ANN-based prediction of shower properties using global observables for validation of Geant4 hadronic models

Sergey Korpachev and Marina Chadeeva

LPI, Moscow

June 30, 2021

Outline

1. Motivation
2. Samples and event selection
3. MC-truth variables and calorimetric observables
4. Implementation of Machine Learning
5. Preliminary results of ANN-based prediction of secondaries

Motivation

Validation of Geant4 hadronic models

- Geant4 simulations predict calorimetric observables quite well, mainly the most important one - measured hadron energy. But we know that discrepancies increase with energy.
- Geant4 uses material properties from thin-sample measurements and theor./phenomenological hadronic models. Additional tuning of models is performed using test beam data.
- Unique opportunities are provided by highly granular hadron calorimeters, which allow detailed study of shower characteristics, such as:
  - position of first inelastic interaction
  - shower radius, longitudinal centre of gravity and shower profiles
  - tracks within a shower
- A comparison of these characteristics between data and simulations give some hints of how to improve the model but no direct answer.

The current study is an attempt to answer the following question: can we move further and compare intrinsic shower properties at secondaries level on an event-by-event basis?

Samples and event selection

CALICE AHCAL data samples as of June 2018

- negative pions, 10–80 GeV
- reconstruction software v04-14
- official PID (Vladimir's BDT)

Run numbers

10 GeV	20 GeV	30 GeV	40 GeV	60 GeV	80 GeV
61265	61272	61378	61275	61262	61279

MC samples

- centrally generated negative pion samples, 10–80 GeV, about 500 kevt / sample
- Geant4 v10.3, physics lists: **FTFP_BERT_HP** and **QGSP_BERT_HP**
- official digitisation, no PID

Reconstruction and event selection

- 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

MC-truth variables

MCParticle collection is used to extract secondaries and their parameters.

Main characteristics under study

- **Number of neutral pions** (some of them might be from η mesons) [*mcNpi0*]
- **Sum of neutral pions energy** [*mcEpi0*]
- **Number of neutrons from interactions** [*mcNnR*]
  except for those that have one parent only that is also neutron (to avoid double counting**[*]**)
- **Sum of kinetic energy of neutrons from interactions** [*mcTnR*]

*** Neutron counting might need improvement and more detailed study**

Additional variables for further studies

- Number of η mesons (except for those that decay to neutral pions)
- Energy sum of η mesons counted above (*in spite of precaution with decay modes, adding up to π^0s' energy results in energy double counting)*
- Total number of neutrons (might include those after e.g. de-excitation, etc.)
- Kinetic energy of all neutrons
- Maximal kinetic energy of all neutrons and of neutrons from interactions

Calorimetric observables

Observables for crosscheck

Number of hits in event and reconstructed energy (see backup)

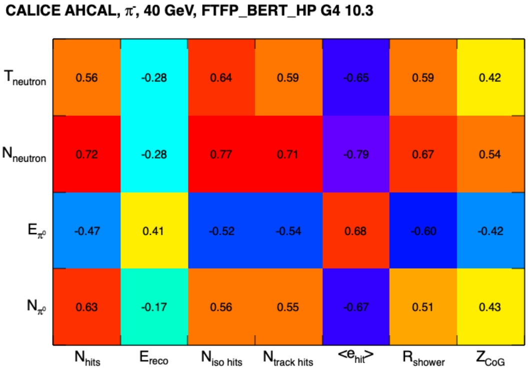
Observables for correlation studies with MC truth

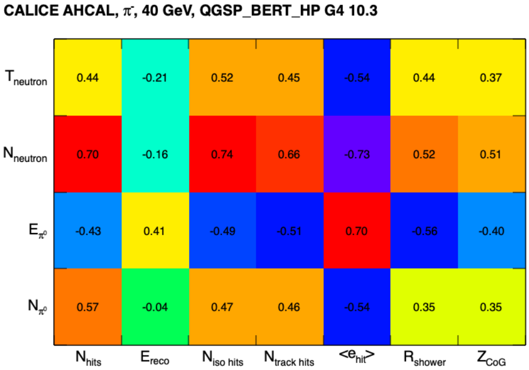
- **Number of isolated hits in a shower** (beyond the found shower start layer)
  isolated hit - 0 neighbours in a cube of 3×3×3 cells around the hit (max 26 neighbours)
  is highly correlated with total number of isolated hits in event due to selection of shower start layer
- **Number of track hits in a shower** (2 in-line neighbours and MIP-like deposition)
- **Mean shower hit energy** (shower hits only)
- **Shower radius** $R = \frac{\sum_{i=1}^{N_{\mathrm{sh}}} e_i \cdot r_i}{\sum_{i=1}^{N_{\mathrm{sh}}} e_i}$, N_{sh} - number of shower hits beyond the found shower start,
  e_i - hit energy, $r_i = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2}$ - hit radial distance from shower axis (x_0, y_0)
- **Longitudinal shower centre of gravity** (in units of $\lambda_{\mathrm{I}}^{\mathrm{eff}}$) $Z0 = \frac{\sum_{i=1}^{N_{\mathrm{sh}}} e_i \cdot (z_i - z_{\mathrm{start}})}{\sum_{i=1}^{N_{\mathrm{sh}}} e_i}$,
  z_i - hit longitudinal coordinate, z_{start} - longitudinal coordinate of shower start
  (for AHCAL, $\lambda_{\mathrm{I}}^{\mathrm{eff}} = 226.5$ mm, $0.118 \cdot \lambda_{\mathrm{I}}^{\mathrm{eff}}$/AHCAL layer)

Correlations between MC-truth and calorimetric variables

Relationship between global observables and MC-truth parameters

investigated by looking at (linear) correlation from 2D histograms
examples of correlation maps for 40 GeV





Few examples of distributions and 2D histograms in backup

Additional calorimetric observables

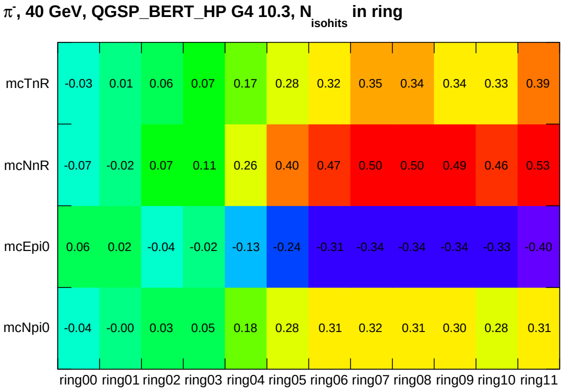
Radial ("ring") observables

Geometry

- 3-cm wide rings around shower axis
- 12 rings in total:
  *ring*00 – innermost
  *ring*11 – outmost
- integrated over longitudinal depth beyond the found shower start layer

Observables

- **Number of hits in a ring**
- **Energy sum in a ring**
  (over all hits in a ring)
- **Number of isolated hits in a ring**
- **Energy of isolated hits in a ring**



Example for 40 GeV pions:

correlation with number of isolated hits in rings

N_{isohits} in outer rings correlates with number of neutrons

Details of correlation studies in the previous talk on AHCAL weekly 02.06.2021

ML-based approach

Goal is to predict parameters of secondaries within a shower

⇒ study correlations of calorimetric observables with MC truth

⇒ use machine learning technique to train regression model

⇒ apply the trained model to data to estimate the characteristic/parameter under study

Input features and targets

Input:

- 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

Target:

- **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

Preprocessing

Sample under study: 40 GeV energy point, about 106k selected events (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 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

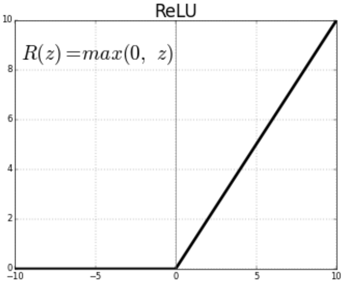
Intermediate goal is to achieve good performance on the training subset

Neural network structure

- Keras library
- Hyperparameters:
  - Layers: 1 input layer, 3 hidden layers and 1 output layer
  - Number of neurons: 29 / 128 / 64 / 32 / 1
  - Neuron activation function: ReLU for hidden layers and linear $(\mathrm{f(y) = y})$ for output layer
- Loss function: MSE
- Learning rate (lr) for optimizer: 0.01, 0.001 (default for Keras) and 0.0001
- Number of epochs: 100
- Batch size (bs): 1, 2, 4, 8, 16 and 32 (default for Keras) $\Rightarrow$ Events / batch: 32k, 16k, 8k, 4k, 2k and 1k events

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

ReLU:



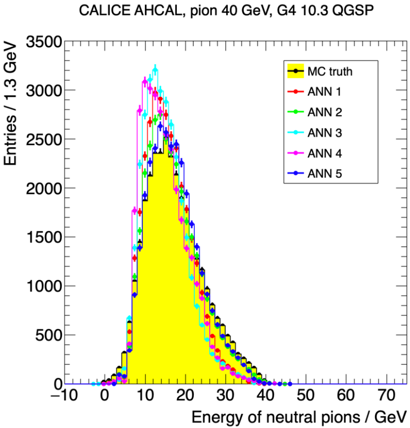
Further plots show results for training subset

Prediction of neutral pions energy: distributions

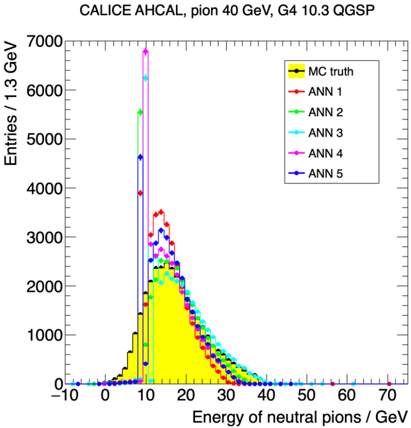
Distribution of true (MC truth) and predicted (ANN) sum of neutral pion energies

Examples of 10 trials with different random seed and batch sizes

lr = 0.001 and bs = 16



lr = 0.001 and bs = 2

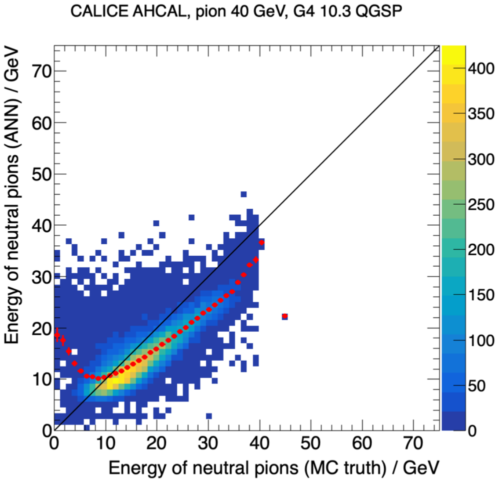


Only few cases demonstrate appropriate performance and almost reproduce the shape.

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

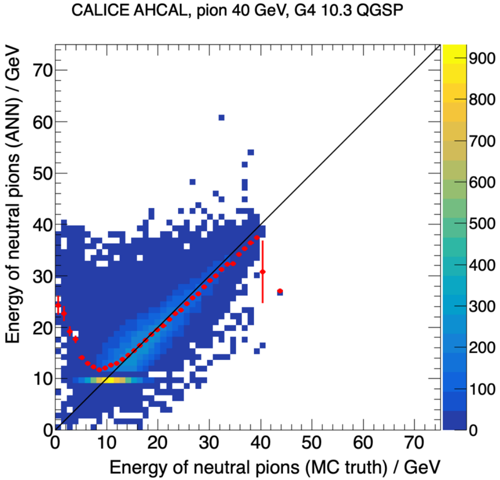
Red points show profile

lr = 0.001 and bs = 16



ANN underestimates E_{pi0} on average

lr = 0.001 and bs = 2

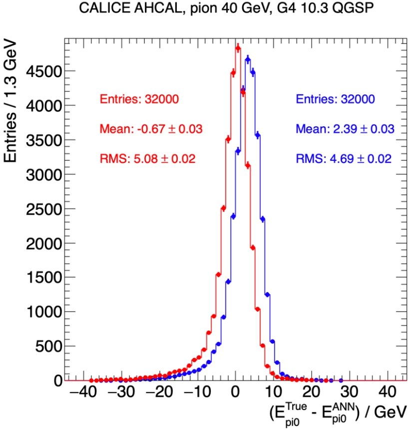


Only tail is reproduced

Prediction of neutral pions energy: ANN performance

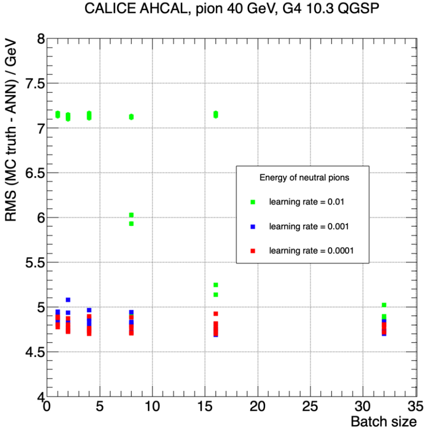
Quantitative estimate: mean and RMS of difference between ANN and MC truth

Distribution of $E_{pi0}^{True} - E_{pi0}^{ANN}$
for bs = 16 and bs = 2, lr=0.001



Both mean and RMS are important

RMS of $(E_{pi0}^{True} - E_{pi0}^{ANN})$
for different learning rates



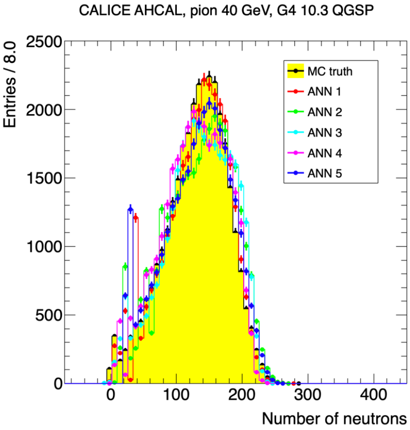
Better RMS with smaller lr

Prediction of number of neutrons: distributions

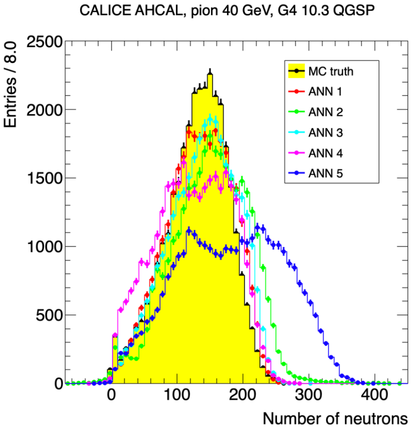
Distribution of true (MC truth) and predicted (ANN) number of neutrons in event

Examples of 10 trials with different random seed and batch sizes

lr = 0.001 and bs = 1



lr = 0.001 and bs = 32

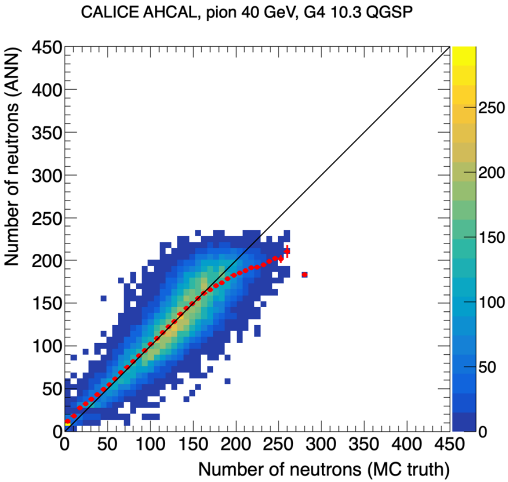


Only few cases demonstrate appropriate performance and almost reproduce the shape.

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

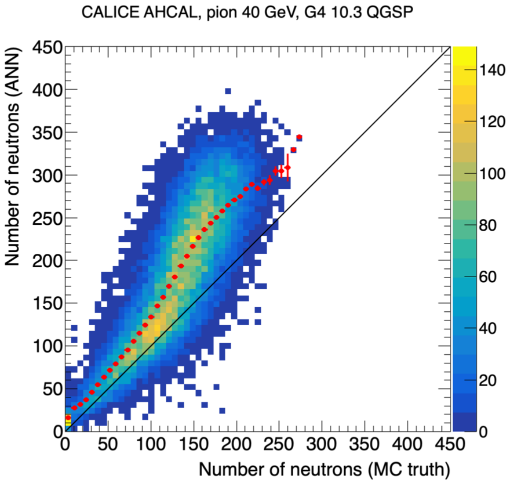
Red points show profile

lr = 0.001 and bs = 1



Good prediction of N_n by ANN except for tail

lr = 0.001 and bs = 32

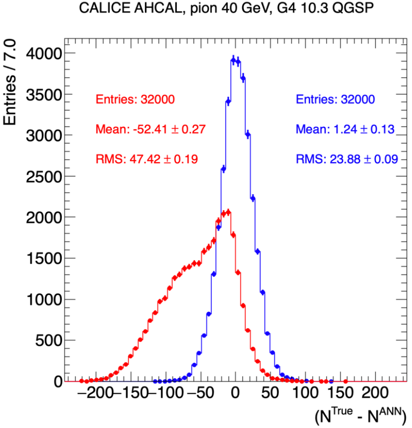


Overestimation of N_n on average

Prediction of number of neutrons: ANN performance

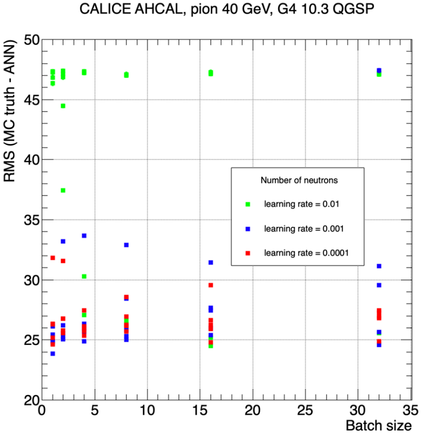
Quantitative estimate: mean and RMS of difference between ANN and MC truth

Distribution of $N_n^{True} - N_n^{ANN}$
for bs $= 1$ and bs $= 32$, lr$=0.001$



Both mean and RMS are important

RMS of $(N_n^{True} - N_n^{ANN})$
for different learning rates



Better RMS with smaller lr

Summary

CALICE AHCAL data contain unique information about hadronic shower development.

Preliminary results

- Correlations were studied between calorimetric observables and parameters of secondaries from Geant4 simulations.
- A neural network from Keras package was trained to **predict, for the first time, energy of neutral pions and number of neutrons using calorimetric observables** from highly granular calorimeter.
- Preliminary results show trend in the right direction, further ANN tuning is necessary.

TO DO

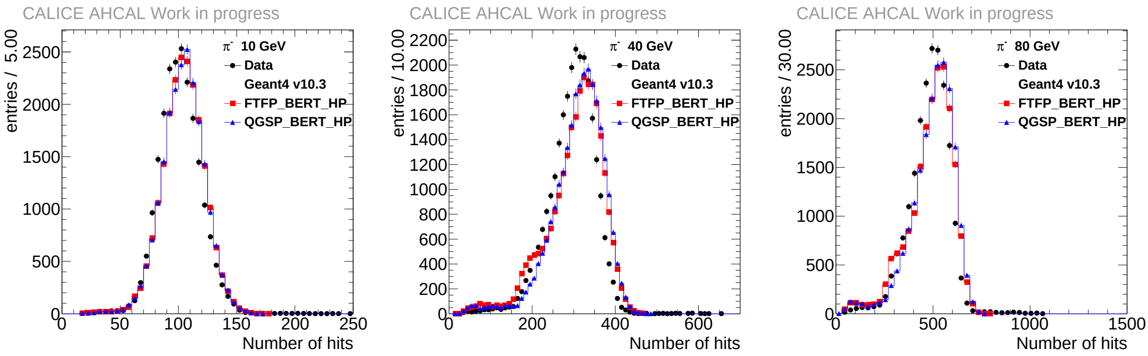
- Optimize hyperparameters of NN
- Define NN performance
  possible candidates: KS or AD test, RMS from 0
- Try to switch to G4 v10.6 as the most interesting case for G4 community.
- Apply to data
- Prepare CALICE Analysis Note

Backup slides

Crosscheck: total number of hits

Legend: **Data**, **FTFP_BERT_HP**, **QGSP_BERT_HP**

For MC-data comparison of calorimetric observables, MC samples are truncated, so that the numbers of selected events are equal in data and MC (∼20 kevt / sample after selections).

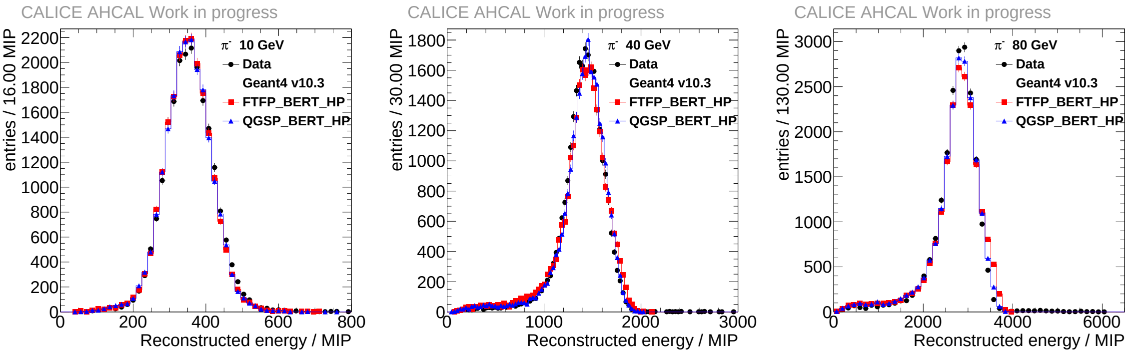


Moderate overestimation of number of hits by simulations, similar for both physics lists
Well pronounced shoulder to low number of hits for FTFP_BERT_HP above 10 GeV

Crosscheck: reconstructed energy

Legend: **Data**, **FTFP_BERT_HP**, **QGSP_BERT_HP**

For MC-data comparison of calorimetric observables, MC samples are truncated, so that the numbers of selected events are equal in data and MC (∼20 kevt / sample after selections).

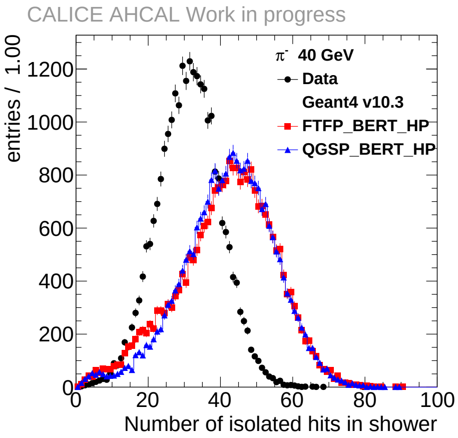


Good agreement between data and simulations
Distributions from simulations above 10 GeV a bit wider than from data

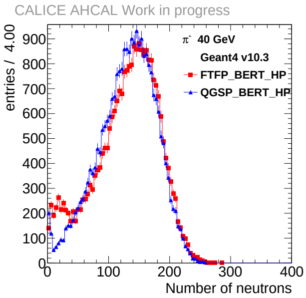
Number of isolated hits vs. number of neutrons at 40 GeV

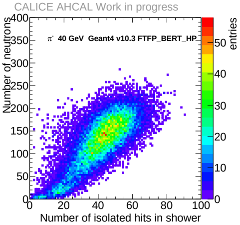
For MC-data comparison of calorimetric observables, MC samples are truncated, so that the numbers of selected events are equal in data and MC (∼20 kevt / sample after selections).

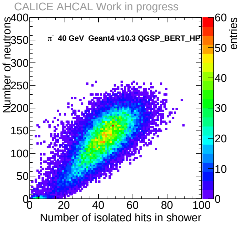
Data, FTFP_BERT_HP, QGSP_BERT_HP



MC truth







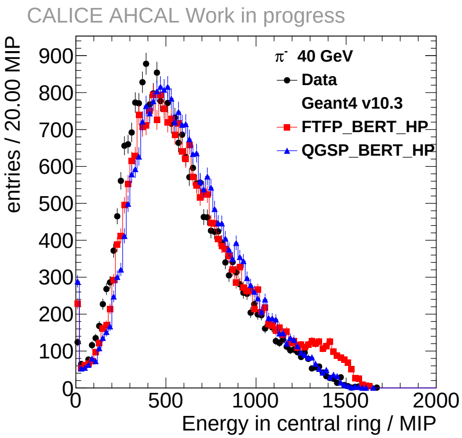
QGSP_BERT_HP: smooth distribution

FTFP_BERT_HP: excess of low N_{n} and iso hits

Energy in innermost ring vs. energy of π^0s at 40 GeV

For MC-data comparison of calorimetric observables, MC samples are truncated, so that the numbers of selected events are equal in data and MC ($\sim$20 kevt / sample after selections).

Data, FTFP_BERT_HP, QGSP_BERT_HP

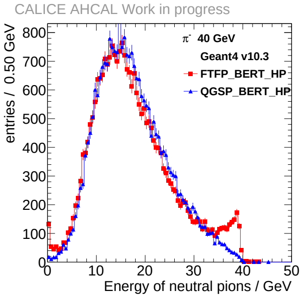


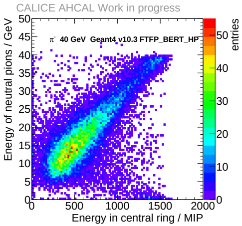
FTFP_BERT model: excess at limit

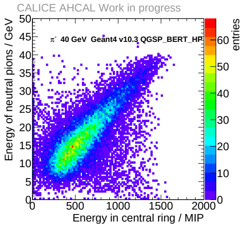
Data: less energy in central ring

Strong correlation with $E_{\pi 0}$

MC truth



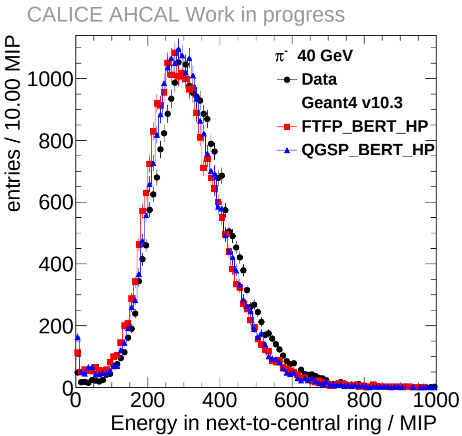




Energy in next-to-central ring vs. energy of π^0s at 40 GeV

For MC-data comparison of calorimetric observables, MC samples are truncated, so that the numbers of selected events are equal in data and MC ($\sim$20 kevt / sample after selections).

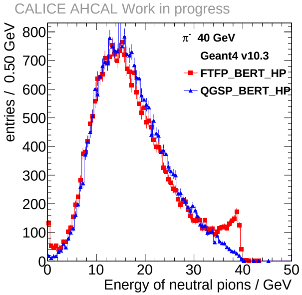
Data, FTFP_BERT_HP, QGSP_BERT_HP

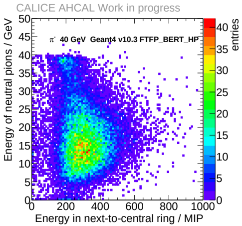


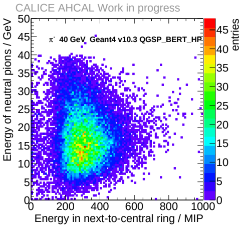
Data: more energy in next-to-central ring

No correlation with E_{π^0}

MC truth



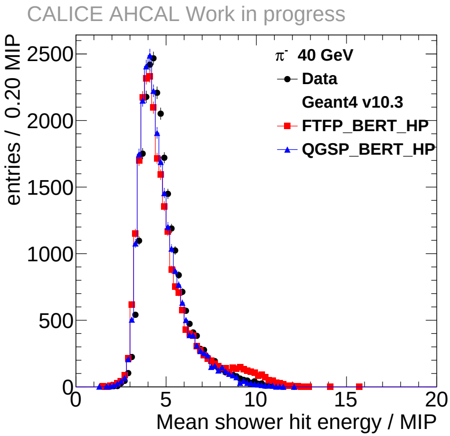




Mean shower hit energy vs. energy of π^0s at 40 GeV

For MC-data comparison of calorimetric observables, MC samples are truncated, so that the numbers of selected events are equal in data and MC ($\sim$20 kevt / sample after selections).

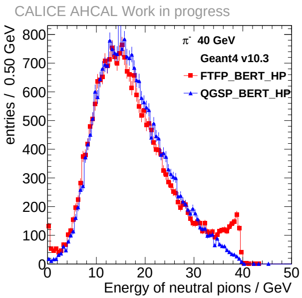
Data, FTFP_BERT_HP, QGSP_BERT_HP

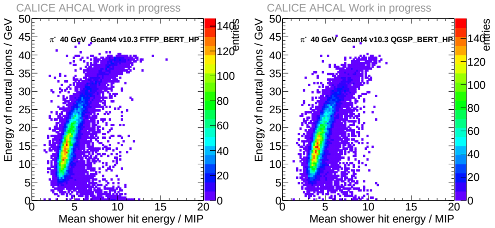


FTFP_BERT model: excess at limit

Slightly higher mean hit energy in data

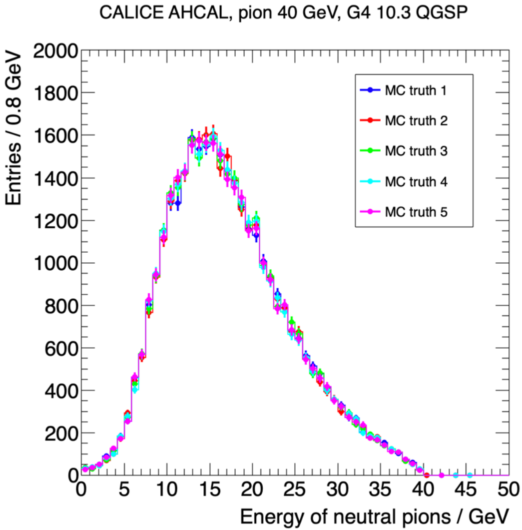
MC truth



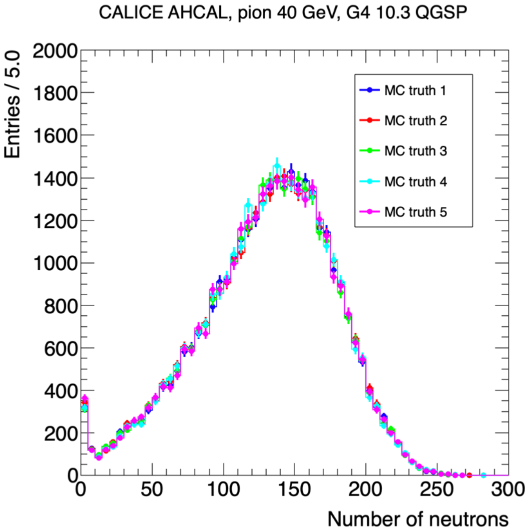


MC truth distributions from different trials

Target: energy of neutral pions



Target: number of neutrons



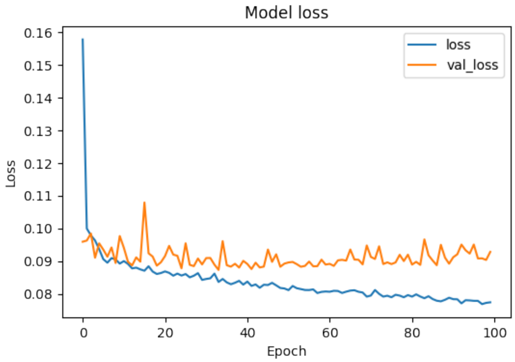
Convergence of loss functions: energy of neutral pions

Loss function

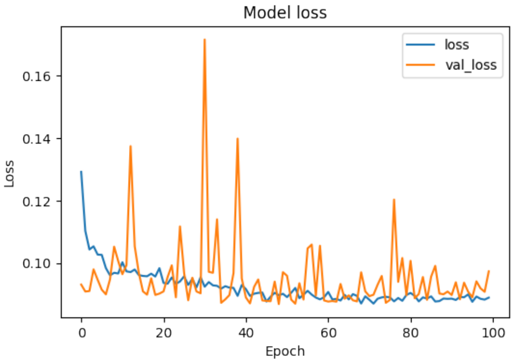
$$Loss = \frac{1}{\mathrm{N}} \cdot \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{w_i} \cdot (\mathrm{Xpred_i} - \mathrm{Xtrue_i})^2, 0 \leq i \leq N,$$

Xpred – prediction, Xtrue – from mc collection
and $\mathrm{w_i}$ – weights from probability density distributions of target variable.

Target: energy of neutral pions (bs $=$ 16)

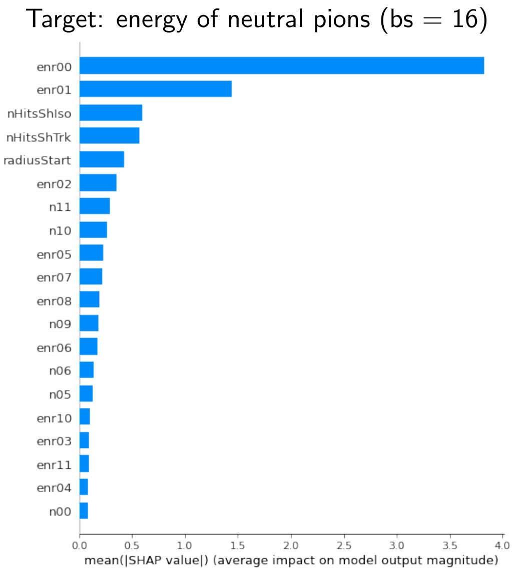


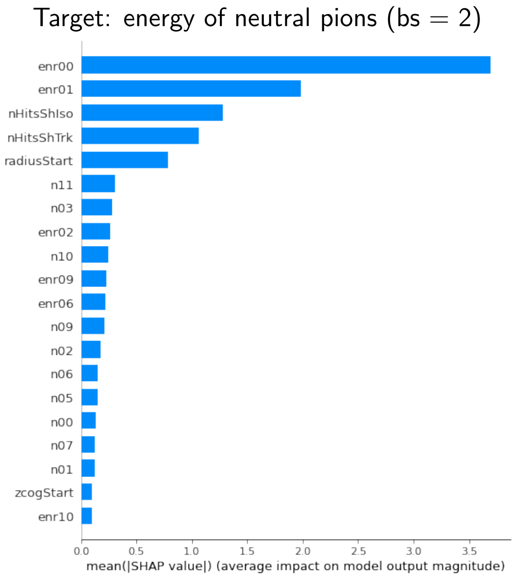
Target: energy of neutral pions (bs $=$ 2)



Significance of input features (SHAP-based): E_{π^0}

20 most significant inputs





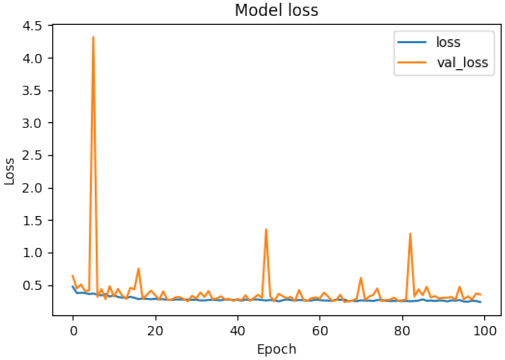
Convergence of loss functions: number of neutrons

Loss function

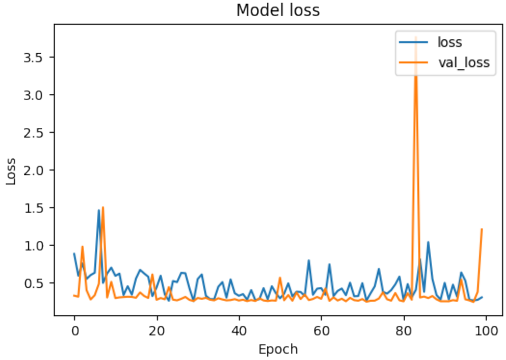
$Loss = \frac{1}{N} \cdot \sum_{i=1}^{N} w_i \cdot (Xpred_i - Xtrue_i)^2, 0 \leq i \leq N,$

Xpred – prediction, Xtrue – from mc collection
and w_i – weights from probability density distributions of target variable.

Target: number of neutrons (bs = 1)



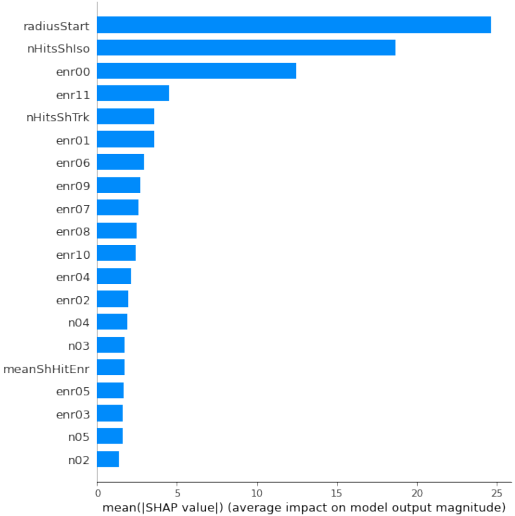
Target: number of neutrons (bs = 32)



Significance of input features (SHAP-based): N_n

20 most significant inputs

Target: number of neutrons (bs = 1)



Target: number of neutrons (bs = 32)