Application of machine learning for validation of Geant4 hadronic models

Sergey Korpachev and Marina Chadeeva

LPI, Moscow

September 10, 2021

Outline

1. Motivation, samples and event selection
2. Implementation of Machine Learning technique
3. Preliminary results of ANN-based prediction of secondaries

MC samples and event selection

Main goal: prediction of shower properties using calorimetric observables

MC samples

- CALICE technological prototype geometry and layout
- centrally generated samples of negative pions, 10–80 GeV, 500 kevt / sample
- Geant4 v10.3, physics lists: FTFP_BERT_HP and QGSP_BERT_HP
- official digitisation, no PID

In the current analysis: 40 GeV pions and QGSP_BERT_HP physics list

Reconstruction and event selection

- reconstruction software v04-14
- official reconstruction chain, 0.5 MIP cut for hits, official start finder algorithm
- for analysis: only events with found start at 3–6 AHCAL layers
- no other constraints, no clustering

MVA approach: algorithm, input features and targets

Goal is to predict distributions of parameters of secondaries within a shower

⇒ regression model realised in Artificial Neural Network

⇒ 29 calorimetric observables as input variables

⇒ one of MC-truth variables as a target

Input features and targets

Inputs

- Number of isolated hits in a shower
- Mean shower hit energy
- Shower radius
- Longitudinal shower centre of gravity
- Number of track hits within a shower
- "Ring" observables: 12 energies (MIP) + 12 numbers of isolated hits

Targets

- **Number of neutrons**
  counted per event except for those, which have one parent only that is also neutron (to avoid double counting)
- **Energy of neutral pions**
  sum of energies of all neutral pions in event

Details on observables and MC-truth in the previous talk by Marina

Preprocessing

Sample under study: 40 GeV energy point, about 106k selected events (from full set)

Features

- 29 input features and 1 target (true from mc collection)
- weighting applied with event weights obtained from the inverse target pdf to get more uniform distribution for target in the loss function calculation
- no normalization

Subsets

- 2/3 (64k) of full set for train/validation (t/v set) and about 1/3 (42k) of full set for test
- train subset is 32k events (50% of t/v set)
- validation subset is 32k events (50% of t/v set)
- events are selected randomly without intersections

Results are shown for the test subsample

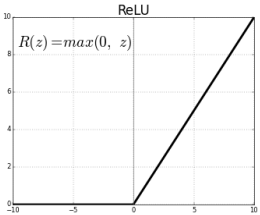
Neural network structure and hyperparameters

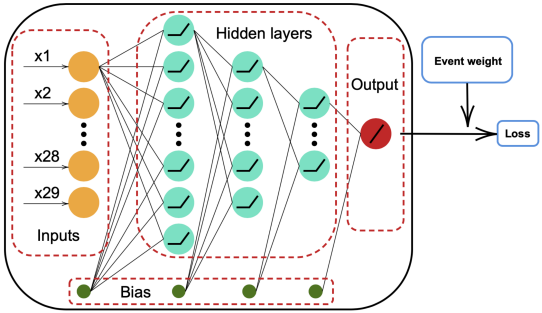
- Keras library
- Architecture:
  - 1 input layer, 3 hidden layers,1 output layer
  - Number of neurons: 29 / 128 / 64 / 32 / 1
  - Activation function: ReLU for hidden layers; linear $(\mathrm{f(y) = y})$ for output layer
- Optimizer: ADAM or NADAM
- Learning rate (lr): from 0.1 to 0.0000001
- Batch size (bs): 1, 2, 4, 8, 16 and 32
  $\Rightarrow$ Events come in batches iteratively
- Number of epochs: 100

Loss function: weighted MSE

$$\mathrm{Loss} = \frac{1}{\mathrm{N}} \cdot \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{W_i} \cdot (\mathrm{Xpred_i} - \mathrm{Xtrue_i})^2, 0 \leq i \leq N,$$

Xpred – prediction, Xtrue – from MC collection and $\mathrm{W_i}$ – weights from pdf of target variable





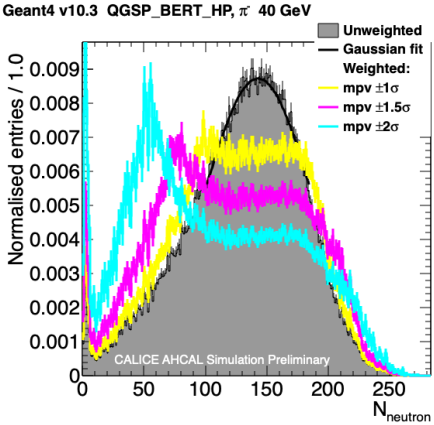
Several launches of the network with different random seed numbers (called *ANN*1... *ANN*5)

Event weights for NN training

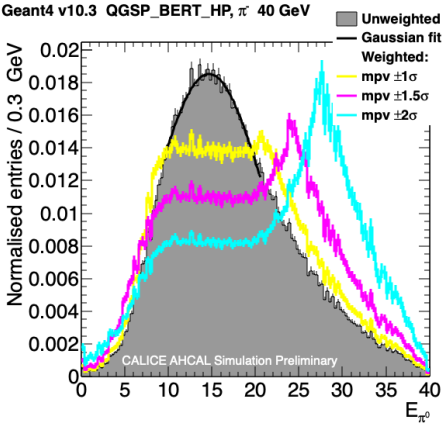
Re-weighting procedure: for i-th event $W_{\mathrm{SR,i}} = 1/\mathrm{pdf}(x_i)$, sigma range (SR) = 1.0, 1.5, 2.0

$$pdf(x) = \begin{cases} Gaus(x, m_{fit}, \sigma_{fit}), & \text{if } m_{fit} - SR \cdot \sigma_{fit} < x < m_{fit} + SR \cdot \sigma_{fit} \\ Gaus(m_{fit} + SR \cdot \sigma_{fit}, m_{fit}, \sigma_{fit}), & \text{otherwise} \end{cases}$$

Number of neutrons



Energy of neutral pions

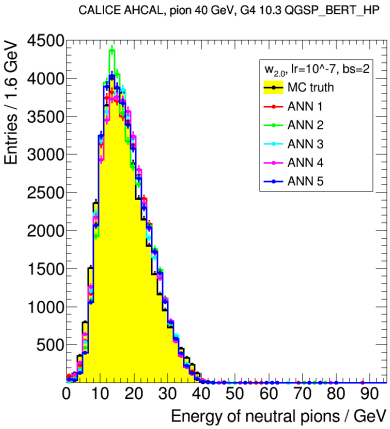


Prediction of neutral pions energy: distributions

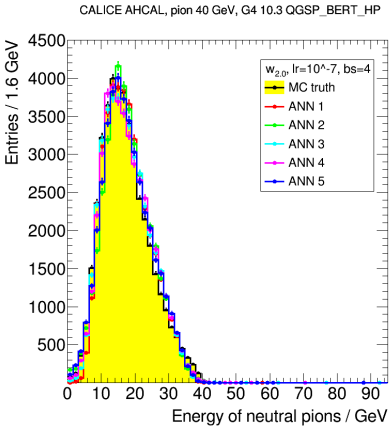
Distribution of true (MC truth) and predicted (ANN) sum of neutral pion energies

Examples of 10 trials with different random seed

$W_{2.0}$, lr $= 10^{-7}$, bs $= 2$, NADAM



$W_{2.0}$, lr $= 10^{-7}$, bs $= 4$, ADAM

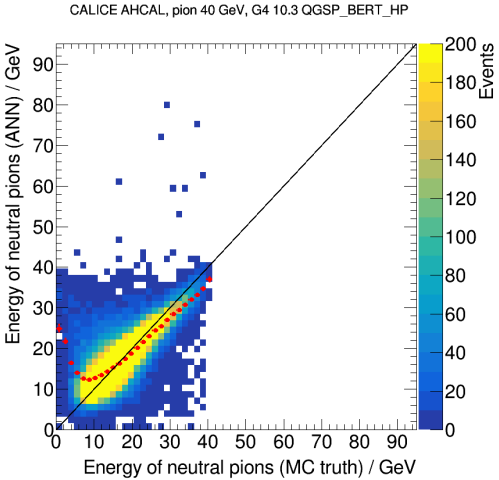


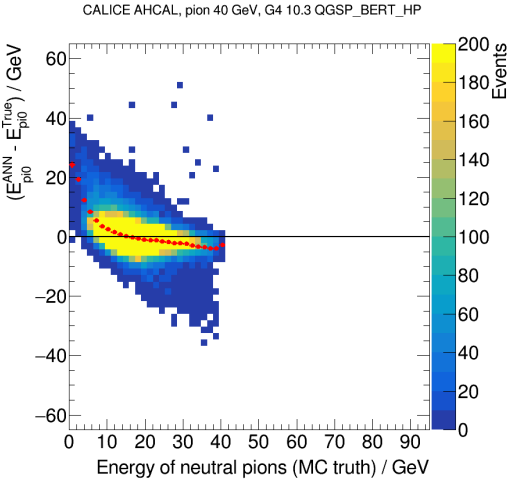
Examples of appropriate performance and well reproduced shape.

Prediction of neutral pions energy: examples of ANN vs MC truth

Red points show profile (means)

$W_{2.0}$, lr $= 10^{-7}$, bs $= 2$, NADAM, ANN5





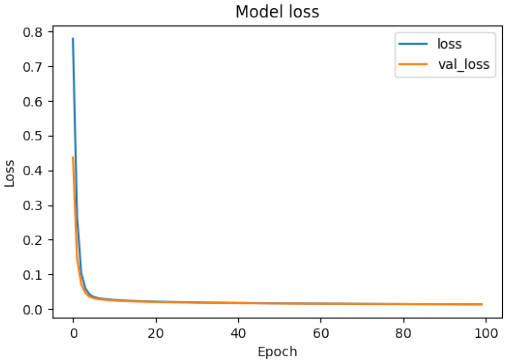
Good agreement above $E_{\pi^0} > 10$ GeV

ANN overestimates small and underestimates large E_{π^0}

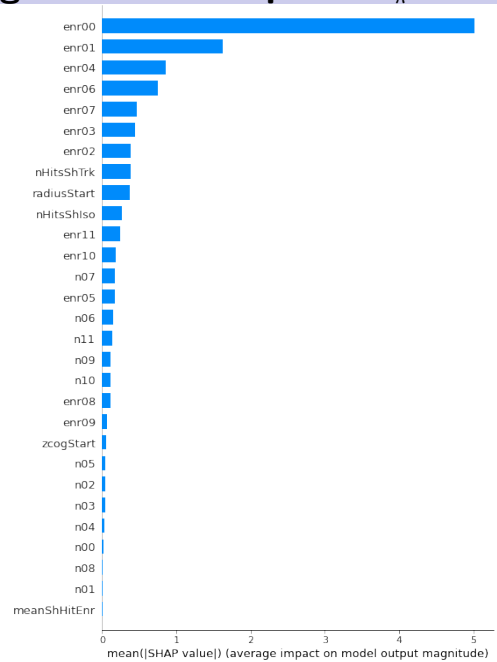
Few outliers

Convergence of loss functions and significance of inputs: E_{π^0}

$W_{2.0}$, lr $= 10^{-7}$, bs $= 2$, NADAM, ANN5



Good convergence of train and validation losses.
The most important features:
energies in central rings.

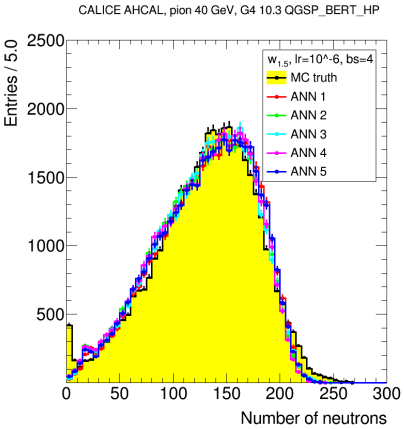


Prediction of number of neutrons: distributions

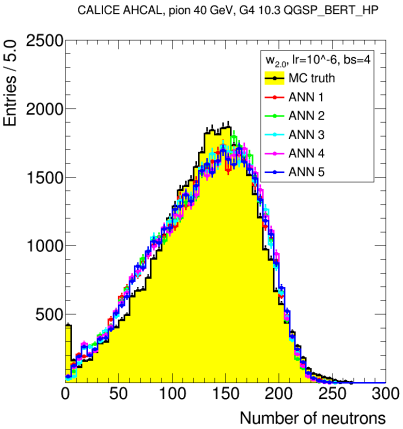
Distribution of true (MC truth) and predicted (ANN) number of neutrons in event

Examples of 10 trials with different random seed

$W_{1.5}$, lr $= 10^{-6}$, bs $= 4$, NADAM



$W_{2.0}$, lr $= 10^{-6}$, bs $= 4$, NADAM

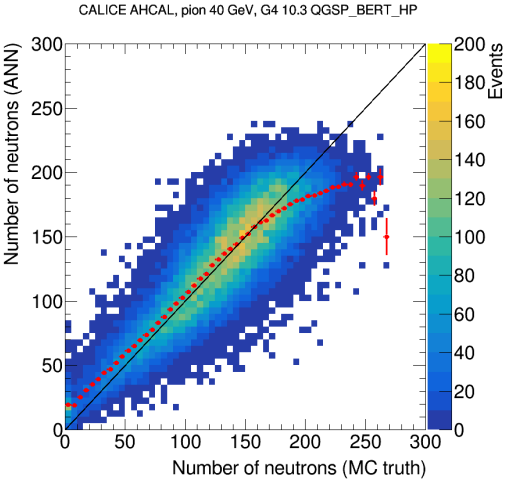


Examples of appropriate performance and well reproduced shape.

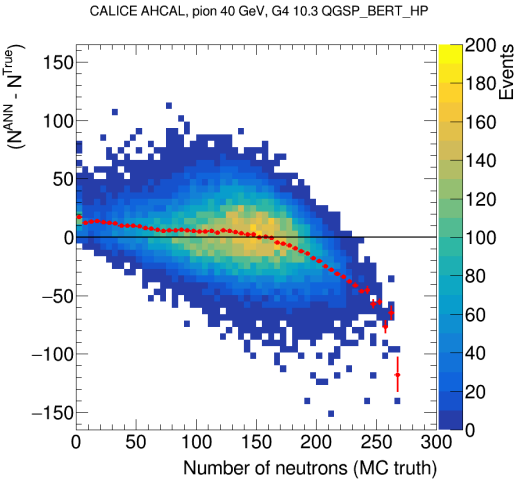
Prediction of number of neutrons: examples of ANN vs MC truth

Red points show profile (means)

$W_{1.5}$, lr $= 10^{-6}$, bs $= 4$, NADAM, ANN5



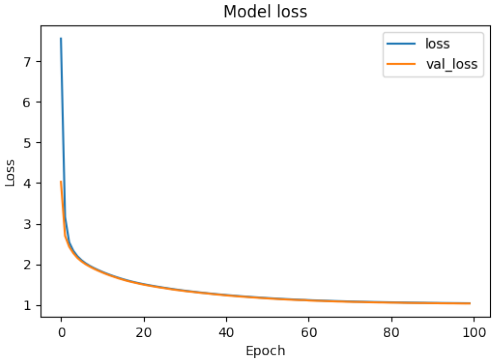
Good prediction of N_n by ANN except for tail



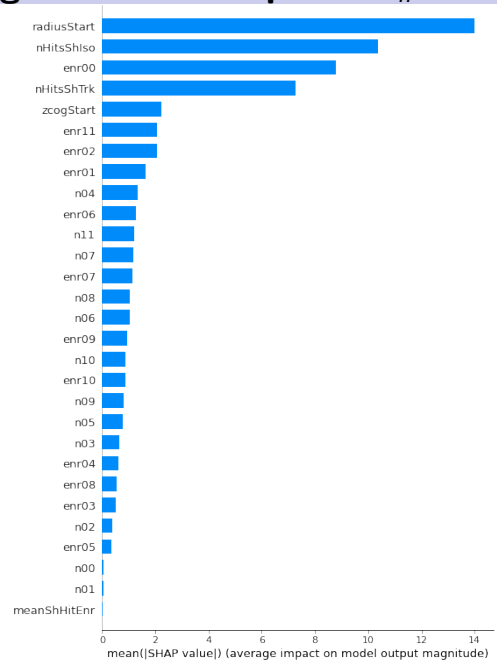
Underestimation of N_n for tail

Convergence of loss functions and significance of inputs: N_n

$W_{1.5}$, lr $= 10^{-6}$, bs $= 4$, NADAM, ANN5



Good convergence of train and validation losses.
The most important features: shower radius and number of isolated hits in a shower.



Summary

Highly granular calorimeters provide unique information about hadronic shower development.

Preliminary results

- An artificial neural network from Keras package was trained to **predict, for the first time, energy of neutral pions and number of neutrons in a hadronic shower using calorimetric observables**.
- Showers induced by simulated pions of 40 GeV in the CALICE AHCAL were studied (G4 v10.3, QGSP_BERT_HP).
- Good predictions of distributions have been achieved of energy sum of neutral pions and number of neutrons within a shower.

Plans

- Define algorithm performance
- Train models for pions with several energies in the range 10–80 GeV
- Prepare conference talk at ILCX2021 event and publication (MC only)
- Apply trained NN to data
- Prepare CALICE Analysis Note involving data