Available on CMS information server

CMS NOTE 2006/043

The Compact Muon Solenoid Experiment

CMS Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

31 March 2006

Tagging b jets with electrons and muons at CMS

P. Demin, S. de Visscher

Université Catholique de Louvain, Louvain-la-Neuve, Belgium

A. Bocci, R. Ranieri

INFN and Università di Firenze, Firenze, Italy

Abstract

The first results of identification of jets from b quarks with soft-lepton tagging algorithms are presented in this note. Jets are built from the energy deposits in the electromagnetic and hadron calorimeters, with an iterative cone algorithm. Electrons and muons are searched for among the reconstructed charged particle tracks associated to these jets with an angular distance criterion. The muon identification is based on standard muon reconstruction algorithms, exploiting the dedicated muon detectors, while electron identification is based on the extrapolation of charged particle tracks into the calorimeter and a detailed analysis of the calorimeter clusters in the region around the track. Jets from b quarks are identified from the kinematic properties of the leptons relative to the jet and the significance of the three dimensional impact parameter of the lepton with respect to the event vertex. The effect of not incorporating the impact parameter significance, as would be necessary for data collected prior to the installation of the silicon pixel tracking detector, is also studied.

1 Introduction

Studies of direct b hadron production and other processes that include the production of bottom quarks are in the Physics programme of the CMS experiment. The decays of top quarks, Higgs boson(s) and supersymmetric particles often involve b quarks in the final state. An efficient technique to select events with b hadrons is therefore desirable, so as to reject the large backgrounds originating from lighter quark and gluon production expected at LHC.

In attempting to identify jets originating from a b quark, several b hadron properties can be exploited:

- The b hadron lifetimes are on the order of 1.6 ps ($c\tau \approx 0.48\ \mathrm{mm}$). At LHC production energies, typically of the order of a few tens of GeV, they travel a significant distance (some millimetres) inside the detector before decaying. Moreover, in most cases b hadrons decay into c hadrons which also have a measurable lifetime, further increasing the travel distance.
- The masses of b hadrons are larger than those of other hadrons, and consequently their decay products have a larger average transverse momentum relative to the initial hadron flight direction than those from lighter hadrons.
- The branching ratio for the direct and cascade decays of b hadrons into electrons and muons is large, $(19.3 \pm 0.5)\%$ for each lepton family [1].

The requirements and performance of a b-tagging algorithm for CMS that relies on identification of leptons within the hadron decay products are presented in this note. In the following, *leptons* will mean either electrons or muons.

A b hadron can decay with the production of a lepton via 3 channels:

- the direct $\mathrm{b} \to \ell$ decay,
$$\mathrm{b} \to \mathrm{W}^{-*}\mathrm{X}, \quad \mathrm{W}^{-*} \to \ell^{-}\overline{\nu}_{\ell}; \tag{1}$$
- the cascade $\mathrm{b} \to \mathrm{c} \to \ell$ decay,
$$\mathrm{b} \to \mathrm{W}^{-*}\mathrm{c}, \quad \mathrm{c} \to \ell^{+}\nu_{\ell}\mathrm{X}; \tag{2}$$
- and the “wrong sign” cascade $\mathrm{b} \to \bar{\mathrm{c}} \to \ell$ decay,
$$\mathrm{b} \to \mathrm{W}^{-*}\mathrm{X}, \quad \mathrm{W}^{-*} \to \mathrm{q}'\bar{\mathrm{c}}, \quad \bar{\mathrm{c}} \to \ell^{-}\overline{\nu}_{\ell}\mathrm{X}. \tag{3}$$

The branching ratio measurements [1] for these decays are $\mathrm{Br}(\mathrm{b} \to \ell^{-}) = (10.70 \pm 0.22)\%$, $\mathrm{Br}(\mathrm{b} \to \mathrm{c} \to \ell^{+}) = (8.02 \pm 0.19)\%$ and $\mathrm{Br}(\mathrm{b} \to \bar{\mathrm{c}} \to \ell^{-}) = (1.62^{+0.44}_{-0.36})\%$. Assuming the cascade decay to be statistically independent from the other two, which are themselves mutually exclusive, the inclusive branching ratio for the decay into at least one lepton is $(19.3 \pm 0.5)\%$. In 1% of the cases both a direct and a cascade decays take place, producing 2 leptons.

In the rest of this work, both cascade and “wrong sign” cascade decays will be referred to simply as *cascade decays*.

The ability of this algorithm to tag b jets is fundamentally limited by this combined branching fraction and further limited by the experimental efficiency for identifying the leptons within these jets. Electron and muon candidates, primarily selected as clusters in the electromagnetic calorimeter and tracks reconstructed in the muon detectors, respectively, are associated to reconstructed tracks in the silicon central tracker to ensure an accurate determination of the lepton momentum and direction. The purity of the b-tagging algorithm, defined by how often lighter quark and gluon jets are tagged as b jets, is limited by light meson (π, K) decays to muons, by photon conversions to $\mathrm{e}^{+}\mathrm{e}^{-}$ pairs, and by the presence of many other charged particles, some of which may satisfy the lepton identification criteria.

This note is organised as follows. In Section 2, a brief description of the detector, the reconstruction algorithms and the simulated event samples used in the analysis is given. The electron and muon identification algorithms are presented in Sections 3 and 4. The performance of tagging b quark jets with these identified leptons is shown in Section 5.

2 Experimental setup and reconstruction algorithms

A detailed description of the CMS detector can be found elsewhere [2]. Existing, standard, reconstruction algorithms for high-level physics objects (jets, charged particle tracks, muons, clusters) are used throughout this note to perform the identification of soft, non-isolated leptons within jets.

Charged particles are detected in the central tracker, equipped with silicon pixel detectors for the innermost layers and silicon strip detectors for the outer part. Charged particle tracks, reconstructed with a Kalman Filter, are required to have at least two hits in the pixel detector, at least five hits in total, and to originate from within a cylinder of length 30 cm and radius 1 mm coaxial with the beam and centred at the nominal interaction point [3].

The energy of electrons and photons is collected in a $\mathrm{PbW0}_4$ crystal calorimeter (ECAL), composed of a barrel section ($|\eta| < 1.479$) and two endcaps ($1.479 < |\eta| < 3.0$), with fine granularity ($\Delta\eta \times \Delta\phi = 0.0175 \times 0.0175$ rad in the barrel) and excellent energy resolution. The clustering algorithm used in the present analysis to reconstruct electron and photon candidates is described in Ref. [4].

The energy of charged and neutral hadrons is further collected in the towers of a brass-scintillator hadron sampling calorimeter (HCAL), with a coarser granularity ($\Delta\eta \times \Delta\phi = 0.0875 \times 0.0875$ rad in the barrel) and a relative energy resolution of $130\%/\sqrt{(E)}$ (E in GeV) [5].

Jets are reconstructed from the energy deposits in these towers and in the corresponding 5×5 crystal matrices, with the Iterative Cone Algorithm and a cone in the (η, ϕ) plane with size $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} = 0.5$ (ϕ in radians). Jet energy calibration is performed, based on comparison with the Monte Carlo parton energy [6].

Charged particle tracks are associated to a jet if their direction is within an angular distance $\Delta R = 0.3$ from the calorimetric jet axis direction. By definition, the flavour of a reconstructed jet is that of the highest energy generated parton within an angular distance $\Delta R = 0.3$ from the calorimetric jet axis direction. The performance of the b-tagging algorithm presented in this note is improved if a charged jet axis is used, determined from the p_T-weighted average of the directions of all charged particle tracks associated to the calorimetric jet, with (electrons) or without (muons) the track associated to the lepton.

The central tracker, the ECAL and the HCAL are immersed in a 4 T axial magnetic field provided by a superconducting solenoid coil. Muons are detected in a muon system hosted in the magnet return yoke of CMS, composed of a barrel part ($|\eta| < 1.2$) and two endcaps ($1.2 < |\eta| < 2.4$). Details of the different components of this system can be found in Ref. [7] and of the standalone muon reconstruction in Ref. [8].

Three data samples were simulated and used throughout the work presented in this note, thereinafter indicated as the signal, background, and flavour enriched samples:

- a sample rich in b jets coming from the leptonic and semi-leptonic decays of $\mathrm{t\bar{t}}$ pairs; a total of 140 000 leptonic and 380 000 semi-leptonic events were used. In both cases the leptonic vertex $\mathrm{t} \to \mathrm{b}\bar{\ell}$ is allowed to decay only into $\ell = \mathrm{e}$ or $\ell = \tau$.
- a sample of QCD proton-proton interactions, simulated with a hard scattering $\hat{p}_\mathrm{T}$ in the range from 30 to 300 GeV/c.
- three samples of QCD di-jet events, with a $\hat{p}_\mathrm{T}$ ranging from 30 GeV/c to over 230 GeV/c. One sample follows a physical distribution of jet flavours and is thus rich in light quark jets, while the other two are artificially enriched in c or b quarks, requiring two jets of the respective flavour to be present in each event. Samples of about 500k b-enriched, 500k c-enriched, and 1.2 million plain QCD events were used.

The detector response was fully simulated with the OSCAR simulation program [9], based on GEANT4 [10], including the pile-up effects expected at low luminosity. The simulated events were subsequently reconstructed with the ORCA program [11]. The electron and muon identification criteria, presented in Sections 3 and 4, were optimised on the $\mathrm{t\bar{t}}$ sample, while potential backgrounds were studied with the QCD sample.

3 Electron identification

3.1 Signal track selection

The algorithm performance was studied on generated electrons with $p_\mathrm{T} > 2$ GeV/c and $|\eta| < 1.2$, with respect to which efficiency and purity numbers are determined.

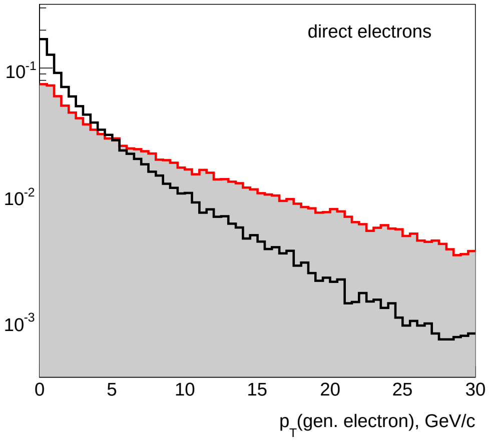

Figure 1: Distribution of the generated p_{T} for direct (shaded) and cascade (solid) electrons.

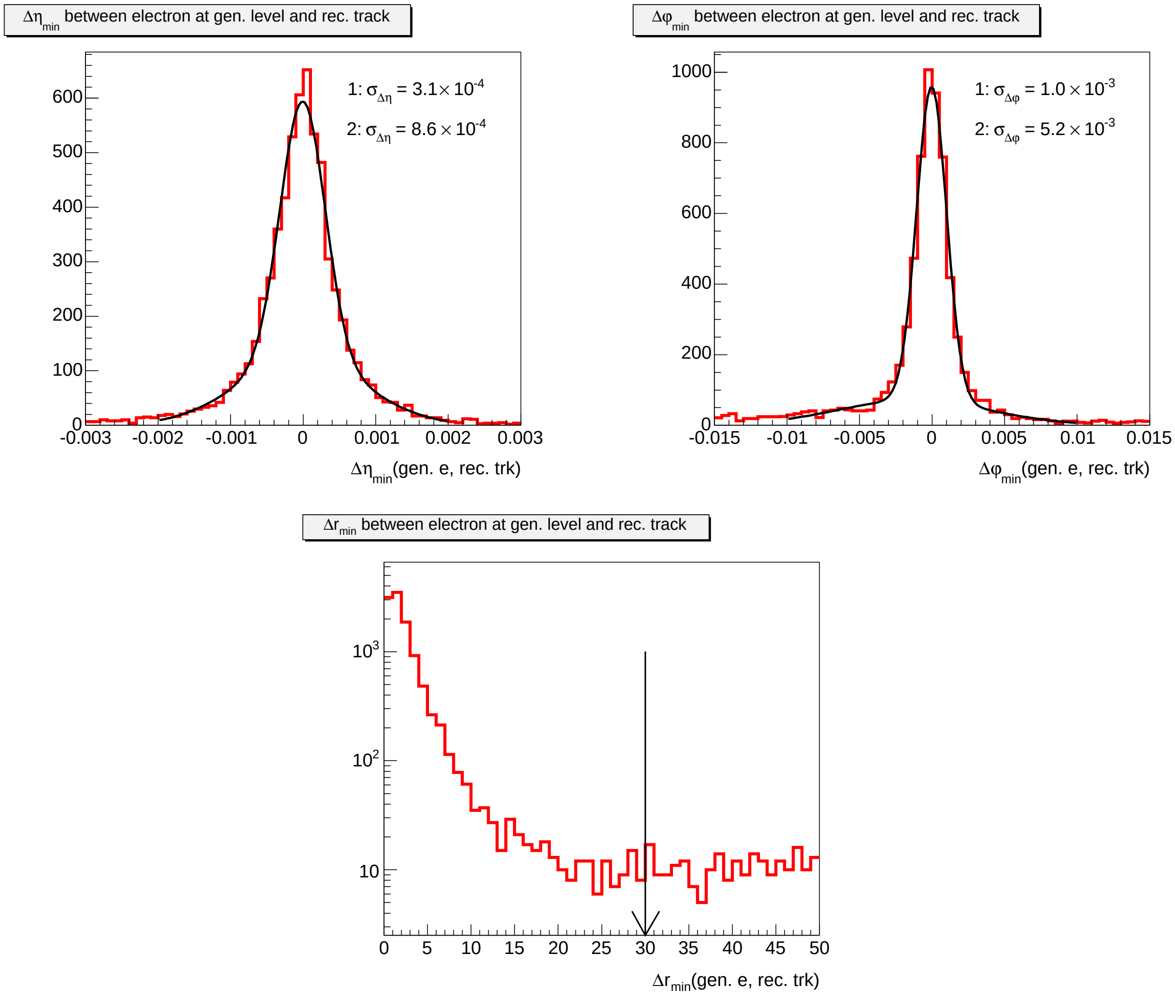

Figure 2: Differences in η (top left) and ϕ (top right) and normalised pseudo-angular distance Δr (bottom) between generated signal electrons and reconstructed tracks. The $\Delta\eta$ and $\Delta\phi$ distributions were individually fitted with two Gaussian functions.

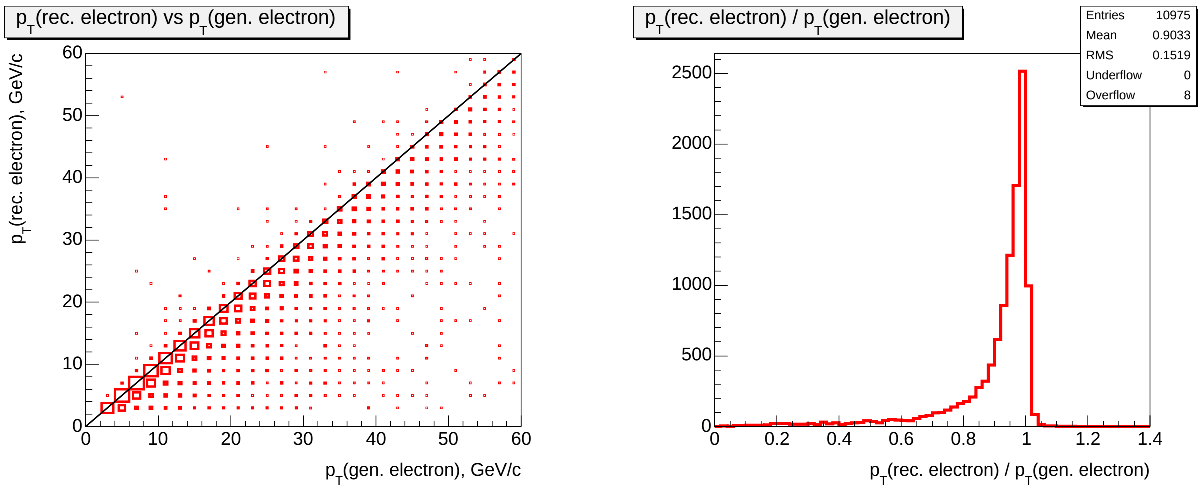

Figure 3: Comparison between the transverse momentum at vertex of the reconstructed track and that of the generated electron.

The generator level p_{T} distributions for electrons in the $\mathrm{t\bar{t}}$ signal samples are shown in Fig. 1 for both direct and cascade produced electrons. Similarly, the reconstructed tracks used for electron identification are required to have a transverse momentum $p_{\mathrm{T}}^{\mathrm{trk}}$ in excess of 2 GeV/c and a pseudo-rapidity $|\eta^{\mathrm{trk}}|$ smaller than 1.2. In addition, a minimum of eight hits in the central tracker are required.

To select tracks associated with electrons, a loose matching between reconstructed tracks and generated signal electrons is performed. The distributions of the η and ϕ differences at vertex are shown in Fig. 2. The Gaussian widths of the cores of the distributions are used to calculate a normalised pseudo-angular distance $\Delta r(e\text{-trk}) = \sqrt{(\Delta\eta/\sigma_{\Delta\eta})^2 + (\Delta\phi/\sigma_{\Delta\phi})^2}$, with $\sigma_{\Delta\eta} = 3.1 \times 10^{-4}$ and $\sigma_{\Delta\phi} = 1.0 \times 10^{-3}$. The distribution of Δr is also shown in Fig. 2. A loose cut $\Delta r < 30$, indicated by the vertical arrow, is chosen to accommodate the bulk of the non-Gaussian tail. For the matched tracks, a comparison between the transverse momentum at vertex of the reconstructed track and that of the generated electron is shown in Fig. 3.

3.2 Electron identification

Electrons are identified by matching an electromagnetic shower in the calorimeter with an associated track in the tracking system. The shower pattern of an electron within the calorimeters depends on its energy and impact position which complicates electron identification. Another difficulty comes from electromagnetic showers of other particles which can mimic the electron shower profile.

The three most common background processes for producing electron-like showers in the calorimeters are the following:

- Charged hadrons with significant energy loss in the electromagnetic section of the calorimeter.
- Neutral pions matched to an unassociated charged particle track in the central tracking detector.
- Photons that convert into an electron-positron pair within the tracking detector material.

In the offline reconstruction, showers in the electromagnetic calorimeters (ECAL) are constructed from energy deposits in groups of neighbouring crystals using standard bump finding algorithms [12].
In order to be matched with reconstructed clusters, tracks are required to satisfy the criteria described in Section 3.1. The tracks are extrapolated to the ECAL using GEANE [13]. The extrapolated track positions are matched to the location of reconstructed clusters. The closest extrapolated track found within $\Delta s(\text{track-cluster}) = \sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2} < 12\ \mathrm{cm}$ is defined to be the matching track for that cluster.

Because the development of electromagnetic and hadronic showers is different for electrons and hadrons, shape information can be used to discriminate between showers originating from these particles. Electrons deposit almost all their energy in the electromagnetic section of the calorimeter, while hadrons are typically much more penetrating. In addition, electromagnetic showers follow a well known teardrop pattern [14], and differences in

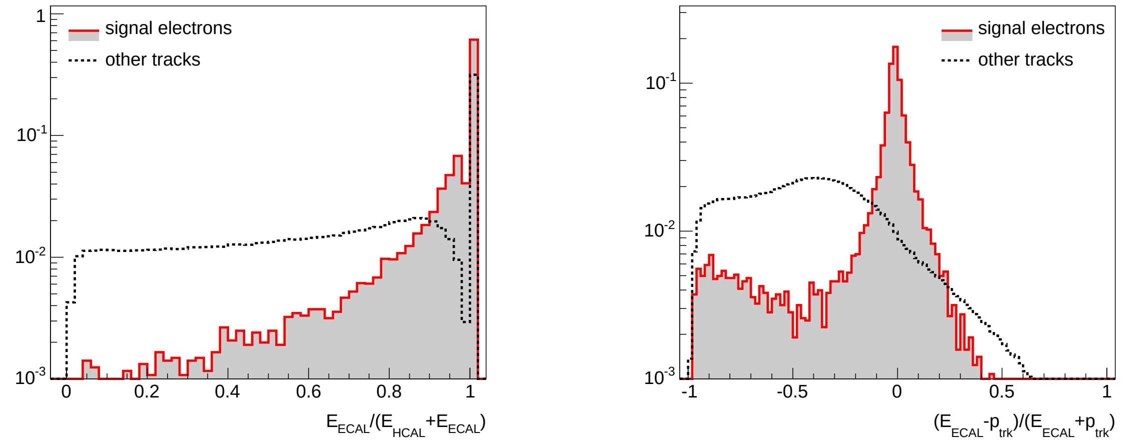

Figure 4: Distribution of $E_{\rm ECAL}/(E_{\rm ECAL}+E_{\rm HCAL})$ (left) and $(E_{\rm ECAL}-p_{\rm trk})/(E_{\rm ECAL}+p_{\rm trk})$ (right) for signal electrons (shaded) and for other charged particle tracks (dashed).

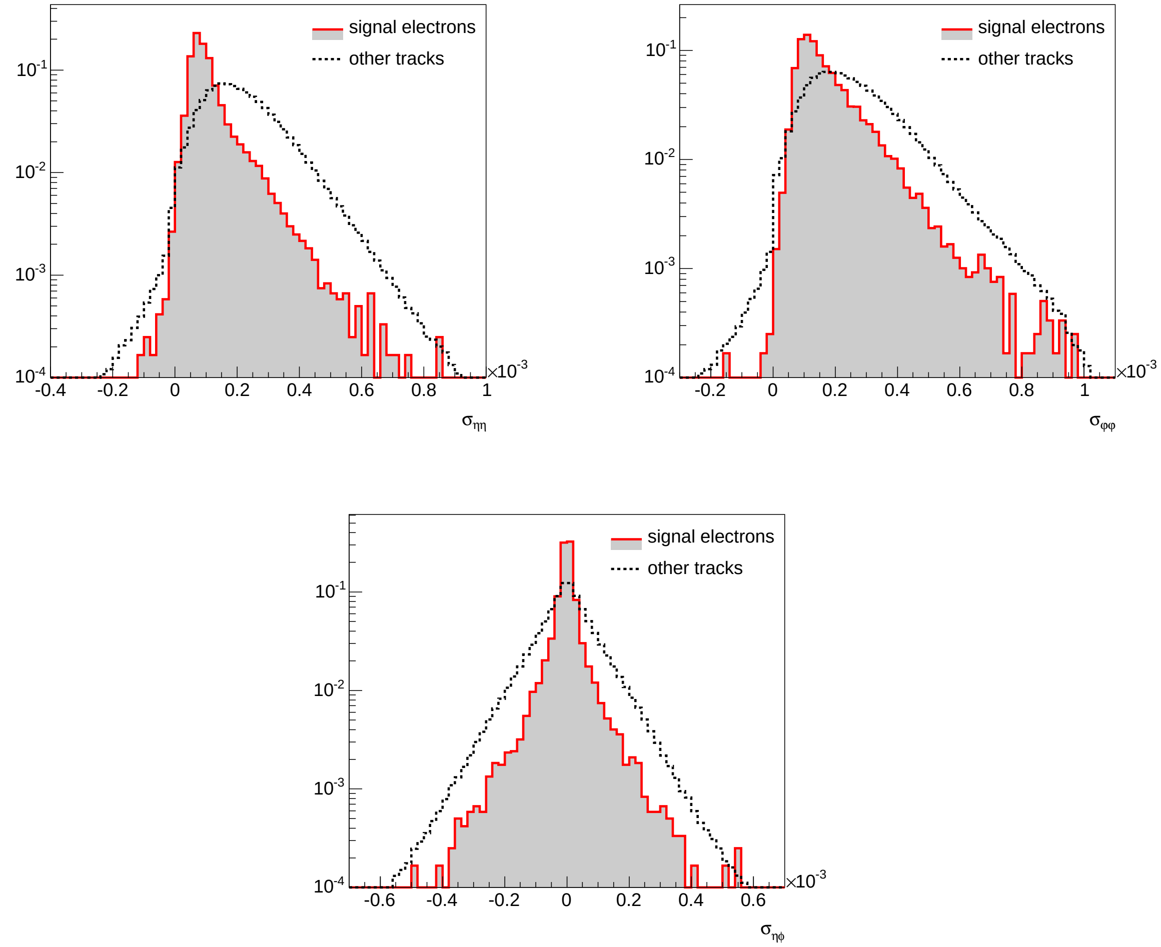

Figure 5: Covariance of the cluster energy distribution $\sigma_{\eta\eta}$ (top left), $\sigma_{\phi\phi}$ (top right), $\sigma_{\eta\phi}$ (bottom) for signal electrons (shaded) and for other charged particle tracks (dashed).

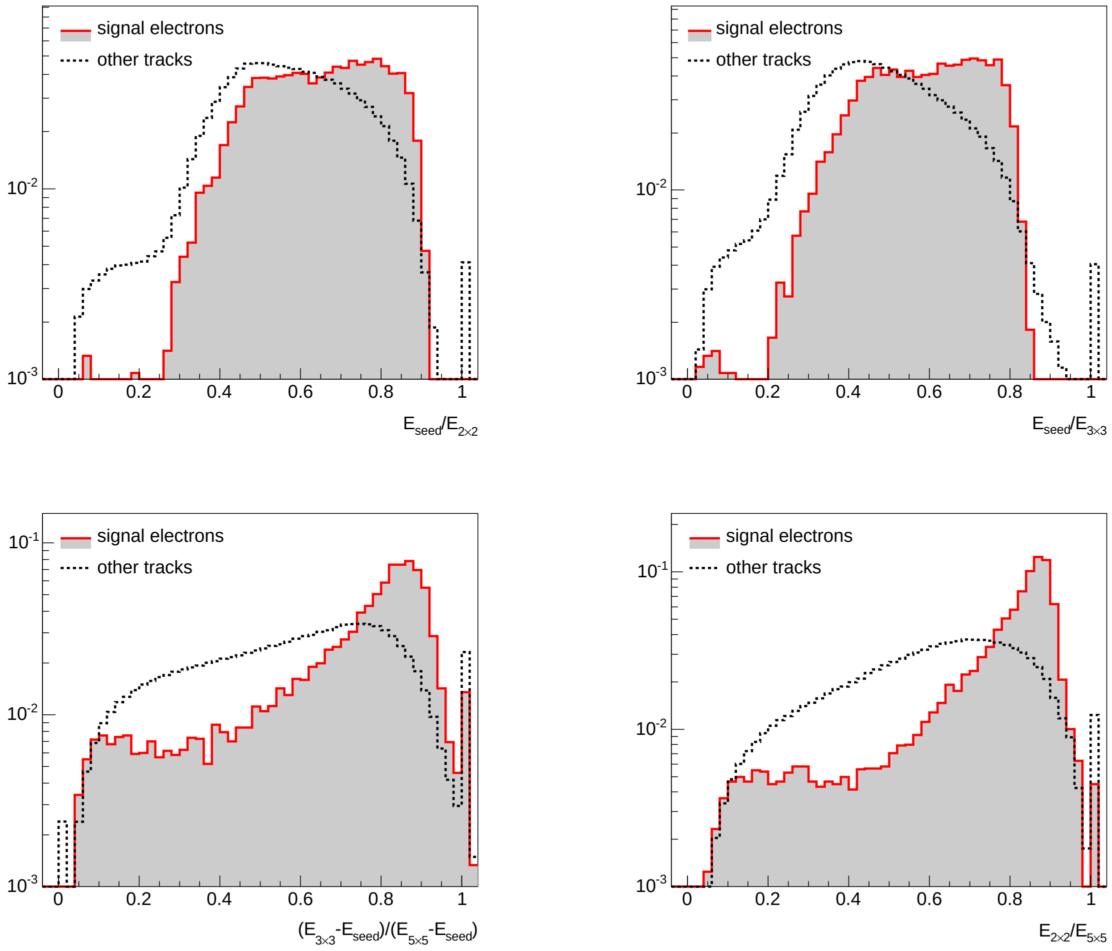

Figure 6: Distribution of cluster energy distribution $E_{seed}/E_{2\times2}$ (top left), $E_{seed}/E_{3\times3}$ (top right), $(E_{3\times3} - E_{seed})/(E_{5\times5} - E_{seed})$ (bottom left), $E_{2\times2}/E_{5\times5}$ (bottom right) for signal electrons (shaded) and for other charged particle tracks (dashed).

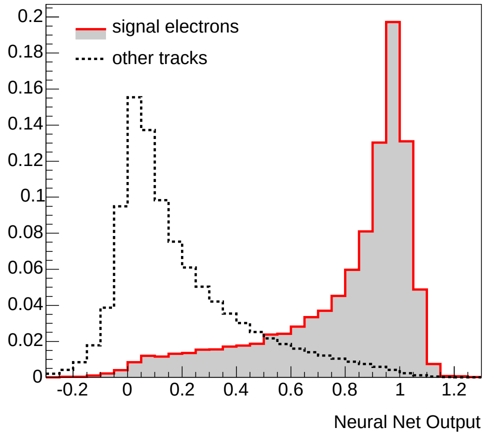

Figure 7: Distribution of the neural network output for signal electrons (shaded) and for other charged particle tracks (dashed).

expected and observed energy deposits for individual crystals within the cluster are expected to be correlated with one other. Therefore, to obtain the best possible discrimination between electrons and hadrons, the electron identification algorithm incorporates variables that describe the cluster transverse shower shape, correlation between energy deposits in crystals within the cluster, and the quality of the match between the cluster and its reconstructed track. An artificial neural network is used to combine these variables, listed in Table 1 and described in more detail below, into a single discriminating variable.

Measurement	variables
covariance of the cluster energy distribution	$\sigma_{\eta\eta}$, $\sigma_{\eta\phi}$, $\sigma_{\phi\phi}$.
distribution of cluster energy	$\frac{E_{seed}}{E_{2\times2}}$, $\frac{E_{seed}}{E_{3\times3}}$, $\frac{E_{3\times3}-E_{seed}}{E_{5\times5}-E_{seed}}$, $\frac{E_{2\times2}}{E_{5\times5}}$, $\frac{E_{\mathrm{ECAL}}}{E_{\mathrm{HCAL}}+E_{\mathrm{ECAL}}}$
cluster energy and track momentum ratio	$\frac{E_{\mathrm{ECAL}}-p_{\mathrm{trk}}}{E_{\mathrm{ECAL}}+p_{\mathrm{trk}}}$

Table 1: List of variables used to identify electrons, where E_{seed}, $E_{N\times N}$ and E_{ECAL} are, respectively, the energy deposit in the cluster seed, the maximal sum of energies within all ECAL crystals in a square of $N \times N$ around the seed and the total energy of the cluster; p_{trk} is the extrapolated track momentum at the ECAL front surface; E_{HCAL} is the sum of energy deposits in HCAL towers next to the ECAL cluster.

The electromagnetic energy fraction of an electron candidate is defined as $E_{\mathrm{ECAL}}/(E_{\mathrm{HCAL}} + E_{\mathrm{ECAL}})$, where E_{HCAL} is the sum of energy deposits in the hadronic calorimeter (HCAL) towers directly behind the crystals contributing to the ECAL cluster and E_{ECAL} is the electromagnetic cluster energy. Figure 4 shows the distribution of $E_{\mathrm{ECAL}}/(E_{\mathrm{HCAL}} + E_{\mathrm{ECAL}})$ for both signal electrons and other charged particle tracks within the simulated $\mathrm{t\bar{t}}$ and QCD di-jet events not found to be associated with electrons.

To describe the correlation between observed energy deposits in crystals within the cluster, the following covariance variables are used: $\sigma_{\eta\eta}$, $\sigma_{\eta\phi}$, $\sigma_{\phi\phi}$. These variables describe the correlation between energy deposits in the ECAL crystals. Distributions of these variables for both signal electrons and other charged particle tracks within the simulated $\mathrm{t\bar{t}}$ and QCD di-jet events are shown in Fig. 5.

To characterise the transverse development of the shower four variables describing cluster energy distribution are used: $E_{seed}/E_{2\times2}$, $E_{seed}/E_{3\times3}$, $(E_{3\times3} - E_{seed})/(E_{5\times5} - E_{seed})$, $E_{2\times2}/E_{5\times5}$, where E_{seed} and $E_{N\times N}$ are respectively the energy deposit in the cluster seed and the maximal sum of energies within all ECAL crystals in a square of $N \times N$ around the seed. Distributions of the variables describing cluster energy distribution for both signal electrons and other charged particle tracks within the simulated $\mathrm{t\bar{t}}$ and QCD di-jet events are shown in Fig. 6.

The variables described above are used as inputs to the neural network used to distinguish electrons from hadrons. A cut on the single variable output of the neural network is used to select showers associated with each particle type. The distributions of the neural network output for signal electrons and other tracks are shown in Fig. 7. Based on the same samples of signal electrons and other charged particle tracks from the simulated $\mathrm{t\bar{t}}$ and QCD samples, both the selection efficiency and misidentification rate of the electron identification algorithm are measured as a function of track p_{T} as shown in Fig. 8.

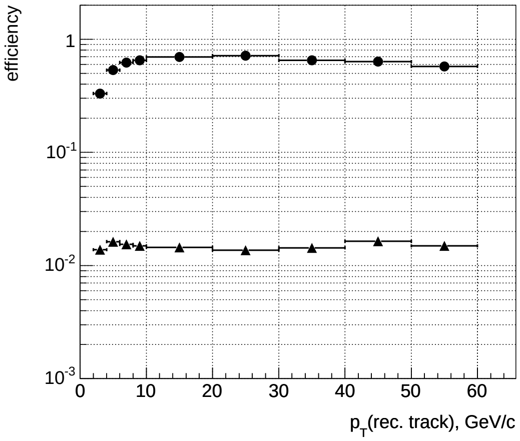

Figure 8: Performance of the soft non-isolated electron identification: efficiency (circles) and misidentification rate (triangles).

p_{T} (GeV/c) vs. $\lvert\eta\rvert$	$0.0\ldots0.9$	$0.9\ldots1.3$	$1.3\ldots1.8$	$1.8\ldots2.4$
2.5…3	0	0	(15.6 ± 2.1)%	(74.2 ± 2.5)%
3…3.5	(1.3 ± 0.5)%	(1.8 ± 0.9)%	(40.3 ± 2.9)%	(83.4 ± 2.1)%
3.5…4	(18.9 ± 1.8)%	(9.7 ± 2.0)%	(61.8 ± 3.0)%	(89.3 ± 1.7)%
4…4.5	(31.0 ± 2.1)%	(35.0 ± 3.3)%	(74.7 ± 2.7)%	(94.3 ± 1.3)%
4.5…5	(53.9 ± 2.2)%	(50.0 ± 3.5)%	(80.7 ± 2.5)%	(95.1 ± 1.2)%
5…6	(74.7 ± 1.4)%	(75.5 ± 2.1)%	(84.0 ± 1.7)%	(96.7 ± 0.7)%
6…7	(89.6 ± 1.0)%	(91.3 ± 1.4)%	(90.3 ± 1.3)%	(98.0 ± 0.5)%
7…8	(95.4 ± 0.7)%	(94.1 ± 1.1)%	(94.5 ± 1.0)%	(98.4 ± 0.5)%
8…9	(97.3 ± 0.5)%	(94.4 ± 1.1)%	(98.2 ± 0.6)%	(98.9 ± 0.4)%
9…10	(97.0 ± 0.6)%	(95.9 ± 1.0)%	(98.0 ± 0.6)%	(98.9 ± 0.4)%
10…20	(98.3 ± 0.1)%	(96.5 ± 0.3)%	(98.3 ± 0.2)%	(98.9 ± 0.1)%

Table 2: Muon reconstruction efficiency for a sample of single muons with flat distributions of transverse momentum $1\ \mathrm{GeV} < p_{\mathrm{T}} < 20\ \mathrm{GeV}$ and pseudorapidity $|\eta| < 2.4$.

p_{T} (GeV/c) vs. $\lvert\eta\rvert$	$0.0\ldots0.9$	$0.9\ldots1.3$	$1.3\ldots1.8$	$1.8\ldots2.4$
2.5…3	(3.2 ± 0.8)%	(2.5 ± 1.1)%	(15.0 ± 2.5)%	(67.1 ± 3.6)%
3…3.5	(3.3 ± 0.8)%	(2.4 ± 1.1)%	(39.9 ± 4.0)%	(80.7 ± 3.4)%
3.5…4	(17.2 ± 1.9)%	(14.2 ± 2.5)%	(54.8 ± 4.0)%	(84.0 ± 3.3)%
4…4.5	(30.7 ± 2.2)%	(28.6 ± 3.5)%	(73.2 ± 3.5)%	(89.2 ± 2.7)%
4.5…5	(46.3 ± 2.5)%	(50.4 ± 4.4)%	(73.1 ± 3.5)%	(90.4 ± 2.6)%
5…6	(57.7 ± 1.9)%	(70.2 ± 2.9)%	(79.6 ± 2.5)%	(91.5 ± 2.0)%
6…7	(72.5 ± 1.8)%	(74.3 ± 2.7)%	(78.6 ± 2.9)%	(96.0 ± 1.4)%
7…8	(74.8 ± 1.9)%	(80.3 ± 2.8)%	(91.2 ± 2.2)%	(93.2 ± 2.1)%
8…9	(80.9 ± 1.9)%	(81.0 ± 2.8)%	(90.2 ± 2.5)%	(98.5 ± 1.0)%
9…10	(81.9 ± 2.0)%	(85.6 ± 2.9)%	(91.2 ± 2.2)%	(96.9 ± 1.5)%
10…20	(83.5 ± 0.8)%	(85.6 ± 1.2)%	(92.5 ± 0.9)%	(97.5 ± 0.6)%

Table 3: Muon reconstruction efficiency for muons inside jets in the signal sample.

4 Muon identification

As mentioned in Section 2, the standard muon reconstruction algorithm is used to select muon candidates. The efficiency of this algorithm for single muons is given in Table 2 as a function of p_{T} and η. The main limitation of this approach is the low efficiency achieved for low p_{T} muons, because of the large bending magnetic field. The efficiency of the algorithm for muons within jets (Table 3) is further reduced due to difficulties in matching tracks in the muon detectors with central tracks contained within the crowded environment of the jets.

In order to calculate the kinematic variables used in the b-tagging algorithm, the globally reconstructed muon track must be matched to a specific reconstructed track within the central tracking detector. This association is made by searching for common hits in the central tracking detector that are attached to both the globally reconstructed muon and one of the reconstructed tracks found within a given jet. To increase the speed of the association algorithm while keeping it as simple as possible, each reconstructed track is only checked against muon tracks sufficiently close in the (η, ϕ) plane, i.e. within 0.1 in η and 0.1 rad in ϕ. To allow different algorithms to be used for global muon and central track reconstruction, hence some flexibility in the track-hit association, the required fraction of common hits between matched tracks is as low as 70%.

Jets originating from light flavour quarks and gluons can be misidentified as b jets in cases where muon candidates are reconstructed within these jets. The most important sources are real muons produced in the decays of light particles, mainly $\pi^{\pm}$ and $\mathrm{K}^{\pm}$ mesons, but also undecayed charged hadrons not fully contained within the calorimeters, which therefore produce hits in the muon detectors. The fraction of muons, electrons, charged pions, charged kaons, and other charged hadrons identified as muons in jets are given in Table 4, for the signal and background samples, as well as for the enriched $\mathrm{b\bar{b}}$ and $\mathrm{c\bar{c}}$ samples.

Generated particle type	μ	e	π	K	other
Enriched $\mathrm{b\bar{b}}$	84.6%	0.2%	10.9%	2.8%	1.4%
Enriched $\mathrm{c\bar{c}}$	78.7%	0.2%	14.6%	4.5%	1.9%
Signal	77.6%	0.3%	16.4%	3.9%	1.8%
Background	44.3%	0.2%	39.0%	11.5%	5.0%

Table 4: Fraction of muons, electrons, and different charged hadrons identified as muons within jets, for different event samples.

5 Tagging algorithm

Once a lepton has been identified in a jet, which already rejects an important fraction of non-b jets, the separation between b and light quark jets can be further improved with the use of lepton and jet kinematic variables as inputs to a feed-forward neural network [15].

5.1 Training samples

Two groups of neural networks were trained for soft muon b-tagging: the first on the signal and background samples, and the second on jets from the flavour-enriched di-jet samples (Section 2). The latter group of networks was found to be less sensitive to the event topology, and was chosen as the primary network for use in the studies reported in this note.

In each case, 10% of the data were used to train the network (training sample) and another 10% for monitoring the training process (test sample). In order to maximise the performance of the network, training was halted when the network performance, as measured on the test sample, started to degrade, as it is a typical symptom of over-training with the original sample. The studies of network performance concerning the same samples as the training one are based on the remaining 80% of the jet sample not used in conjunction with network training.

5.2 Tagging variables

The kinematic parameters of the lepton and jet used as inputs to the neural network are as follow.

- The lepton transverse momentum p_{T}^{rel} relative to the charged jet axis.
- The significance $S_{\mathrm{IP}}^{\mathrm{3D}}$ of the distance of closest approach of the lepton track to the event reconstructed primary vertex (*impact parameter*).
- The pseudo angular distance ΔR in the (η, ϕ) plane between the lepton and the charged jet axis.
- The ratio of the lepton momentum, as measured from the reconstructed track, to the calorimetric jet energy.

To account for differences in detector response at different jet energies and in different regions of the calorimeter, the neural network for soft muon b-tagging also uses the the calibrated calorimetric jet energy and the calorimetric jet pseudorapidity.

Figures 9 and 10 show the distributions of the four common discriminating variables for b, c, light and gluon jets (signal and background samples) for electrons and muons, respectively.

5.3 Performance

The distributions in Fig. 11 show the output of the soft electron b-tagging neural network for different flavours of jets contained within the signal and background samples. Figure 12 shows the efficiency for tagging b jets versus the mistagging efficiencies for each of the lighter jet flavours determined by making a series of progressively tighter cuts on the neural network output variable.

The distributions in Fig. 13 show the output of the soft muon b-tagging neural network for the different jet flavours, from the flavour-enriched samples (left) and the signal and background samples (right).

The distributions obtained from the two samples are similar but have noticeable differences. The resulting b jet tagging efficiencies for the two data samples are plotted as a function of the mistagging efficiencies for each non-b jet flavour in the top half of Fig. 14. The bottom half of Fig. 14 shows resulting soft muon b-tagging purity versus efficiency distributions for the two samples.

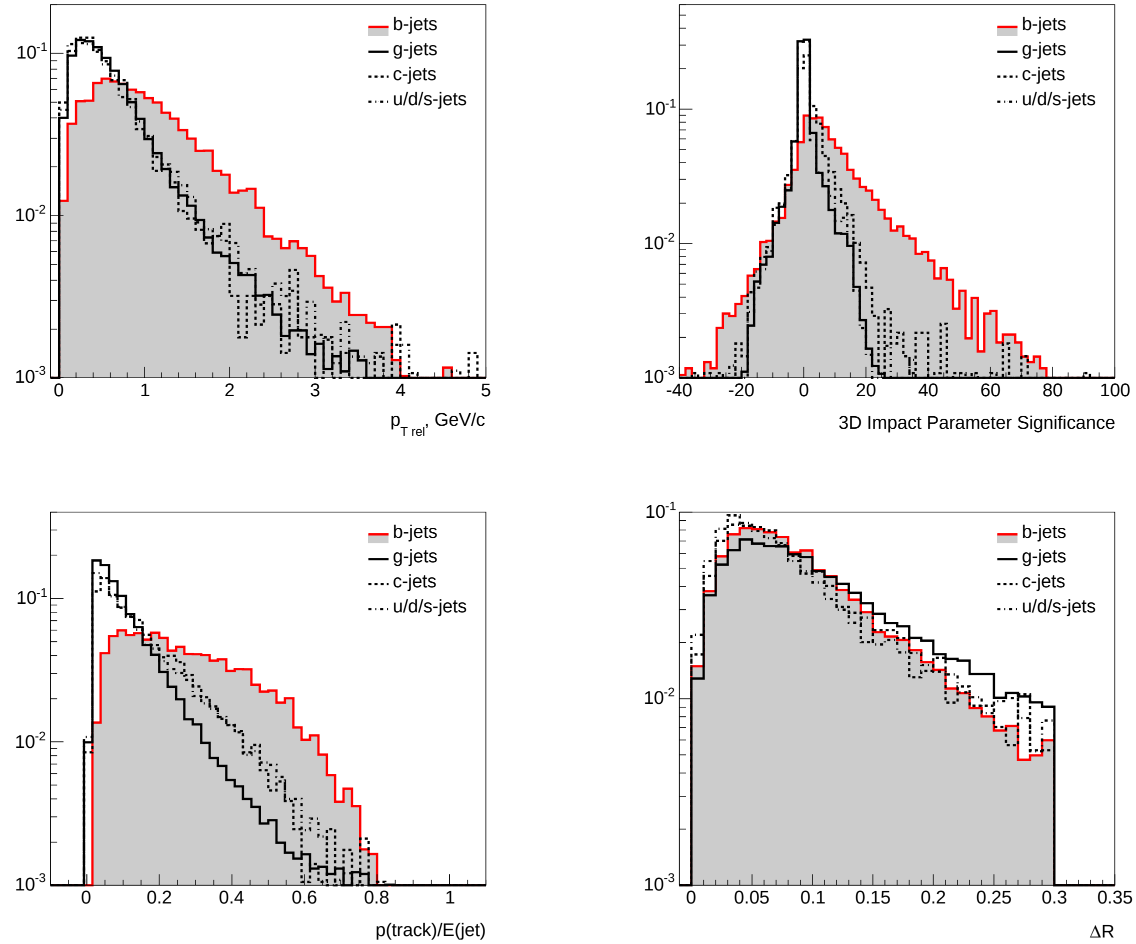

Figure 9: Distribution of the four discriminating variables used for tagging b jets with electrons: electron transverse momentum relative to the charged jet axis (top left), $S_{\rm IP}^{\rm 3D}$ of the electron track (top right), ratio of the electron momentum to the calorimetric jet energy (bottom left), ΔR between the electron track and the charged jet axis (bottom right). The distributions are shown separately for b jets (shaded), c jets (dashed), light jets (dash-dotted) and gluon jets (solid) and are obtained for jets found in the signal and background samples in the barrel region of the detector ($|\eta| < 1.4$).

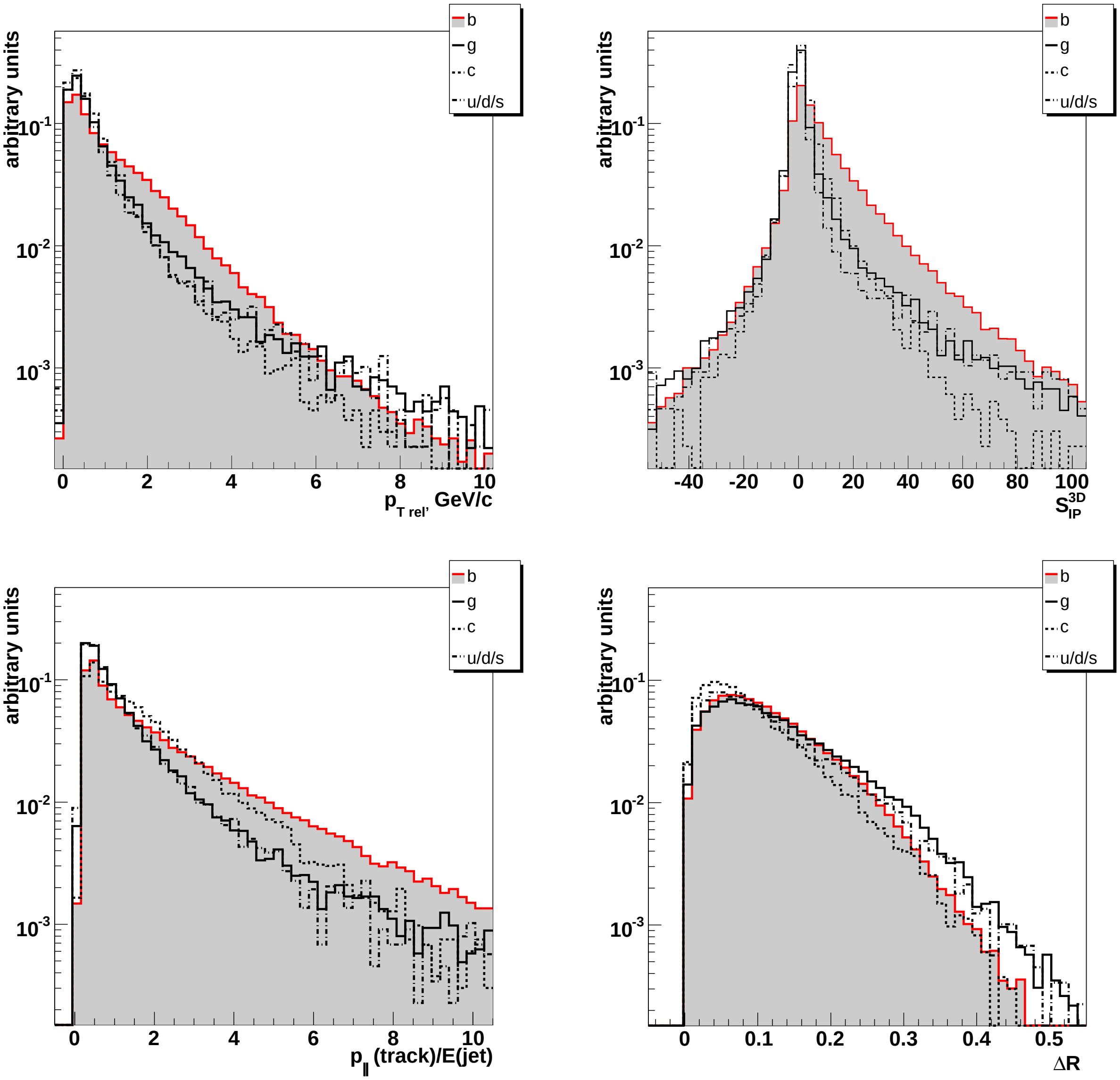

Figure 10: Distribution of the four discriminating variables used for tagging b jets with muons: muon transverse momentum relative to the charged jet axis (top left), $S_{\mathrm{IP}}^{\mathrm{3D}}$ of the muon track (top right), ratio of the muon momentum to the calorimetric jet energy (bottom left), ΔR between the muon track and the charged jet axis (bottom right). The distributions are shown separately for b jets (shaded), c jets (dashed), light jets (dash-dotted) and gluon jets (solid) and are obtained for jets found in the signal and background samples in in the whole detector acceptance ($|\eta| < 2.4$).

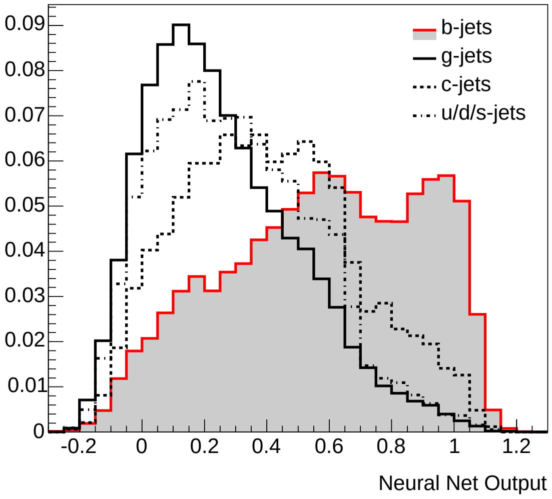

Figure 11: Distribution of the neural network output for the electron based b jet identification algorithm, shown separately for b jets (shaded), c jets (dashed), light jets (dash-dotted) and gluon jets (solid) as obtained for jets found in the signal and background samples in the barrel region of the detector ($|\eta| < 1.4$).

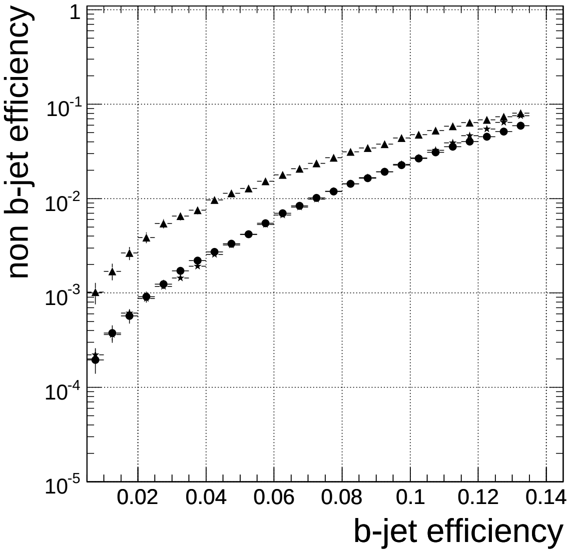

Figure 12: Performance of the electron based b jet identification algorithm. Non-b jet mistagging efficiency versus b jet tagging efficiency is shown separately for c jets (triangles), light jets (circles) and gluon jets (stars) and is obtained for jets found in the signal and background samples in the barrel region of the detector ($|\eta| < 1.4$) .

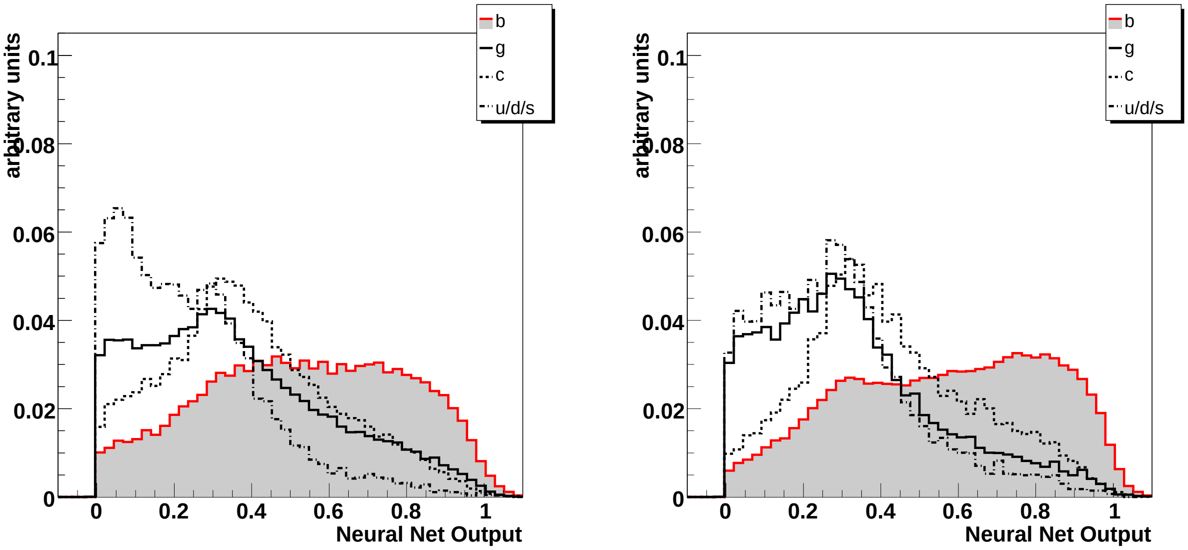

Figure 13: Distribution of the neural network output for the muon based b jet identification algorithm, shown separately for b jets (shaded), c jets (dashed), light jets (dash-dotted) and gluon jets (solid) as obtained for jets found in the flavour-enriched samples (left) and signal and background samples (right) in the whole detector acceptance ($|\eta| < 2.4$).

5.4 Tagging without vertex information

Preliminary studies on the application of the soft muon b-tagging in the case where no track impact parameter information is available have also been performed. This information would not be available for the first data collected with the CMS detector in the scenario where the installation of the silicon pixel detector is staged beyond first LHC collisions. Therefore, a neural network has been trained that doesn't make use of such variable. The output of this network for the signal and background samples is shown in Fig. 15, and its overall performance over the same samples is shown in Fig. 16. A comparison with Fig. 14 shows that at a given b tagging efficiency, the mistagging rate increase is of roughly a factor 1.5 for charm quark jets, 3 for light quarks jets and 2 for gluon jets.

6 Conclusions

This note describes an algorithm for tagging b jets based on the identification of electrons and muons within the jet. While the reconstruction of jets, muons and charged particle tracks uses standard tools, a dedicated electron identification has been developed and tuned specifically for this purpose.

Different tagging variables related to the lepton and jet parameters have been studied, as well as possible combinations thereof. The results achieved with non-linear neural network techniques are presented in this note, and show that a clear separation of jets from b quark from jets from lighter quarks is possible.

A separate study has also been performed for tagging b jets based on the muon identification within the data collected prior to the installation of the silicon pixel tracking detector, and thus with a limited spatial resolution for the reconstruction of the interaction primary vertex. The b-tagging performance in this scenario is slightly worse, yet still provides a useful tool for jet flavour discrimination.

Further improvements are possible both to the soft lepton b tagging algorithm itself, requiring a more detailed study of the distribution of the tagging variables for the signal and background events, and to the reconstruction algorithms that this work relies on, such as better tracking for the electrons and muons, the use of calorimetric deposits for muon identification, the reconstruction of jets using more advanced algorithms.

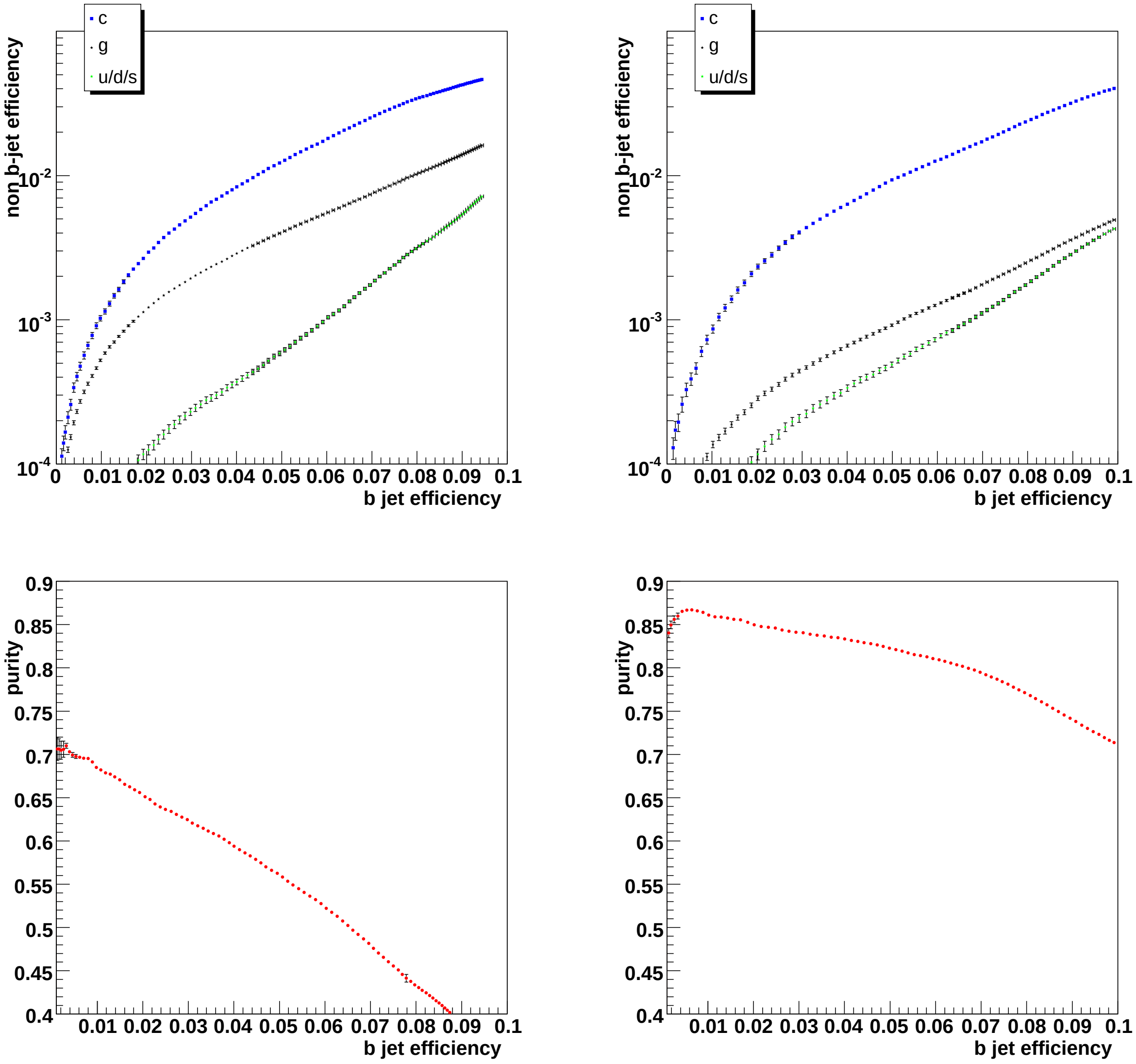

Figure 14: Performance of the muon based b jet identification algorithm, for jets found in the flavour-enriched samples (left) and the signal and background samples (right) in the whole detector acceptance ($|\eta| < 2.4$). Top: mistagging of charm (top), light quark (middle) and gluon (bottom) jets as a function of the tagging efficiency; bottom: b jets identification purity vs. efficiency.

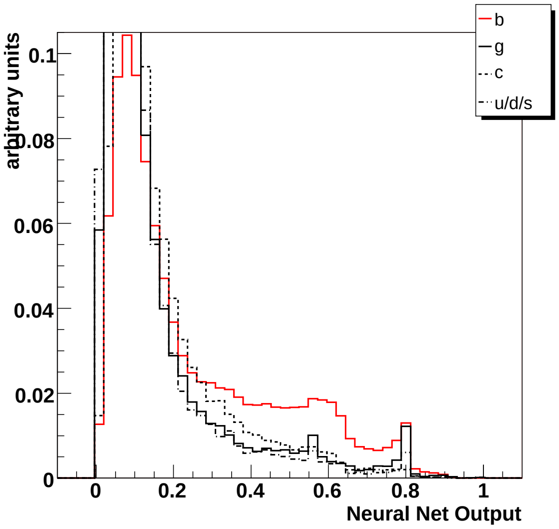

Figure 15: Distribution of the neural network output for b jets (solid, grey), c jets (dashed), light jets (dash-dotted) and gluon jets (solid) as obtained from a neural network that doesn't use the lepton impact parameter, for jets found in the signal and background samples in the whole detector acceptance ($|\eta| < 2.4$).

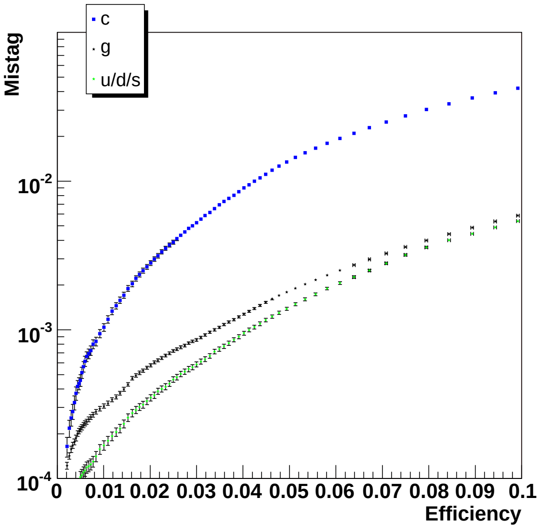

Figure 16: Mistagging rate for soft muon b-tagging without impact parameter significance, for jets found in the signal and background samples in the whole detector acceptance ($|\eta| < 2.4$).

References

[1] The ALEPH Collaboration, the DELPHI Collaboration, the L3 Collaboration, the OPAL Collaboration, the SLD Collaboration, the LEP Electroweak Working Group, the SLD Electroweak and Heavy Flavour Groups, "Precision Electroweak Measurements on the Z Resonance", CERN-PH-EP/2005-041, SLAC-R-774, 2005.

[2] CMS Collaboration, "CMS, the Compact Muon Solenoid: Technical proposal", CERN/LHCC-94-38, 1994.

[3] CMS Collaboration, "CMS Physics Technical Design Report Volume I: Software and Detector Performance", CERN/LHCC-2006-001, 2006.

[4] CMS Collaboration, "CMS: The Electromagnetic Calorimeter. Technical Design Report", CERN/LHCC-97-33, 1997.

[5] CMS Collaboration, "CMS: The Hadron Calorimeter Technical Design Report", CERN/LHCC-97-31, 1997.

[6] A. Heister *et al.*, "Measurement of Jets with the CMS Detector at the LHC", CMS NOTE 2006/036, 2006.

[7] CMS Collaboration, "CMS: The Muon Project. Technical Design Report", CERN/LHCC-97-32, 1997.

[8] E. James, Y. Maravin, M. Mulders, N. Neumeister, "Muon identification in CMS", CMS NOTE 2006/010, 2006.

[9] S. Abdoulline *et al.*, "An Object-Oriented Simulation Program for CMS Analysis", in "Proceedings of the Computing in High Energy Physics 2004 Conference", CHEP 2004, Interlaken, Switzerland 27 September - 1 October 2004,

[10] S. Agostinelli *tet al.*, "GEANT4: A simulation toolkit", Nucl. Instrum. Meth. A **506**, p. 250, 2003.

[11] CMS Collaboration, "CMS OO Reconstruction – User's Guide and Reference Manual", site located at `http://cmsdoc.cern.ch/orca`.

[12] E. Meschi, T. Monteiro, C. Seez, P. Vikas, "Electron Reconstruction in the CMS Electromagnetic Calorimeter", CMS NOTE 2001/034, 2001.

[13] V. Innocente, M. Maire and E. Nagy, "GEANE: Average Tracking and Error Propagation Package", CERN Program Library, IT-ASD W5013-E, 1991.

[14] T. Ferbel, "Experimental Techniques in High-Energy Nuclear and Particle Physics", Second Edition, World Scientific Publishing Company, 1991.

[15] C. M. Bishop, "Neural Networks for Pattern Recognition", Oxford University Press, 1995.