red) and perpendicular to the shower axis (black arrow); (b) Crystals traversed by the shower axis (red cubes). These crystals are selected to calculate energy ratios.

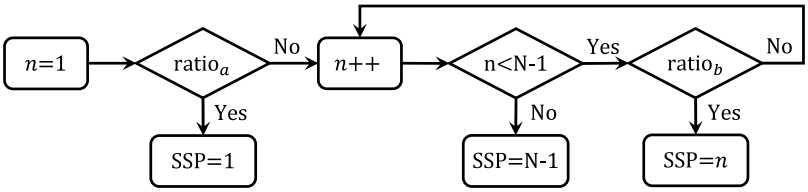


Fig. 5. Flowchart of the energy-ratio method for SSP identification.

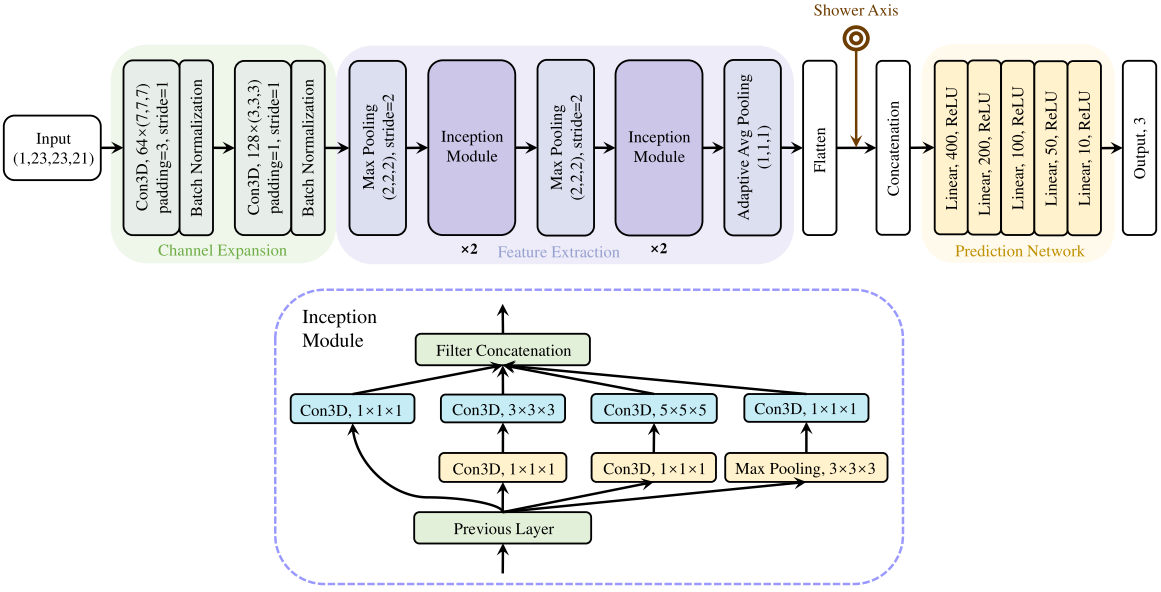


Fig. 6. Top: Architecture of IncepCNN model for SSP reconstruction, comprising a channel expansion block, a feature extraction block and a prediction network. **Bottom**: Structure of the Inception module.

Table 1
Hyperparameters used to train, model parameters (in millions), training time (in hours) for 30 epochs, memory footprint (in GB) and minimum MSE loss ($\times 10^{-2}$).

Model	Learning rate	Batch size	Params [M]	Time [h]	Memory [GB]	MSE_{min} [$\times 10^{-2}$]
IncepCNN	0.001	128	8.6	4.5	73.1	80.0
Transformer	0.001	128	4.3	11.0	98.0	77.2

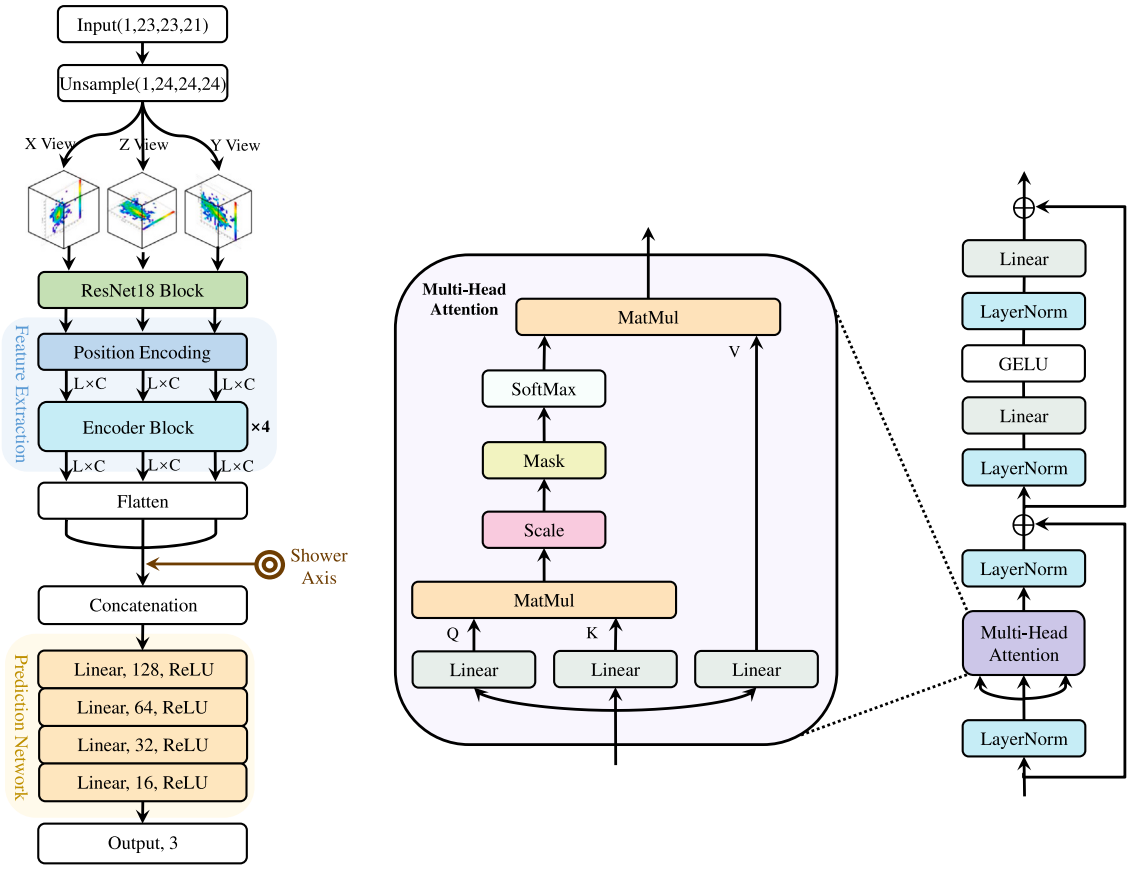


Fig. 7. Left: Architecture of Transformer for SSP reconstruction, comprising a ResNet18 block, a position encoding, several encoder blocks and a prediction network. **Right**: Encoder block used in the Transformer.

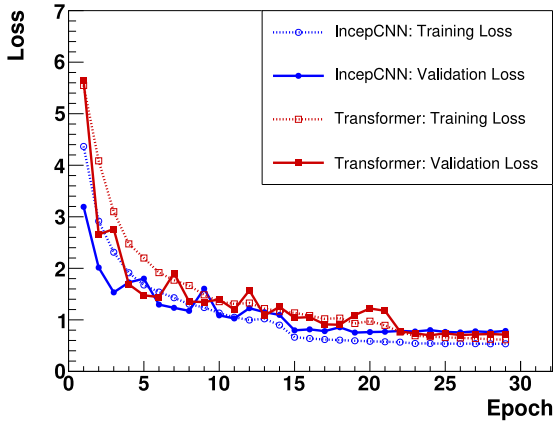


Fig. 8. Loss curves for training and validation datasets. Blue and red lines represent IncepCNN and Transformer models respectively.

4.4. Transformer method

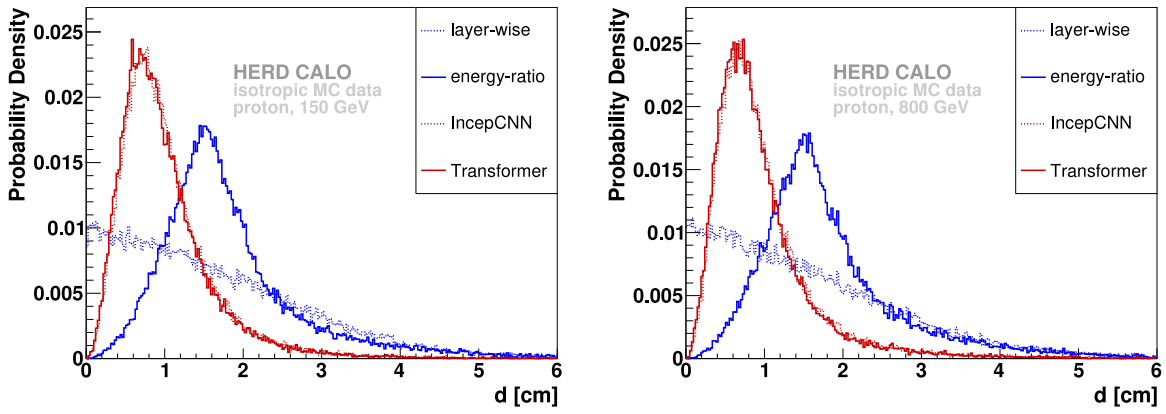
The field of computer vision is currently undergoing a paradigm shift in backbone architectures transitioning from CNNs to Transformers. This trend came with the proposal of Vision Transformer (ViT) [21], which employs a standard Transformer encoder to model global spatial dependencies between image patches. Building upon ViT's success in 2D vision, researchers have extended Transformer architectures to 3D tasks [22].

However, the attention mechanism's computational complexity scales quadratically with image size [30], making it computationally prohibitive to feed the 3D image directly into the Transformer model. To address this problem, we adopt a multi-view training strategy [31, 32] that decomposes the 3D image into 2D slices along three views (x, y and z), processing each view independently. Our results demonstrate that this approach significantly reduces parameters compared to direct 3D processing without compromising performance, as the 3D spatial relations in the shower development can be captured by fusing the extracted features from all three views to form a complete 3D representation.

The architecture of our Transformer model, illustrated in Fig. 7, is designed to implement this strategy and consists of four main stages: a ResNet18 block, positional encoding, encoder blocks, and a prediction network. The 3D image is first upsampled into dimensions of (1,24,24,24) to standardize its size for multi-view processing. Each view is independently processed by a ResNet18 block, which projects the 3D image into an embedded slice with arbitrary dimension (denoted as C). To ensure the model understands the order of slices within a view and can distinguish between the different views, we apply a combined positional encoding. This positional encoding merges a sinusoidal encoding (to preserve slice order) with a learnable encoding (to identify the view type).

Next, these slice embeddings are processed by a stack of encoder blocks. The multi-head attention mechanism within these blocks allows the model to weigh the importance of different slices, effectively capturing both local and global inter-slice dependencies. Finally, to fuse cross-view information, we concatenate the flattened features with an additional token representing the shower axis, and then feed the result into the prediction network.

The Transformer model, as well as the IncepCNN, is developed in `PyTorch` version 2.0.0 and trained on an NVIDIA H100 GPU. We employ the Mean Squared Error (MSE) as the loss function to calculate



(a) (b)

Fig. 9. Normalized distribution of d for simulated isotropic protons at (a) 150 GeV and (b) 800 GeV. The layer-wise method constrains the reconstructed SSP to the shower axis. Since the true SSP also lies almost entirely on this axis, the layer-wise method yields a most probable value at $d = 0$. In contrast, ML methods place the SSP anywhere in CALO, resulting in a non-zero peak. Similarly, the energy-ratio method fixes the SSP to crystal centers, also producing a non-zero peak.

Table 2
Reconstruction accuracy, offset and spatial resolution for various energies.

Method	Energy [GeV]												
	30	50	150	200	250	330	400	500	600	700	800	900	1000
Layer-wise													
Accuracy [%]	60.9	60.8	62.4	62.8	62.6	63.6	63.6	63.6	65.1	65.0	64.6	63.9	64.2
Offset [cm]	1.63	1.64	1.59	1.58	1.57	1.56	1.55	1.55	1.53	1.53	1.53	1.56	1.54
d_{68} [cm]	2.07	2.10	2.05	2.02	2.02	1.98	1.98	1.99	1.94	1.94	1.94	1.98	1.95
Energy-ratio													
Accuracy [%]	48.1	52.0	54.4	53.3	55.0	55.5	55.7	54.2	55.7	53.7	55.1	55.5	53.5
Offset [cm]	2.00	1.97	1.93	1.96	1.93	1.93	1.93	1.92	1.91	1.94	1.92	1.93	1.93
d_{68} [cm]	2.60	2.36	2.28	2.30	2.26	2.27	2.26	2.27	2.22	2.28	2.25	2.25	2.29
IncepCNN													
Accuracy [%]	62.8	79.8	90.1	90.7	91.6	91.9	91.5	90.6	91.2	91.0	89.6	89.7	89.6
Offset [cm]	1.69	1.32	1.03	1.01	0.98	0.98	0.96	0.99	0.97	0.97	1.01	1.02	1.02
d_{68} [cm]	1.99	1.46	1.13	1.12	1.09	1.07	1.06	1.07	1.06	1.06	1.08	1.10	1.09
Transformer													
Accuracy [%]	64.6	79.8	91.1	91.0	91.9	92.5	92.2	91.7	91.7	92.5	91.0	91.1	90.7
Offset [cm]	1.67	1.29	1.00	0.99	0.96	0.95	0.93	0.95	0.94	0.94	0.96	0.97	0.98
d_{68} [cm]	1.93	1.44	1.10	1.09	1.07	1.03	1.02	1.04	1.03	1.03	1.03	1.05	1.05

the difference between the true and predicted SSP. The AdamW optimizer is used to minimize the loss function and estimate the model's parameters. The learning rate is dynamically adjusted using a `ReduceLROnPlateau` scheduler, which monitors the validation loss with a patience of 4 epochs and a factor of 0.1. Table 1 summarizes the model parameters, training time, memory footprint, and training hyperparameters. The Transformer's parameters are reduced by 0.5 compared to those of the IncepCNN, but it requires a longer training time and a larger memory footprint, primarily due to its slice encoding process and the computational cost of attention mechanisms.

As depicted in Fig. 8, the MSE loss of both the Transformer and the IncepCNN converges after 25 epochs. The losses on the training and validation datasets are similar. We select the epoch with the minimum validation loss for further analysis. With the optimized models, we evaluate their performance on test samples, as described in Section 5.

5. SSP reconstruction results

The distances (d) between the reconstructed and true SSP are calculated using the four methods. As an example, for simulated isotropic protons with 150 GeV and 800 GeV, the d distribution is shown in Fig. 9. The following metrics are defined to evaluate performance of reconstruction methods:

offset The mean value of the d distribution.

spatial resolution The d_{68} value, where 68 % of events satisfy $d < d_{68}$.

accuracy The fraction of events with the difference between reconstructed and true SSP in the x, y and z directions less than ±1.5 cm, divided by the total number of events, as specified in Eq. (11).

$$\text{Accuracy} = \frac{N_{|\Delta_{x,y,z}| \le 1.5\,\text{cm}}}{N_{\text{total}}}\ [\%] \tag{11}$$

Fig. 10 shows the correlation between the reconstructed and true SSP for isotropic 150 GeV protons in the x, y and z directions. The high density of entries along the diagonal indicates that SSP is generally well reconstructed by all methods. We note that the energy-ratio method yields discrete values, corresponding to crystal center coordinates. Fig. 11 compares the accuracy and spatial resolution of four methods. Overall, both metrics show a flat dependence on energy. At energies above 150 GeV, IncepCNN and Transformer consistently outperform the other two methods, achieving approximately 30 % improvement in accuracy and 1.0 cm better spatial resolution.

For the two ML methods, the observed decline in accuracy and d_{68} at 30 GeV and 50 GeV can be attributed to two factors. First, the number of hits in low-energy showers is very small, resulting in sparse 3D images that make it challenging for the models to capture shower characteristics effectively. Second, the training dataset contained only 150 GeV and 330 GeV events, without lower-energy samples. This configuration ensures excellent high-energy performance at the cost of

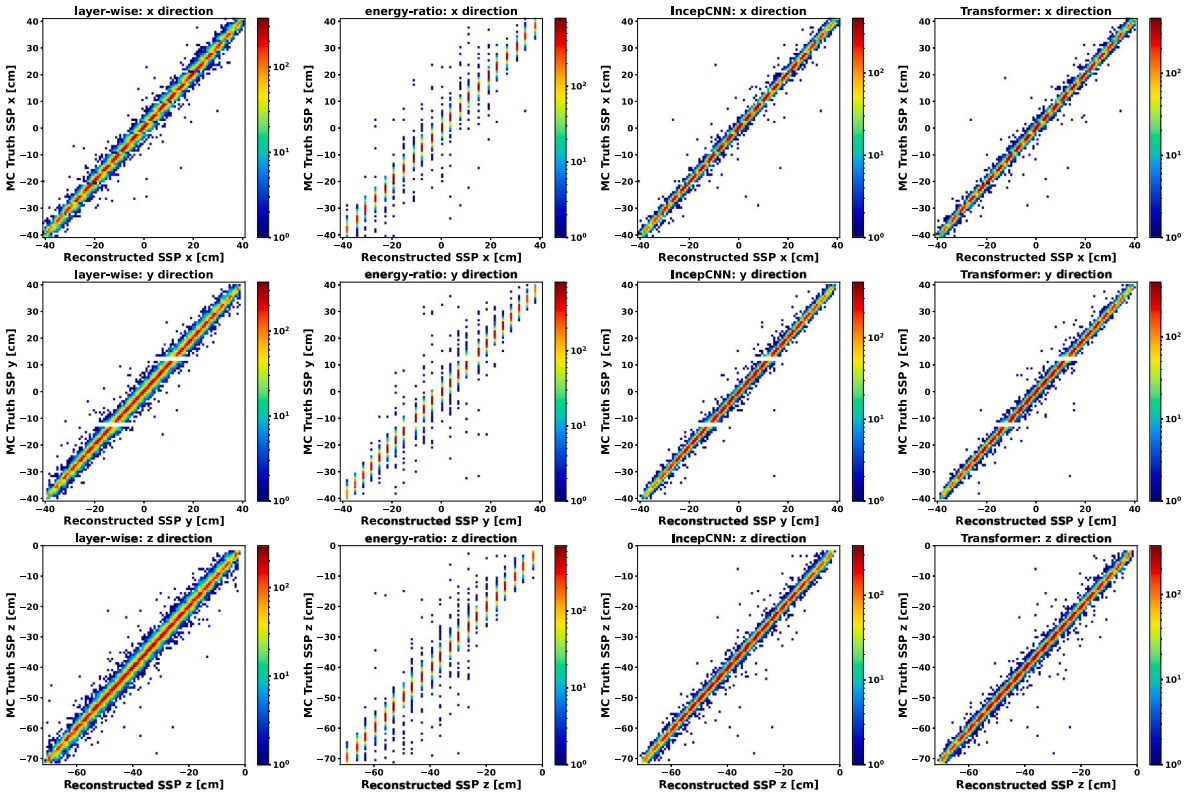
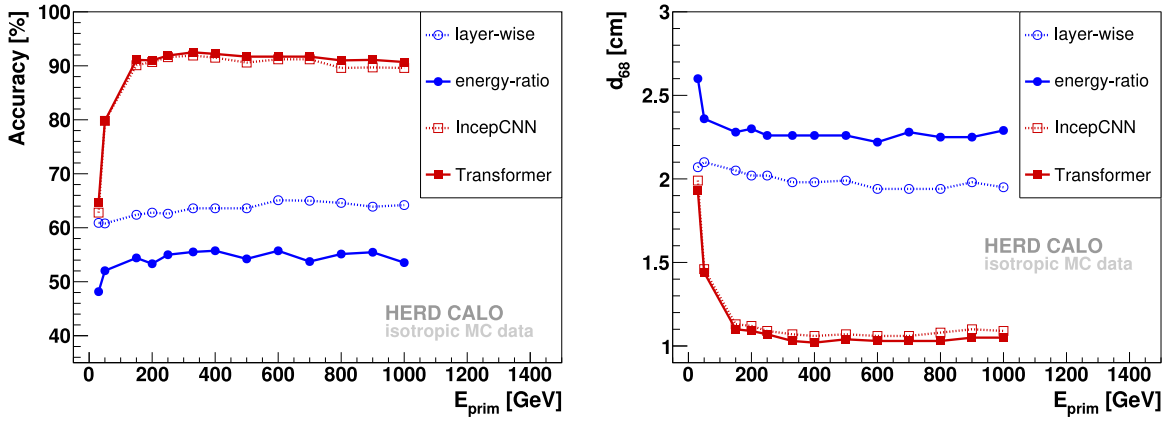


Fig. 10. Correlation between reconstructed and true SSP for 150 GeV protons. Each row represents spatial dimensions (x, y, z), with columns showing different reconstruction methods (layer-wise, energy-ratio, IncepCNN, Transformer). Color intensity indicates event frequency. The discontinuous distribution of the true SSP in the y direction is due to the gaps between individual crystals.



(a) (b)

Fig. 11. The accuracy (a) and spatial resolutions (b) for simulated isotropic protons.

some performance loss at low energies. This trade-off is considered acceptable, as the subsequent section focuses on energy reconstruction for high-energy particles, where precise SSP reconstruction at high energies is the main scope of this work.

Table 2 summarizes the reconstruction results. The advancement of ML methods demonstrates their generalization capability and robustness, confirming that models can be trained with limited simulation data to extrapolate reconstruction to energies not used for training.

6. Application of SSP to leakage correction

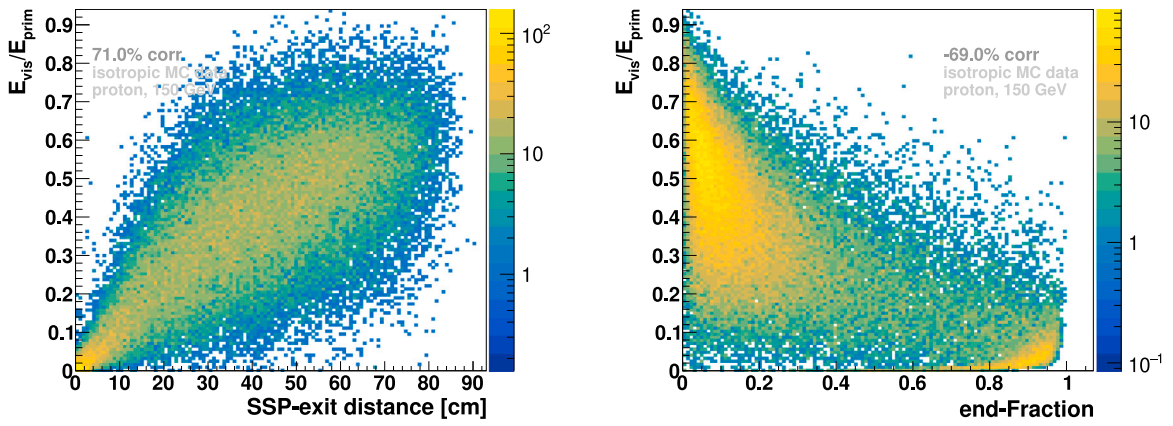
In this section, we present an application of SSP in leakage correction. SSPs are obtained from four methods (layer-wise, energy-ratio, IncepCNN and Transformer) that have been optimized in Section 5. To avoid using primary energy information in the correction, we replace the primary energy (E_{prim}) with the visible energy (E_{vis}) for threshold determination in the layer-wise and energy-ratio methods.

6.1. Correction algorithm

Typically, the energy of an incoming particle is reconstructed from the visible energy (E_{vis}) by multiplication with a suitable weight (ω) [33]:

$$E_{\text{rec}} = \omega \cdot E_{\text{vis}}\,[\text{GeV}], \tag{12}$$

In more complicated schemes, ω can depend on two observables (the SSP-exit distance and the end-fraction) and several free parameters [6].



(a) (b)

Fig. 12. Correlation plots from simulated 150 GeV isotropic proton showers: (a) SSP-exit distance versus normalized measured energy, (b) end-fraction versus normalized measured energy. The correlation coefficients are 71 % and −69 % respectively. Events with an end-fraction between 0.8 and 1 correspond to very late-starting showers that have a small SSP-exit distance. In these events, the shower development is concentrated in the last few crystal layers and only a very small amount of energy is deposited.

The SSP-exit distance refers to the distance from the SSP to the exit point where the shower axis intersects the CALO boundary. The end-fraction is defined as the energy fraction deposited in the last 10 % of the shower development in CALO. In Fig. 12, the left distribution shows the correlation between the SSP-exit distance and the normalized measured energy, while the right distribution presents the corresponding correlation for the end-fraction. The Pearson correlation coefficients of 71 % and −69 % respectively confirm their strong sensitivity to energy leakage.

We divide the end-fraction distribution into eight equal-probability bins (i.e. a 12.5 % probability for a given end-fraction to be found in any one of the bins). For events with end-fraction in the bth bin, the weight ω in Eq. (12) is defined as a quadratic function of the SSP-exit distance with three free parameters, $p_0^{(b)}$, $p_1^{(b)}$, $p_2^{(b)}$, shown in Eq. (13):

$$\omega(x) = p_0^{(b)} + p_1^{(b)} \cdot x + p_2^{(b)} \cdot x^2, \tag{13}$$

where x = (SSP-exit distance)$/S$, S = 70 cm is a scale factor. Eq. (13) introduces 24 parameters (3 free parameters × 8 bins). The event energy can then be reconstructed by combining Eqs. (12) and (13).

All parameters are tuned using simulated proton events with a uniform spectrum in the energy range from 100 GeV to 1 TeV, considering isotropic and vertical incidence directions separately. The optimization simultaneously considers all energies in the spectrum to find a set of general parameters that minimize the relative deviations $|(E_{\text{rec}} - E_{\text{prim}})/E_{\text{prim}}|$.

We use all selection cuts specified in Section 3.3, except for $E_{\text{kin}}^{\text{gen}}/E_{\text{prim}} \geq 1/3$. Additionally, events are selected by requiring an SSP-exit distance greater than 31 cm to remove those with severe leakage, which are challenging to reconstruct.

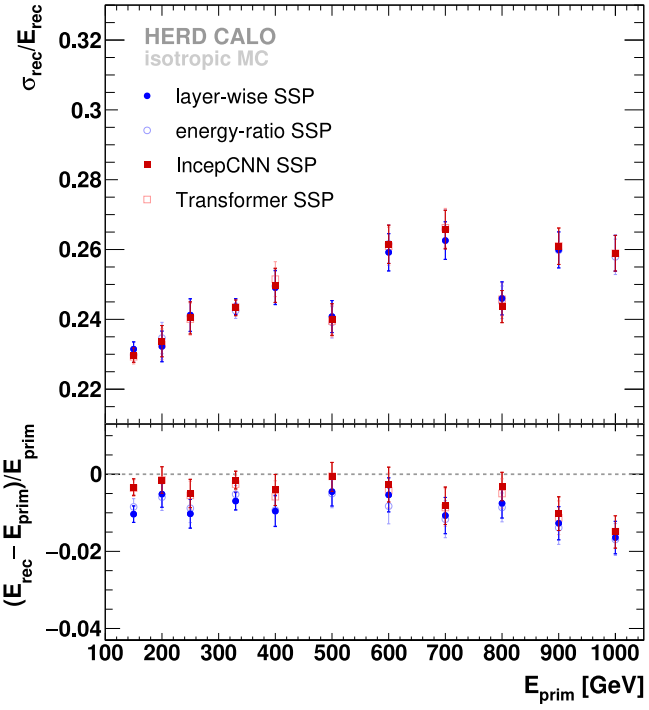


Fig. 13. CALO energy resolution (top) and bias (bottom) for simulated isotropic protons. Blue and red indicate results using SSPs obtained from traditional and ML methods, respectively.

6.2. Correction on isotropic Monte Carlo samples

The correction algorithm, optimized using isotropic MC samples with a uniform spectrum from 100 GeV to 1 TeV, is applied to 11 energy points (150 GeV, 200 GeV, 250 GeV, 330 GeV, 400 GeV, 500 GeV, 600 GeV, 700 GeV, 800 GeV, 900 GeV and 1000 GeV) to test its performance. As shown in Fig. 13, the weighting technique yields reconstructed energies with a bias of about 1 % and a resolution ranging from 23 % to 27 %. Compared to SSPs obtained from traditional methods, the ML-based SSPs reduce the energy bias by ∼0.4 % at low energies and by 0.1–0.2 % at higher energies. However, a gradual deterioration in energy resolution is observed with increasing energy. This trend can be attributed to the impact of energy leakage, as the visible energy distribution (E_{vis}) exhibits a skewed "leakage tail" that deviates from the expected Gaussian distribution. As the primary particle energy increases, energy leakage becomes more severe, amplifying this leakage tail. The correction algorithm that employs a weighting factor effectively corrects the mean of E_{vis} distribution but fails to fully restore its symmetry.

6.3. Correction on test beam samples

Experimental data were collected with the CALO prototype during the 2024 test beam campaign at CERN. The CALO prototype, comprising a 7 × 7 × 21 array of LYSO cubes, was tested with a 350 GeV proton beams at the CERN Super Proton Synchrotron (SPS). In this analysis, only the well-calibrated central 3 × 3 × 21 crystals are used.

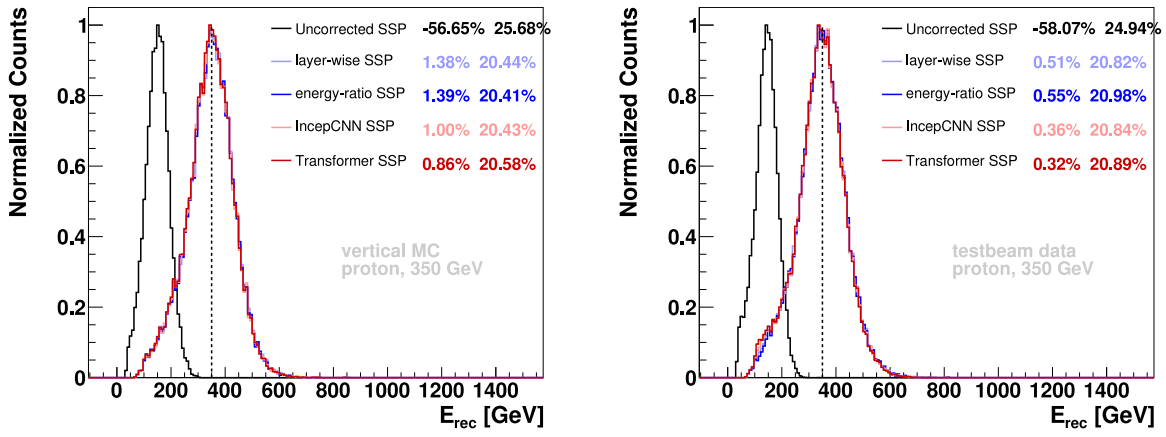


Fig. 14. Reconstructed energy spectra of 350 GeV protons: (a) simulation and (b) 2024 test beam data. Black line indicates uncorrected calorimeter response (E_{vis}), while blue and red lines indicate corrected results using SSPs obtained from traditional and ML methods, respectively. The energy bias and resolution are tabulated in the upper right corner.

As mentioned previously, the correction weight (ω) for experimental data differs from that used for isotropic simulations due to the smaller crystal array size. Accordingly, the algorithm is re-optimized using simulated proton samples with a uniform energy spectrum from 100 GeV to 1 TeV and vertical incidence. In event selection, the threshold for the visible energy cut is reduced to 10 % from the previously used 20 % to adapt to the smaller array size.

Fig. 14 presents the reconstructed energy spectra for 350 GeV test beam data as well as a simulation thereof. To improve the agreement between simulation and test beam data, a selection on the center of gravity of energy deposition in the x-y plane ($< 1 \times 1$ cm^2) is applied to both. No significant difference is observed after this selection. The ML-based SSPs achieve the lowest energy bias: 0.9 % bias with 20.6 % resolution for simulation, and 0.3 % bias with 20.9 % resolution for test beam data. These results indicate that the implemented leakage correction works effectively for both simulation and data.

6.4. Discussion

The fundamental challenge in proton energy reconstruction stems from the complexity of hadronic showers and the practical constraints of the CALO. The electromagnetic fraction of a hadronic shower fluctuates stochastically from event to event and varies with the primary energy, leading to a non-linear calorimeter response and fundamentally limiting the energy resolution. Additionally, part of the energy deposited by a hadronic shower cannot be detected and is called "invisible energy" (e.g., from neutrinos and nuclear binding energy losses), and its fluctuating fraction further degrades the resolution. Moreover, the compact design of the CALO, with a depth of only 3 nuclear interaction lengths, results in significant energy leakage for showers above several hundred GeV, especially given the isotropic incidence of cosmic rays. This leakage becomes more severe with increasing energy, causing the measured energy distribution to develop a pronounced non-Gaussian tail and leading to the observed resolution degradation. Consequently, the development of hadronic showers cannot be fully characterized by only two observables like the shower starting point and the end-fraction, necessitating more representative observables from the full 3D shower profile or more advanced reconstruction techniques to solve these effects.

7. Conclusion

Precise SSP reconstruction is of importance for the data analysis of HERD CALO. We have presented two ML models (IncepCNN and Transformer) to reconstruct SSP, starting from the training with the simulation through the validation with the test beam data. Compared to traditional methods, both ML methods demonstrate superior performance, achieving a 30 % higher accuracy and 1.0 cm better spatial resolution. These results indicate that a well-designed ML model can overcome the limitations of traditional techniques, and generalizes well to energies beyond the training range.

During model training, the configuration of training dataset should be designed for specific applications. For energy reconstruction of high-energy particles, we use a dataset that prioritizes high-energy performance. Conversely, for applications like calorimeter on-orbit calibration where low-energy protons are dominant, a training dataset with a power-law spectrum would be more appropriate. Furthermore, training specialized models for different energy ranges might bring about further improvements.

When applied to energy leakage correction, the ML-based SSPs reduce the energy bias to about 1 % for isotropic simulations and 0.3 % for test beam data. However, the improved precision in SSP reconstruction achieved by ML approaches do not yield a significant advantage in energy correction. This can be attributed to the insensitivity of current energy correction algorithms to SSP precision or the dominance of other factors, such as shower development models and detector response. Nevertheless, SSP remains a fundamental parameter in hadronic shower analysis, and its accurate determination is critical for subsequent physical measurements. The 30 % improvement in accuracy demonstrates the potential of ML methods to enhance the reliability of this key step, which may benefit other aspects of shower analysis beyond energy correction, such as particle identification, heavy ion reconstruction and shower shape characterization. Future work could explore how to better leverage the improved SSP precision in energy reconstruction or apply it to areas where SSP accuracy plays a more decisive role.

CRediT authorship contribution statement

X.F. Tang: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Data curation. **Z. Quan:** Writing – review & editing, Software, Investigation, Data curation, Conceptualization. **Y.W. Dong:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **T.W. Bao:** Resources. **C.L. Liao:** Writing – review & editing, Investigation. **X. Liu:** Resources. **J.Y. Sun:** Writing – review & editing, Investigation. **J.J. Wang:** Writing – review & editing, Investigation. **R.J. Wang:** Resources. **Z.G. Wang:** Resources. **Q. Wu:** Writing – review & editing, Investigation. **M. Xu:** Writing – review & editing, Software, Investigation, Data curation, Conceptualization. **X.G. Yang:** Writing – review & editing, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Key R&D Program of China (2021YFA0718401, 2021YFA0718402, 2021YFA0718403, 2021YFA0718404), the National Natural Science Foundation of China (Grant No. 12027803).

Data availability

Data will be made available on request.

References

[1] N. Akchurin, R. Wigmans, Hadron calorimetry, Nucl. Instrum. Meth. A 666 (2012) 80–97, http://dx.doi.org/10.1016/j.nima.2011.10.035.
[2] B. Liu, I. Laktineh, Q. Shen, et al., Particle identification using boosted decision trees in the semi-digital hadronic calorimeter, J. Instrum. 15 (05) (2020) C05022, http://dx.doi.org/10.1088/1748-0221/15/05/c05022.
[3] J. Zang, G. Chen, J. Bian, et al., A study of hadronic shower development in the ECAL of the alpha magnetic spectrometer II, Chin. Phys. C 35 (8) (2011) 763–768, http://dx.doi.org/10.1088/1674-1137/35/8/012.
[4] O. Adriani, G. Barbarino, G. Bazilevskaya, et al., A statistical procedure for the identification of positrons in the PAMELA experiment, Astropart. Phys. 34 (1) (2010) 1–11, http://dx.doi.org/10.1016/j.astropartphys.2010.04.007.
[5] Q. Wu, Z. Quan, T. Bao, et al., Shower-enabled on-orbit calibration algorithm for HERD 3D imaging calorimeter using high-energy cosmic-ray protons, Radiat. Detect. Technol. Methods (2025) http://dx.doi.org/10.1007/s41605-025-00564-2.
[6] S. Lu, Shower leakage correction in a high granularity calorimeter, 2012, arXiv:1201.6260.
[7] Y. Dong, S. Zhang, G. Ambrosi, Overall status of the high energy cosmic radiation detection facility onboard the future China's space station, in: Proceedings of 36th International Cosmic Ray Conference — PoS(ICRC2019), vol. 358, 2019, p. 062, http://dx.doi.org/10.22323/1.358.0062.
[8] F. Gargano, The high energy cosmic-radiation detection (HERD) facility on board the Chinese space station: hunting for high-energy cosmic rays, in: Proceedings of 37th International Cosmic Ray Conference — PoS(ICRC2021), vol. 395, 2021, p. 026, http://dx.doi.org/10.22323/1.395.0026.
[9] D. Kyratzis, Latest advancements of the HERD space mission, Nucl. Instrum. Meth. A 1048 (2023) 167970, http://dx.doi.org/10.1016/j.nima.2022.167970.
[10] X. Liu, O. Adriani, X. Bai, et al., Double read-out system for the calorimeter of the HERD experiment, in: Proceedings of 38th International Cosmic Ray Conference — PoS(ICRC2023), vol. 444, 2023, p. 097, http://dx.doi.org/10.22323/1.444.0097.
[11] D. Kyratzis, F. Alemanno, C. Altomare, et al., The plastic scintillator detector of the HERD space mission, in: Proceedings of 38th International Cosmic Ray Conference — PoS(ICRC2023), vol. 444, 2023, p. 140, http://dx.doi.org/10.22323/1.444.0140.
[12] G. Silvestre, The silicon charge detector of the high energy cosmic radiation detection experiment, J. Instrum. 19 (03) (2024) C03042, http://dx.doi.org/10.1088/1748-0221/19/03/C03042.
[13] C. Dai, H. Liu, X. Liu, et al., Development of Transition Radiation Detector for the High Energy cosmic-Radiation Detection Facility, in: Proceedings of 38th International Cosmic Ray Conference — PoS(ICRC2023), vol. 444, 2023, p. 113, http://dx.doi.org/10.22323/1.444.0113.
[14] O. Adriani, S. Albergo, L. Auditore, et al., CaloCube: An isotropic spaceborne calorimeter for high-energy cosmic rays. Optimization of the detector performance for protons and nuclei, Astropart. Phys. 96 (2017) 11–17, http://dx.doi.org/10.1016/j.astropartphys.2017.10.002.
[15] B. Bilki, J. Repond, L. Xia, et al., Pion and proton showers in the CALICE scintillator-steel analogue hadron calorimeter, J. Instrum. 10 (04) (2015) P04014, http://dx.doi.org/10.1088/1748-0221/10/04/P04014.
[16] B. Acar, G. Adamov, C. Adloff, et al., Performance of the CMS high granularity calorimeter prototype to charged pion beams of 20–300 GeV/c, J. Instrum. 18 (08) (2023) P08014, http://dx.doi.org/10.1088/1748-0221/18/08/P08014.
[17] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, in: Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 25, 2012, pp. 1097–1105, http://dx.doi.org/10.1145/3065386.
[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2015, arXiv:1409.1556.
[19] K. He, X. Zhang, S. Ren, et al., Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.
[20] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, vol. 30, 2017, URL https://arxiv.org/abs/1706.03762.
[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2021, arXiv:2010.11929.
[22] J. Chen, Y. He, E. Frey, et al., ViT-V-Net: Vision transformer for unsupervised volumetric medical image registration, 2021, arXiv:2104.06468.
[23] A. Aurisano, A. Radovic, D. Rocco, et al., A convolutional neural network neutrino event classifier, J. Instrum. 11 (09) (2016) P09001, http://dx.doi.org/10.1088/1748-0221/11/09/P09001.
[24] R. Hashmani, E. Akbaş, M. Demirköz, A comparison of deep learning models for proton background rejection with the AMS electromagnetic calorimeter, Mach. Learn. Sci. Tech. 5 (4) (2024) 045008, http://dx.doi.org/10.1088/2632-2153/ad7cc0.
[25] N. Akchurin, C. Cowden, J. Damgov, et al., On the use of neural networks for energy reconstruction in high-granularity calorimeters, J. Instrum. 16 (12) (2021) P12036, http://dx.doi.org/10.1088/1748-0221/16/12/P12036.
[26] C. Liao, Z. Quan, Y. Dong, et al., Application of machine learning method for energy reconstruction on space based high granularity calorimeter, Exper. Astron. 58 (3) (2024) 12, http://dx.doi.org/10.1007/s10686-024-09957-5.
[27] A. Tykhonov, A. Kotenko, P. Coppin, et al., A deep learning method for the trajectory reconstruction of cosmic rays with the DAMPE mission, Astropart. Phys. 146 (2023) 102795, http://dx.doi.org/10.1016/j.astropartphys.2022.102795.
[28] X. Yang, Z. Quan, Y. Dong, et al., Application of a deep learning method for shower axis reconstruction in a 3D imaging calorimeter, Nucl. Instrum. Meth. A 1066 (2024) 169571, http://dx.doi.org/10.1016/j.nima.2024.169571.
[29] C. Szegedy, W. Liu, Y. Jia, et al., Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 1–9, http://dx.doi.org/10.1109/CVPR.2015.7298594.
[30] F. Keles, P. Wijewardena, C. Hegde, On the computational complexity of self-attention, 2022, arXiv:2209.04881.
[31] E. Jun, S. Jeong, D. Heo, et al., Medical transformer: Universal encoder for 3-D brain MRI analysis, IEEE Trans. Neural Netw. Learn. Syst. 35 (12) (2024) 17779–17789, http://dx.doi.org/10.1109/TNNLS.2023.3308712.
[32] S. Yan, X. Xiong, A. Arnab, et al., Multiview transformers for video recognition, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 3323–3333, http://dx.doi.org/10.1109/CVPR52688.2022.00333.
[33] Y. Kulchitsky, M. Kuzmin, V. Vinogradov, The e/h method of energy reconstruction for combined calorimeter, 1999, arXiv:hep-ex/9912014.