Development of ML Algorithms for the study of Astro-particles

- Luca Tabarroni
- Università di Roma Tor Vergata
- luca.tabarroni@roma2.infn.it

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

ASI Agenzia Spaziale Italiana

SSDC SPACE SCIENCE DATA CENTER

AMS experiment

The Alpha Magnetic Spectrometer is a particles detector operating on the International Space Station used to study the composition of Cosmic Rays.

Search for Anti-Helium

Development of a Machine Learning algorithm for the identification of possible Anti-Helium events in the Cosmic Rays.

Dataset:

- Signal: Monte Carlo simulations of the Helium. Helium differs from Anti-Helium only for the charge.
- Background: Monte Carlo simulations of events easily misinterpreted as Anti-Helium.

Algorithm:

- BDT: Boost Decision Tree trained to distinguish between signal and background.
- AEC: Auto-Encoder trained to reconstruct the signal. The error in the reconstruction is used as selector.

HEPD

The High-Energy Particle Detector is a scientific instrument designed to measure changes in the flow of high-energy particles.

Classification of Protons vs Electrons

Development of a Machine Learning algorithm for the classification of Protons and Electrons crossing the detector.

Dataset:

The dataset is composed of Monte Carlo simulations of events of high energy Protons and Electrons. Each event is characterized by the thirty-two numbers of the Scintillator Counters and by the nine numbers of the Lyso Cristals.

Algorithm:

- BDT: trained to distinguish between Protons and Electrons.
- FCNN: Fully Connected Neural Network whose output is the probability taht each event belongs to one of the two classes.

HERD Calorimeter

HERD is a high-energy cosmic-ray detector based on a deep three-dimensional electromagnetic calorimeter, proposed to be installed on the Chinese Space Station. The main scientific objectives of HERD include detecting dark matter particles, studying cosmic ray composition, and observing high energy gamma rays.

Classification of Protons vs Electrons

The study is focused on the classification task of simulated electrons and protons detected by the HERD Detector done with Machine Learning algorithm.

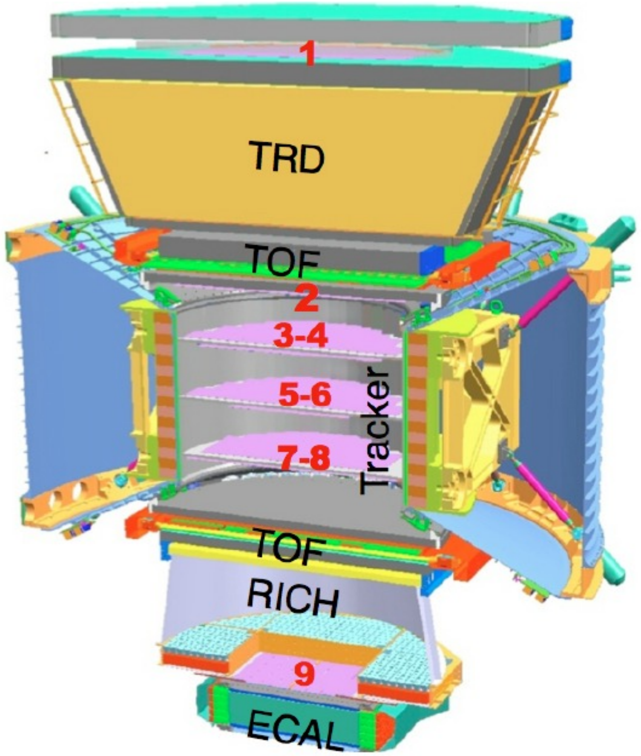
Dataset:

Our classification task is based on data from Monte Carlo simulations of proton and electron particle showers in the HERD electromagnetic calorimeter, with energies ranging from 100 GeV to 20 TeV. We have two datasets, one composed of three-dimensional images, and the other from their two-dimensional projections.

Algorithm:

Our approach is inspired by the Inception neural network, a very deep convolutional neural network that achieved state-of-the-art performance in the ImageNet Large Scale Visual Recognition Challenge 2015 when combined with residual connections.

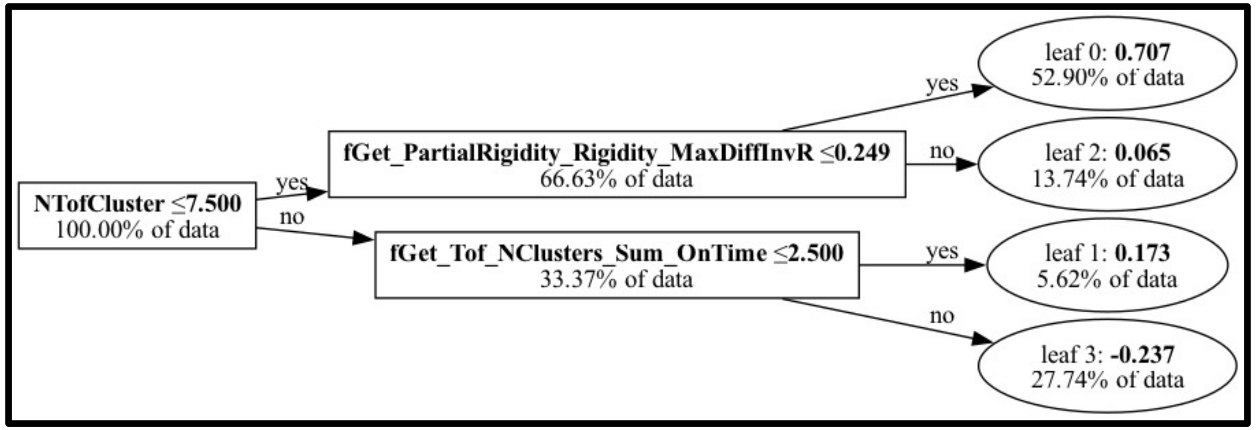
AMS
Anti-Helium research

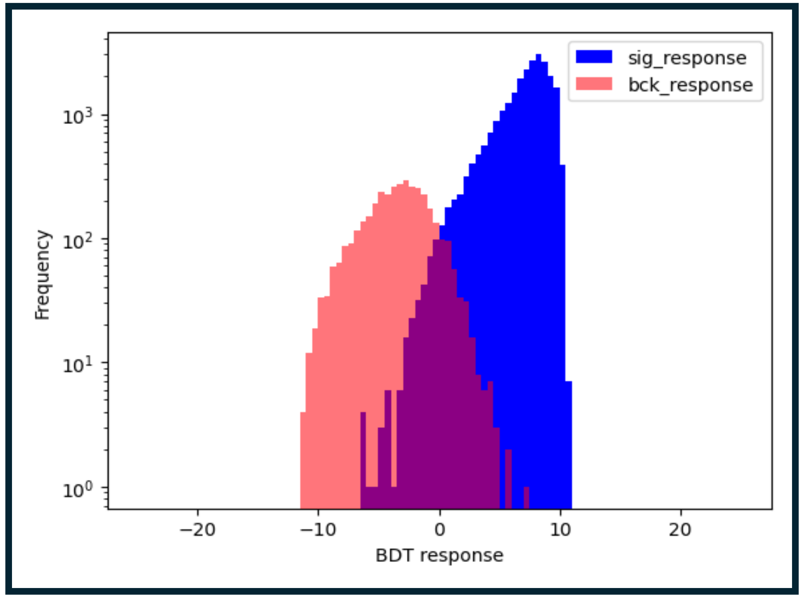


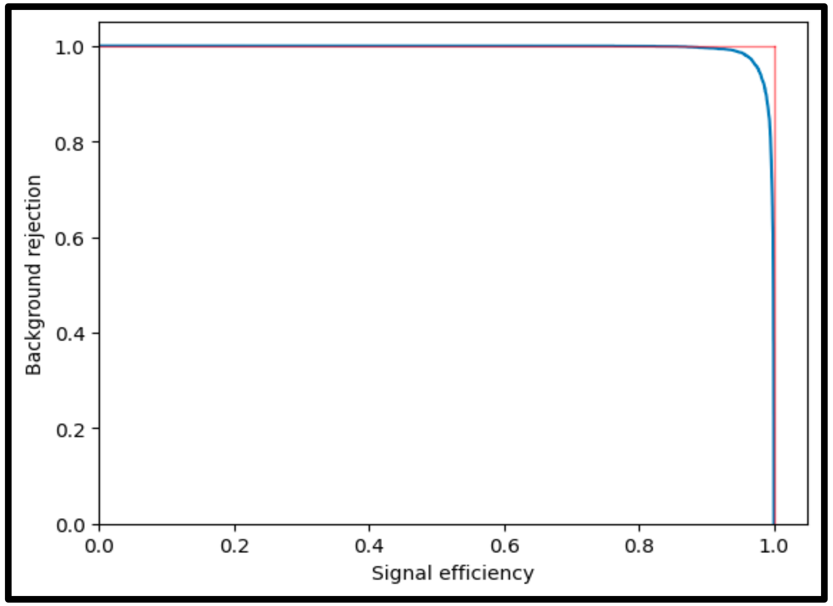
Boost Decision Tree

In our case, we implemented an algorithm with:

- One-hundred trees
- Stopping conditions:
  - 4 leaves
  - No depth conditions
  - No purity conditions
- Gradient Boosting
- Learning rate of 0.25



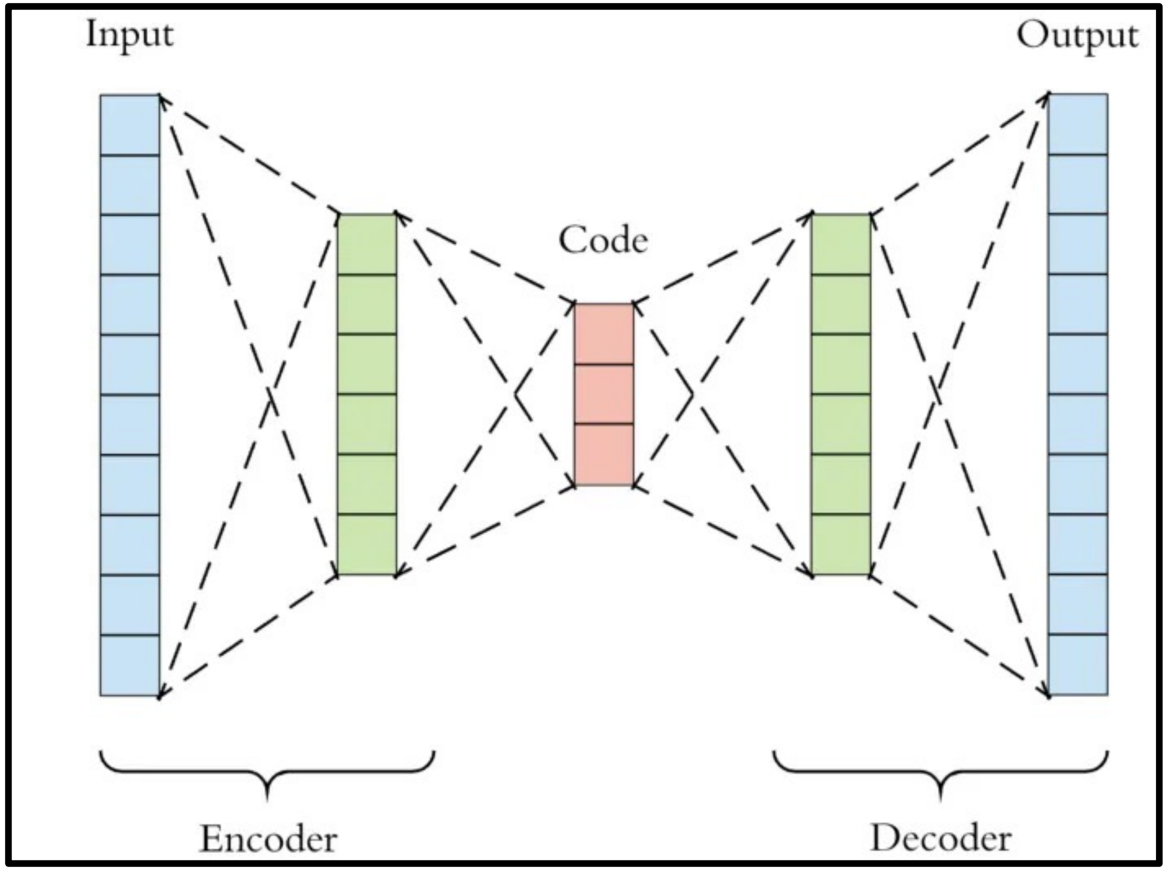




Auto-Encoder General Introduction

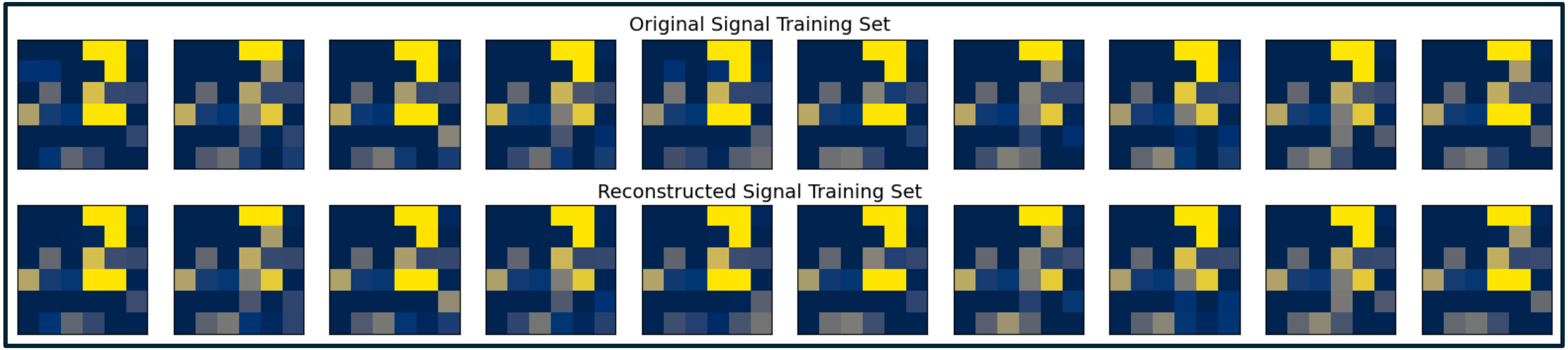
Image compression

- Image reduction to a lower dimension object: Code space
- Image reconstruction from the Code space back to the original dimension
- Convolutional layers
- Max-Pooling layers
- Up-Sampling layers



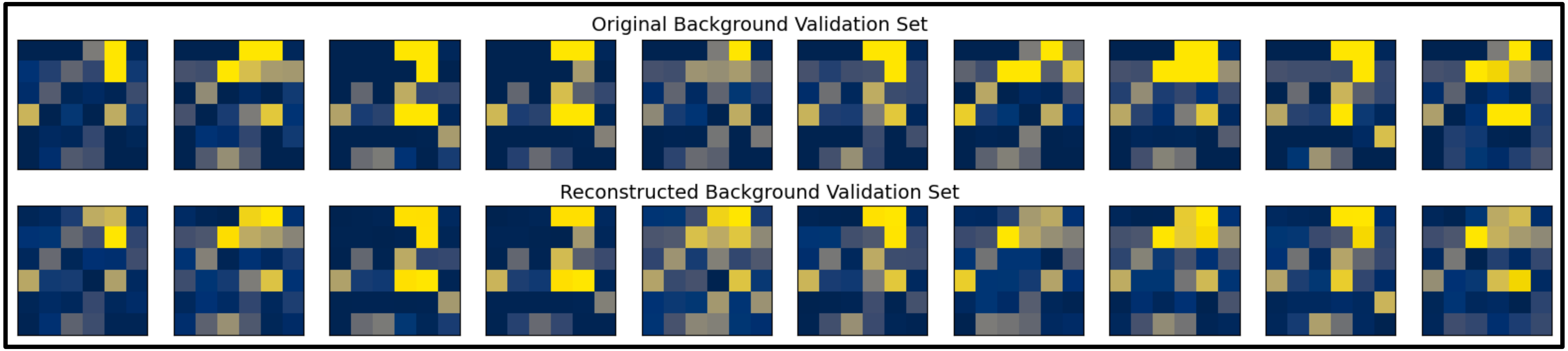
Auto-Encoder Training

- We submit to the AEC a training set of images taken from the original signal dataset
- The AEC will perform its training confronting the output images with the input ones
- The AEC goal is to reach the maximal accordance between the input and the output



Auto-Encoder Validation

- We submit to the AEC a validation set of images taken from the original background dataset
- The AEC has never seen this set of images
- The AEC is expected to perform poorly on the background validation set since it has been trained on the signal



Auto-Encoder Error in Reconstruction

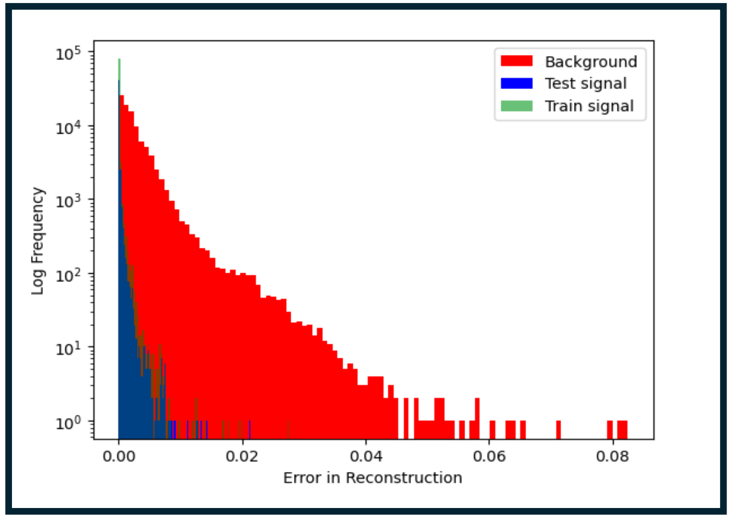
- We can evaluate the difference between an original and a reconstructed image, pixel by pixel:

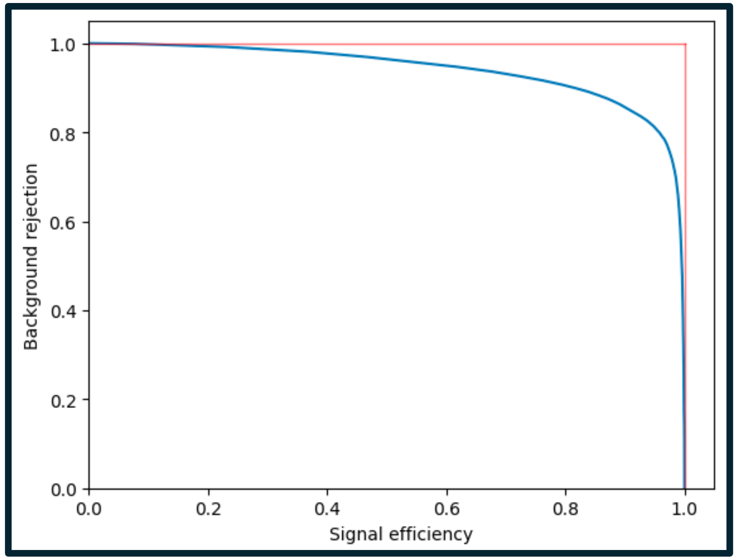
$$E_1 = \frac{(R_1 - O_1)^2}{36}$$

$$\vdots$$

$$E_{36} = \frac{(R_{36} - O_{36})^2}{36}$$

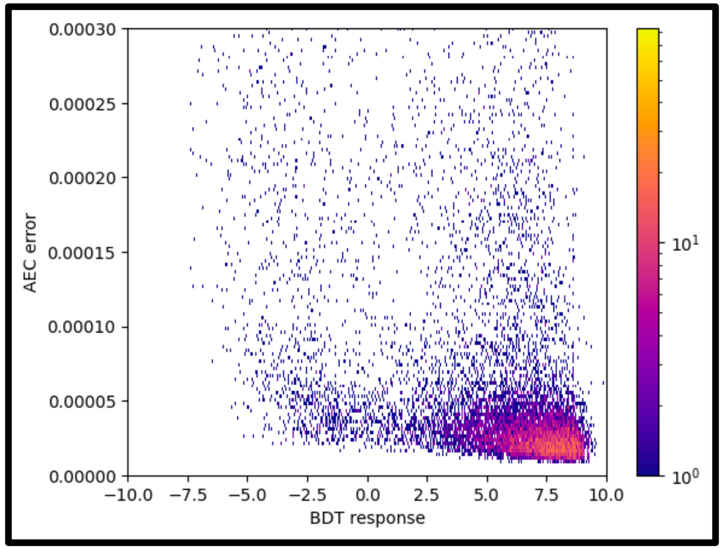
- The total error will be the AEC selector

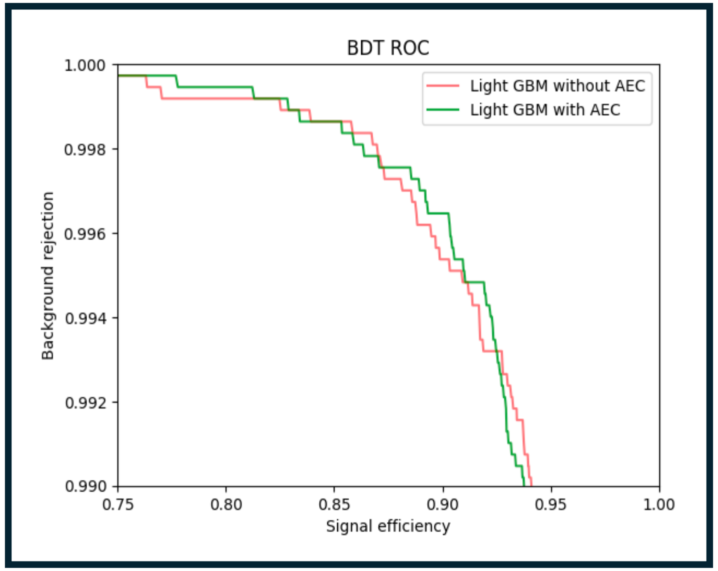




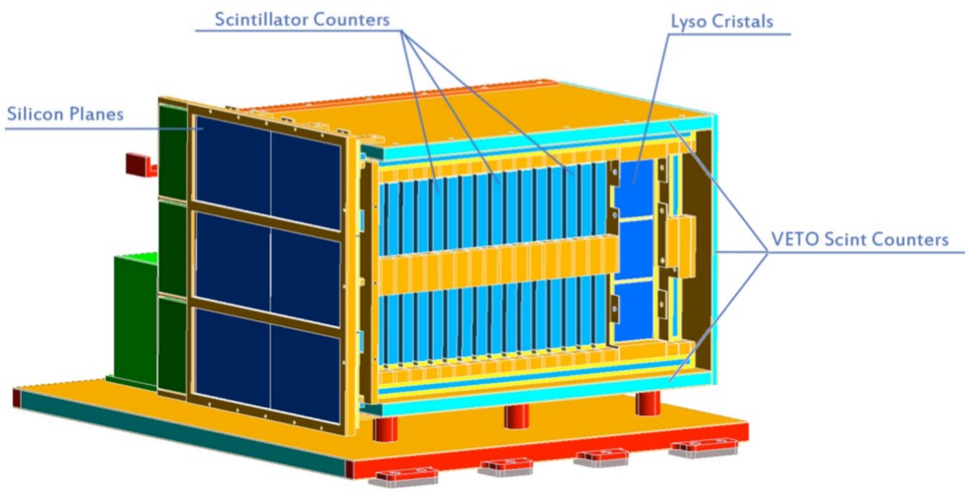
Auto-Encoder AEC Error vs BDT Response

- The signal is characterised by a lower AEC error and a higher BDT response
- The background is characterised by a lower BDT response but no actual range for the AEC error
- The two selectors, the BDT response and the AEC error, are not tightly correlated





HEPD
Electrons/Protons



159329 Protons

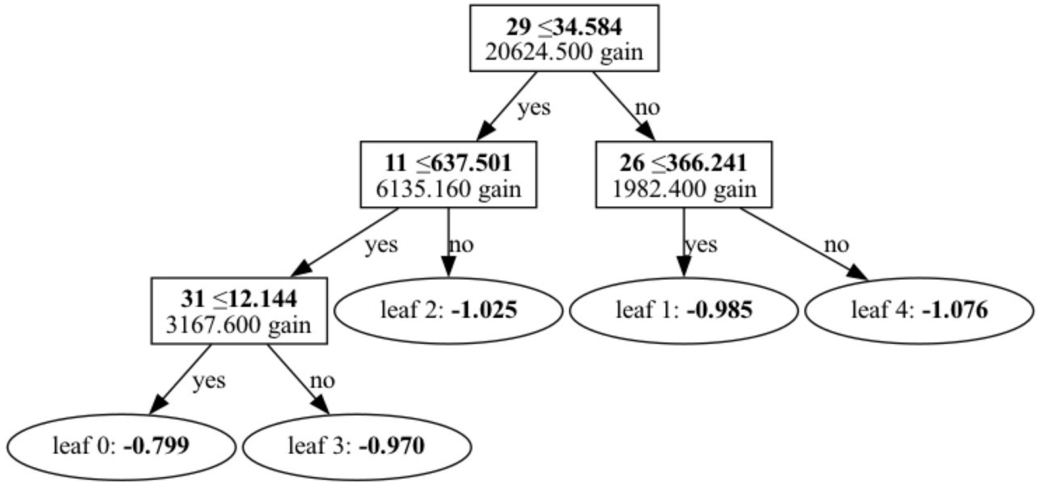
61243 Electrons

31 numbers from the Scintillator counter

9 from the Lyso

BDT parameters

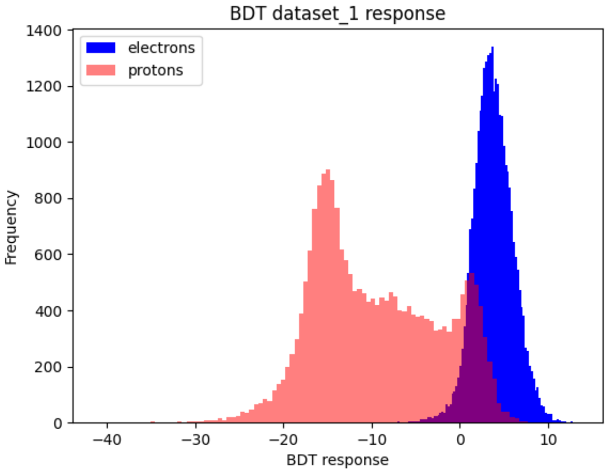
- No maximum depth
- Number of leaves: 30
- Number of estimators: 2500
- Learning rate: 0.05

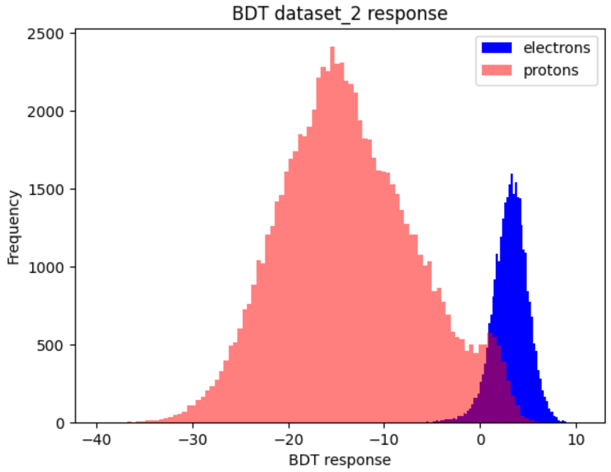


HEPD

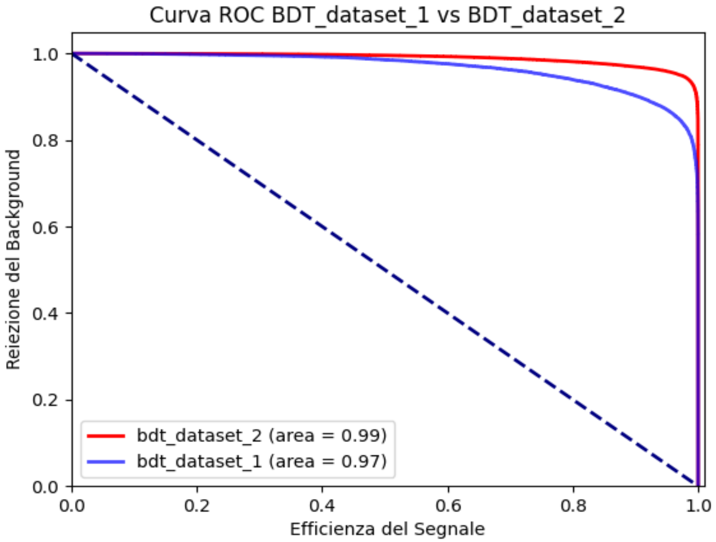
Machine learning algorithm for Protons vs Electrons classification.

BDT trained on two different datasets of different size.

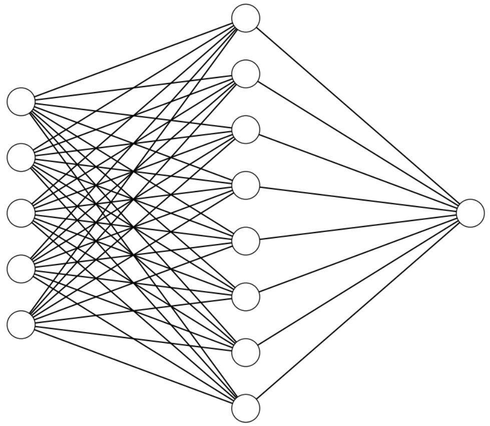




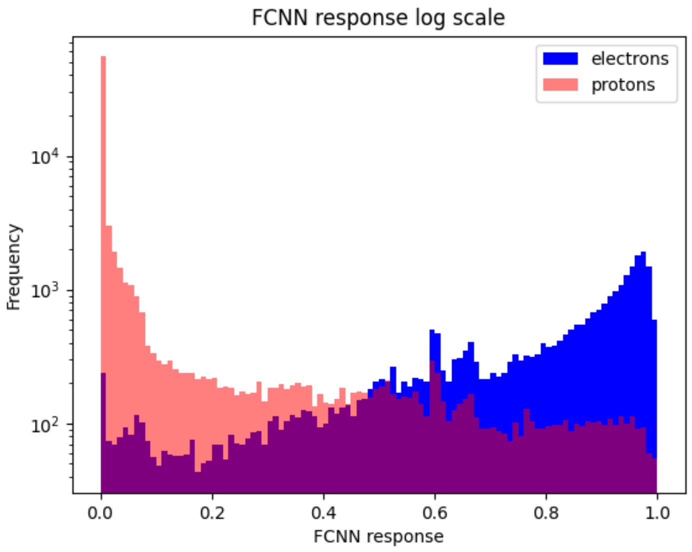
Signal Efficiency	Background Rejection
0.9	0.9690
0.75	0.9849
0.5	0.9957
0.4	0.9973

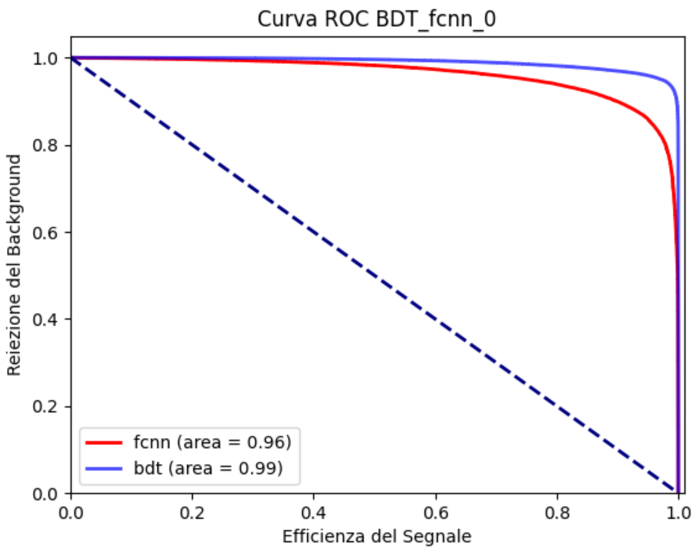


Fully Connected Neural Network



- Forty-dimensional input (31 ADC+ 9 Lyso)
- Deep fully connected neural network
- One-dimensional output, used for classification



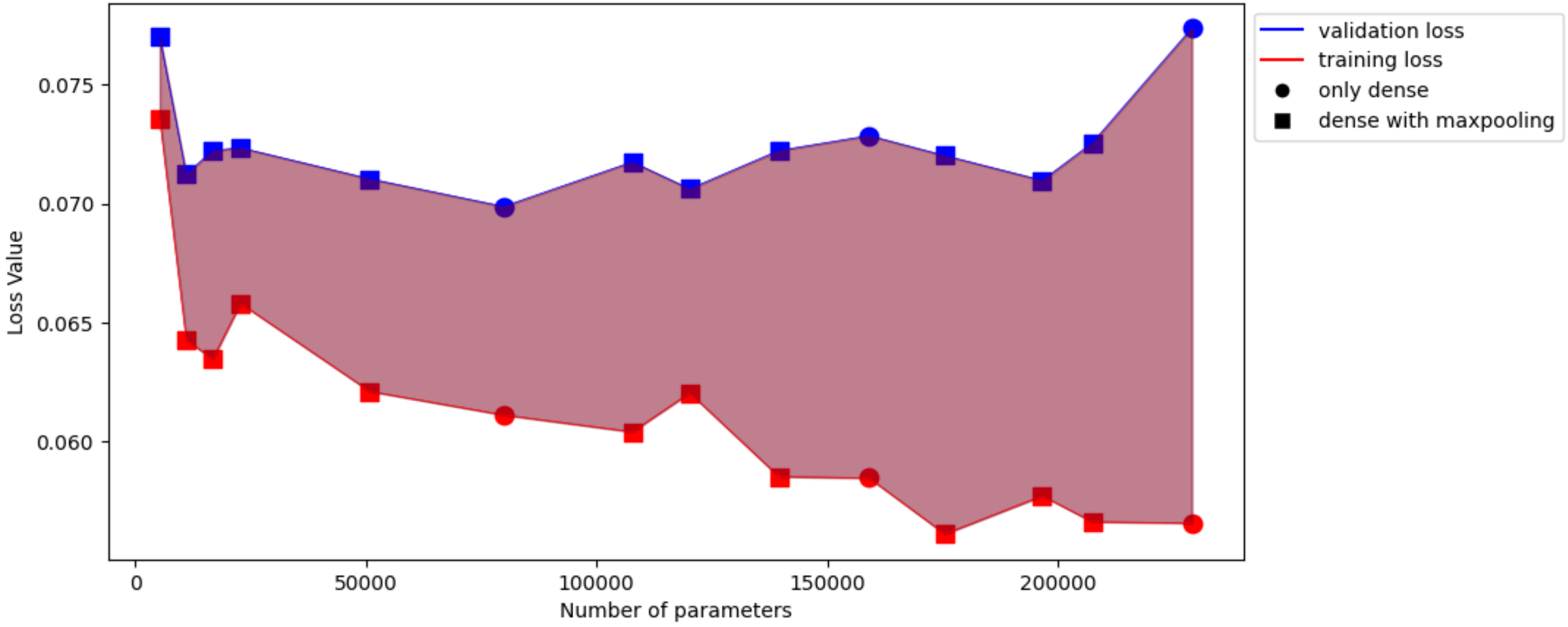


Systematic Study

We are looking for the best Neural Network:

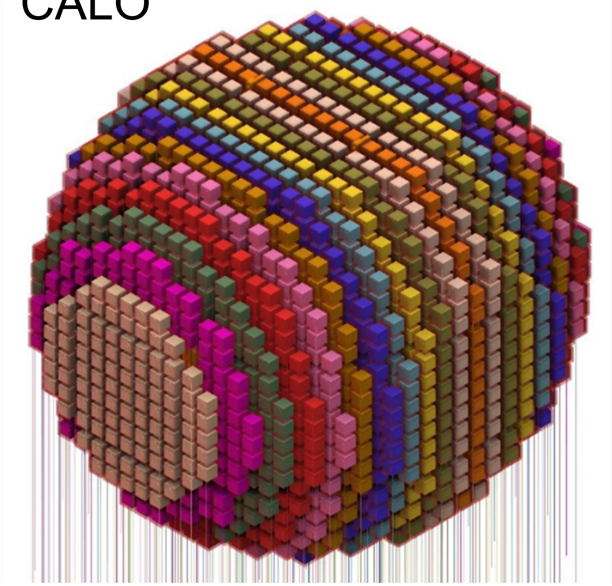
- Complex enough to classify the data

- Simply enough to avoid ’’overfitting’’



HERD
Electrons/Protons

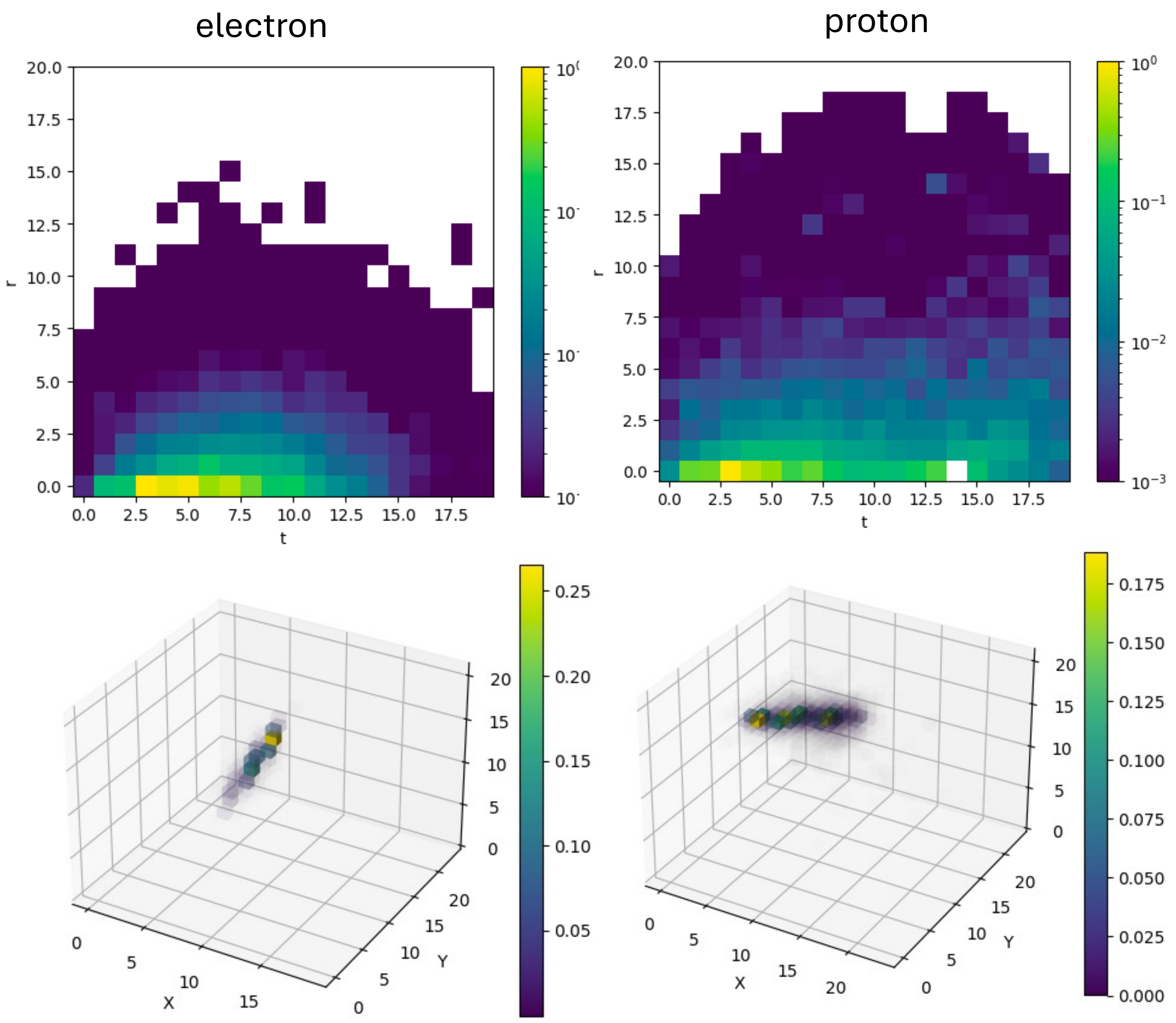
CALO



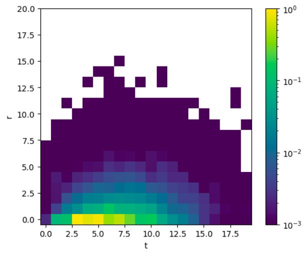
- ~7500 LYSO cubes
- 3x3x3 cm each
- spherical shape to accept events from any direction

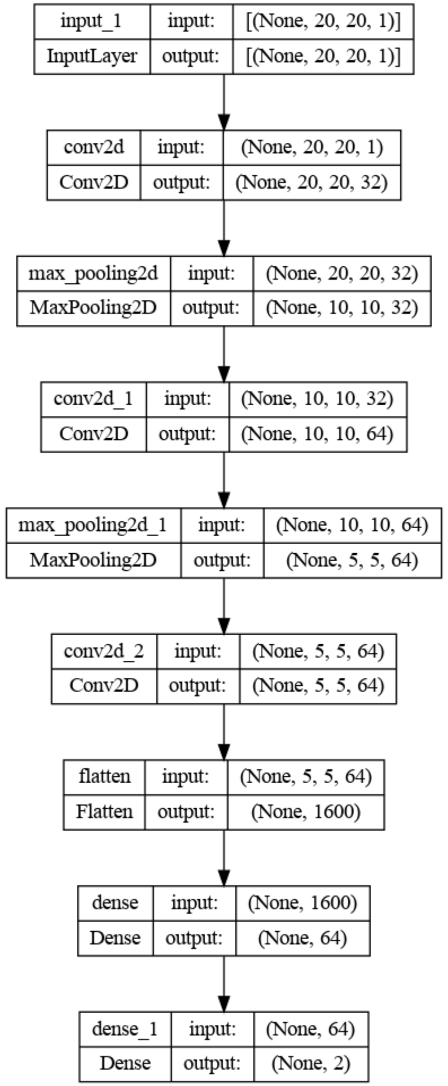
DATASET

Proton and electron particle showers in the HERD electromagnetic calorimeter, with energies ranging from 100 GeV to 20 TeV.

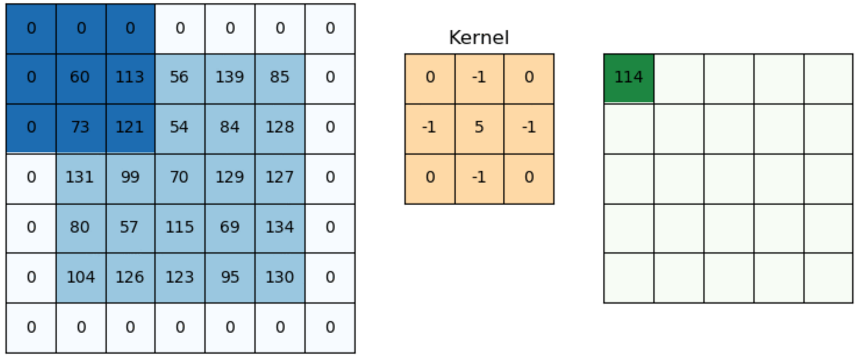


- Electron events: 284515
- Proton events: 160171

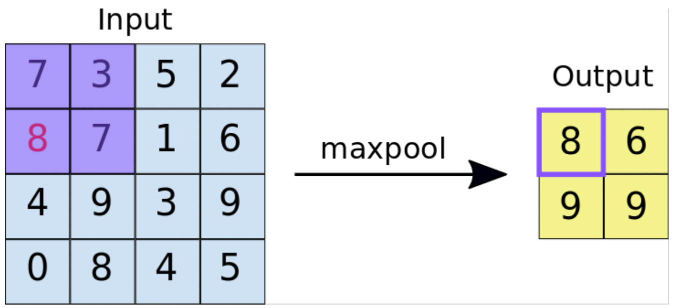


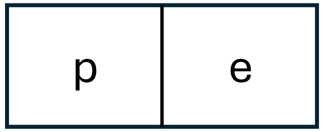


Convolutional layer



Max pooling layer





p ''protoness''

e ''electroness''

p>e result: proton

p<e result: electron

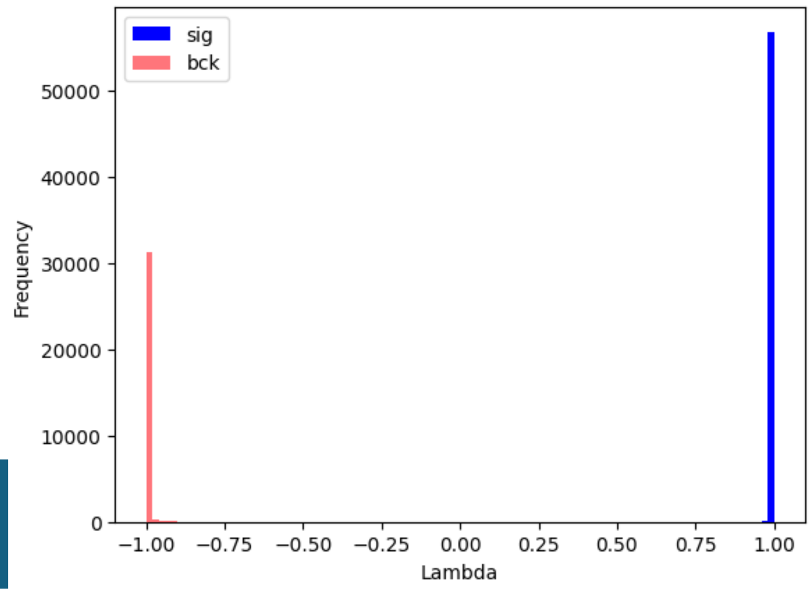
$$\lambda = \frac{e-p}{e+p}$$

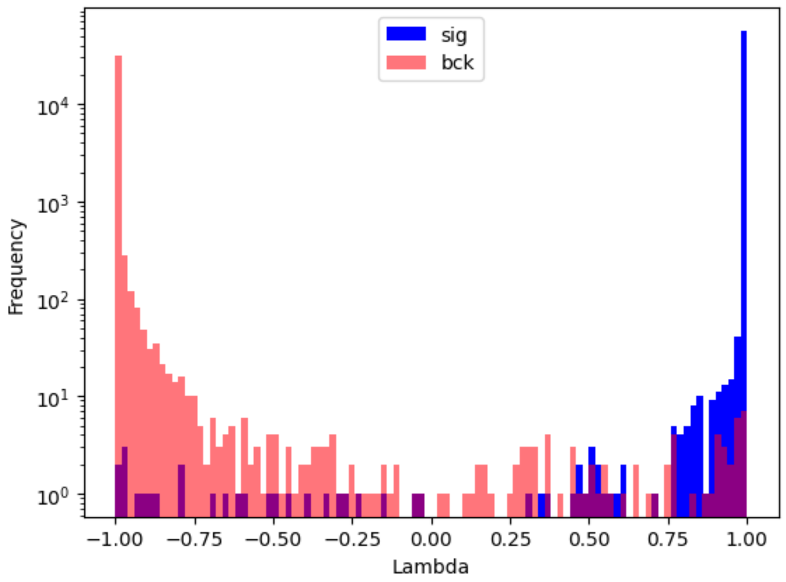
Results

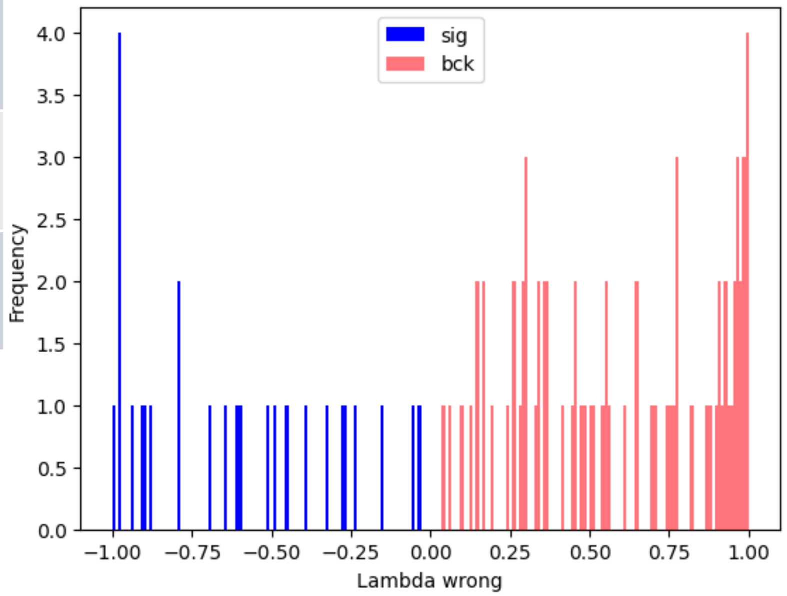
Training/Test split:

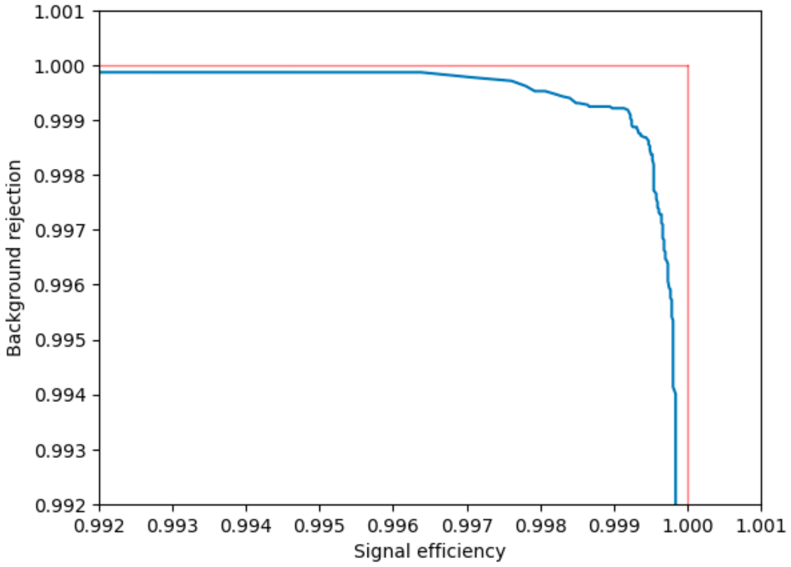
80-20 : misclassified protons $\backsim$ 20 over 32000
$(6 * 10^{-4})$

Signal efficiency	Background rejection
1.0	0.9766
0.9999	0.9906
0.9898	1.0



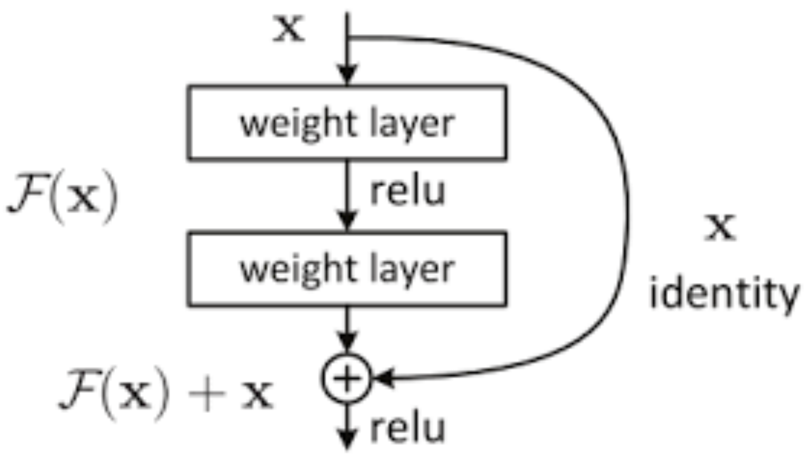






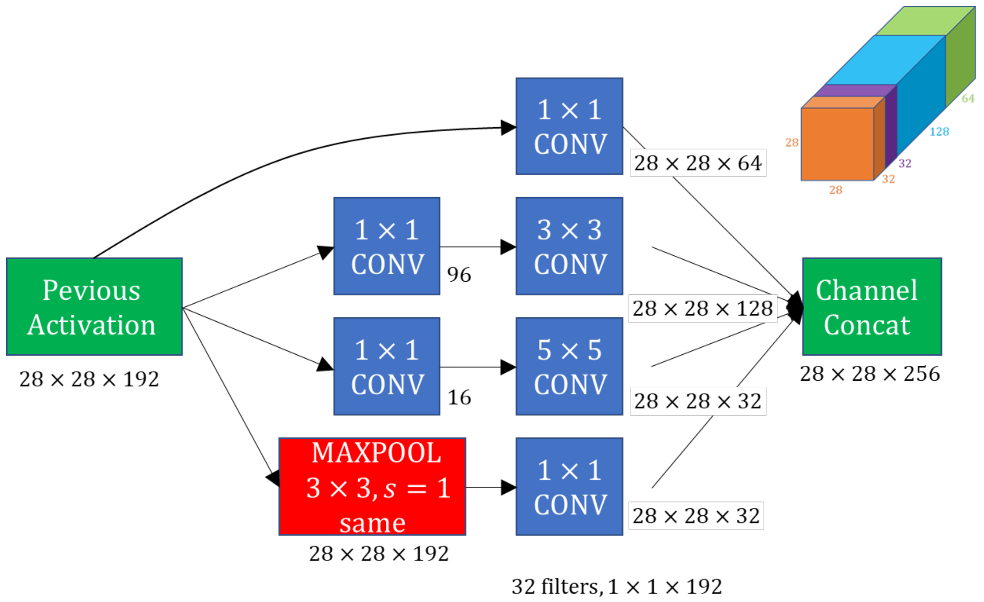
Current developments

ResNet



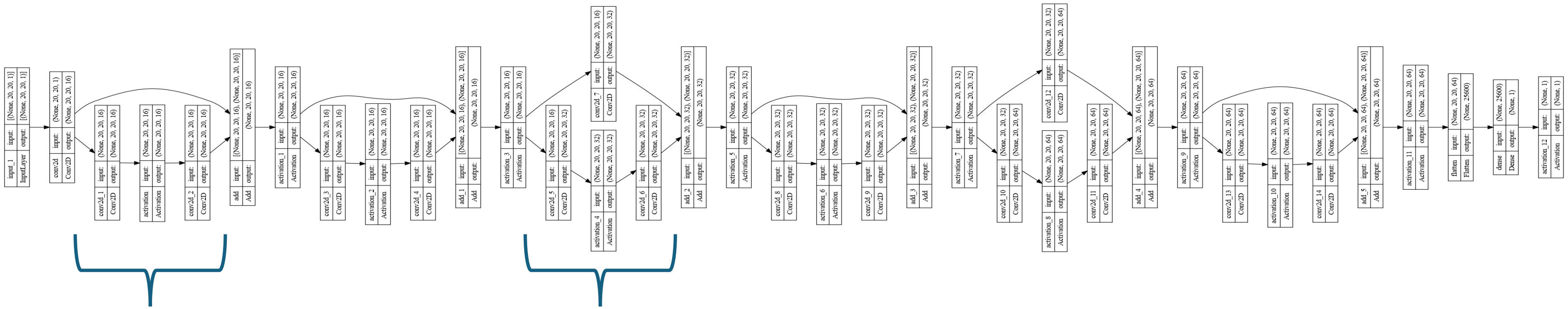
Improve the training efficiency maintaining
the correlation between the gradients
at each step

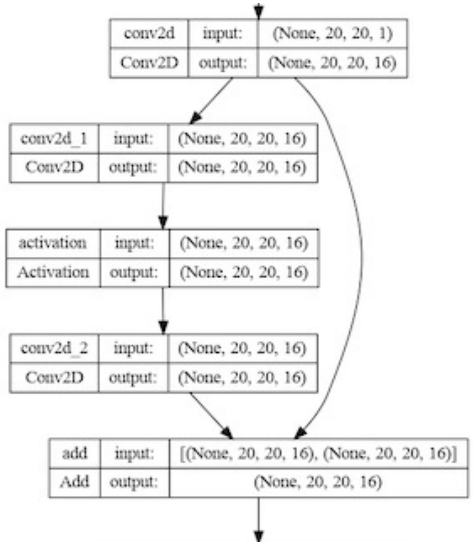
Inception

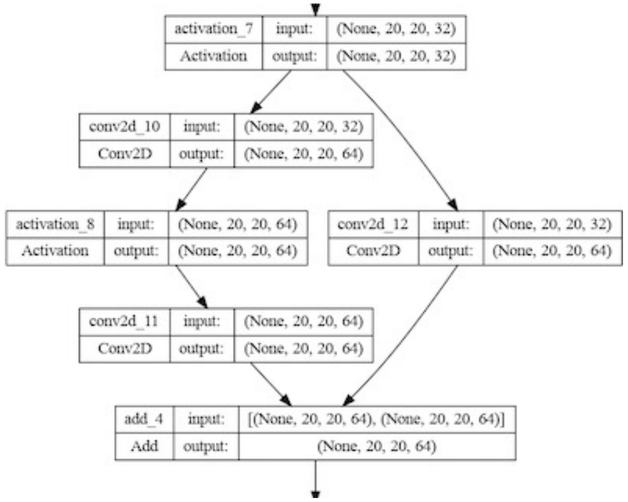


Improve the classification power
by capturing the correlation of different regions
of the image

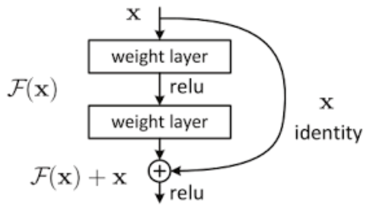
ResNet models







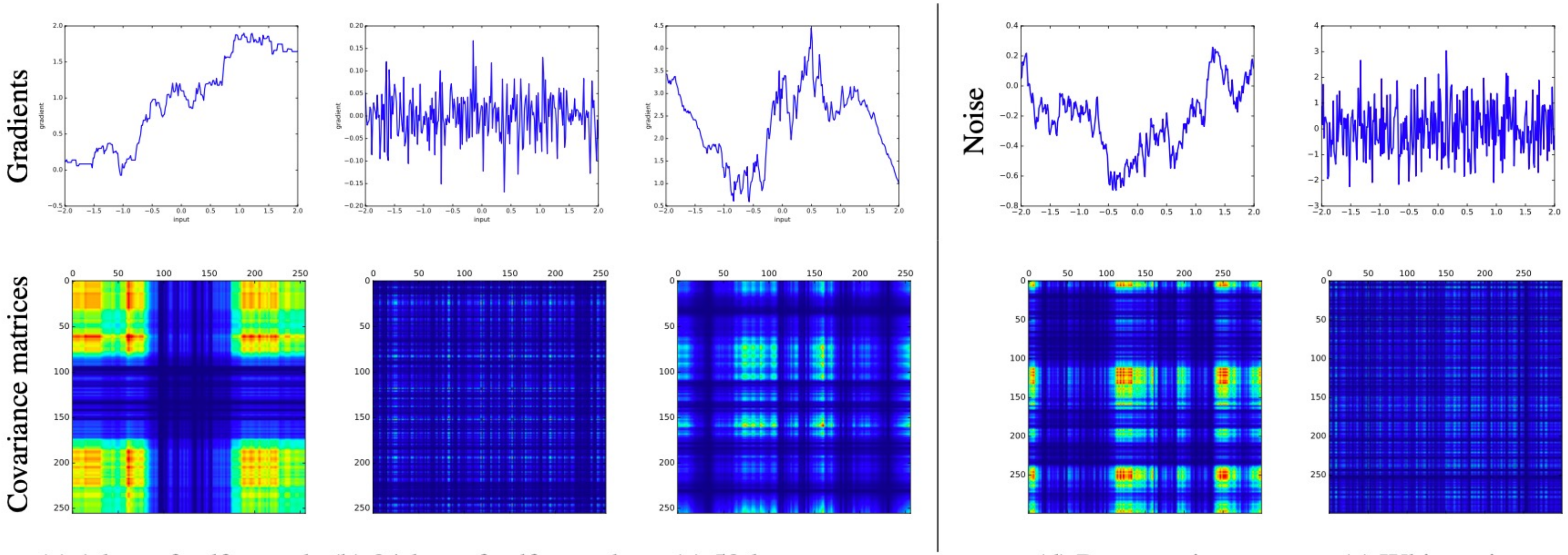
- ResNet is a key feature of very deep convolutional models for classification.
- ResNet adds the input back to the output after passing through a series of convolutional layers.



Why ResNet?

The Shattered Gradients Problem: If resnets are the answer, then what is the question? *Balduzzi et al.* 1702.08591v2

The Shattered Gradients Problem



(a) 1-layer feedforward. (b) 24-layer feedforward. (c) 50-layer resnet. (d) Brown noise. (e) White noise.

Keeping the gradients correlated as the depth of the network increases is crucial for efficient training and improving the classification power of complex models.

Inception models

- The use of 1x1 convolutions reduces the number of parameters and computational operations, allowing for deeper and more complex networks without an exponential increase in computational costs.
- Inception modules can capture features at different scales by combining filters of various sizes. This allows the network to be more flexible and powerful in capturing relevant information from images.
- As the input proceeds through the network, the result of the convolutions loses the local correlations. Therefore, the ratio between large and small kernels should increase with depth.