Image Calorimetry: Status and Challenges

Josh Bendavid (Caltech/LPC)
on behalf of the CMS Collaboration

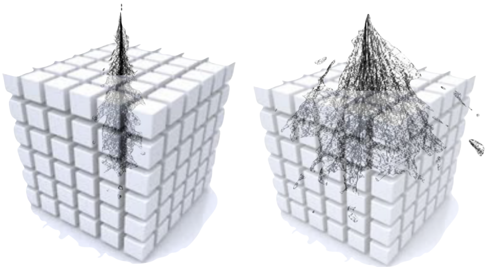
Caltech

CMS

May 8, 2017
DS@HEP 2017, FNAL

Introduction

- Purpose of calorimetry is to measure the **energy** of incoming particles
- Calorimeters typically constructed from dense material $\rightarrow$ incoming particles interact and initiate a **shower** of successively lower energy particles, with $\sim$ all of the incoming energy ultimately deposited in the calorimeter material through further interactions
- In "active" material, the deposited energy can be measured by e.g. electronic or optical means
- Calorimeters may contain different arrangements of "active" and "passive" material, may be segmented longitudinally (1D), transversely (2D) or both (3D)



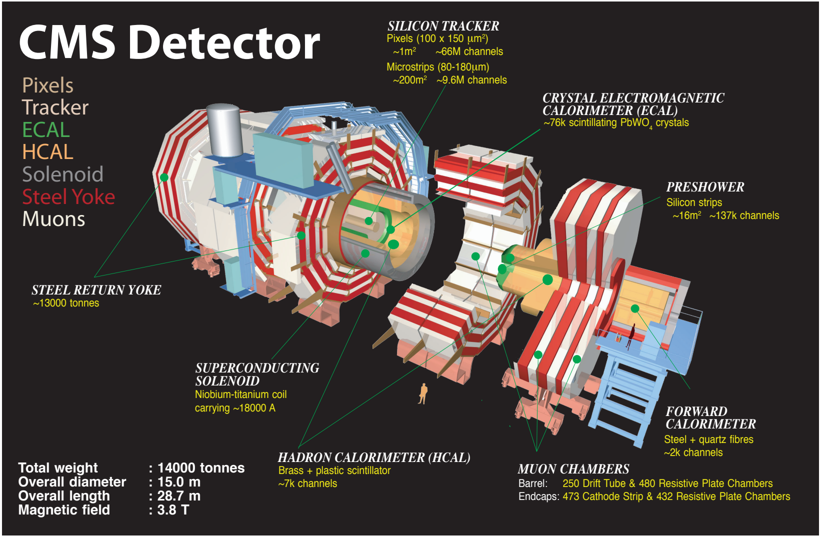
Introduction

- **Some general considerations for a transversely segmented calorimeter:**
  - Number and geometry of cells to include in energy sums
  - **Local containment:** Loss of energy outside of included cells or in gaps/cracks
- For a longitudinally segmented calorimeter: additional degree of freedom: Optimal weighting of layers
- In a hadronic environment: **Contamination of showers** by additional activity (hadrons or photons from neutral meson decays)
- In case of significant upstream material and/or magnetic field: **Global containment:** Loss of energy to interactions with material, low momentum charged particles or bremsstrahlung at wide angles

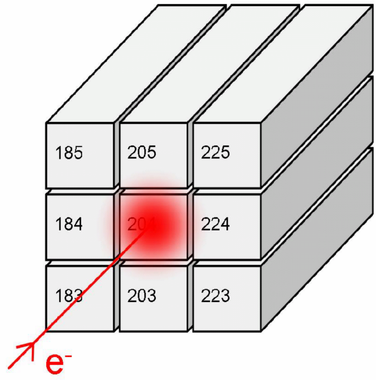
Introduction

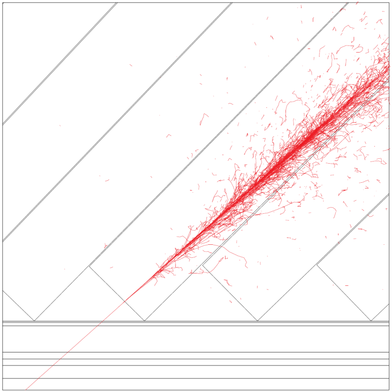
- **For ultimate energy resolution need to take care of two related questions:**
    - How to optimally combine information from different channels
    - How to optimally correct the energy for known losses/overmeasurement/nonlinearity
    - In a machine learning context, this is a prototypical **regression** problem
- One is often also interested in the **type** of incoming particle (single γ vs $\pi^0 \to \gamma\gamma$ vs $e^{\pm}$ vs $\pi^{\pm}$
    - Discriminating power is contained in the spatial distribution of the showers
    - Typical binary or multi-**classification** problem
- Will start with some examples using the (current) CMS Electromagnetic calorimeter

The CMS Detector



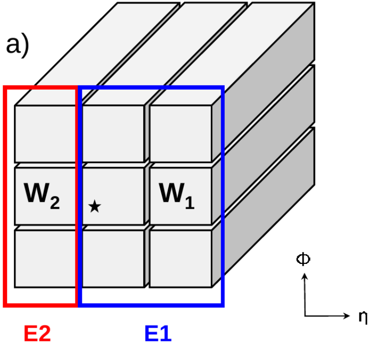
Electromagnetic shower in ECAL

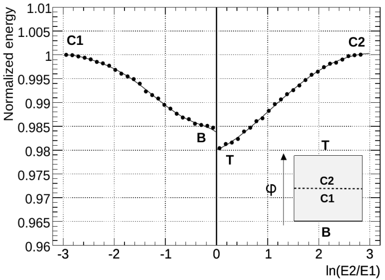




- Transversely segmented (2D) calorimeter, longitudinally homogeneous (no segmentation or passive layers)
- Crystals are approximately $1R_M^2$ at front face $\rightarrow \sim 95\%$ containment in 3x3 grid, $> 99\%$ in 5x5 grid

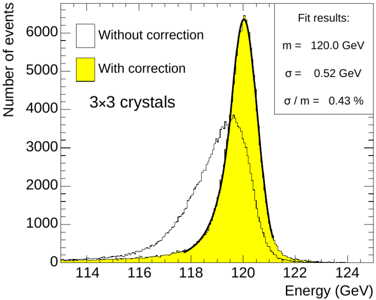
Parametrized Local Containment Corrections

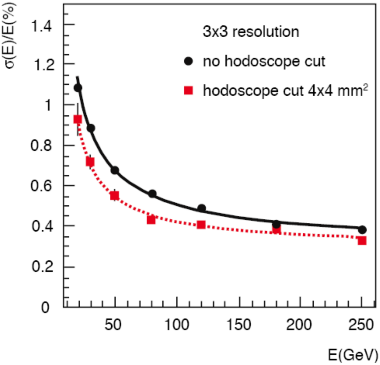




- Appropriate energy ratios are a good proxy for impact location relative to crystal center → Local containment losses can be parametrized in-situ

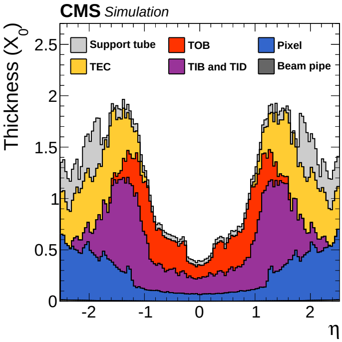
Parametrized Local Containment Corrections





- Significant improvement on energy resolution, but even after corrections, resolution is better for central incidence

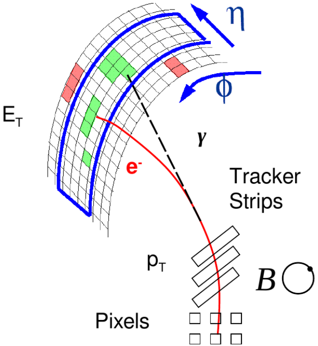
Material Interactions and Global Containment

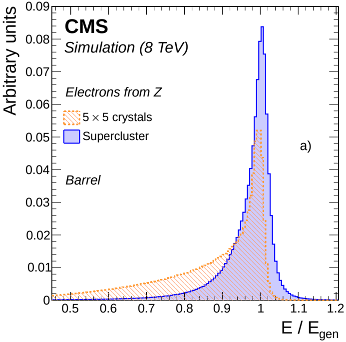


(a) Tracker Material Budget

- Lots of material in front of the ECAL, significant probability of photon conversions and bremsstrahlung from electrons

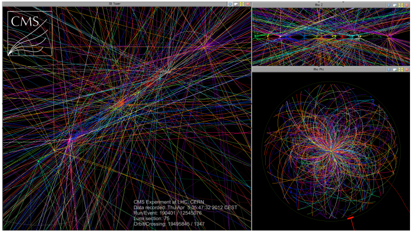
Material Interactions and Global Containment



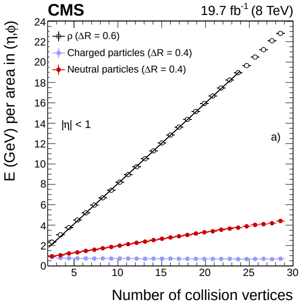


- Reconstruction forms Superclusters extended in ϕ to collect conversion legs/bremsstrahlung spread out by magnetic field
- Soft conversion legs and associated bremsstrahlung may not reach calorimeter or arrive too far to be included in Supercluster

Pileup in CMS



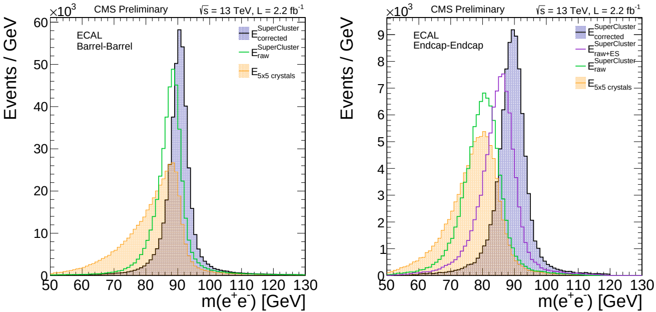
(a) Event with 29 vertices



(b) Energy Density vs pileup

- Additional energy from pileup contaminates isolation cones, but also the shower itself, (8 TeV data collected with an average of 20 pileup interactions per crossing)
- **Contamination depends on the size of the cluster**

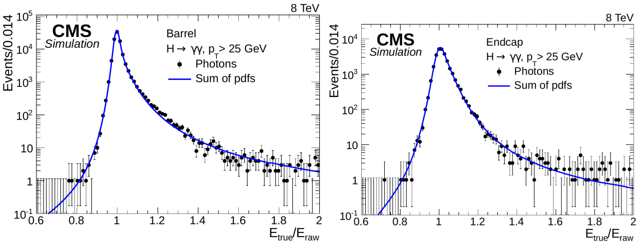
Photon/Electron Energy Reconstruction/Regression



(a) Barrel (b) Endcap

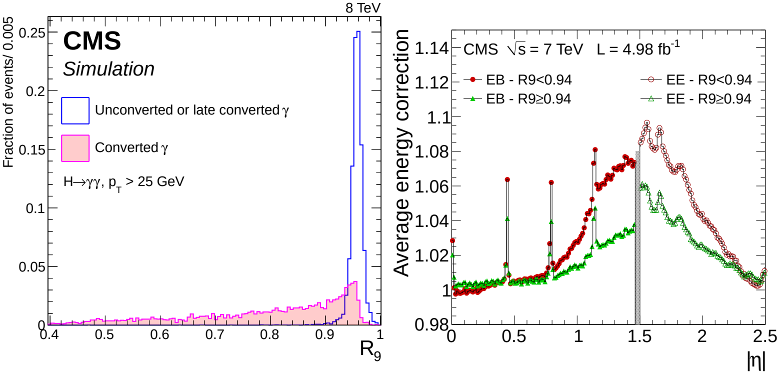
- High-level corrections to Super Cluster energy with BDT regression based on ∼high level features (shower shape variables, different energy ratios, cluster location, pileup energy density, etc)
- Reconstructed Z mass in data with different levels of energy reconstruction and corrections
- Progression clearly visible even with 2.5 GeV natural Z width

Energy Regression: Predicted Response Distribution



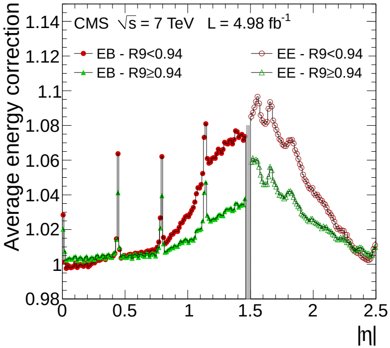
- Semi-parametric regression provides a prediction for the full lineshape (here showing simulation vs regression-prediction for target variable E_{True}/E_{Raw})
- Total predicted pdf is given by sum of predicted lineshape for each simulation event

Inside the corrections

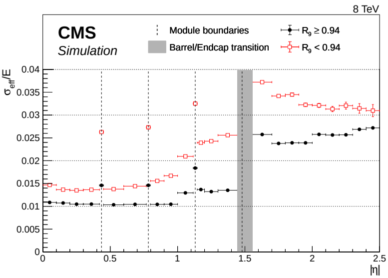


- $R_9 = E_{3x3}/E_{SC}$ is an effective, but not 100% pure conversion tagging variable (electrons and photons treated separately, no explicit converted vs unconverted distinction)
- Correction vs η has a non-trivial correlation with R_9 (and other shower profile variables)

Per-photon Resolution Estimate



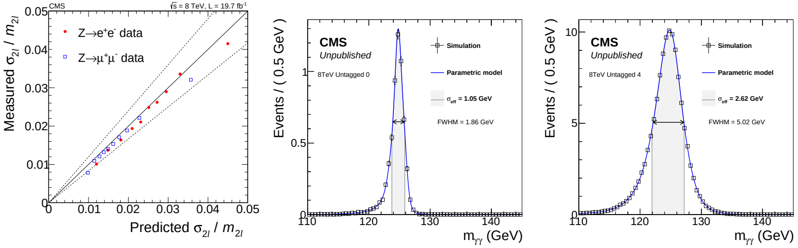
(a) Correction



(b) post-correction resolution

- Strong, but non-trivial relationship between size of correction and post-correction resolution (size of effect vs photon-to-photon fluctuations)
- Per-photon resolution estimate mapped with the full granularity of the multidimensional space used to derive the corrections

Per-photon Resolution Estimate



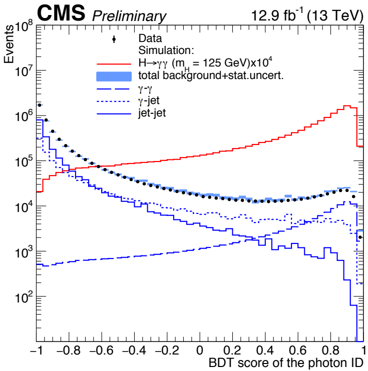
(a) Observed vs predicted σ_m

(b) $H \to \gamma\gamma$ Best Category

(c) $H \to \gamma\gamma$ Worst Category

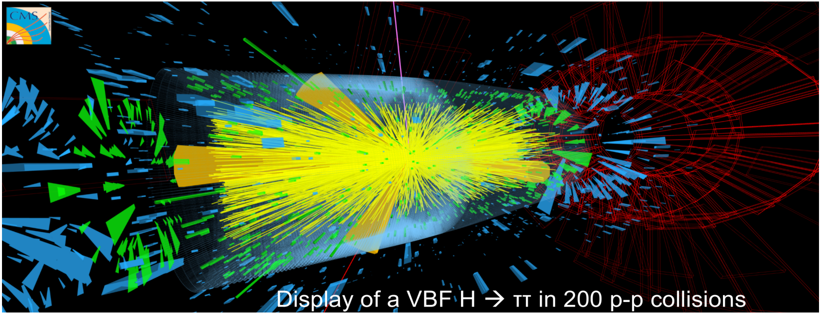
- In a resonance search, per-photon resolution estimate can be used to construct a per-event mass resolution estimate $\frac{\sigma_m}{m_{\gamma\gamma}} = \frac{1}{2}\sqrt{\frac{\sigma_{E1}^2}{E_1^2} + \frac{\sigma_{E2}^2}{E_2^2}}$
- Can be used to select or categorize events to make optimal use of highest resolution events (two unconverted photons in the center of the detector, incident on the center of the crystal, far from module boundaries)
- In-situ estimate of the **accuracy** is important

Photon Identification



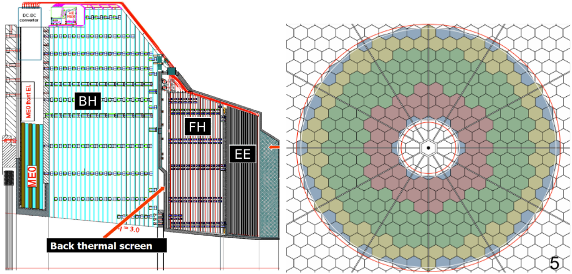
- Particle identification at the shower level so far mostly cast as a binary classification problem
- Main background for photons is boosted/collinear $\pi^0 \to \gamma\gamma$ decays inside jets
- BDT-based classifiers exploit high level shower width variables and energy ratios as well as isolation

Calorimeters for HL-LHC



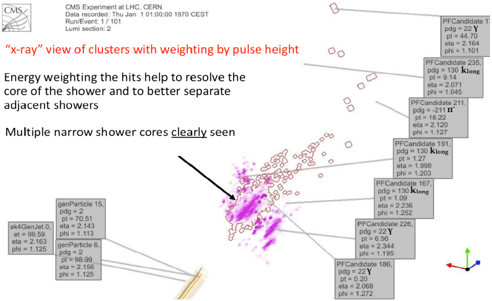
- HL-LHC implies a large increase in beam intensity and pileup→up to 140-200 interactions per crossing

Calorimeters for HL-LHC



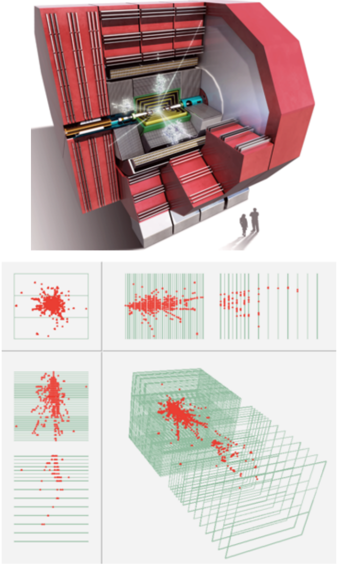
- CMS forward calorimeter to be replaced with High Granularity Calorimeter including 28+12 high granularity silicon layers (+absorber)
- High granularity provides more information to disentangle nearby or overlapping showers
- **Much more difficult to construct high level features by hand summarizing relevant info** $\rightarrow$ deep learning on lower level inputs strongly motivated
- Closely related to silicon calorimeter designs for future linear collider experiments

CMS HGCAL



- Calorimetry starts to converge towards a 3D image-like problem
- Clustering/separating overlapping particles in high pileup environments is non-trivial → machine learning may be useful also for pattern recognition (moving beyond traditional regressions and classifiers)
- Beyond model accuracy, computational complexity of inference is critical (possible large gain over traditional algorithms with combinatorics)

Linear Collider Calorimeters

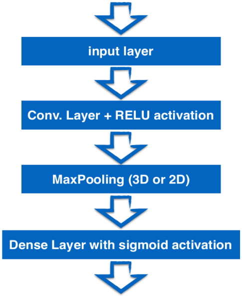


- High granularity silicon + absorber design also employed for proposed linear collider detectors (for example the "LCD Calorimeter" CLIC detector design shown here)
- Simpler test case for machine learning techniques in several ways
  - Lower density environment → more easily factorize clustering/pattern recognition from regression and classification-type problems
  - Rectangular geometry
  - Fewer restrictions on use of simulation outside of collaborations → see Amir's talk

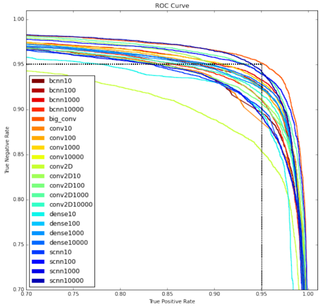
Machine Learning on Image Datasets

- Extensive work on 2D image recognition, generative models, etc in the literature
- In most cases extendable to 3D images a la high granularity calorimeter showers

Simple Example for LCD Calorimeter



(a) Example Architecture

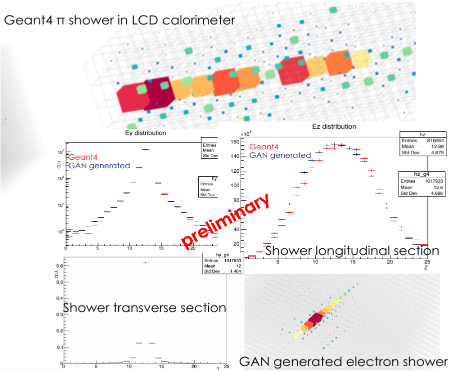


(b) ROC Curves

N. Howe

- "Simple" binary classification problem (photon vs π^0)
- Different network architectures tested including dense, 2D, and 3D convolutional layers
- More recent progress to be discussed during hands-on session this afternoon

Applications of Generative Models



S. Vallecorsa et. al. `https://indico.cern.ch/event/627852/contributions/2538513/attachments/1450473/2236405/MLsimulation_GeantV.pdf`

- Calorimeter Showers can be simulated accurately with detailed detector simulation and per-particle tracking (GEANT), but this is extremely cpu intensive
- HL-LHC physics program requires billions of simulated events → trillions of simulated calorimeter showers
- Generative deep learning models provide a possible path to fast, but high accuracy simulations
- See Sofia's talk Wed. Morning

Outlook

- Development of deep learning techniques for calorimetry closely linked to development of new detector designs and associated reconstruction and simulation
- Existing techniques from 2D image processing/recognition/generation serve as an important foundation
- Some challenges:
  - Optimization of network architectures
  - Extension of convolutional techniques to non-rectangular geometries
  - Incorporation of precision timing measurements within the shower

Backup

Regression Energy Corrections

- Photon energy reconstruction in CMS:

$$E_{e/\gamma} = F_{e,\gamma}(\bar{x}) \times \sum_{i}^{N_{crystals}} G(GeV/ADC) \times S_i(t) \times c_i \times A_i$$

- Two main components to photon energy resolution which at least partly factorize:
  1. Crystal level calibration (ADCtoGEV, Intercalibration, transparency corrections)
  2. Higher level reconstruction (**local containment, global containment, PU contamination**)
- Shower containment is complex and not clear if/how different contributions factorize
- Best performance is obtained with multivariate regression using BDT with cluster η, ϕ, shower shape variables, local coordinates, and number of primary vertices/median energy density as input
- Regression is trained on real electrons/photons in Monte Carlo, using the ratio of the generator level energy to the raw cluster energy, also provides a per photon estimate of the energy resolution

Evolution of Regression Energy Corrections in CMS

- Photon energy regression in CMS initially trained using TMVA BDT implementation
- Physics performance was ok, but serious problems with size on disk and memory consumption (1GB xml files!)
- CMS has an in-house BDT storage format, persistable in root file or conditions database, disk/memory/cpu efficient (tree structure represented in flattened arrays, one inlined while loop for evaluation). Can convert weights from TMVA or produce with native BDT training tool written to exploit parallelization, speed up training with large datasets, produce more compact trees
- Later CMS moved to "semi-parametric" regression

Evolution of Regression Energy Corrections in CMS: "Traditional" Regression

- Multivariate techniques used in general to overcome lack of knowledge of multidimensional likelihood using finite event samples
- Traditional regression as used so far based on minimization of Huber loss function for target prediction $F(\bar{x})$ given target variable $y = E_{True}/E_{Raw}$ for a set of input variables $\bar{x}$ (in our case cluster position, shower profile and pileup variables)

$$L = \begin{cases} \frac{1}{2}(F-y)^2 & |F-y| \leq \delta \\ \delta\left(|F-y| - \delta/2\right) & |F-y| > \delta \end{cases}$$

- Minimized the square deviation out to some cutoff (by default $\pm 1\sigma$) and the linear deviation beyond that
- No built-in estimate of the per-photon resolution, accomplished with a second training on an independent subset of the training sample with target $y = |E_{Cor}/E_{Raw} - E_{True}/E_{Raw}|$

Semi-parametric Regression

- Start with ansatz that in any infinitesimal slice of phase space in $\bar{x}$, the energy response distribution is given by a double crystal ball (ie gaussian core with power law tails on both sides)
- In terms of E_{True}/E_{Raw} the **right** tail (undermeasurement of the energy) corresponds to the usual radiative losses, etc, whereas the **left** tail (overmeasurement of the energy) comes from pileup, etc.

$$p(y|\bar{x}) = \text{DoubleCrystalBall}\left(y|\mu(\bar{x}), \sigma(\bar{x}), \alpha_{left}(\bar{x}), n_{left}(\bar{x}), \alpha_{right}(\bar{x}), n_{right}(\bar{x})\right)$$

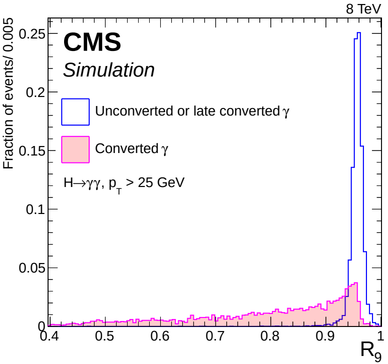
Semi-parametric Regression

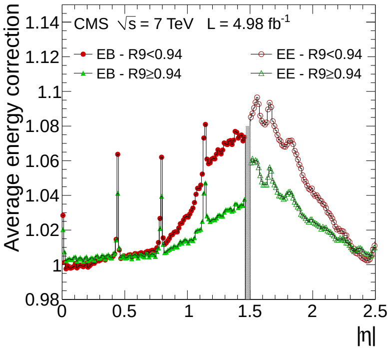
- The log likelihood ratio for a training sample can be written simply as

$$L = -\sum_{MCPhotons} \ln p(y|\bar{x})$$

- Minimize this loss function directly with gradient boosting, where $\mu(\bar{x}), \sigma(\bar{x}), n_{left}(\bar{x}), n_{right}(\bar{x})$ are regression outputs estimated by BDT's (using RooFit-based bdt-training tool, which ensures proper pdf normalization, etc)
- This gives a simultaneous estimate for energy correction and resolution among other things

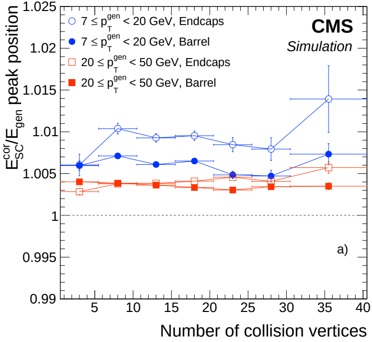
Inside the corrections



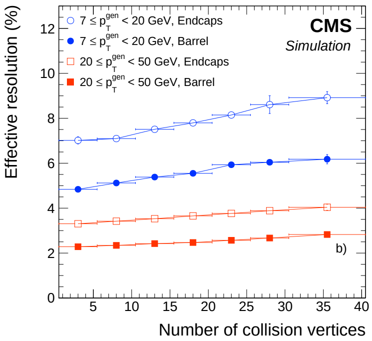


- Correction is parametrized and plotted with respect to the supercluster energy, but the corrected energy can also be considered a non-trivial weighting of supercluster, 3x3, 5x5, and other energy sums/ratios in input (dynamic noise/pileup vs containment tradeoff as a function of shower energy, inferred impact position, etc)

Pileup Contamination



(c) Scale vs Pileup

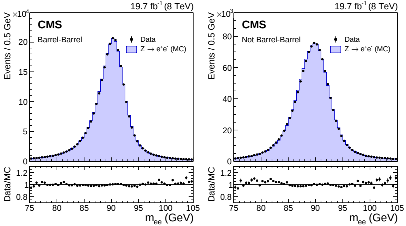


(d) Resolution vs Pileup

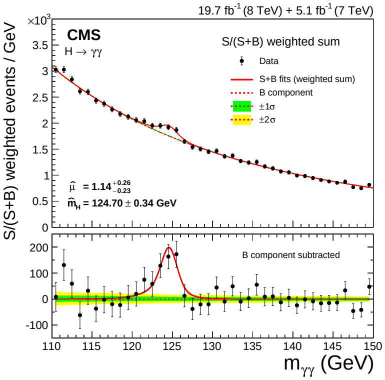
- After corrections, scale is flat vs pileup, resolution only modestly degraded

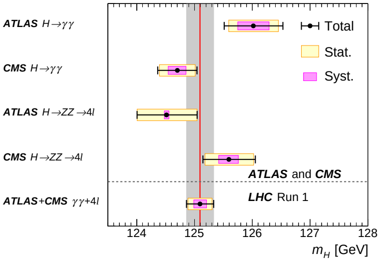
Energy Scale and Resolution

- Photon Energy Scale and Resolution in data measured with $Z \rightarrow ee$ events, applying either final photon-trained regression corrections, or equivalent electron-trained version
- Monte Carlo is smeared to match data resolution
- Data energy scale is adjusted to match Monte Carlo
- Energy scale is determined very precisely from (millions of) $Z \rightarrow ee$ events, remaining systematic uncertainties from electron-photon extrapolation and extrapolation in energy
- Overall systematic uncertainty on higgs mass measurement (dominated by energy scale uncertainty) 0.12% (but per-photon energy scale uncertainty varies according to detector region and photon quality)



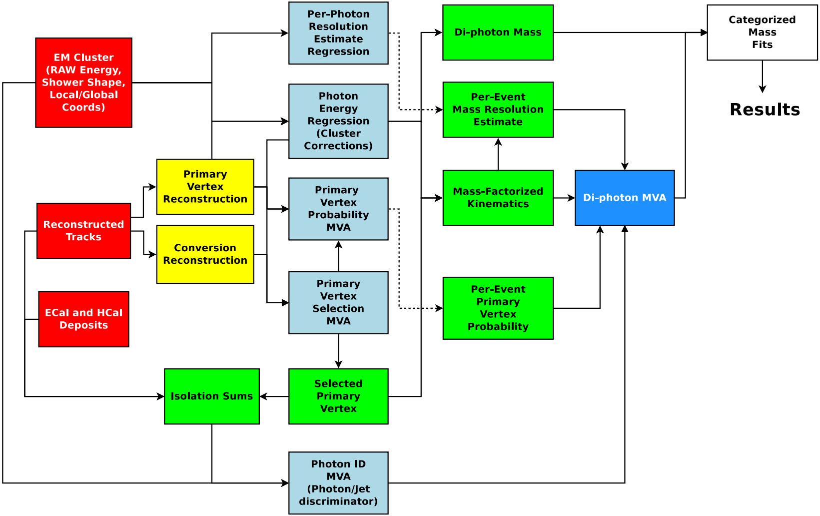
Result





- Already a precision measurement of the mass, CMS $H \to \gamma\gamma$ most precise single measurement
- Multivariate corrections and use of per-photon resolution in categorization play a crucial role

Higgs$\rightarrow \gamma\gamma$: All Together



- Strategy: Process available information into quantities with straightforward physical interpretations in order to combine per-event knowledge of expected mass resolution and S/B into a single "Diphoton MVA" variable