CMS DN-11-000

2012/07/31

Improving the Identification of Electromagnetic Showers in the CMS Forward Hadron Calorimeter at the LHC

C. Frye, K. Klapoetke, and J. Mans
University of Minnesota, Minneapolis, United States of America

Abstract

The CMS Forward Hadron Calorimeter (HF) lies in a pseudorapidity region which is not covered by the inner tracking system, and we can rely only on the shapes of showers that hit the detector to determine whether they are due to electromagnetic particles or jets. We review the current method of distinguishing these two types of showers in the HF, and we mention a drawback that will become present as the luminosity of the LHC increases and creates a need for tighter shower-shape cuts. We provide a method to correct this drawback, and we analyze the effectiveness of various tight cuts at isolating signal from background. We utilize data from proton-proton collisions collected at CMS as well as Monte Carlo simulations to aid in the analysis. We also comment on the agreement of summer 2011 simulation, which utilized HF GFlash, and 2011 data.

1 Introduction

We present improvements to current methods of distinguishing electromagnetic showers and jets in data collected using the Forward Hadron Calorimeter (HF) of the Compact Muon Solenoid (CMS) detector at the Large Hadron Collider. Current methods work well for the present LHC setup, but as upgrades increase beam luminosity, an increased amount of pileup (i.e., uninteresting secondary interactions) will require the use of tighter cuts to isolate electromagnetic showers from the background. These tighter cuts will expose flaws in our current methods, including a cut effectiveness that depends on shower energy. We correct this drawback and consider techniques to maximize the effectiveness of tighter shower-shape cuts.

The ability to better reconstruct events in the HF will increase our accuracy in measurements such as $Z \to e^+e^-$ rapidity and forward-backward asymmetry, in detection of events such as $W^+ \to e^+\nu$, and in the search for $H \to ZZ^{(*)} \to e^+e^-\ell^+\ell^-$.

In [1], we presented our first results on electromagnetic identification in the HF using early 2010 proton-proton collision data from CMS and Monte Carlo produced in spring, 2010. Here we provide an update on those efforts, showing improvements made to the methods described there. We use 2011 data from proton-proton collisions that occurred at center-of-mass energies of 7 TeV and Monte Carlo produced in the summer of 2011 in this analysis. This is the first set of Monte Carlo produced with the HF GFlash upgrade, and we compare this simulation to the data to evaluate its integrity.

We begin by briefly describing the design of the HF and move on to give more detail about the samples on which we performed the analysis presented here. We then give an overview of the HF event reconstruction process developed in [1]. Next, we explain the methods by which we currently identify electromagnetic showers in the HF, and we present our main results concerning improvements to these methods.

2 The Forward Hadron Calorimeter

The rate of particle production as a function of pseudorapidity in CMS is roughly constant. However, the surface area of the detector per unit pseudorapidity, $|dA/d\eta|$, decreases considerably at high values of $|\eta|$. Also, average energy per particle increases as $|\eta|$ becomes large. The HF, located in the region $3 < |\eta| < 5$, thus experiences showers with high energy density during collisions. Because of the high dose of radiation it receives, the HF was built from more radiation-resistant materials than other parts of the hadronic calorimeter in CMS.

The HF consists of steel absorber plates containing quartz fibers which run parallel to the beamline and relies on Cerenkov light produced in these fibers to detect showers. (See Figure 1 for the layout of the Forward Hadron Calorimeter.) The quartz fibers are of two different lengths. Long fibers extend 1.65 meters from the rear of the HF to its inner face, while short fibers extend 1.43 meters from the rear and leave a 0.22 meter gap filled with steel between their ends and the inner face of the detector.

Each of the two HF calorimeters (one for positive η, one for negative η) is divided into 18 wedges, each containing 24 towers (see Figure 2). Each tower (also referred to as a cell below) is a "square" of side 0.171 in $\eta\phi$-space, except the towers in the innermost and outermost rings of the detector, as can be seen in the figure. The outermost cells have a shorter η dimension because, being in the shadow of the end-cap calorimeters, they serve only to measure the transverse shower leakage from the rest of the HF.

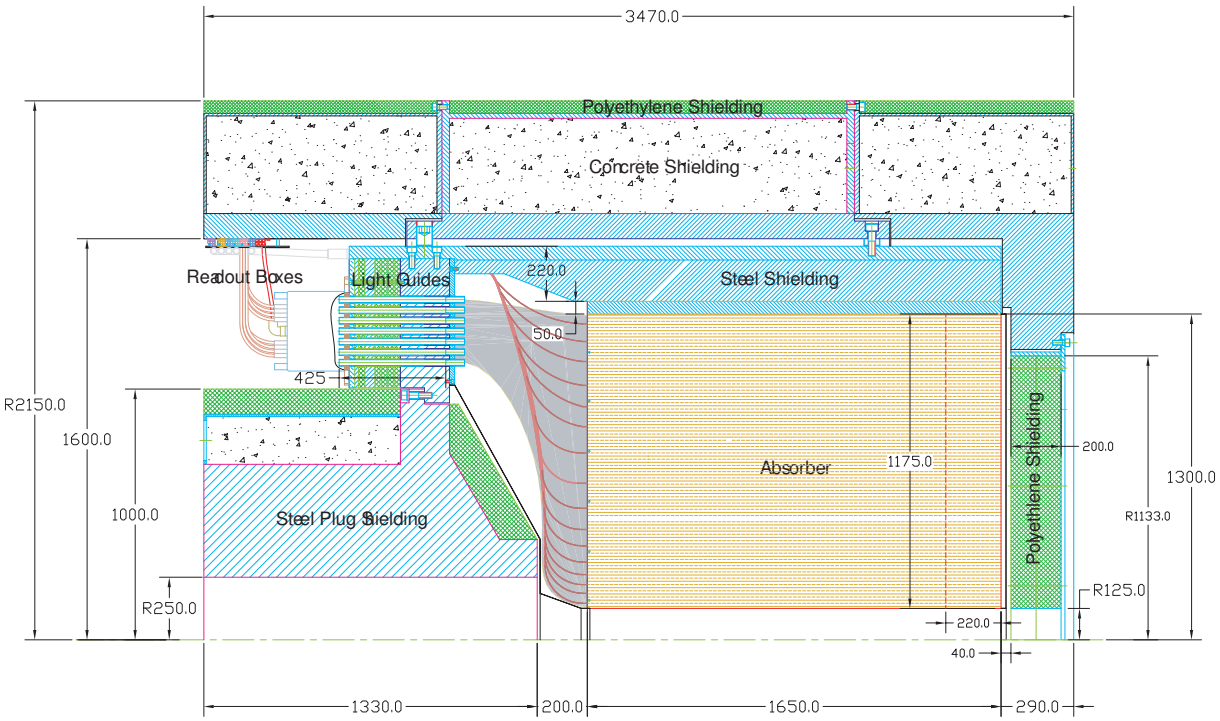


Figure 1: Cross-sectional view of the HF, where r increases vertically and z increases to the left (i.e., showers enter from the right). Steel absorber plates and other shielding is shown. Photomultipliers are located in the readout boxes. The vertical dashed red line marks the end of the short fibers.

Each HF tower connects to two photomultiplier tubes (PMTs). One PMT reads the energy absorbed by the long fibers of that tower while the other PMT is associated with the short fibers. In Section 4 we provide a discussion of hit reconstruction beginning at the stage of the PMTs. For documentation of the performance of the HF detector in test beam runs, see [2].

3 Data and Monte Carlo Samples

We use 1.01 fb^{-1} of `DoubleElectron` data from `Run2011A`, specifically,

- `Run2011A/DoubleElectron/RECO/05Jul2011ReReco-HF`.

In our analysis, we predominantly use events that passed the high-level trigger,

- `HLT_Ele17_CaloIdL_CaloIsoVL_Ele15_HFL`,

(referred to as the HLT below), which requires a loose 17 GeV electron to be received in the Electron Calorimeter (ECAL) and a loose[1] 15 GeV electron to be received in the HF.

To obtain signal from this data, we apply a skim, which requires the ECAL electron to pass WP80 requirements (see [3]) and the event to have an invariant mass in the Z window, 70-120 GeV/c^2. We subsequently require the signal to pass a selection which applies a series of loose cuts (similar to those in the skim) on the shower-shape variables of the HF electron introduced in Subsection 5.1 as well as further cuts on the ECAL electron. We will refer to this sample as

[1] A "loose" electron here is one that passed the lateral containment cut $E_{9/25} \geq 0.92$, see equation (2), and the 2d cut in equation (3) with $C_{2d} = 0.2$.

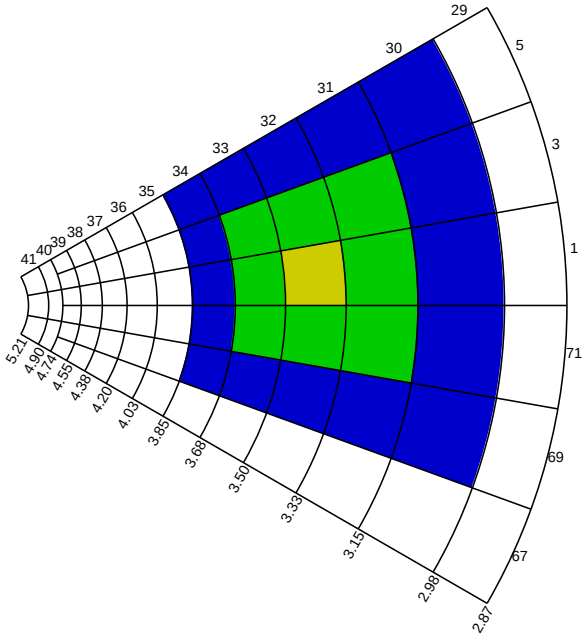


Figure 2: Three adjacent wedges of HF, each spanning $\Delta\phi = 20^\circ$, shown in $r\phi$-space. Indices on the upper and outer edges are called i_η and i_ϕ, respectively. The values along the lower edge denote the η-values of the cell boundaries. As introduced in Section 4.3, the shading depicts the seed cell (in yellow), 3×3 region (green + yellow), and 5×5 region (blue + green + yellow) of an electromagnetic cluster.

"Data: Signal" below in the figures.

We get background from this data by taking all events that fail the selection mentioned above. We also apply a cut on the invariant mass of background events, requiring 70 GeV/c^2 $< m <$ 120 GeV/c^2. We refer to this sample as "Data: Background" below. By imposing these requirements, we use background similar to the pileup that will be difficult to reject as the LHC luminosity increases.

We also use simulated signal Monte Carlo from `Summer11`, specifically,

- `Summer11/DYToEE_M-20_TuneZ2_7TeV-pythia6/GEN-SIM-RECO/PU_S3_START42_V11-v2`.

This is the first set of CMS Monte Carlo that was produced using HF GFlash, which creates simulation up to ten thousand times faster than Geant4 which was previously used. We require this to pass the same selection imposed on "Data: Signal," and in the figures we call this sample "Monte Carlo: Signal."

In one of the studies below, we compare the performance of summer Monte Carlo to that produced in the spring. We use

- `Spring11/DYToEE_M-20_CT10_TuneZ2_7TeV-powheg-pythia/GEN-SIM-RECO/PU_S1_START311_V1G1-v1`.

We subject the spring simulation to the same selections imposed on summer simulation, and we will refer to this sample below as "Spring MC: Signal."

4 Reconstruction Algorithms for HF Hits

We have made no change to the reconstruction algorithms since the release of [1], and our methods are described there in detail. We give a summary of that information here for completeness.

4.1 Local Reconstruction

Photomultiplier tubes are used to convert the Cerenkov light produced in the quartz fibers of the HF into electric charge, which is subsequently integrated and digitized by the Charge Integrator Encoder Application-specific Integrated Circuit (QIE ASIC). Separate pedestal runs allow for correction of the pedestal offset. Then the following gain factors are applied to the received charge: a base gain determined using test-beam data and a response correction determined using measurements of ϕ-symmetry from collision data. Early collision data was also used to align the integration timing of the HF.

4.2 PMT Hit Rejection

One type of HF background is created when particles (usually muons) pass through HF and directly into a photomultiplier tube, generating Cerenkov light and causing a fake signal. The HCAL DPG has developed several variables that can be used to remove PMT hits as well as other types of anomalous background. These techniques are documented in [4], and we use the "Version 6 Cleaning" algorithm described there in our reconstruction. This algorithm was developed to provide strong anomalous hit rejection while still retaining signals from isolated photons.

When a PMT hit occurs in the same location as a jet, however, the combination could pass electron identification cuts; this type of PMT hit will not be removed by the above algorithm. This usually occurs in low-η regions of the detector and would likely be eliminated in standard background removal for any further physics search.

4.3 Clustering the Remaining Hits

So that noise from the electronics does not contribute to the data, we only accept hits with $E > 4\,\text{GeV}$ in the long fibers, which corresponds roughly to $E_T > 400\,\text{MeV}$ (long fibers) at $|\eta| = 3$.

With the HF hits that pass this energy requirement and survive the PMT hit rejection described above, we form clusters as follows. We declare HF hits with $E_T > 5\,\text{GeV}$ to be seed candidates. We consider all hits meeting this criterion in order of decreasing E_T, removing as candidates those cells that lie within the 5×5 tower matrix surrounding a seed candidate with higher E_T. In this way, we avoid overlapping clusters.

For each of the surviving clusters, we define the raw cluster energy to be

$$E_{\text{raw}} = \sum_{i \in 3\times 3} L_i,$$

where L_i is the energy absorbed by the long fibers in the i-th tower of the cluster matrix. We use the 3×3 matrix here because of its complete containment of electromagnetic showers. Even

though the transverse size of electromagnetic showers is actually even smaller than one tower, we cannot use 1×1 or 1×2 matrices to measure total shower energy because of hits near a seed cell edge or corner. We define the core of a cluster to be the seed cell along with its highest-energy-absorbing neighbor, where this neighbor must absorb at least half as much energy as the seed cell. If no such neighbor exists, the cluster core reduces to the seed cell alone.

The (η, ϕ)-position of a hit is calculated as the mean of the (η, ϕ)-coordinates of the tower centers in the cluster, each weighted by the natural logarithm of the tower's absorbed energy; specifically,

$$\eta_{\text{raw}} = \frac{\sum_{i \in 3\times 3} \left[\log(L_i \,/\, 1\,\text{GeV})\right] \eta_i}{\sum_{i \in 3\times 3} \log(L_i \,/\, 1\,\text{GeV})},$$

$$\phi_{\text{raw}} = \frac{\sum_{i \in 3\times 3} \left[\log(L_i \,/\, 1\,\text{GeV})\right] \phi_i}{\sum_{i \in 3\times 3} \log(L_i \,/\, 1\,\text{GeV})},$$

where (η_i, ϕ_i) are the coordinates of the center of the i-th tower in the cluster. We weight the cells in this way because the logarithm of the energy absorbed in a hit is proportional to the penetration depth of the corresponding hit.

4.4 Position Corrections

The coordinates (η_i, ϕ_i) in the previous section give the location of the inner face of the i-th tower in a cluster. Since the HF towers are parallel to the beam-line, we will be slightly off in determining the position of a hit when the particle enters the tower some distance behind its inner face. Because of this, η tends to be underestimated. Although this effect is dependent on shower depth and thus varies with energy, we use a constant additive correction C_{bias} which is sufficient to remove this η-bias over the energy range typical in the HF.

A second type of position bias occurs as a result of our inability to determine the location of a hit within a tower. Since, as mentioned before, the transverse size of a tower exceeds that of an electromagnetic shower, systematic error develops when we calculate η_{raw} and ϕ_{raw}. In studying this bias, we introduce variables to quantify relative location inside a cell,

$$\eta_{\text{cell}} = \frac{\eta_{\text{raw}} - |\eta|_{\text{min}}}{|\eta|_{\text{max}} - |\eta|_{\text{min}}},$$
$$\phi_{\text{cell}} = \frac{\phi_{\text{raw}} - \phi_{\text{min}}}{\phi_{\text{max}} - \phi_{\text{min}}},$$

where $|\eta|_{\text{min}}, \phi_{\text{min}}$ (respectively $|\eta|_{\text{max}}, \phi_{\text{max}}$) are the smallest (respectively largest) $|\eta|, \phi$ values found in the seed cell of a cluster. In [1] we showed that the difference between Monte Carlo truth and η_{raw} (respectively ϕ_{raw}) depends sinusoidally on η_{cell} (respectively ϕ_{cell}).

Comparing data to Monte Carlo truth, we use the following position corrections (see [1] for further analysis of their impact). For reconstructed η we use

$$\eta_{\text{reco}} = \eta_{\text{raw}} + \text{sgn}(\eta_{\text{raw}}) \cdot [A_\eta \sin(2\pi\eta_{\text{cell}}) + C_{\text{bias}}],$$
$$A_\eta = 0.00683 \pm 0.00013,$$
$$C_{\text{bias}} = 0.00938 \pm 0.00009,$$

and for reconstructed ϕ we use

$$\phi_{\text{reco}} = \phi_{\text{raw}} + A_\phi \sin(2\pi\phi_{\text{cell}}),$$
$$A_\phi = 0.00644 \pm 0.00010.$$

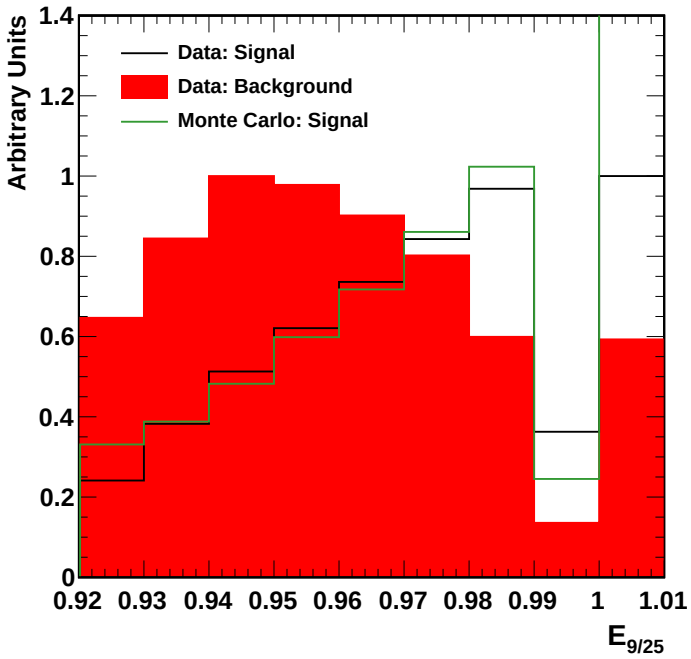


Figure 3: Histogram depicting $E_{9/25}$ for data and simulation. We see reasonable agreement between data and simulation, disregarding the rightmost bin. The Monte Carlo sample has many more events with $E_{9/25} = 1$ than does "Data: Signal" because we have not re-weighted the simulation, which was produced with less-pileup.

4.5 Energy Corrections

Since we only use energy absorbed by the long fibers in calculating E_{raw}, we must scale this value to match the true energy of Monte Carlo hits,

$$E_{\text{reco}} = \omega_{i_\eta} G E_{\text{raw}}, \qquad (1)$$
$$G = 1.352 \pm 0.001,$$

where G is a global correction factor obtained specifically from $Z \to e^+e^-$ Monte Carlo simulations. This factor makes up for our neglect of the energy absorbed by the short fibers and also corrects for energy leakage effects. The factor ω_{i_η} depends on η and removes the effects of dead material in the front of the HF. For small values of η, ω_{i_η} is significantly different from unity. (See [1] for further analysis of these corrections.)

5 Identification of Electromagnetic Showers

5.1 Identification Variables

We rely on the way that different shower types spread out laterally as they move through CMS as well as how deeply they penetrate into the HF detector to distinguish electromagnetic and hadronic hits. In analyzing these characteristics of the HF events, we introduce variables that quantify the lateral containment of showers as well as their transverse and longitudinal shape.

To measure lateral containment, we use the variable $E_{9/25}$ which is defined as the ratio of the total energy absorbed (by the long and short fibers) in the 3×3 matrix surrounding the seed

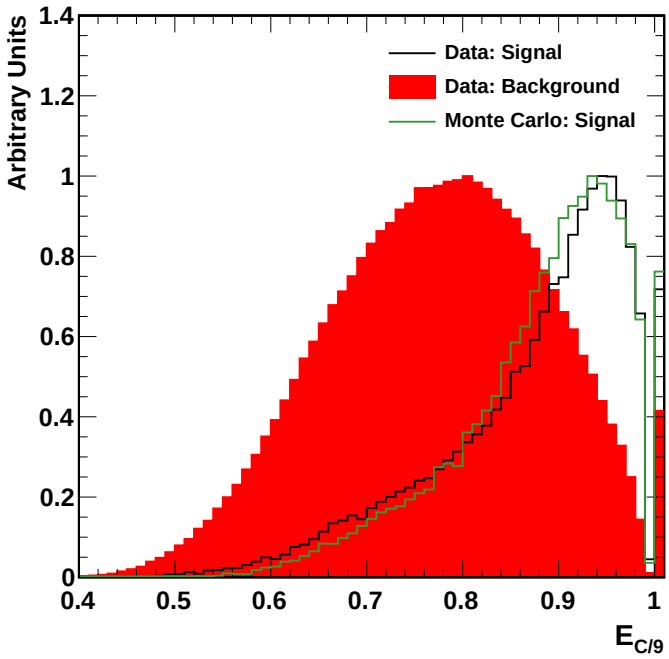


Figure 4: Histogram depicting $E_{C/9}$ for data and simulation. Signal and background differ a good deal even after both were required to pass the HLT. Real signal and simulated signal match quite well.

cell to that absorbed in the 5×5 matrix, i.e.,

$$E_{9/25} = \frac{\sum_{i \in 3\times3}(L_i + S_i)}{\sum_{i \in 5\times5}(L_i + S_i)} \,. \tag{2}$$

Electromagnetic showers, which are composed of lighter particles than the hadrons in jets, remain laterally dense as they propagate through CMS while jets tend to spread out. Because of this, electromagnetic showers are virtually completely contained in the 3×3 matrix surrounding the seed cell and have an $E_{9/25}$ value of nearly unity; conversely, the value of $E_{9/25}$ for jets is appreciably lower. In current methods of signal isolation [1], we use a cut $E_{9/25} > 0.94$, which has very high signal efficiency, to begin separating signal from background.

To quantify transverse shape, we use $E_{C/9}$ which is defined as the ratio of the energy absorbed by the long fibers in the core matrix to that absorbed by the long fibers in the 3×3 matrix of the cluster, i.e.,

$$E_{C/9} = \frac{\sum_{i \in \text{core}} L_i}{\sum_{i \in 3\times3} L_i} \,.$$

For the same reasons mentioned above, the value of $E_{C/9}$ tends towards unity for electromagnetic showers and is significantly lower for jets (see Figure 4).

Finally, we measure the longitudinal shower shape with $E_{S/L}$, defined as the ratio of the energy absorbed by the short fibers in the 3×3 matrix of the cluster to that absorbed by the corresponding long fibers, i.e.,

$$E_{S/L} = \frac{\sum_{i \in 3\times3} S_i}{\sum_{i \in 3\times3} L_i} \,,$$

where S_i is the energy absorbed by the short fibers in the i-th cell of the cluster. As mentioned in the introduction, electromagnetic showers must travel through 12.5 radiation lengths

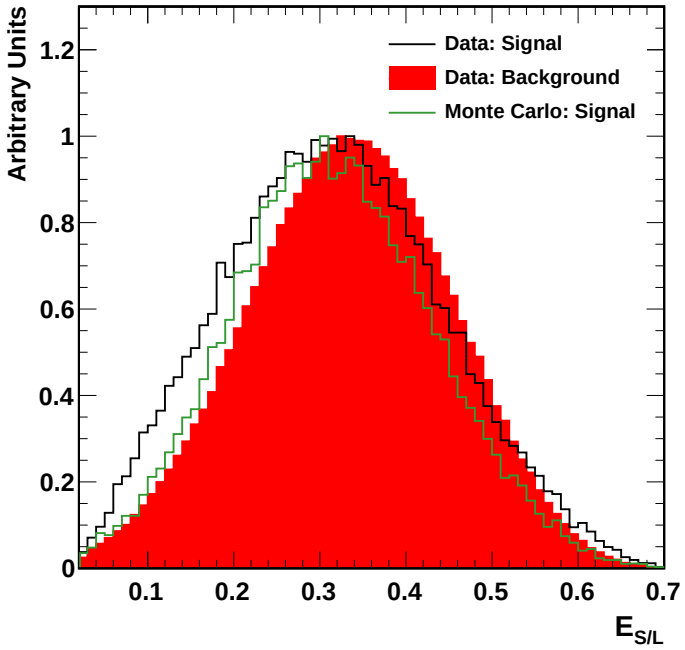


Figure 5: Histogram comparing distributions of $E_{S/L}$ for data and simulation. Signal and background peaks vary only slightly here, as a result of the HLT. Signal from real data and from Monte Carlo match well.

of steel before reaching the short fibers of the HF, while this distance is only about 1.3 interaction lengths for jets. Thus, electromagnetic showers penetrate to relatively shallow depths, giving values of $E_{S/L}$ lower than those for jets (see Figure 5).

We combine the information about the events given by $E_{C/9}$ and $E_{S/L}$ into a two-dimensional cut to further separate signal from background. Currently (see [1]) we require

$$E_{C/9} - 1.125 \cdot E_{S/L} > C_{2d}, \tag{3}$$

for some cut-value C_{2d}, and we use $C_{2d} = 0.2$ in the HLT. On a plot of $E_{C/9}$ vs. $E_{S/L}$, this cut corresponds to drawing a diagonal line and keeping all events that lie above it (see Figure 13a for a related plot). With the inclusion of the coefficient 1.125 (which is the slope of the cut-line) more weight is placed on restricting $E_{S/L}$ than $E_{C/9}$. This pre-selection effect gives a partial explanation for the close similarity of signal and background in Figure 5, which were considerably more distinct in [1].

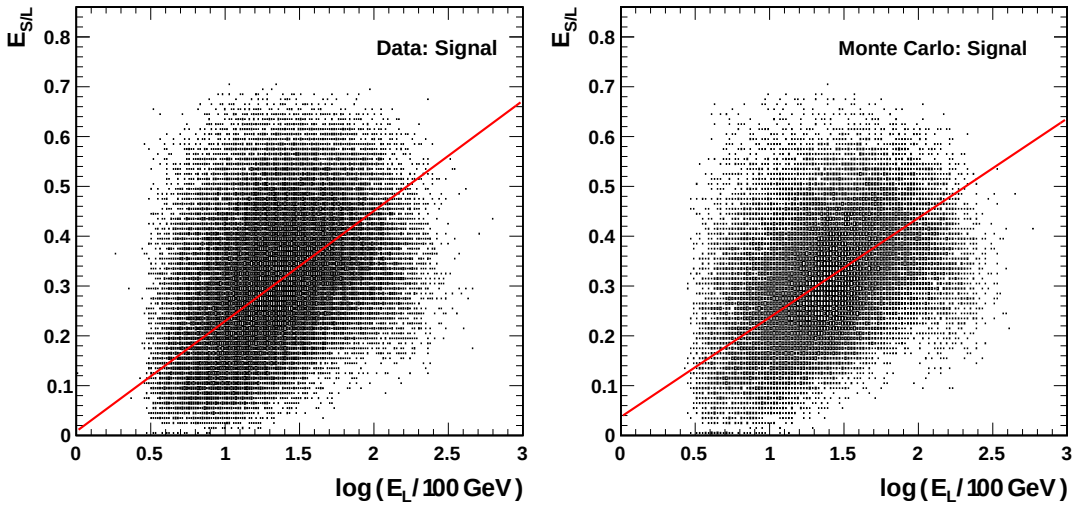
5.2 Transforming $E_{S/L}$

The depth that a shower penetrates into the detector depends on the total energy of the shower. Defining E_L as the energy absorbed in the long fibers of the 3×3 matrix surrounding the seed cell[2],

$$E_L = \sum_{i \in 3\times 3} L_i,$$

we can study this effect. As displayed in equation (1), E_L is proportional to the total energy of the hit. Plotting $E_{S/L}$ vs. $\log(E_L \,/\, 100\,\text{GeV})$ for signal events (see Figure 6) shows that $E_{S/L}$

[2]This value was called E_{raw} in Subsection 4.3.



(a) (b)

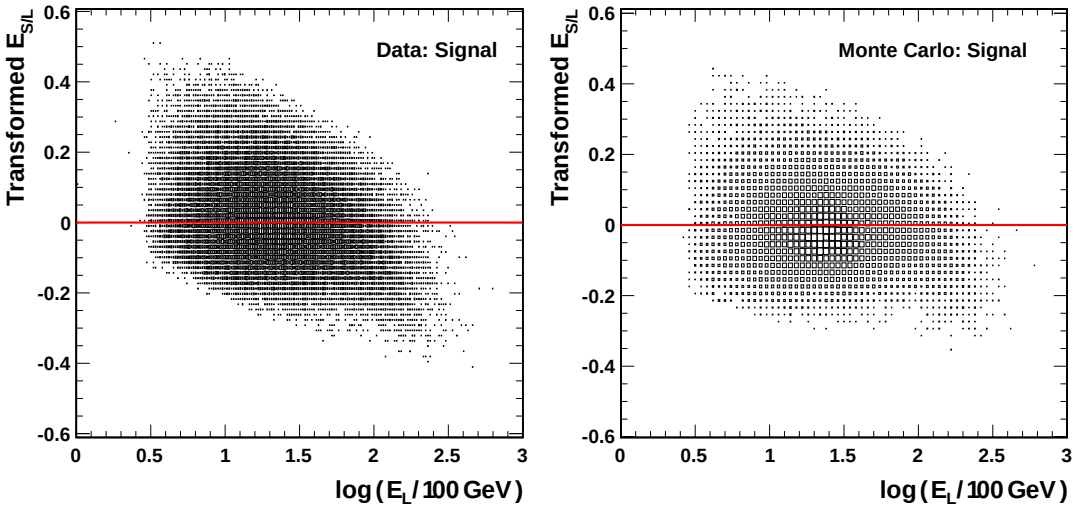
Figure 6: Plot showing $E_{S/L}$ vs. $\log(E_L \,/\, 100\,\text{GeV})$ for (a) real signal and (b) simulated signal. The fit lines were obtained using the method described in the text.

increases with total shower energy for electromagnetic hits. Thus, if we use $E_{S/L}$ to make a cut on a sample, we may remove high-energy electromagnetic events because their penetration depth (measured by $E_{S/L}$) disguised them as jets.

To remove the dependence of $E_{S/L}$ on E_L for signal, we make a transformation that produces a new variable, $E_{S/L}^{\text{cor}}$, which we call "Transformed $E_{S/L}$" in the figures. An algorithm performs the transformation using the $E_{S/L}$ vs. $\log(E_L \,/\, 100\,\text{GeV})$ plot as follows. (In the paragraphs below, we refer to the horizontal axis of this plot as the x-axis and the vertical axis as the y-axis for simplicity.)

First, we fit the points in the plot to a line. To do this we let x_i step bin-by-bin along the horizontal axis of the histogram, in each step considering all events having $x = x_i$ (these make up a vertical strip). We fit the frequency of the events in this strip to a Gaussian function of y which has mean y_i, and we collect these points (x_i, y_i) for each step in the process. We then fit the points (x_i, y_i) to a line, taking into account the errors, thus obtaining two constants m, b which are the slope and y-intercept of the line, respectively. To neglect outliers, we consider only those vertical strips that contain at least 1% of the total entries in the histogram. Second, we choose an arbitrary point x_0 on the horizontal axis, and we translate all points in the plot so that $(x_0, mx_0 + b)$ becomes the origin. Third, we rotate the entries in the plot clockwise by angle $\tan^{-1}(m)$ after which the fit line coincides with the x-axis. We then call the new y-value of each point $E_{S/L}^{\text{cor}}$.

This transformation sends each point (x, y) in the plot to the new point (x, y'), and the trans-



(a) (b)

Figure 7: Plot showing $E_{S/L}^{\text{cor}}$ vs. $\log(E_L \,/\, 100\,\text{GeV})$ for (a) real signal and (b) simulated signal. After the transformation the distributions are flat, and a line fit to the dense regions of events would coincide with the x-axis.

formation can be written $y \mapsto y' = \alpha + \beta x + \gamma y$, where

$$\alpha = \frac{-b}{\sqrt{1+m^2}},$$
$$\beta = \frac{-m}{\sqrt{1+m^2}},$$
$$\gamma = \frac{1}{\sqrt{1+m^2}},$$

and b, m are the parameters of the fit line mentioned above. This is equivalent to sending $(x, y) \mapsto (x, y')$ where y' is the perpendicular distance from the point (x, y) to the line $y = mx + b$.

The fit lines are shown in Figure 6, and we compare all the constants involved in this transformation in the tables shown in Figure 8. We attribute the differences in the values of b, α for data and simulation to the slight difference in the mean $\log(E_L \,/\, 100\,\text{GeV})$-value for the two samples (shown as $\langle x \rangle$ in the table), which has the same order of magnitude. (The error in $\langle x \rangle$ originates from the finite bin width in the histogram.) Also, in the tables we include the transformation constants for "Spring MC: Signal," in order to show the improvement made by summer Monte Carlo in simulating the data. When we plot $E_{S/L}^{\text{cor}}$ vs. $\log(E_L \,/\, 100\,\text{GeV})$ for signal (see Figure 7), we see that the entries are "flat," i.e., that the dependence of the longitudinal variable on the total energy of the corresponding shower is removed. We transform "Data: Background" with the same constants obtained for "Data: Signal." We expect the background events to be located farther from the fit line than the signal events; hence, values of $E_{S/L}^{\text{cor}}$ should be greater for background than for signal, which will allow us to make a useful cut with this variable.

	b	m	$\langle x \rangle$
Data: Signal	0.008 ± 0.0042	0.221 ± 0.0030	1.34 ± 0.00084
Monte Carlo: Signal	0.037 ± 0.0032	0.200 ± 0.0022	1.37 ± 0.00084
Spring MC: Signal	-0.062 ± 0.0064	0.242 ± 0.0043	1.41 ± 0.00085

	α	β	γ
Data: Signal	-0.008 ± 0.0041	-0.216 ± 0.0028	0.9764 ± 0.00061
Monte Carlo: Signal	-0.036 ± 0.0031	-0.196 ± 0.0021	0.9806 ± 0.00042
Spring MC: Signal	0.061 ± 0.0063	-0.235 ± 0.0039	0.9719 ± 0.00095

Figure 8: Parameters used for data and simulation in the transformation $E_{S/L} \mapsto E_{S/L}^{\text{cor}}$. We show values for spring Monte Carlo to compare how well spring and summer simulation match the data. Summer Monte Carlo simulates data more accurately than does spring Monte Carlo, as we see from the values of $b, \alpha, \langle x \rangle$.

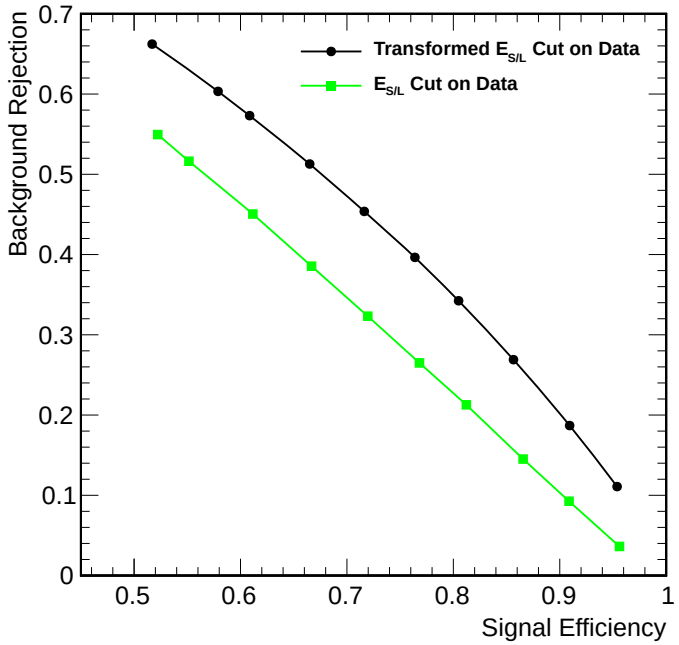


Figure 9: Comparison of the $E_{S/L}$ and $E_{S/L}^{\text{cor}}$ cuts. For each value of signal efficiency, the $E_{S/L}^{\text{cor}}$ gives greater background rejection.

5.3 Improving the Cuts

We performed a simple analysis to compare the effectiveness of the transverse, longitudinal, and two-dimensional shower shape cuts available to us using the variables $E_{C/9}$, $E_{S/L}$, $E_{S/L}^{\text{cor}}$. Since our signal and background have already been required to pass the cuts described in Section 3, the goal of this analysis is to determine where small improvements can be made. Since our goal is to determine the optimal set of cuts to perform to isolate signal from background, we will focus mostly on the relative effects of the different cuts and less on there absolute effectiveness.

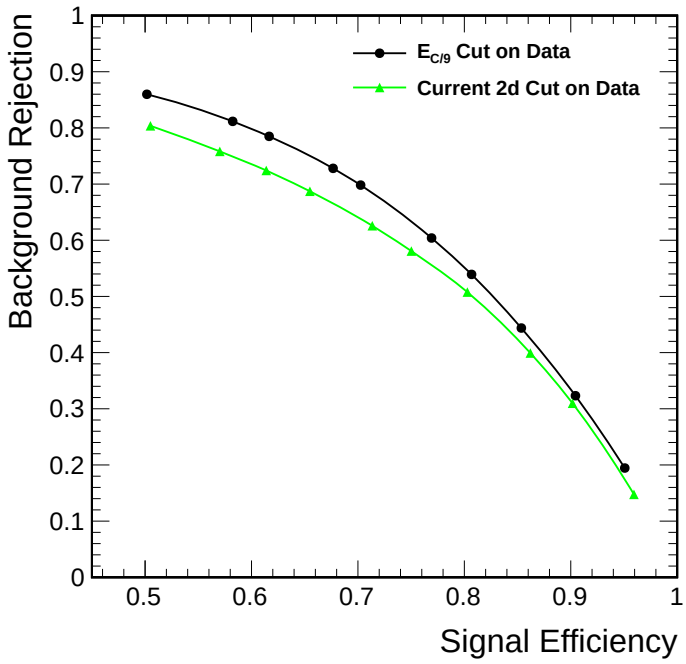


Figure 10: Comparison of the transverse and current two-dimensional cuts. For the tight cuts shown here, $E_{C/9}$ is more effective than the cut in equation (3).

5.3.1 The Longitudinal Cut

First, we evaluate the improvement made by transforming the longitudinal shower shape variable. We consider a cut on the data requiring $E_{S/L} < C_\ell$ and alternatively a cut requiring $E^{\text{cor}}_{S/L} < C'_\ell$, for some cut-values C_ℓ, C'_ℓ. In analyzing these cuts we consider

$$\text{signal efficiency} = \frac{\text{\# of signal events that survive cut}}{\text{total \# of signal events}}$$

and

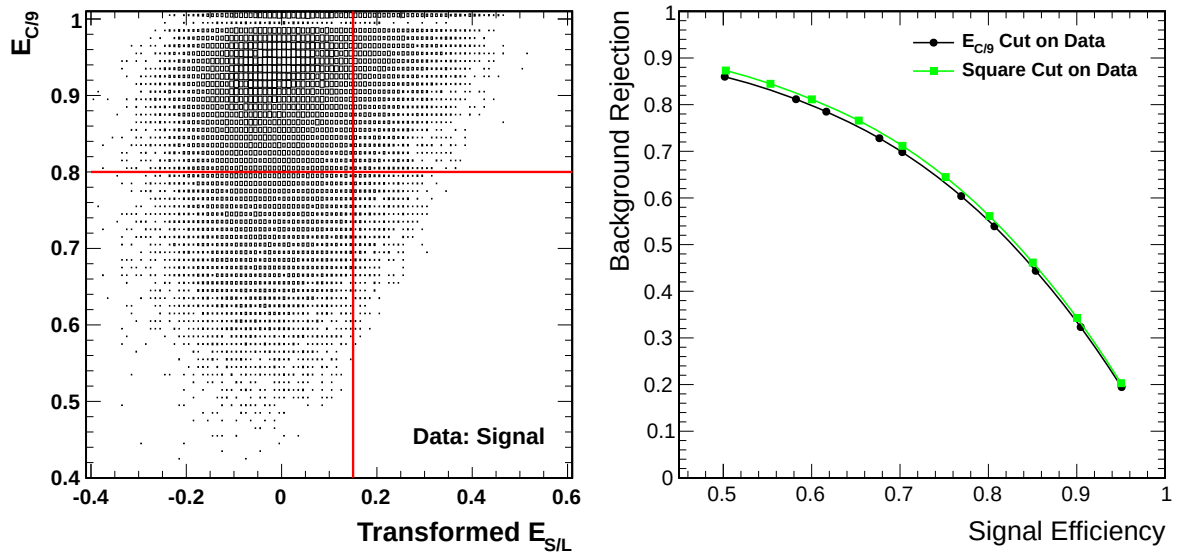
$$\text{background rejection} = \frac{\text{\# of background events that fail cut}}{\text{total \# of background events}}.$$

We employ an algorithm that, using our data set, computes the background rejection as a function of signal efficiency for each of the longitudinal cuts. We compare the $E_{S/L}$ and $E^{\text{cor}}_{S/L}$ cuts by plotting this information in Figure 9. We see in the figure that the $E^{\text{cor}}_{S/L}$ cut performs better than the $E_{S/L}$ cut for any desired signal efficiency. Thus, the $E_{S/L}$ cut should be replaced with the Transformed $E_{S/L}$ cut in all cases.

5.3.2 The Transverse Cut

It is already known that the $E_{C/9}$ cut is quite effective [1]. In fact, for tight selections, the $E_{C/9}$ cut is more effective than the two-dimensional cut in equation (3), which is currently in use. Indeed, we find in Figure 10 that for any given signal efficiency, the transverse cut rejects more background than the current 2d cut.

We now evaluate the benefits of combining this transverse shower shape cut with the longitudinal cut $E^{\text{cor}}_{S/L}$. One way to combine these cuts is to perform them independently in succession, i.e., to require $E_{C/9} > C_t$ and $E^{\text{cor}}_{S/L} < C_\ell$ for some values of C_t, C_ℓ. On a plot of $E_{C/9}$ vs. $E^{\text{cor}}_{S/L}$,



(a) (b)

Figure 11: (a) $E_{C/9}$ vs. $E_{S/L}^{\text{cor}}$ with example square cut-lines for "Data: Signal." (b) Comparison of the square cut and the $E_{C/9}$ cut on data, which turn out to have nearly identical effects.

this corresponds to drawing a vertical line and a horizontal line and keeping only those events in the top-left corner of the plot (see Figure 11a). Hence, we call this the square cut. In using this cut to obtain a certain signal efficiency, there are many combinations of C_t, C_ℓ that will work. We use a simple algorithm to perform the square cut with those values of C_t, C_ℓ that give the desired signal efficiency and maximize the background rejection.

In Figure 11b, we compare the square cut to the $E_{C/9}$ cut by plotting background rejection versus signal efficiency for each of these cuts when performed on the data. We see that the two graphs almost coincide, suggesting that in choosing the optimal values of C_t and C_ℓ the square cut becomes the $E_{C/9}$ cut. In Figure 11a, this corresponds to the vertical line being pushed all the way to the right, making the square cut effectively a one-dimensional cut on $E_{C/9}$. Thus, it is not effective to use the transverse and longitudinal cuts independently, and to simplify coding we should always use the simpler $E_{C/9}$ cut instead of the square cut.

5.3.3 The Two-dimensional Cut

Though $E_{C/9}$ is quite effective on its own, we would like to gain some benefit from our knowledge of $E_{S/L}^{\text{cor}}$. Hence, we introduce an optimized two-dimensional cut in the style of equation (3), where the transverse and longitudinal cuts are not performed independently. We require

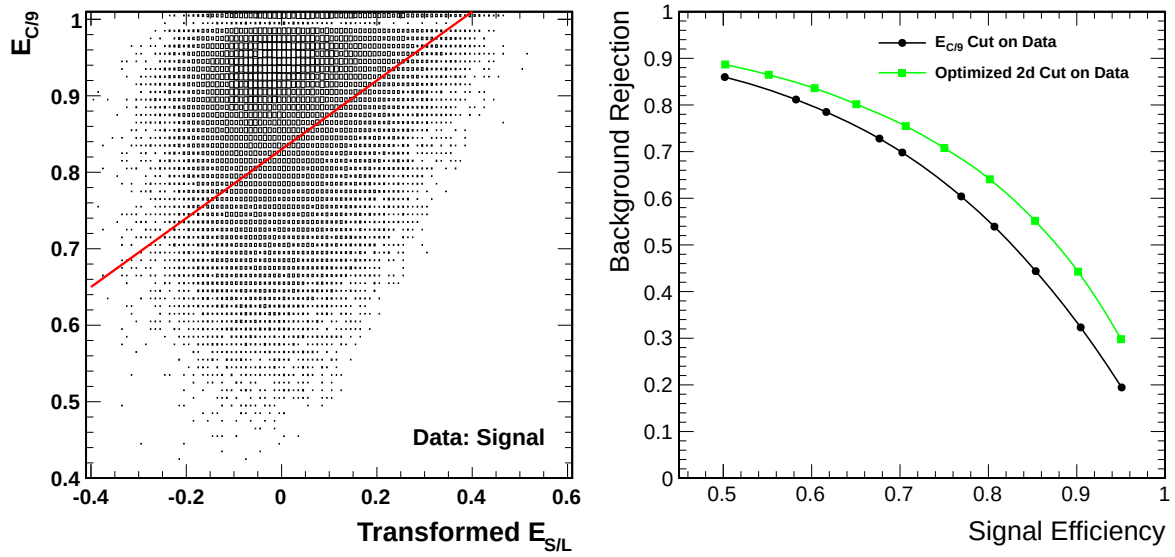
$$E_{C/9} - m \cdot E_{S/L}^{\text{cor}} > C_{2d}$$

for some values of m, C_{2d}. Whereas the square cut corresponded to drawing two perpendicular lines on the plot in Figure 11a and requiring events to be located in the top-left corner, this two-dimensional cut draws one slanted line on the plot (see Figure 13a) with slope m and y-intercept C_{2d} and requires events to be above this line. As in the square cut, we use an algorithm that chooses values of m and C_{2d} which give the desired signal efficiency while optimizing background rejection.

In Figure 13b, we graph background rejection versus signal efficiency to compare this new two-dimensional cut to the $E_{C/9}$ cut. The new two-dimensional cut dominates over the transverse

Efficiency	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%
C_{2d}	0.90	0.88	0.87	0.85	0.83	0.81	0.78	0.74	0.71	0.64
m	0.26	0.30	0.33	0.43	0.35	0.48	0.50	0.65	0.60	0.80

Figure 12: Optimized C_{2d}, m values used in the new two-dimensional cut on data. Step size in the optimization algorithm determines the error on these parameters: $\Delta C_{2d} = 0.01, \Delta m = 0.025$.



(a) (b)

Figure 13: (a) $E_{C/9}$ vs. $E^{\text{cor}}_{S/L}$ with an example 2d cut-line for "Data: Signal." (b) Comparison of the new 2d cut and the $E_{C/9}$ cut on data, where it is clear that the new 2d cut offers an improvement in background rejection.

cut alone for all signal efficiencies considered. Thus, it will create improvement in signal isolation to use the new two-dimensional cut as the cut of choice in distinguishing electromagnetic showers from jets. The optimized parameters C_{2d}, m used in the new two-dimensional cut on data are shown in Figure 12 for each value of signal efficiency.

6 Future Work

We can further investigate the bias of the shower-shape cuts by studying the dependency of the signal efficiency of the $E_{C/9}$ cut on total shower energy. Such a study will tell us whether or not we need to perform a transformation of the transverse shower-shape variable to enhance effectiveness in tighter cuts.

We have begun this investigation using the "Data: Signal" sample. We produced two plots of transverse momentum of the HF electron versus the invariant mass of the double-electron event, one using the entire sample and the other using only those events that passed a cut $E_{C/9} \geq 0.845$. This transverse shower-shape cut keeps roughly 70% of "Data: Signal" events. Since we are interested in pure signal efficiency, we attempted to disregard the background still present in our sample as follows.

In our p_T vs. m_Z plots, we considered intervals Δp_T of transverse momentum and fit the events

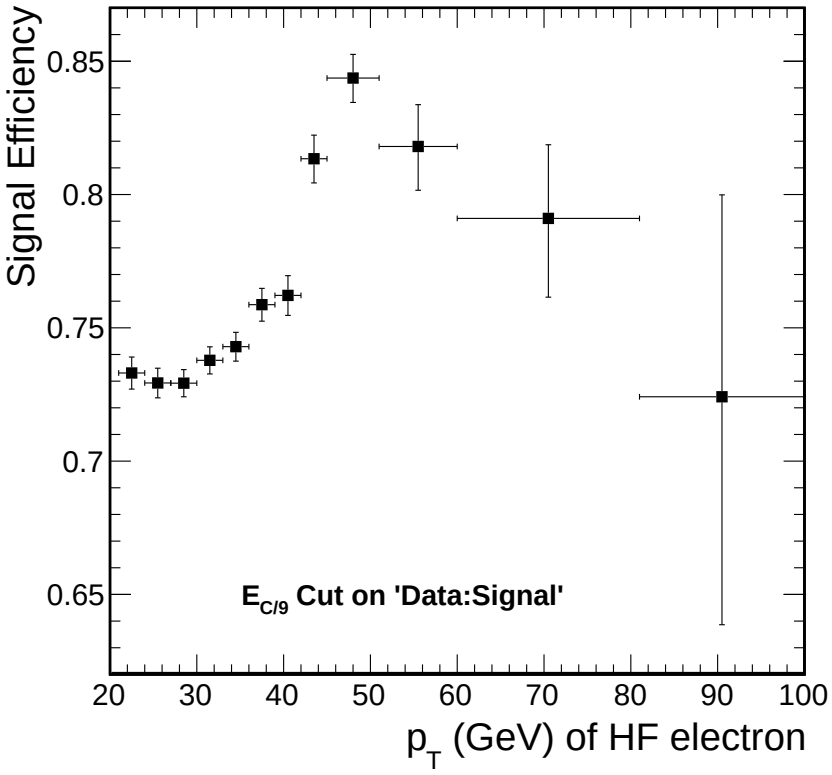


Figure 14: Plot of modified signal efficiency versus transverse momentum of the HF electron. The efficiency depends of p_T, showing that an improvement can be made in the transverse shower-shape variable.

in each range to a function $f(m_Z)$ of invariant mass. We required $f(m_Z)$ to be a Gaussian superimposed on a decaying exponential, where the Gaussian (with its peak in the Z-mass range and its width required to be the width of the Z-mass curve) corresponds to signal and the decaying exponential corresponds to background. We then integrated the Gaussian part of $f(m_Z)$ to determine the number of signal events that fall in the interval Δp_T. (In this process we varied the size of Δp_T in order to accept enough events to obtain statistically significant results.)

Performing this algorithm produced two histograms displaying number of signal events versus p_T, one using the entire "Data: Signal" sample and the other using those events that passed the $E_{C/9}$ cut. From these two histograms we can determined the signal efficiency of this $E_{C/9}$ cut as a function of p_T which scales with total shower energy. We used Bayesian statistics to determine the errors in efficiency. See Figure 6 for the results.

From the figure we see that the efficiency of the $E_{C/9}$ cut does vary with transverse momentum, peaking at roughly $p_T = 50$ GeV. This implies that some transformation of the transverse shower-shape variable could produce a cut that accepts events without a shower-energy bias.

A similar analysis on the $E_{S/L}^{\text{cor}}$ cut shows signal efficiency increasing with transverse momentum, suggesting that the transformation we described in Subsection 5.2 may have over-corrected the longitudinal shower-shape variable. If we perform a more careful transformation $E_{S/L} \mapsto E_{S/L}^{\text{cor}}$, perhaps rotating by a smaller angle, we might be able to improve more the effectiveness of this cut and further remove the dependency of its efficiency on shower energy.

7 Conclusion

We have begun to eliminate the shower-energy dependence from the cut on the longitudinal shower shape, and we have introduced a new two-dimensional cut that gives increased background rejection to aid in signal isolation. We also investigated which shower-shape cuts are most effective and which we should abandon. This will eliminate unnecessary computing and increase the purity of future signal samples. We have also commented on the similarity of summer Monte Carlo and 2011 data, validating the integrity of GFLash at simulating HF events. Lastly, we have discovered that further improvements can be made to our shower-shape analysis if we eliminate the dependence of the $E_{C/9}$ cut efficiency on shower energy.

References

[1] G. Franzoni, J. Haupt, K. Klapoetke et al., "Identification of Electromagnetic Particles in the CMS Forward Hadron Calorimeter with LHC data collected at $\sqrt{s} = 7$ TeV", *CMS Analysis Note* (2011).

[2] G. Baiatian and A. Sirunyan, "Design, Performance and Calibration of CMS Forward Calorimeter Wedges", *Eur. Phys. J. C53:139-166* (2008).

[3] S. Chatrchyan, V. Khachatryan, A. M. Sirunyan et al., "Measurement of the Inclusive W and Z Production Cross Sections in pp Collisions at $\sqrt{s}$ = 7 TeV", *Which Journal?* (2011).

[4] F. Chlebana, I. Vodopiyanov, V. Gavrilov et al., "Optimization and Performance of HF PMT Hit Cleaning Algorithms Developed Using pp Collision Data at $\sqrt{(s)}$=0.9, 2.36 and 7 TeV", *CMS Note* (2010).