Calorimetric Measurement of Multi-TeV Muons via Deep Regression

Jan Kieseler, Giles Strong,
Filippo Chiandotto, Tommaso Dorigo, Lukas Layer

ML4JETS, Online - 07/07/21

arXiv:2107.02119 [physics.ins-det]

MOTIVATION

- Muons traditionally used as a clean probe for searches & measurements:
  - Minimal EM losses = easily distinguishable from background objects
  - Energy measurements made via magnetic bending in tracker (+ possible muon system)
- But: radius of curvature increases with energy
  - Poorer resolution at high energy - relative uncertainty scales linearly with E
  - Limits usefulness of muons for new-physics searches

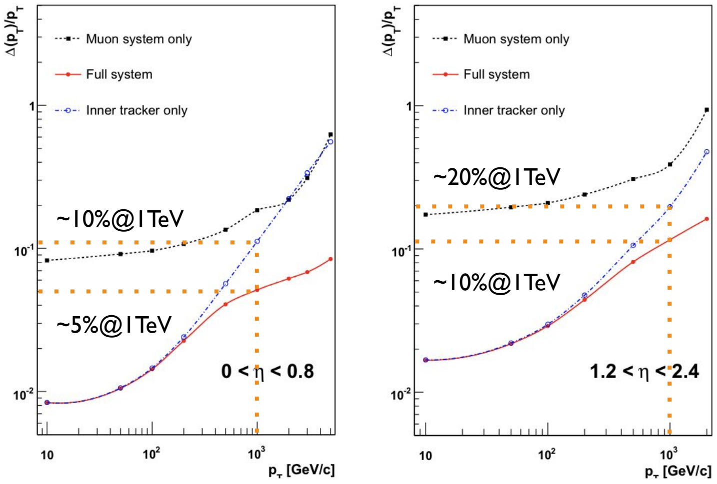


Image source: CMS JINST 3 S08004

MOTIVATION

- At high energy, muons no longer behave as minimum ionising particles
    - Radiative losses increase with energy
- By recording these low-energy photons in a granular calorimeter we gain a complementary measurement of muon energy
- But radiation is still low, e.g. 0.23%@1TeV

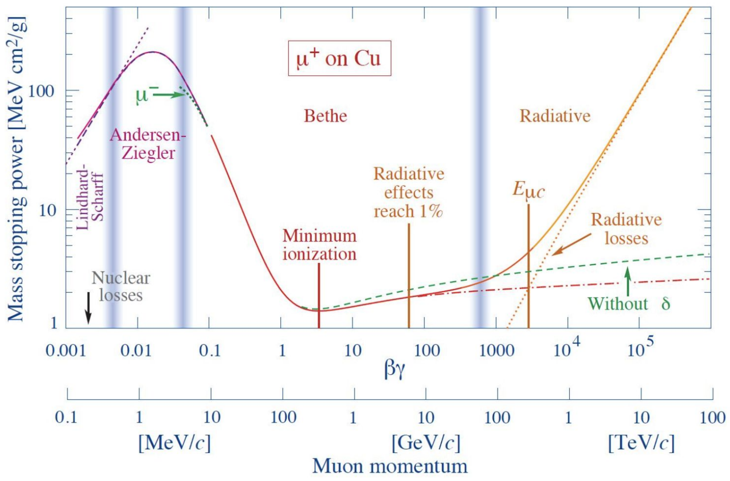


Image source: PDG Phys. Rev. D 98, 030001

APPROACH

- Reasonable to assume that the total deposited energy provides information the muon energy:
  - But radiation due to stochastic processes
  - Cannot precisely rescale reco energy to true energy
  - ATLAS does demonstrates that the E-sum is still useful in Nikolopolous et al. 2007
- Can the spatial info of the deposits help?
  - Could summarise deposits with high-level features
  - Relies on our domain-knowledge
- Can we work with the raw-hit data and learn to predict the muon energy?
  - 3D CNN could be used to run over the calo cells and learn to extract better HL features
  - Can still use HL-feature as additional inputs

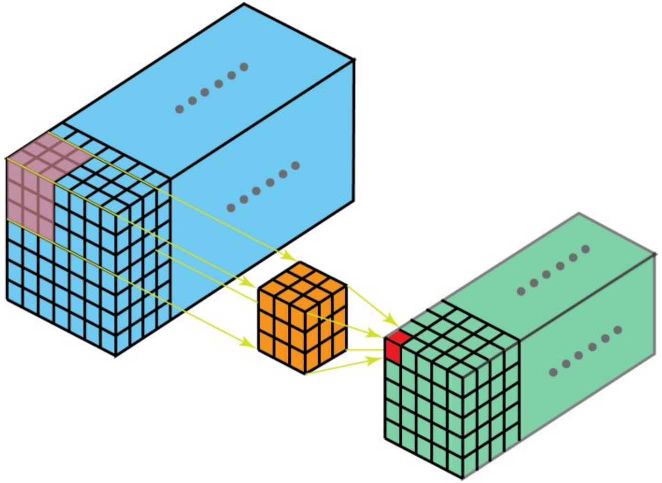


Image source: ashenoy ai.stackexchange.com

DATA

CALORIMETER

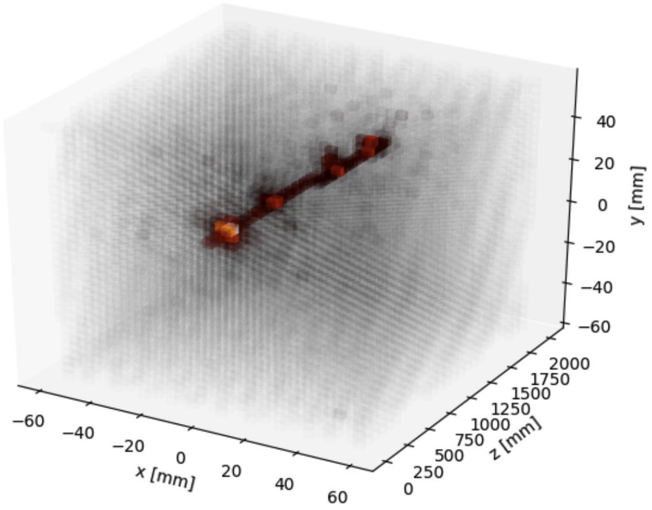
- Lead tungstate block
  - (2032x120x120 mm (z,x,y))
  - Subdivided into 39.6x3.75x3.75 mm cells
  - 50x32x32 (z,x,y) cells = 51200 cells total
- Immersed in uniform 2T magnetic field (along y-axis)

MUONS

- Generated with:
  - Uniform energy, $P = p_z$
  - Uniform x & y in [-20,+20] mm
- 887k training + validation muons, continuous energy in [50,8000] GeV
- 430k testing muons, discrete energy in 400 GeV steps in [100,4100] GeV

TYPICAL MUON

- Energy deposits concentrated along flight path
  - Occasional high concentrations
- Magnetic bending slightly visible
- Large number of low-energy deposits
- Relatively sparse hits



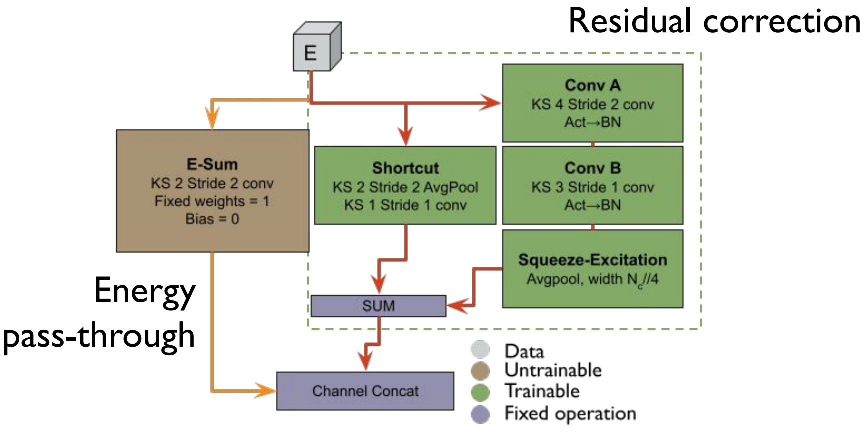
HL-FEATURES

- Most basic approach: sum up energy deposits and learn linear correction to true energy
  - But doesn't include multiplicity information
- Better to sum up energies according to cell energy, i.e. sum of energy in cells with:
  - 0-0.01 GeV
  - 0.01-0.1 GeV
  - >0.1 GeV
- These are referred to as the E-sum features

- 24 additional HL features computed
  - Energy measured from bending in calo
  - Energy spread for range of z depths
  - Features extracted from clustering of deposits
  - Energy 1st moments in x & y
  - MET

ARCHITECTURE - CNN BLOCK

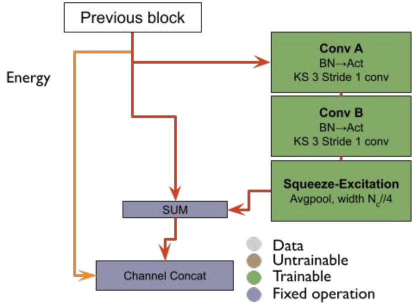
- Custom 3D CNN arch aims to learn small corrections to reco. Energy
  - Reco. energy summed up by 2x2 filter
  - Correction learnt by residual convolutional layers
  - Summed reco. energy is concatenated to output, so always available to later layers
- Running BatchNorm, helps with data sparsity
  - Applies running average during training, rather than batchwise stats
- Swish-1 used as activation function
- Squeeze-excitation block further improves performance



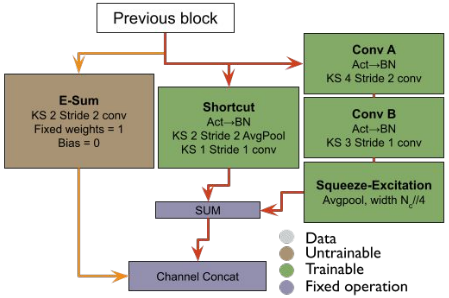
ARCHITECTURE - FULL MODEL

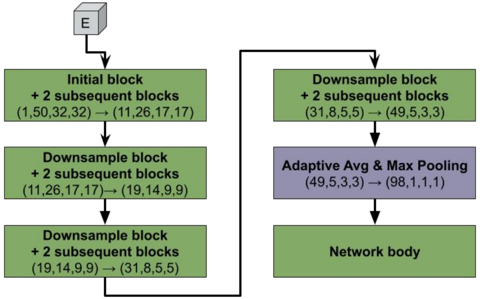
- Can build deeper networks by not downsampling the grid
- Further downsampling uses pre-activation layout
- Full CNN contains 12 blocks, followed by mean and max aggregation
  - 51,200 inputs → 98 features
- CNN head outputs combined with HL features and fed through 3 FC layers

Subsequent



Downsample





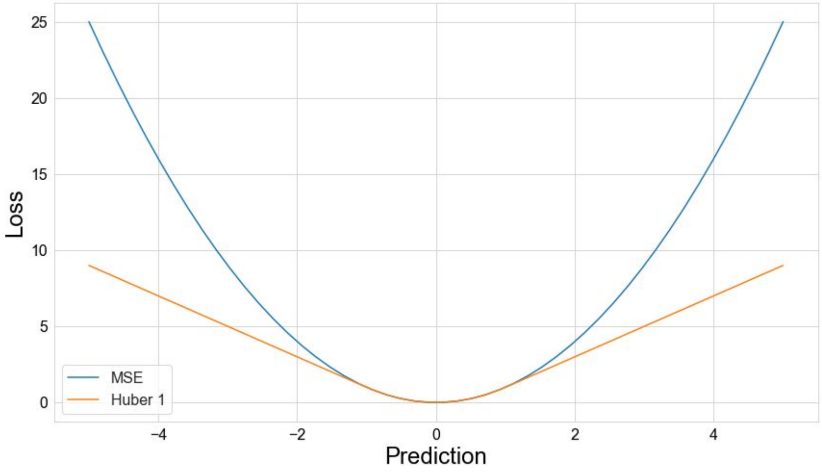
LOSS FUNCTION

- MSE places too much emphasis on high-energy
  - 5% error @ 100 GeV = 25 SE
  - 5% error @ 1000 GeV = 2500 SE
- Due to higher radiation levels, high energy muons easier to predict
- Instead use calo-inspired mean fractional squared error (MFSE)
  - 5% error @ 100 GeV = 0.25 SE
  - 5% error @ 1000 GeV = 2.5 SE
  - Lower dependence on true energy

$$\mathcal{L}(y, \hat{y}) = \frac{1}{N} \sum_{n=1}^{N} \frac{(y_n - \hat{y_n})^2}{y_n},$$

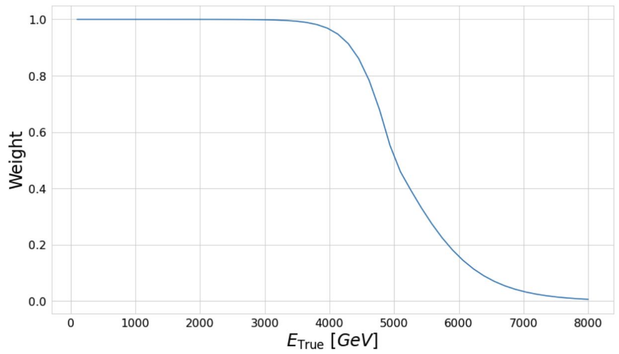
HUBERISED LOSS

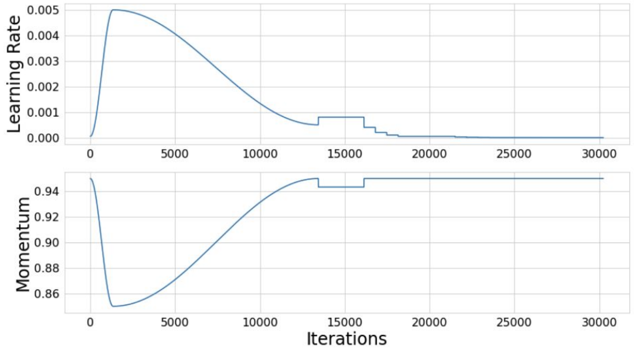
- Radiative losses are a stochastic process
  - Can have low energy muons with lots of radiation, & vice versa
  - Very easy to have outliers in data, whose poor predictions dominate the loss
- Huber loss transitions from squared error to absolute error
  - Diminishes the effect of outliers
  - We actually use multiple, adaptive transition points according to true energy (see backups)
- For MFSE instead compute Huber loss then divide by true energy



TRAINING

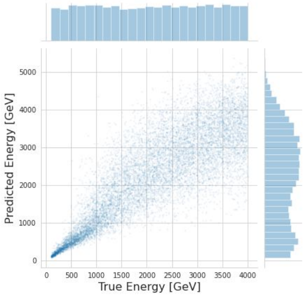
- Hard boundaries in training data → biassed regression response
  - Train on 50-8000 GeV muons
  - Evaluate on 100-4000 GeV muons
  - Slightly compensate training by down-weighting muons above 4 TeV
- Batchsize = 256
- Train ensemble of 5 models
- Each model trained according to:
  - 1-cycle schedule with cosine annealing for 20 epochs
  - Upto 30 epochs step-decay

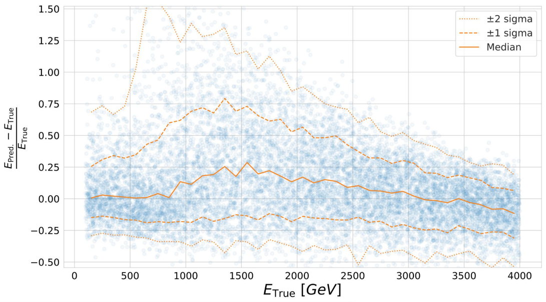


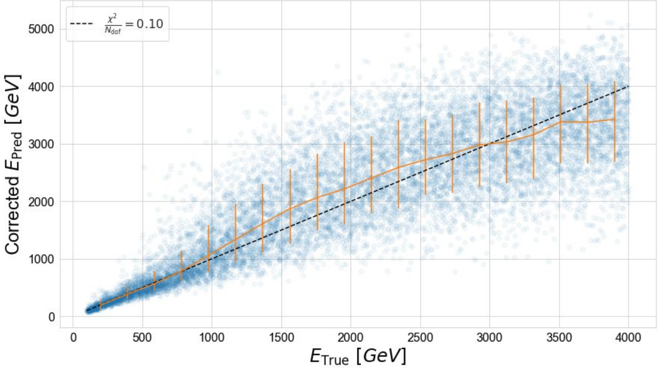


PREDICTIONS

- Regression predictions are fairly linear with true energy
  - But tend to slightly over-predict medium energy and under predict high energy
- Correct bias via linear fit to predictions in bins of true energy
  - Inversion of fit provides lookup-function mapping prediction to true energy (corrected predictions)
  - Still some residual bias
- The correction is fixed using validation data and then applied to testing data

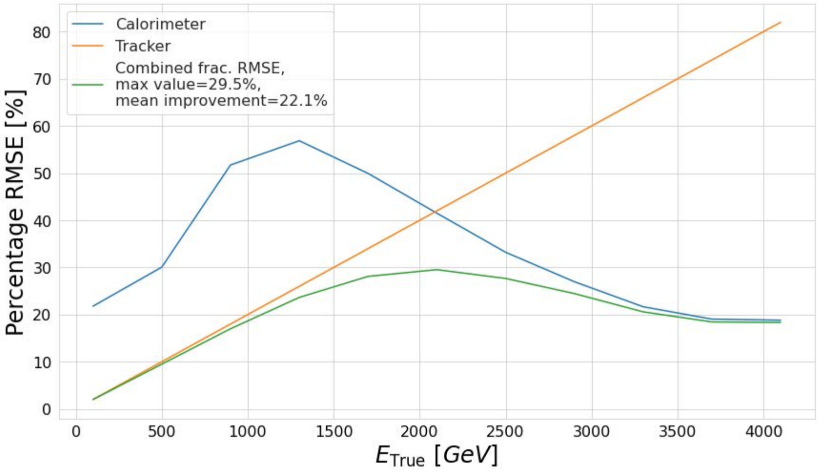






RESOLUTION

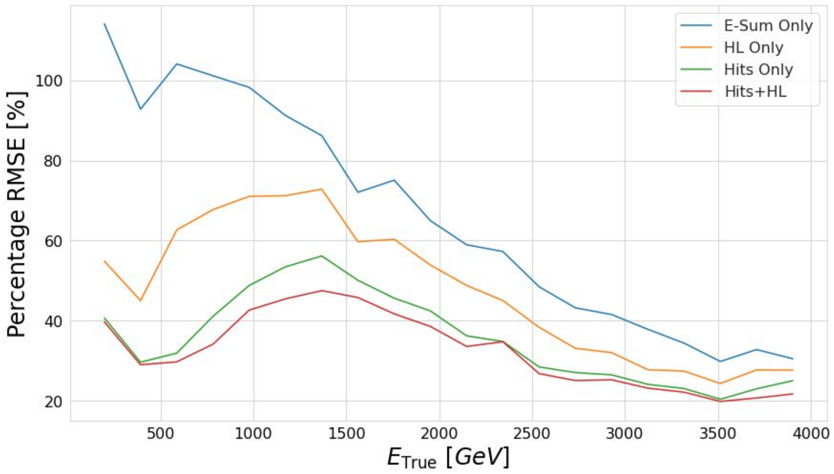
- Compute resolution as RMSE
  - Quadrature sum of variance and residual bias
- Differing energy dependence means calo and tracker measurements are complementary
  - Best resolution is a weighted average of both measurements
  - Assume here a linear tracker resolution with 20% @ 1 TeV
  - The improvement over tracker due to calo regression is useable as an optimisation metric



Final results on testing data

INPUT COMPARISON

- Do the HL features help?
- Comparing on validation data:
  - HL feats help significantly over using just the sum of recorded hits
  - HL feats still beneficial at low energy when CNN is used



CONCLUSIONS

- Demonstrated that radiative losses in calorimeters can provide muon-energy measurements in collider experiments
- To our knowledge this is the first such demonstration in collider context to fully utilise the raw hit information
  - IceCube previously measured muons but with much larger calo (20x in rad-lengths)
    - Resolutions 10x poorer than ours in our energy range (although their focus is more on higher-energy muons)
    - Abasi et al., 2013 & Aartsen et al., 2014
  - ATLAS also demonstrated a calorimeter-based method
    - But only sums up the deposited energy in cones, rather than considering the spatial structure of hits
    - Nikolopolous et al. 2007

- Calo measurements improve with energy and are entirely complementary to existing tracker measurement approaches
  - Allows good resolution across the full muon-energy spectrum
  - Can improves the sensitivity and reach of searches for New Physics
- Preprint: arXiv:2107.02119
- Data and code will be released soon

BACKUPS

ENVIRONMENT

SOFTWARE

- Networks & training: LUMIN+PyTorch
- Data generation: Geant 4
- HL-features: ROOT
- Data formats: HDF5 (final datasets), ROOT (raw data)

HARDWARE

- Data generation & feature processing: batch system
- Training & inference: Nvidia V100S
- Requirements:
  - Training: 5GB VRAM, 23GB RAM, 23*5 hours
  - Inference: 0.1*5 seconds per batch of 256 muons

ADAPTIVE HUBERISED LOSS

- Suitable transition points from SE to AE also depend on true energy
  - Batch divided by true energy into 5 bins
  - Transition point chosen to be 68th percentile of SE in each bin
- Problem: expect 51 muons per bin - transition point can fluctuate significantly
  - But: want thresholds to vary as training progresses - shouldn't fix thresholds
  - Solution: Loss class tracks running average of thresholds per bin

$$\mathcal{L}_{\mathrm{Huber},i} = t + \left(2\sqrt{t}\left(|y_i - \hat{y}_i| - \sqrt{t}\right)\right),$$

where *i* are indices of the data-points in a given bin with a squared-error loss greater than the threshold *t* for that bin.

The threshold for bin *j* evolves per minibatch as:

$$t_j \leftarrow 0.9 t_j + 0.1 \mathcal{L}_{\mathrm{SE},j,68^{\mathrm{th}}}.$$

ABLATION STUDY

LOSS

- MFSE and down-weighting both provide large benefits
- Huberisation provides indication of improvement
  - When used, multiple bins should be used with either fixed, or averaged thresholds

Ablation	MI	Change in MI [%]
Default	19.42 ± 0.08	N/A
Single bin	19.14 ± 0.08	-1.5 ± 0.6
Batchwise thresholds	19.25 ± 0.04	-0.9 ± 0.5
Non-Huberised loss	19.36 ± 0.06	-0.4 ± 0.5
Fixed thresholds	19.39 ± 0.05	-0.2 ± 0.5
MSE loss	18.43 ± 0.06	-5.1 ± 0.5
No down-weighting	16.5 ± 0.2	-15.12 ± 1.03

ENSEMBLING & DATA USAGE

- Regressors benefit from larger training datasets
- Ensembling recommended if inference time is not a concern

Ablation	Dataset size	Times		MI
		Training [h]	Inference [second per batch]	
Full ensemble	862 085	113.4	0.47	20.72
Full singles	862 085	22.7	0.091	20.29 ± 0.04
Unique ensemble	862 085	23.3	0.47	19.83
Unique singles	197 048	4.7	0.0.091	19.37 ± 0.08

CNN ARCH

- CNN much better than flattening raw hits
- Running BN improves performance and stability
- ResNet layout is useful
- Other additional components provide indications of improvement

Ablation	MI	Change in MI [%]
Default	19.42 ± 0.08	N/A
No BN	18.5 ± 0.3	-5 ± 1
No identity path	18.72 ± 0.08	-3.6 ± 0.6
Nominal BN	19.2 ± 0.2	-1.1 ± 0.9
No E-pass	19.30 ± 0.05	-0.6 ± 0.5
No SE	19.33 ± 0.09	-0.5 ± 0.6
No pooling	19.4 ± 0.1	-0.4 ± 0.7
No CNN	17.45 ± 0.09	-10.2 ± 0.6

OTHER ABLATIONS

- 1-cycle + step-decay provides a 5.2±0.6 % increase in performance and quicker convergence
  - Compared to fixed optimiser hyper-parameters and same number of maximum possible epochs
- Whilst minor, the bias correction improves performance by 2.1%