Thomas Reitenspiess (ETHZ) on behalf of the CMS collaboration

CMS detector simulation tuning through machine learning

Detector Simulation Tuning

Motivation

- The $H \rightarrow \gamma\gamma$ analysis takes fully differential input variables to e.g. measure differential cross-section

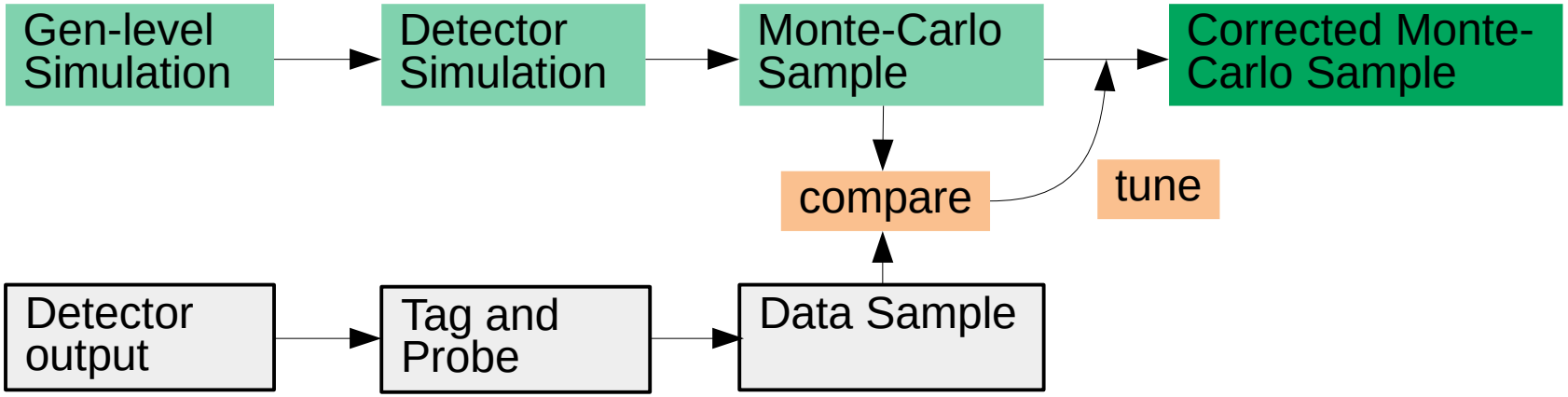
Differential detector simulation correction using $Z \rightarrow e^+e^-$ tag and probe sample

- Measuring $H \rightarrow \gamma\gamma$, this requires a good understanding of the detector response to photons
  - Photons produce showers in CMS ECAL
    - ➔ **Shower shape variables**
  - Account for additional object in the detector
    - ➔ **Isolation variables**

- Using electrons from $Z \rightarrow e^+e^-$ to measure and fine-tune the ECAL response to isolated electromagnetic objects
  - Electrons leave very similar trace as photons in ECAL and can be reconstructed as such

Detector Simulation Tuning

General Strategy



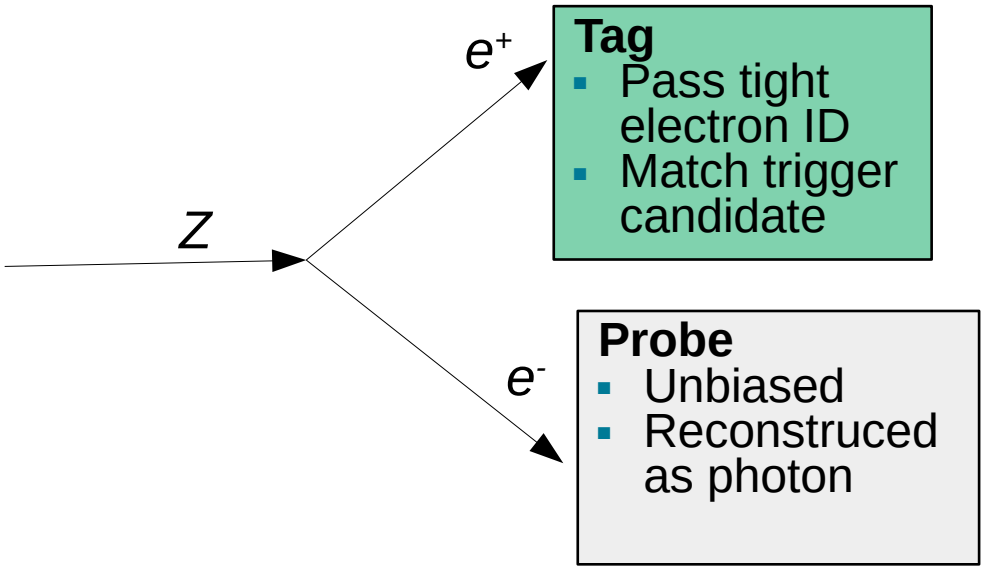
- Using the decay $Z \rightarrow e^{+}e^{-}$ as a clean standard candle, since its signature in the detector is very well understood
- Using detector related variables
- Therefore: remaining differences come from not simulating detector perfectly
- This strategy allows effective tuning of simulation

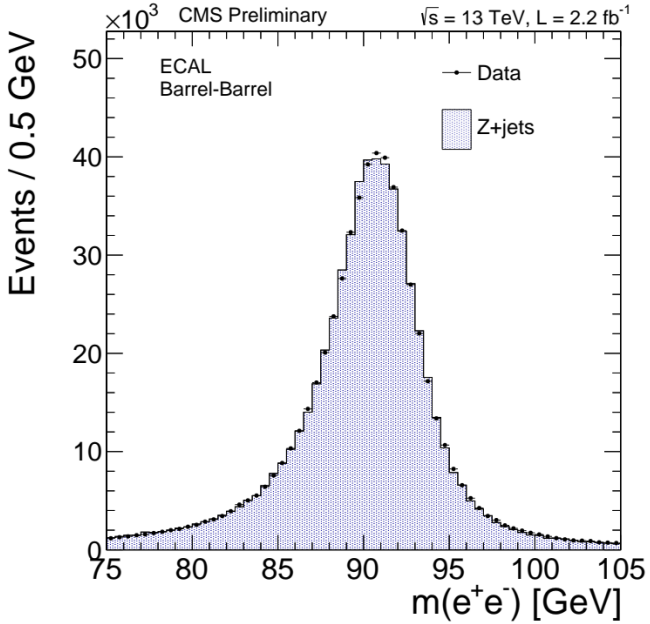
Detector Simulation Tuning

Tag and Probe

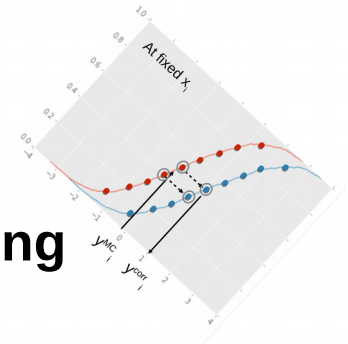
Tag and Probe

- One electron is identified very tightly and used as tag
- Other electron is used as unbiased probe to test detector response and derive corrections
- Electrons leave very similar trace in ECAL as photons, can therefore be used to derive correction for photons
- Very large sample – $O(10^7)$ – gives the ability to fine-tune detector response



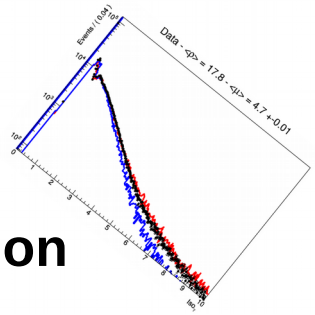


Detector Simulation Tuning
Methods



Quantile Morphing

- Shift simulation to match data according to cumulative distribution functions
- Cdf can be obtained differentially in kinematics and event-energy-density using quantile regression
- Method works only on continous distributions
- Generally applicable effective correction method

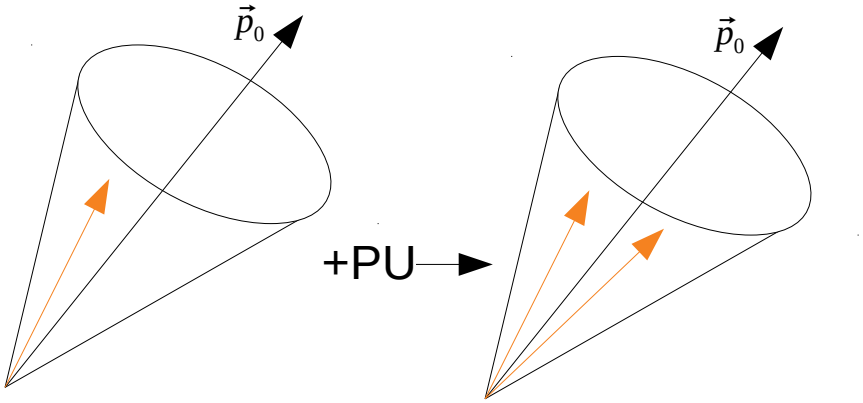


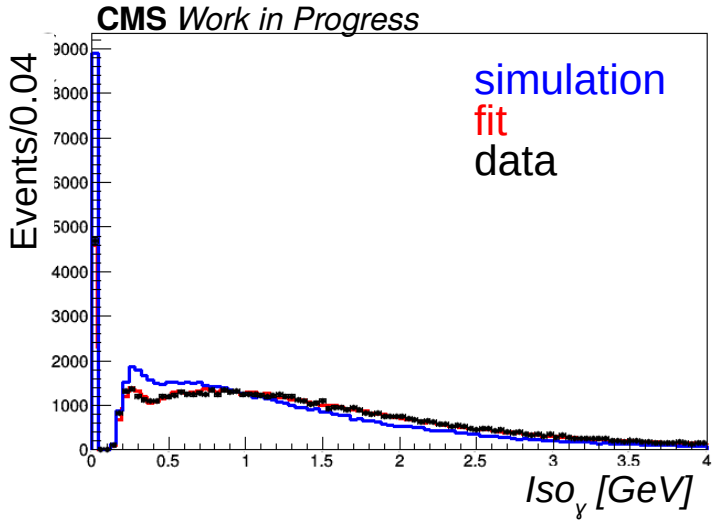
Stochastic Correction

- Works for discontious variables
- Developed for isolation variables
- Stochastic element needed to account for discontinuity caused by detector properties
- Effective correction method for effects caused by mismatch of number of objects in isolation cone

Detector Simulation Tuning

Stochastic Correction



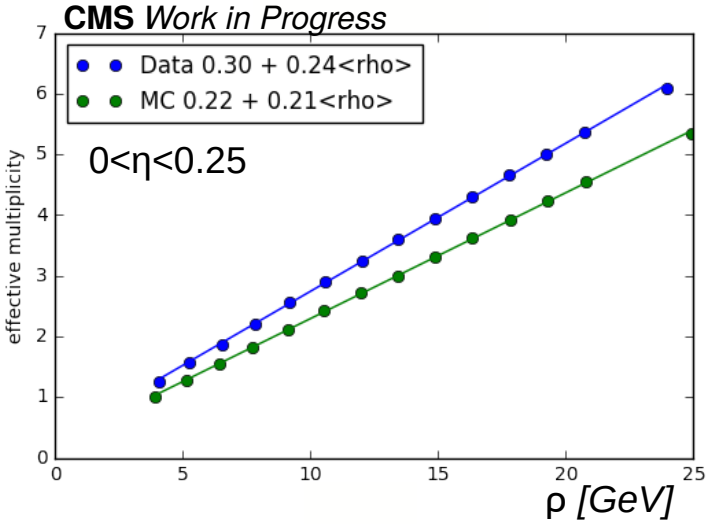


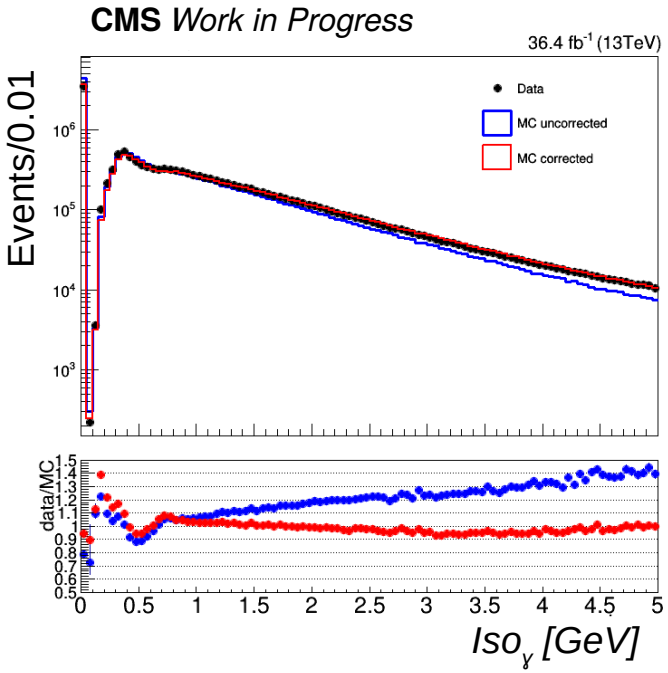
- The highger the event-energy-density ρ the more objects will populate the isolation cone
- The low pile-up behaviour is simulated well
- Therefore the simulated isolation distribution for low pile-up is taken as starting point

- The low pile-up isolation distribution is resampled μ times
- Between distributions for integer numbers of μ an interpolation is performed to get distribution continous in μ
- Likelyhood fit to data and simulation of resampled distribution is performed with μ as free parameter, binned in η and ρ

Detector Simulation Tuning

Stochastic Correction





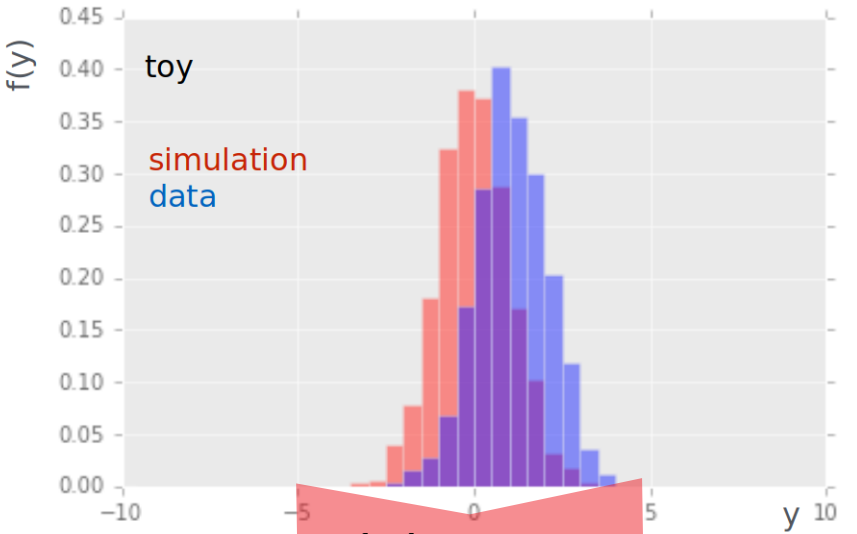
- In every η-bin, get μ for data and simulation for different values of ρ
- Perform linear interpolation between them
- Add transverse momentum to the cone in simulation according to the difference in μ, depending on η and ρ for every event

- Result shows good agreement between data and corrected simulation
- First bin (Iso_{γ}=0) is corrected very well
- Method catches also non-trivial detector effects

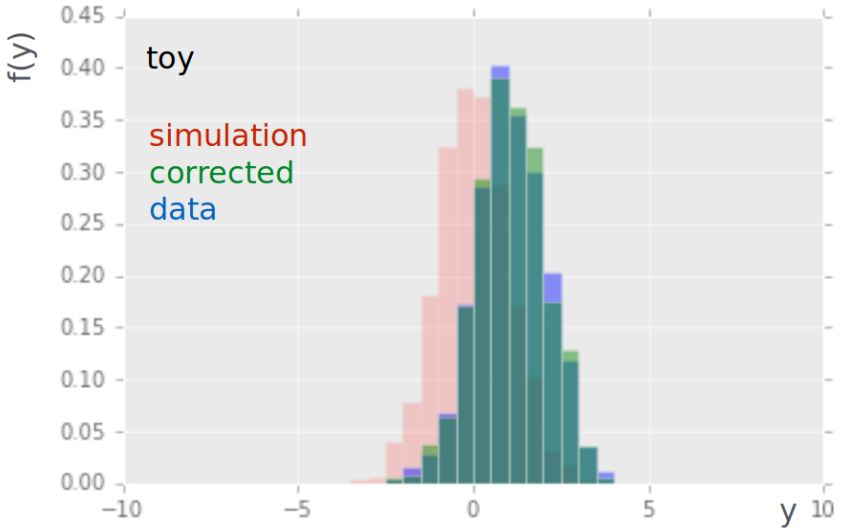
Detector Simulation Tuning

Quantile Morphing

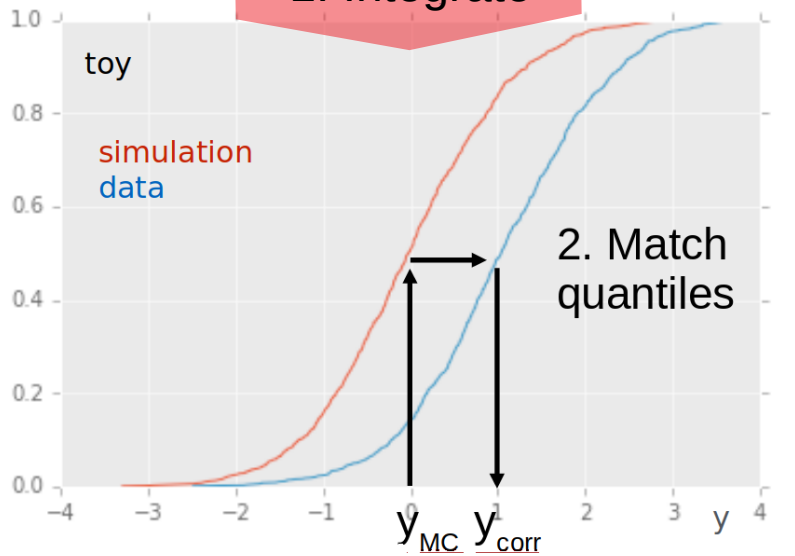
- Requirements:
  - Continuous distribution
  - Availability of pure control sample



1. Integrate



3. Transform



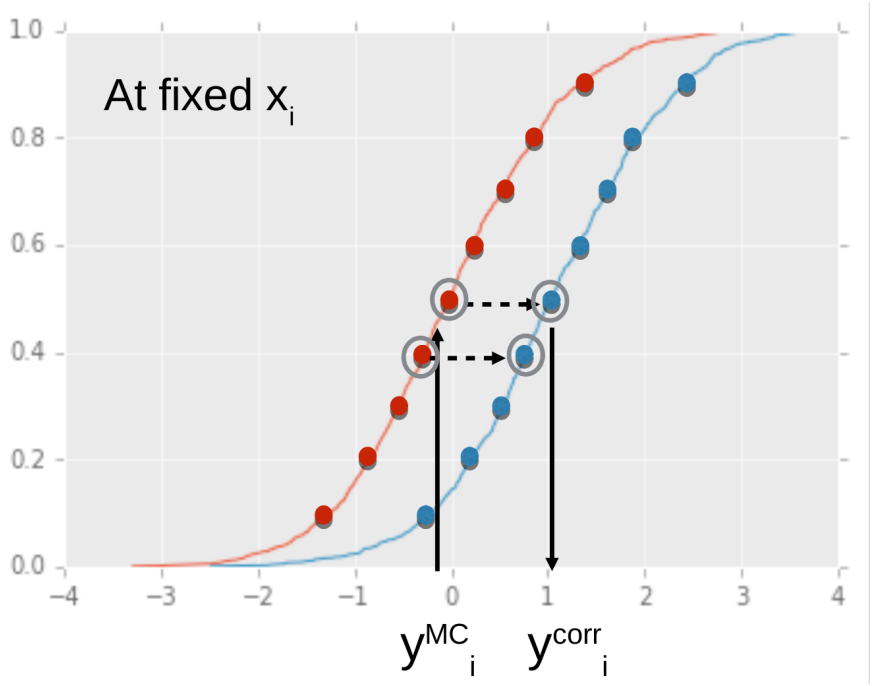
Detector Simulation Tuning

Quantile Regression

Goal: Differential Corrections

Measuring differential fiducial cross-sections, therefor it is crucial to have a well corrected simulation in every region of the phase space

1. Train n_q BDTs per variable to predict conditional shape of cdf depending on p_t, η, ϕ, ρ for data and simulation
2. Find two q_t aorund y w.r.t $x_i = [p_t, \eta, \phi, \rho]$ for data and simulation
3. Use linear interpolation between the two points $(q_\tau,\tau)_i$ to find $cdf(y|X)$ for data and simulation
4. Correct simulation by matching cdf_{data} and cdf_{mc}

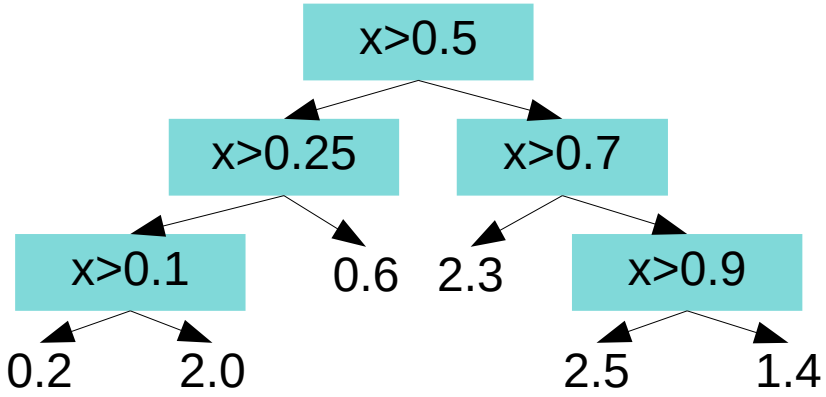


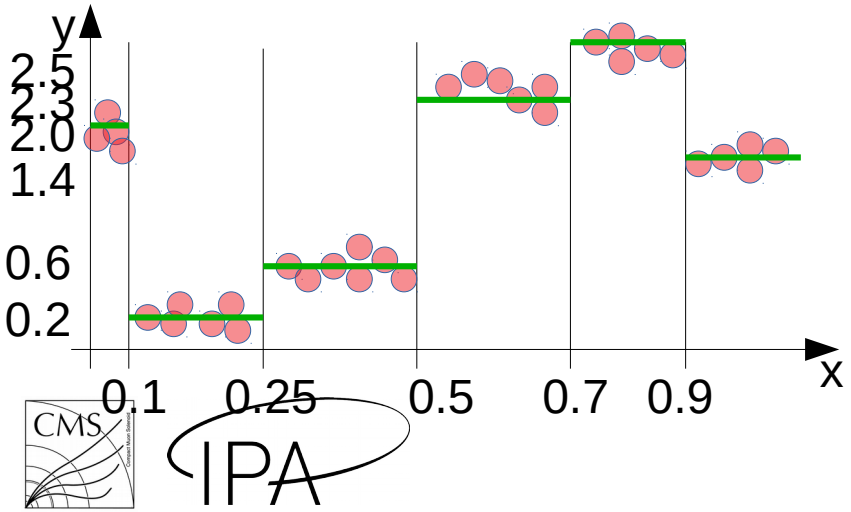
Detector Simulation Tuning

Boosted Decision Trees

Training Data Set $(x_i, y_i), i = 1, \ldots, n$

Decision Tree splits dataset into regions in $x = [p_t, \eta, \phi, \rho]$ in a binary fashion, to minimize Loss function



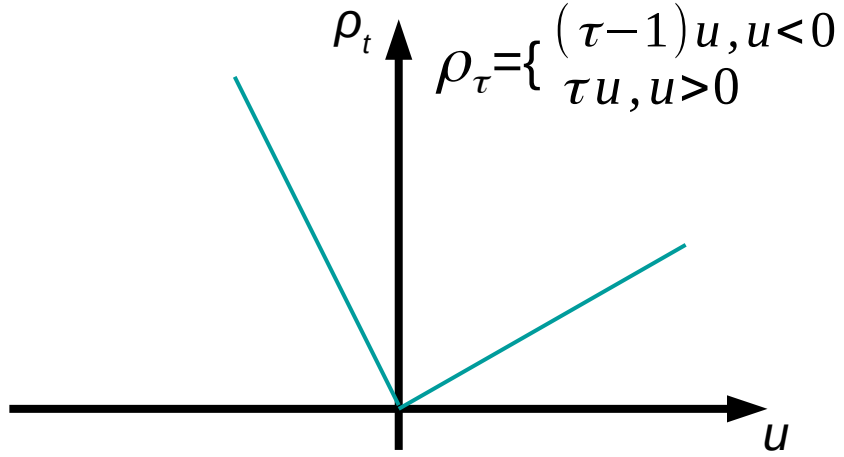


In practice

- Using gradient boosting
- Minimizing quantile loss function
- Catch dependence of corrected variables to $x = [p_t, \eta, \phi, \rho]$
- Input variables not limited to these, could add e.g. time

Quantile Loss fucntion

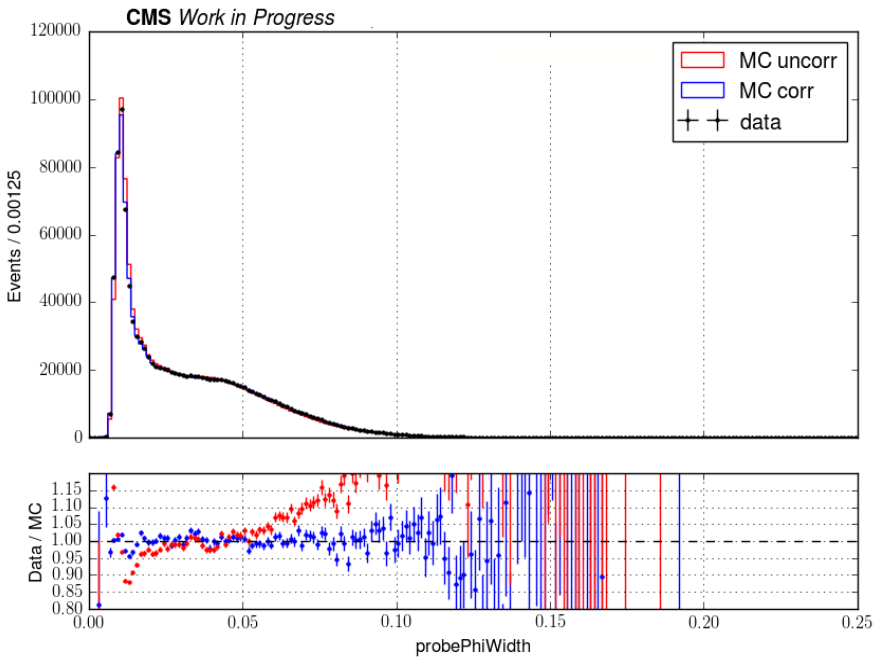
$$L(y_i, q_t(x_i)) = \sum_i \rho_\tau (y_i - q_t(x_i))$$



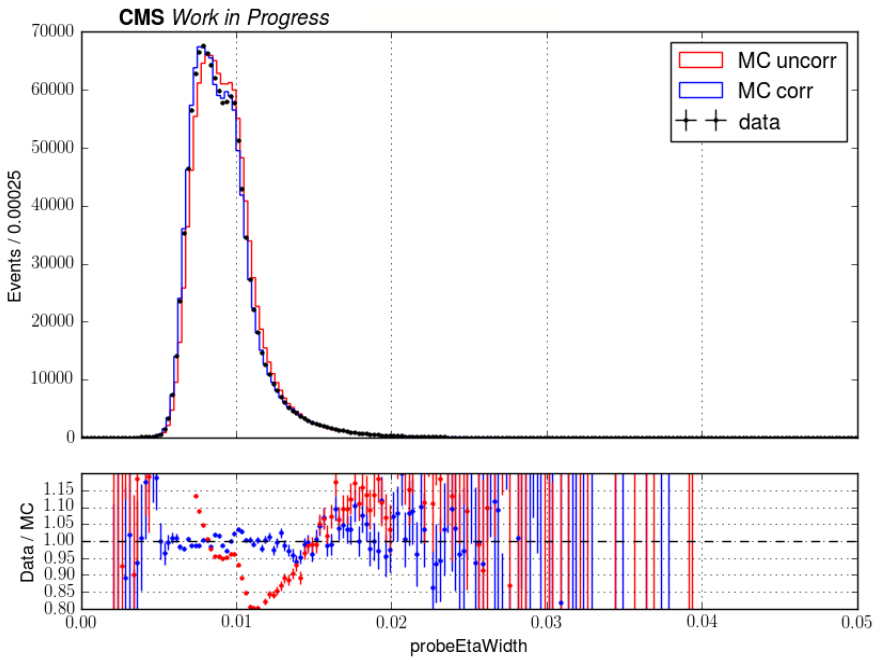
Detector Simulation Tuning
Quantile Regression

Some Results

σ_ϕ



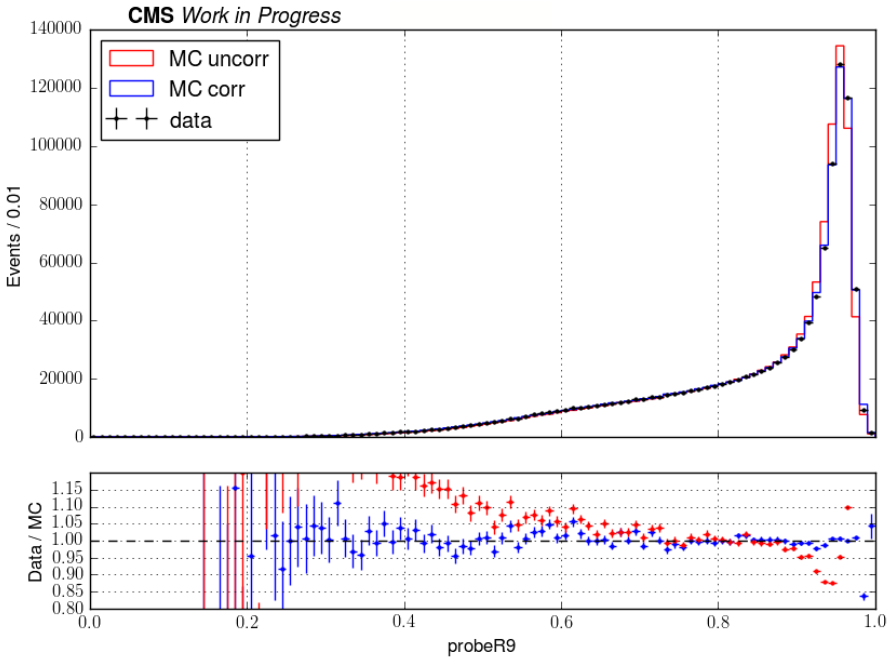
σ_η



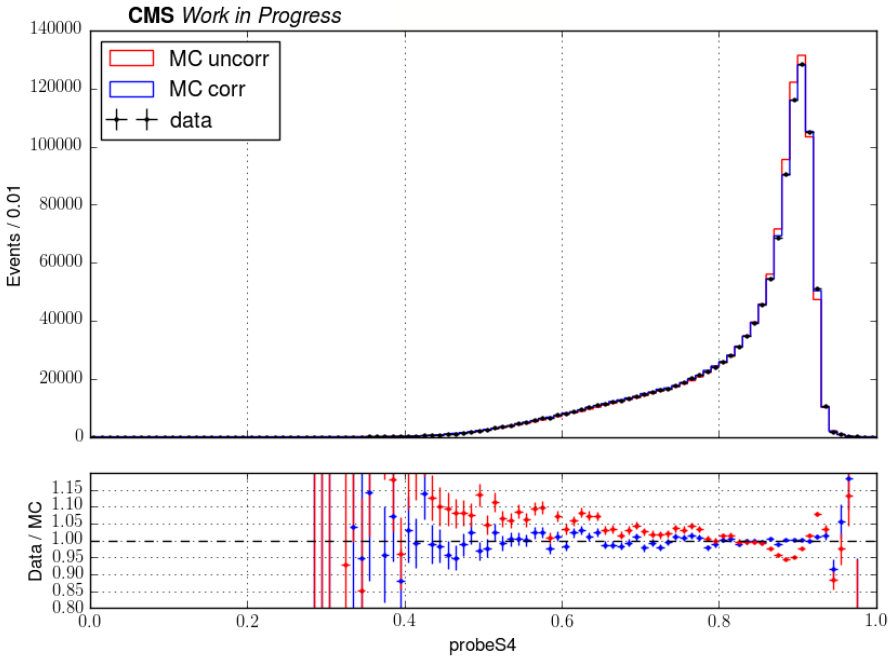
Detector Simulation Tuning
Quantile Regression

Some Results

R_9



S_4



Detector Simulation Tuning

Quantile Regression

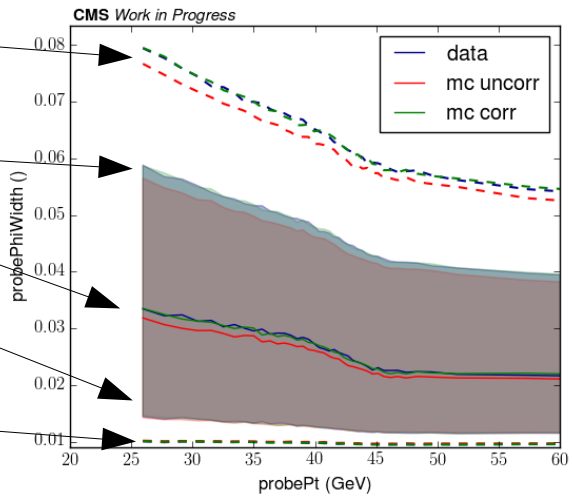
90%

75%

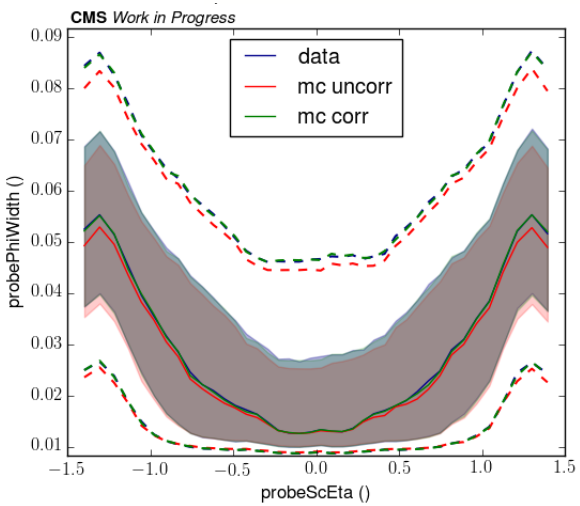
50%

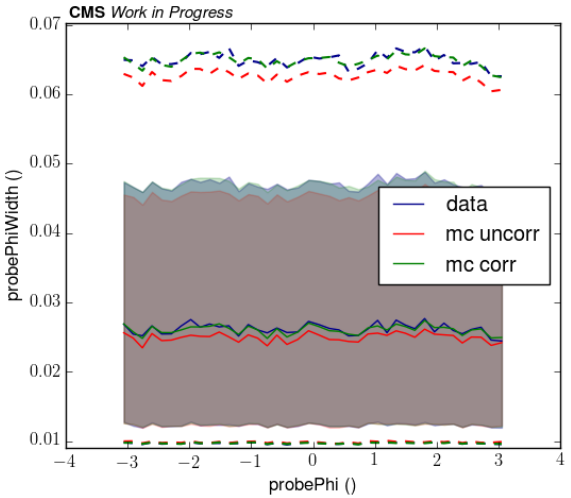
25%

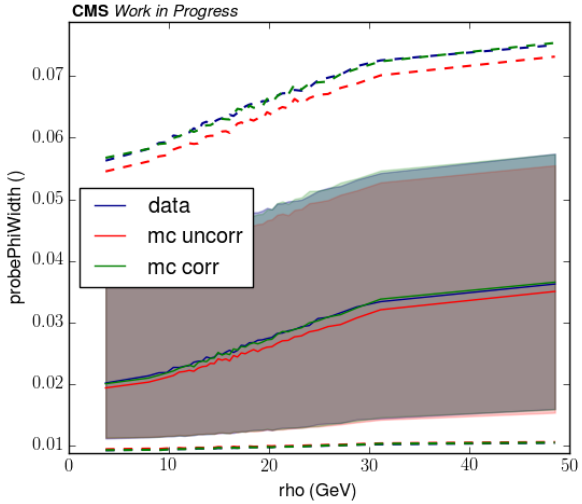
10%



σ_ϕ



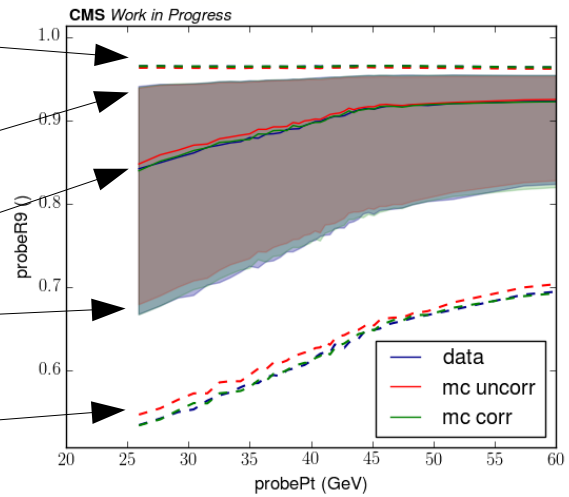




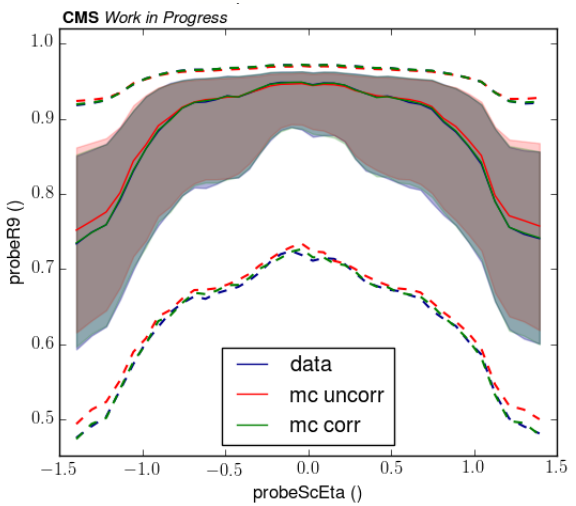
Detector Simulation Tuning

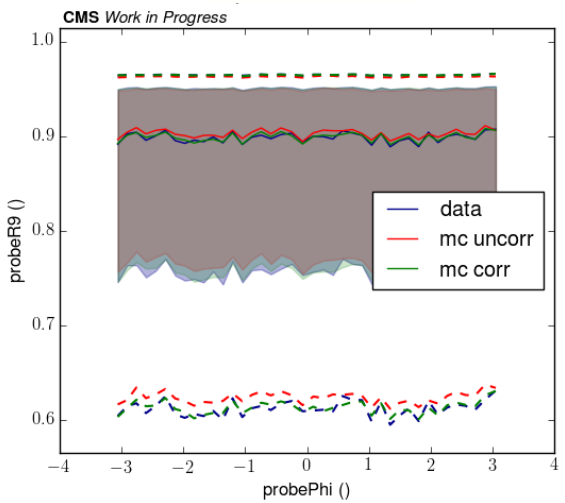
Quantile Regression

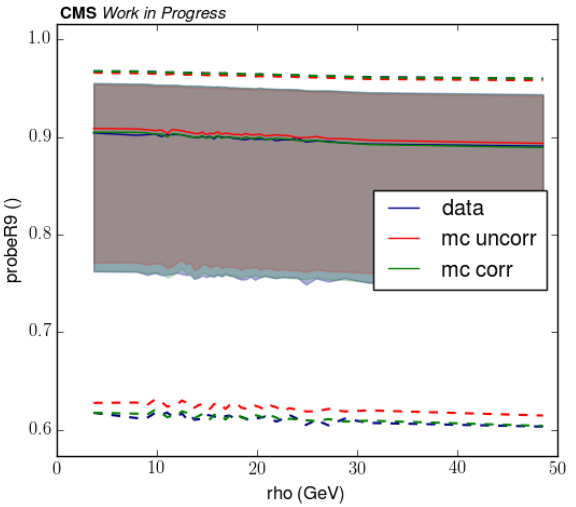
90%
75%
50%
25%
10%



R_9







Conclusion

Introduced two methods to tune detector simulation

Differentially tune simulation using machine learning techniques

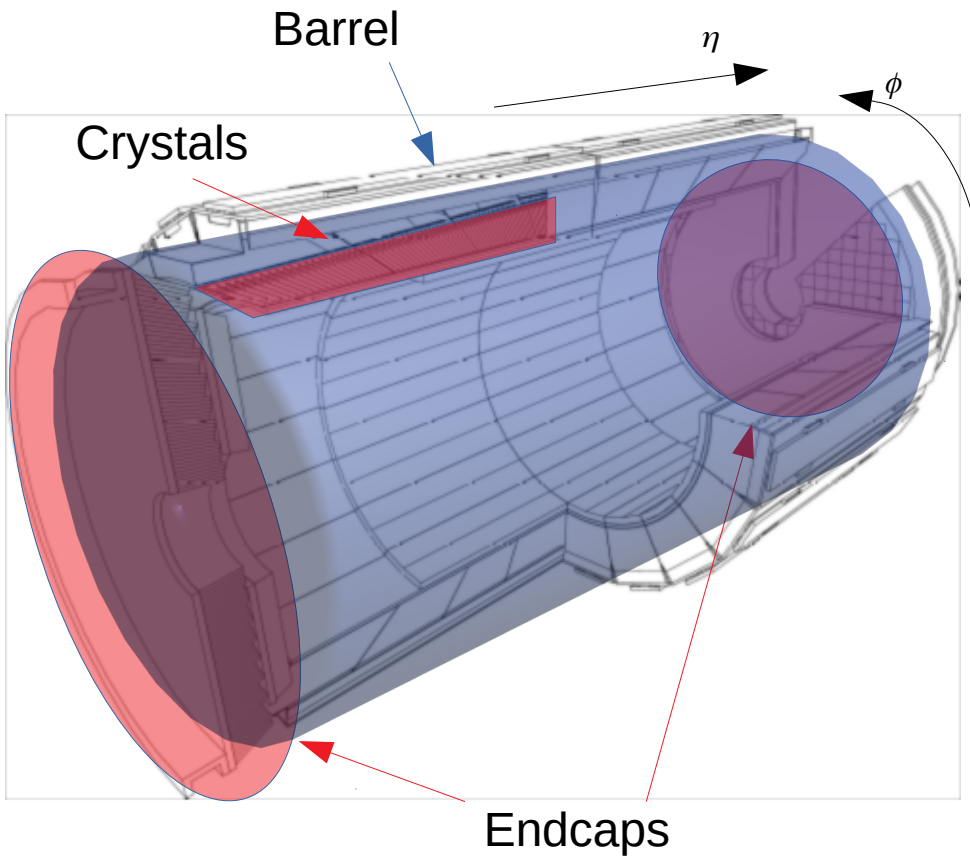
- Based on quantile regression applying boosted decision trees
- Ability to fine-tune simulation dependent on kinematics and event-energy-density
- Method is scalable to more input variables
- Method can be easily applied elsewhere

Tune discontinuous isolation variables

- Applying stochastic techniques to account for discontinuity caused by detector properties
- Differential in pseudorapidity and event-energy-density
- Method can track non-trivial detector effects very well

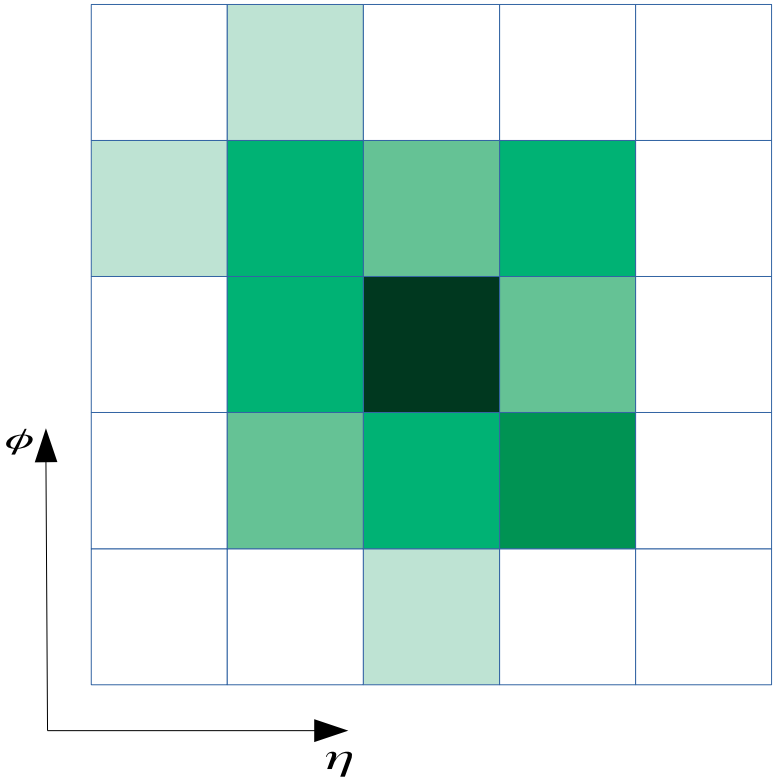
Backup

CMS ECAL



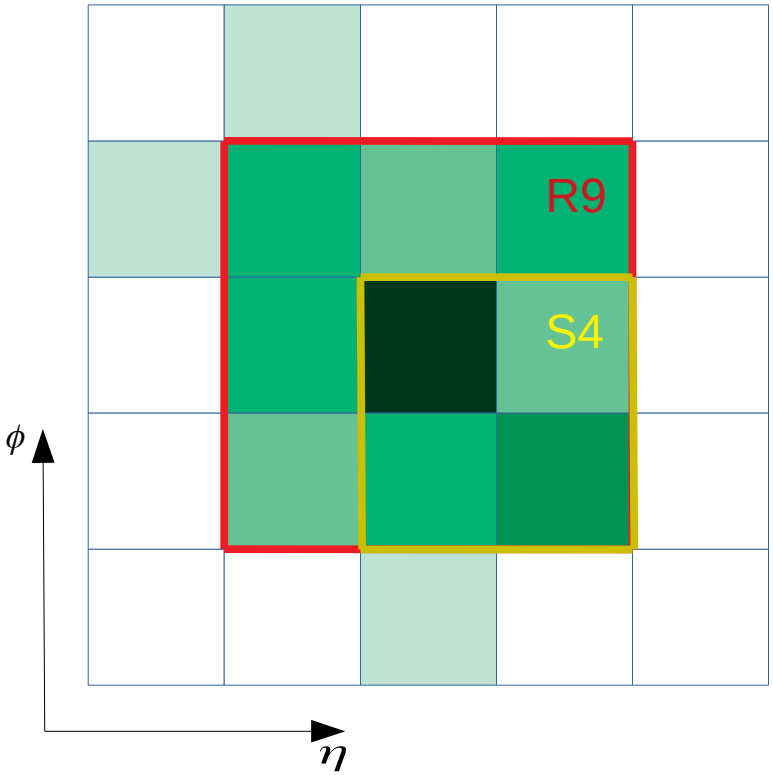
- CMS ECAL is a homogeneous calorimeter
- Barrel consists of 36 supermodules with 1700 crystals each, 61200 in total
- Edcaps consist of 4 “Dee’s” with 3662 scitillaring PbWO4 crystals each, 14648 in total
- Mounted inside the 3.8T Magnet
- The crystals are alligned quasi-projectively

ECAL SuperCluster



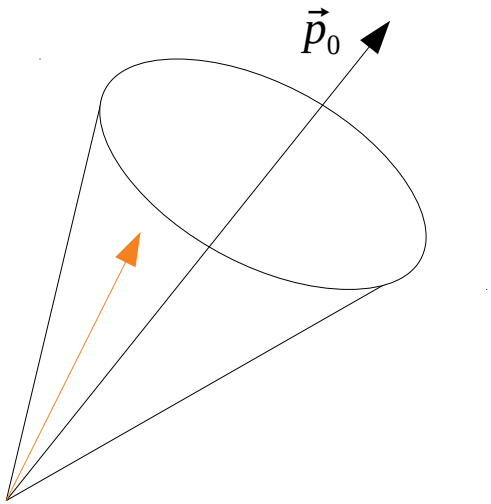
- For unconverted photons, a supercluster results to be formed of the 5x5 crystals centered around the crystal with the highest transverse energy deposit
- More complicated for converted photons, including more 5x5 blocks in ϕ direction
- In endcaps, 5x5 blocks can overlap

ECAL SuperCluster Variables



- Six Variables defined in the SuperCluster are used in photon identification:
  - $R_9 = \frac{\sum_{3x3} E_i}{\sum_{5x5} E_i}$
  - $S_4 = \frac{\sum_{2x2} E_i}{\sum_{5x5} E_i}$
  - $\sigma_\eta = \sqrt{\sum_{SC} (\bar{\eta} - \eta_i)^2 E_i}$
  - $\sigma_\phi = \sqrt{\sum_{SC} (\bar{\phi} - \phi_i)^2 E_i}$
  - $\sigma_{i\eta i\eta} = \sqrt{\sum_{5x5} (\bar{\eta} - \eta_i)^2 w_i}$
  - $\sigma_{i\eta i\phi} = \sqrt{\sum_{5x5} (\bar{\eta} - \eta_i)(\bar{\phi} - \phi_i) E_i}$

Isolation Variables



- Isolation defined as amount of transverse momentum in cone with ΔR around the reconstructed photon coming from other photons, charged or neutral objects
- Isolation variables are independent of SuperCluster
- Photon isolation and charged isolation are used in photon identification
- Isolation varibales are discontinous since only hits over energy threshold are recorded