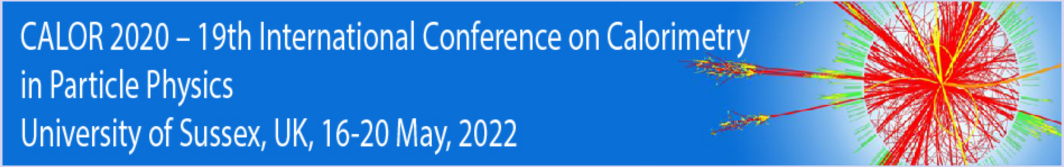


A novel approach to understanding hadronic showers using machine learning technique

Marina Chadeeva and Sergey Korpachev

1. Motivation and goal
2. Highly granular detector model and calorimetric observables
3. MC-truth variables
4. Correlations between parameters of secondaries in a shower and calorimetric observables
5. Deep Neural Network architecture and training
6. Preliminary results of application of DNN-based regression model

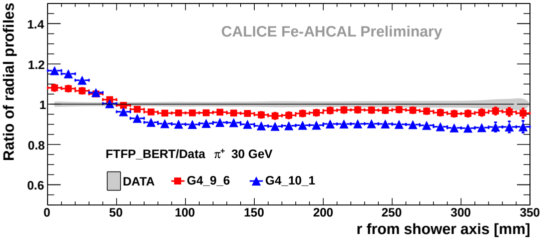
Motivation

Geant4 package: prediction of standard calorimetric observables

- **ionisation losses and electromagnetic showers**: percent-level precision
- **hadronic showers**: few percent precision of measured hadron energy for different materials
  discrepancies increase with energy and depend on physics list

Topological observables for hadronic showers

- important for particle flow reconstruction and fast simulation (shower libraries)
- depend on first interaction simulation and secondaries production
- Geant4 validation with test beam data:
  discrepancies increase with hadron energy and are above 10% for some observables (e.g. shower transverse size)



[CALICE-PUB-2014-003, CAN-040b]

G4 validations with the first-generation CALICE prototypes:
[JINST 10 (2015) P04014], *[JINST 10 (2015) P12006]*

Goal

Improvement of validation of simulations

- understanding of "hidden" hadronic shower properties, which cannot be directly measured
- prediction of properties of secondaries produced in hadronic showers

A novel approach combines

- unique calorimetric observables available in highly granular devices
- machine learning technique

Focus on simulations in this talk

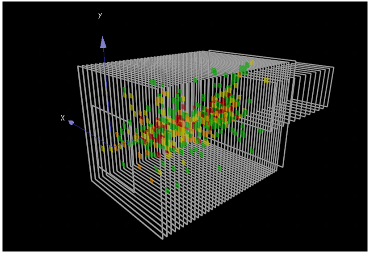
- pion-induced showers in the scintillator-steel hadron calorimeter (CALICE AHCAL)
- single negative pions @ 10–80 GeV, about 500 kevt / energy point (centrally produced by CALICE DESY group)
- **Geant4 version 10.3**, physics lists:
  - **FTFP_BERT_HP** – currently recommended
  - **QGSP_BERT_HP** – previously recommended

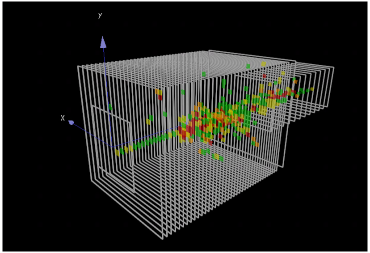
HP (High Precision) – precise neutron models and cross sections for 20 MeV and below

Detector model for simulations: CALICE AHCAL

Scintillator-SiPM analog hadron calorimeter with stainless steel absorber (see details in backup)

long. depth: ∼4.3 nucl. int. length
transverse size: 72×72 cm^2





Simulated 80-GeV pion shower

Calibration and MC digitisation

Data

- MIP calibration with test beam muons: ADC-to-MIP factor, temperature corrections
- cell response equalised at MIP level
- pedestals and SiPM saturation treated at cell level

MC

- MIP calibration with MC muons: GeV-to-MIP factor
- no simulation of light collection and photon detection by SiPM, pixelisation and saturation emulated in digitisation
- digitisation tuned with MC-to-Data comparisons for muons and electrons

Reconstruction chain and event selection

- cell signals above 0.5 MIP threshold — hits
- shower start finder algorithm tuned on MC
  for analysis: only events with found shower start at 3–6 AHCAL layers
- no clustering, no hadron energy scale calibration

Calorimetric observables in highly granular calorimeters

Counting observables

- **Total number of hits**, N_{hits}
- **Number of isolated hits**, N_{iso} [isolation – 0 neighbours in a cube of 3×3×3 cells around the hit]
- **Number of track hits**, N_{trk} [defined as having 2 in-line neighbours and MIP-like deposition]

Amplitude observables (e_i - energy of hit with coordinates x_i, y_i, z_i; N_{sh} - number of shower hits)

- **Reconstructed energy**, E_{reco} (sum of hit energies)
- **Mean shower hit energy**, $\langle e_{\text{hit}} \rangle$
- **Shower radius** $R_{\text{sh}} = \frac{\sum_{i=1}^{N_{\text{sh}}} e_i \cdot r_i}{\sum_{i=1}^{N_{\text{sh}}} e_i}$, $r_i = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2}$ - hit radial distance from shower axis (x_0, y_0)
- **Longitudinal shower centre of gravity** $Z_{\text{CoG}} = \frac{\sum_{i=1}^{N_{\text{sh}}} e_i \cdot (z_i - z_{\text{start}})}{\sum_{i=1}^{N_{\text{sh}}} e_i}$,
  z_i - hit longitudinal coordinate, z_{start} - longitudinal coordinate of shower start

Additional "ring" observables (integrated over longitudinal depth)

3-cm wide rings around shower axis, consistent with cell transverse size; 12 rings in total

- **number of hits in a ring**, $N_{\text{all}}^{\text{ring}}$; **number of isolated hits in a ring**, $N_{\text{iso}}^{\text{ring}}$
- **energy sum in a ring**, $E_{\text{all}}^{\text{ring}}$; **energy of isolated hits in a ring**,$E_{\text{iso}}^{\text{ring}}$

MC-truth variables

Parameters of secondaries at generator level are extracted from MCParticle collection

Neutral pions

All π^0s are counted independently of their parents (some of them might be from η mesons)

- **Number of neutral pions in an event**, N_{π^0}
- **Sum of the energies of neutral pions**, E_{π^0}

The dominated contribution to electromagnetic fraction within a hadronic shower comes from gammas produced in neutral pion decays.

Neutrons

Not all neutrons are counted. The neutrons are excluded, which have one parent only, which is also neutron (to avoid double counting)

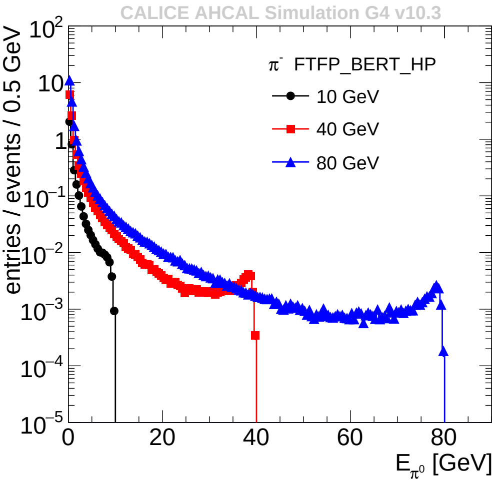
- **Number of neutrons from interactions**, $N_{neutron}$
- **Sum of kinetic energies of neutrons from interactions**, $T_{neutron}$

Neutron counting might need improvement and more detailed study with advices from G4 team.

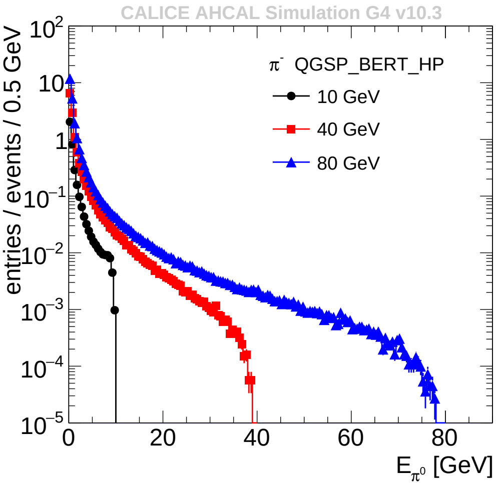
MC-truth: energy spectra of neutral pions in hadronic shower

Legend: **10 GeV**, **40 GeV**, **80 GeV** (∼100 kevt / sample after selections)

FTFP_BERT_HP



QGSP_BERT_HP

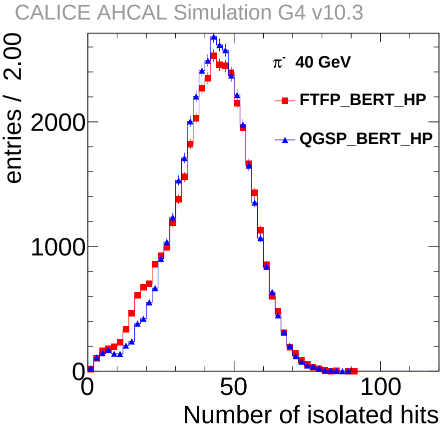


Very different spectra from FTFP and QGSP models above 10 GeV.
Similar behaviour at 10 GeV due to the same Bertini model (BERT) in this energy range.

Number of isolated hits and number of neutrons in a shower at 40 GeV

Legend: **FTFP_BERT_HP**, **QGSP_BERT_HP** (same number of selected events)

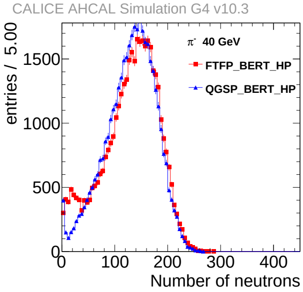
Calorimetric observable: number of isolated hits

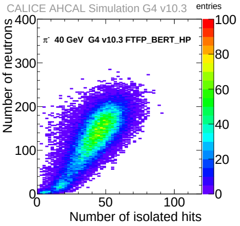


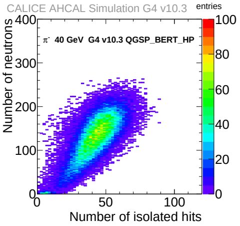
Difference between models for both N_{iso} and N_{neutron}

Visible correlation between N_{iso} and N_{neutron}

MC truth: number of neutrons



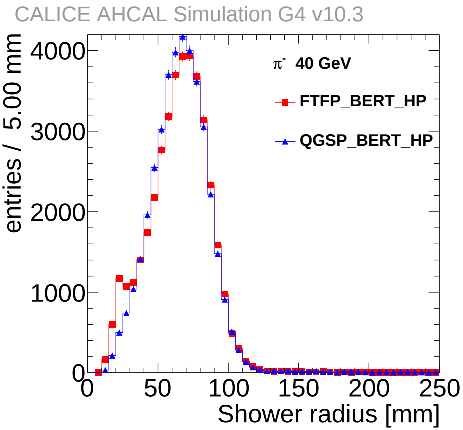




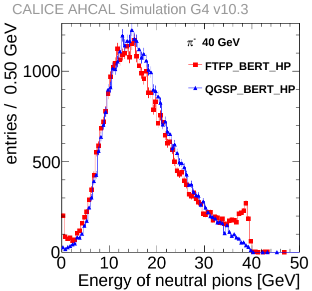
Shower radius and energy of π^0s in a shower at 40 GeV

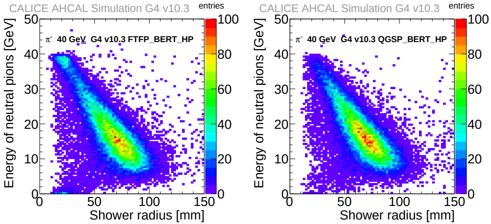
Legend: **FTFP_BERT_HP**, **QGSP_BERT_HP** (same number of selected events)

Calorimetric observable: shower radius



MC truth: energy of π^0s





Difference between models for both R_{sh} and E_{π^0}

Anticorrelation between R_{sh} and E_{π^0}

Deep Neural Network architecture for regression model

Goal is to predict parameters of secondaries within a shower using calorimetric observables

Regression model with 29 inputs (calorimetric obsevables)

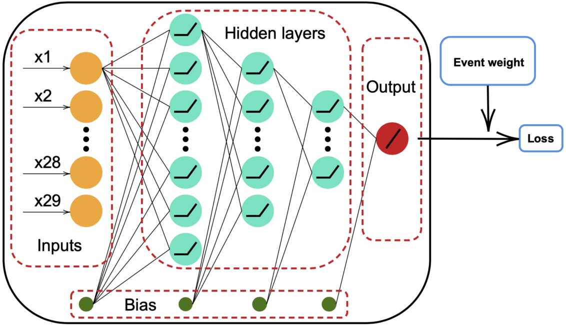
- number of isolated hits in a shower
- mean shower hit energy
- shower radius
- longitudinal shower centre of gravity
- number of track hits within a shower
- 12 energies ($E_{\mathrm{all}}^{\mathrm{ring}}$) + 12 numbers of iso hits ($N_{\mathrm{iso}}^{\mathrm{ring}}$)

Target: parameter of secondaries, e.g. number of neutrons or energy of neutral pions

Network architecture

TensorFlow library, Keras framework, scikit-learn

- number of layers:
  1 input, 3 hidden, 1 output
- number of neurons:
  29/128/64/32/1
- activation function:
  ReLU for hidden,
  linear $(\mathrm{f(y)} = \mathrm{y})$ for output
- bias neurons and weighted loss
- **supervised learning**



Deep Neural Network training and optimisation

Hyperparameters

no optimisation, just tests of several options

- optimiser: ADAM or NADAM,
  both show similar performance
- learning rate(lr): from 0.1 to 0.0000001,
  better performance with lr $\leq 10^{-6}$
- batch size(bs): 1, 2, 4, 8, 16 and 32
  $\Rightarrow$ events come in batches iteratively
  bs $= 8$ selected as compromise
- number of training epochs: about 15
  stable behaviour w/o overtraining

Training, validation and test subsamples

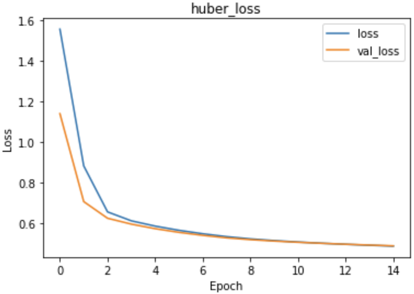
full sample after selections: $\sim$100 kevt
$\sim$60% for training
$\sim$20% for validation
$\sim$20% for test

Huber loss function with weighting

$$L_i = \begin{cases} 0.5 \cdot X_i^2, & |X_i| \leq 1; \\ |X_i| - 0.5, & |X_i| > 1; \end{cases}$$

$$X_i = Y_i^{\mathrm{predicted}} - Y_i^{\mathrm{true}}, \quad \mathrm{Loss} = \frac{1}{N} \cdot \sum_{i=1}^{N} W_i \cdot L_i$$

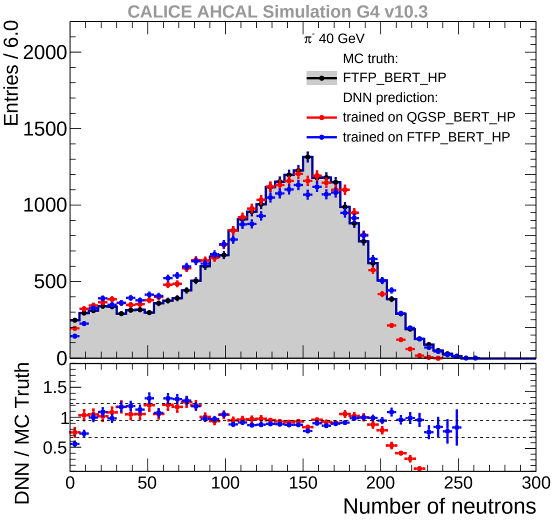
N - number of events for training
$\mathrm{W_i}$ – event weights from density-based weighting



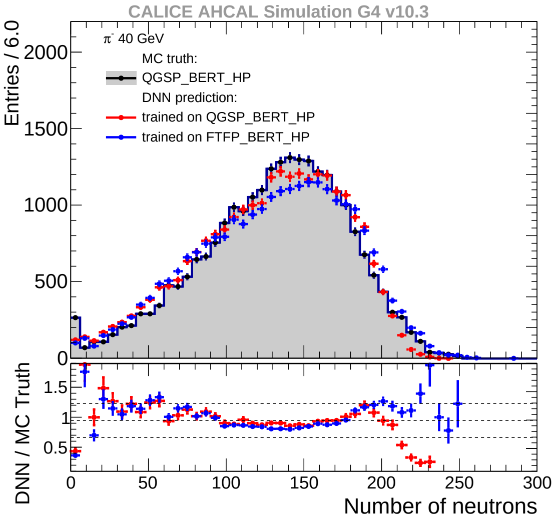
Number of neutrons in a shower from 40 GeV pion: distributions

Predictions from DNN trained on QGSP_BERT_HP or trained on FTFP_BERT_HP

MC truth from **FTFP_BERT_HP** test sample



MC truth from **QGSP_BERT_HP** test sample

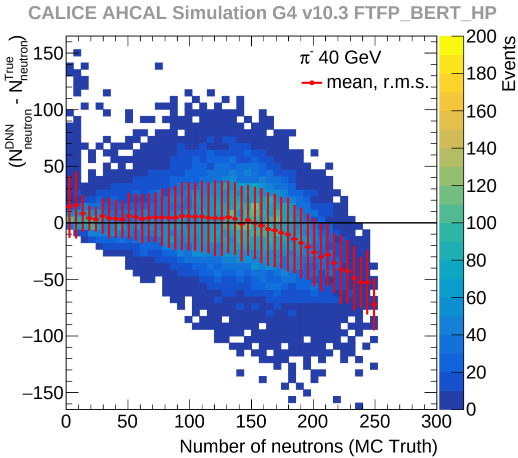


Mean and r.m.s. of the distributions reproduced at few percent level, reasonable generalisation ability.

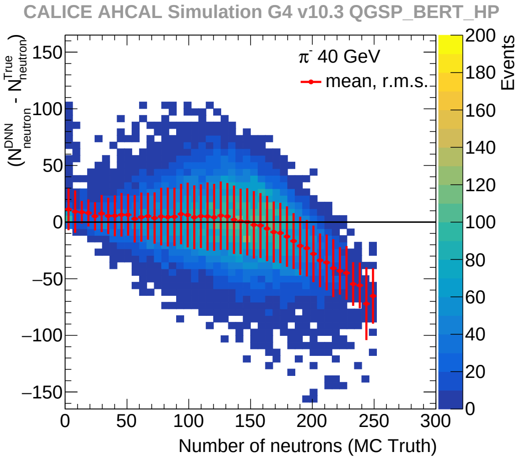
Number of neutrons in a shower from 40 GeV pion: event-by-event comparison

Difference between predicted and true values vs MC truth

MC truth from **FTFP_BERT_HP** test sample
trained on FTFP_BERT_HP



MC truth from **QGSP_BERT_HP** test sample
trained on QGSP_BERT_HP

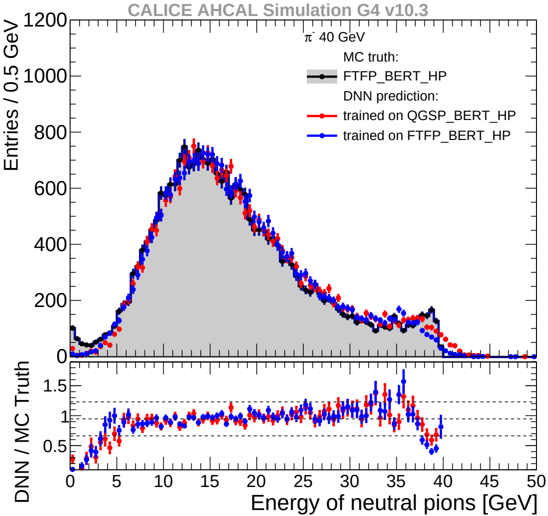


Three most important observables from feature importance studies:
energy in the most central ring around shower axis, number of isolated hits in a shower, shower radius.

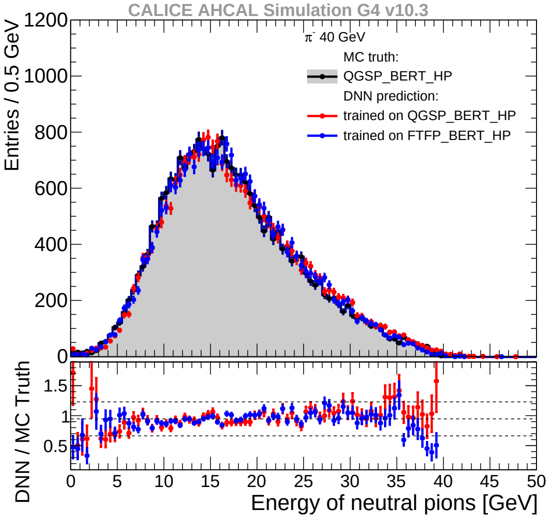
Energy of π^0s in a shower from 40 GeV pion: distributions

Predictions from DNN trained on QGSP_BERT_HP or trained on FTFP_BERT_HP

MC truth from **FTFP_BERT_HP** test sample



MC truth from **QGSP_BERT_HP** test sample

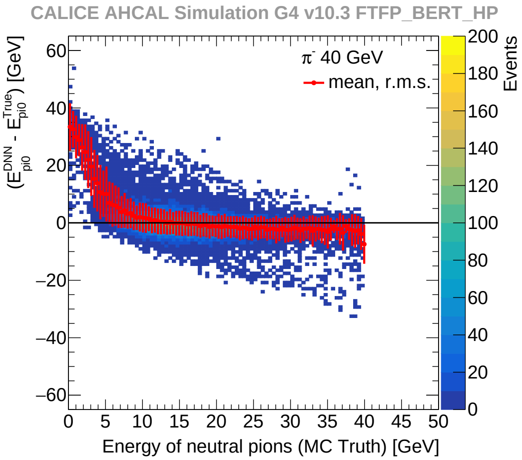


Mean and r.m.s. of the distributions reproduced at few percent level, reasonable generalisation ability. Both DNN models do not reproduce tails of FTFP_BERT_HP distribution.

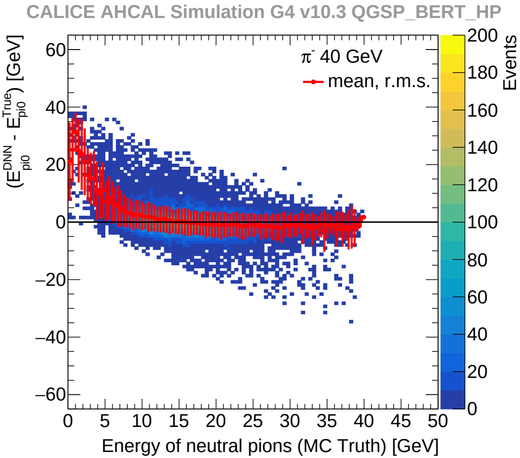
Energy of π^0s in a shower from 40 GeV pion: event-by-event comparison

Difference between predicted and true values vs MC truth

MC truth from **FTFP_BERT_HP** test sample trained on FTFP_BERT_HP



MC truth from **QGSP_BERT_HP** test sample trained on QGSP_BERT_HP



Three most important observables from feature importance studies:
energy in the most central and next-to-central rings around shower axis, shower radius.

Summary

A novel approach is proposed to extract properties of secondaries in hadron-induced showers

- **Technique:** regression model in Deep Neural Network trained using supervised learning
- **Inputs:** 29 calorimetric observables from the highly granular CALICE AHCAL
- **Target:** **number of neutrons** or **energy of neutral pions**
- **Preliminary results of DNN training on 40 GeV pion showers**
  simulated using FTFP_BERT_HP and QGSP_BERT_HP physics lists from Geant4 v10.3:
  - few percent accuracy in prediction of mean, r.m.s. and asymmetry of distributions
  - good model-to-model generalisation ability
  - reasonable performance in event-by-event predictions
- **Possible applications:**
  - validation of simulation
  - software compensation

Plans

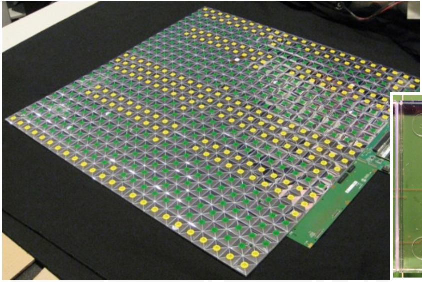
- test the most recent Geant4 versions
- test different pion energies
- test generalisation ability by combining different energies
- **apply trained model to data and provide feedback to Geant4 developers**

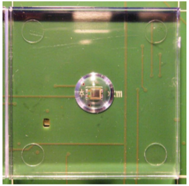
Backup slides

Scintillator-SiPM analog hadron calorimeter with stainless steel absorber

AHCAL technological prototype

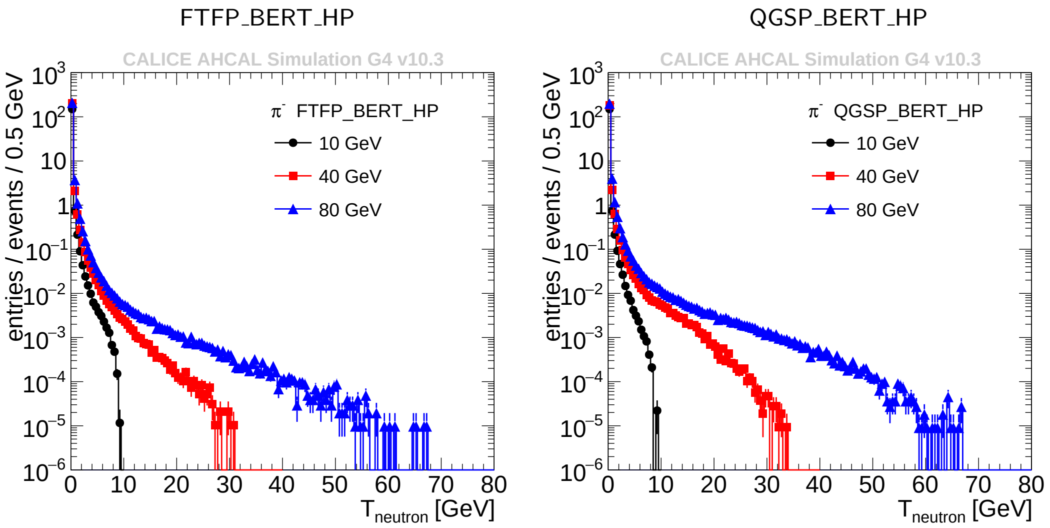
- Materials and layout
  - active: scintillator tiles w/dimple wrapped in foil and read-out by SiPM
  - absorber: 2-cm thick steel plates
  - $\sim$22000 channels, embedded electronics
- Longitudinal segmentation and depth
  - 38 active layers
  - depth $\sim$4.3 nucl. int. length
- Transverse segmentation and size
  - tile size $3\times3\times0.3$ cm^3
  - plane size 72×72 cm^2





MC-truth: spectra of neutron kinetic energy in hadronic shower

Legend: **10 GeV**, **40 GeV**, **80 GeV** (∼100 kevt / sample after selections)



Different shape of high energy tails from FTFP and QGSP models. Neutron energy cut is set to 1 MeV.
Similar behaviour at 10 GeV due to the same Bertini model (BERT) in this energy range.