CMS Draft Analysis Note

The content of this note is intended for CMS internal use and distribution only

2013/04/10
Head Id: 180541
Archive Id: 180539:180543M
Archive Date: 2013/04/10
Archive Tag: trunk

New gluon/quark jet discriminator

A.C. Marini, F. Pandolfi[1], S. Bansal, T. Cornelis, N. van Remortel[3], P. Azzurri[4], and D. del Re[5]

[1] Institute for Particle Physics, ETH Zürich, CH-8093 Zürich, Switzerland
[2] European Organization for Nuclear Research, CERN
[3] University of Antwerp, Antwerpen, Belgium
[4] INFN Sezione di Pisa, Italy
[5] University and INFN Rome, ??? , Italy

Abstract

A new discriminator has been developed to distingish jets originating from quarks and gluons, that extends the separation capabilities to jets in the forwad region, up to $|\eta| = 5$, and increases the overall performances, in particular for jets of low trasverse momentum, down to $p_{\mathrm{T}} = 20\,\mathrm{GeV}$. Discriminating observables have been optimized and chosen based on performances established with Monte Carlo simulated QCD events. Two alternative multivariate quark-gluon jet taggers are presented with slightly different performances. Comparisons of 2012 dijet and Z plus jets data and Monte Carlo simulations for the input discriminating variables and the output tagger distributions have been performed. Finally a recipe to evaluate systematic uncertainties related to the use of either tagger is given, based on the observed data versus Monte Carlo differences.

Contents

1 Introduction

It is known, both from first principles and a large collection of experimental measurements [1–4], that hadronic jets initiated from gluons exhibit differences with respect to jets from light flavor quarks. In jet fragmentation, final state hadrons are produced in a cascade process of subsequent bremsstrahlungs, followed by hadronisation setting in at a scale of $\Lambda_{QCD} \sim 200$ MeV. Bremsstrahlung is directly proportional to the coupling of the radiated gluon to the radiator, in this case a gluon or light quark. The probabilities are therefore directly proportional to the color factor $C_A = 3$ for gluons and $C_F = 4/3$ for quarks. Neglecting gluon splitting to quark-antiquark pairs, the ratio of the number of gluons radiated from a gluon to that from quarks is therefore expected the approach $9/4$ at leading order and for infinite phase-space. Corrections to this ratio can can be calculated analytically up to NLO with the inclusion of energy (but not momentum) conservation. Numerical approaches exist as well, which are typically implemented in Monte Carlo generators, and have the advantage of taking the phase space limits more accurately into account, but by introducing ambiguities in the scale at which the evolution of the gluon multiplicities radiated from a parton are being evaluated. Regardless of the approach, a larger number of softer particles are expected to be produced in a gluon jet and the gluon jet is expected to be wider than a quark jet of the same energy.

Many experimental measurements of these fundamental properties have been made. Albeit with ambiguities in the gluon-jet separation, the measurements are in qualitative agreement with the analytical expectations and are described quantitatively by most recent Monte Carlo models and have been used to extract the ratio C_A/C_F which is compatible with the SU(3) expectation. The most accurate experimental results originate from LEP, supplemented with more data from hadron colliders such as the Tevatron. The following main experimental observations were made:

- The charge multiplicity is higher in gluon jets than in light quark jets, independent of the particle species;
- the fragementation function of gluon jets is considerably softer than that of a quark jet, which can be explained by the higher multiplicity of soft gluons radiated off a gluon and double string formation necessary before hadronization in case of $g \to q\bar{q}$ splitting;
- gluon jets are less collimated than quark jets, an observation that can be quantified in the jet broadness or the major and minor axis of the jet cone.

In this study we will exploit these differences to construct a probability tagger capable of discriminating jets initiated by light quark partons from those initiated by gluons. The use of Particle Flow jet reconstruction allows us to access particle-level information during jet reconstruction, exploiting the full granularity of the CMS detector. We shall show how this advantage allows us to define a broad range of discriminating variables.

2 Motivations

Some analyses rely on the identification of an exclusive final state that includes a fixed number of hadronic jets that generally originate from light or heavy quarks. Some examples are hadronic W, Z and top decays, the production of vector bosons or Higgs via vector boson fusion, associated vector boson and Higgs production and cascade decays of squarks. In many cases these hadronic final states suffer from overwhelming backgrounds from multi-jet QCD production or from electroweak backgrounds with hard initial or final state gluon radiation.

The tagging of quark jets is also useful in the mass reconstruction of hadronically decaying objects, where the resolution is generally degraded due to combinatorial backgrounds.

In addition, there is also the increasing amount of pile-up that can result in the reconstruction of artificially created jets by clustering energy deposits and/or tracks originating from different proton-proton interactions in teh same bunch crossing. These fake jets will have characteristics that are different from showering gluons or quarks and can therefore also be discriminated.

Developing a general tool that suits all purposes is difficult, as many analyses use different kinds of jet objects, work in different kinematic regimes and/or use loose or stringent detector acceptances and quality requirements.

The need for this tool arose from an analysis dedicted to the measurement of a Higgs boson, produced via the vector-boson-fusion process and subsequently decaying to b-quarks. This process is characterized by a four jet final state, of which two are b-tagged and two forward going light quark jets with a large rapidity separation. The multi-jet QCD backgrounds for this process are overwhelming, resulting in signal-to-noise ratios of 10^{-5}, even after using dedicated triggers. A gluon/quark tagging tool will undoubtedly help in suppressing these types of background. When developing the tool, we will try to cover the most generally used jet objects and cover a large pseudorapidity and transverse momentum range. The approach can also serve as a guideline for those who want to develop their own discriminators when using more specialized jet definitions or quality cuts.

3 Object Definition

The CMS software framework supports a susbtantial variety of jet objects, some dedicated to QCD physics, some to electroweak analyses and searches. In addition, there are several levels of jet energy scale corrections and procedures to subtract pile-up effects. In this study we use jets clustered with the Anti-kt algorithm with a distance parameter $R = 0.5$ using charged and neutral particles that are reconstructed with the Particle Flow algorithm. Residual jet energy corrections are applied in order to make the response as linear as possible with unit slope. In this study we have used the 53X(V7) jet energy corrections. The same clustering algorithm is applied to stable generator particles in simulated events, thus defining generator jets.

All jets are further subjected to a set of basic quality requirements:

- they should originate from an event with at least one good quality vertex with a minimum of four degrees of freedom and a maximal distance to the nominal interaction point along the beam axis of 24 cm.
- a tight jet identification, according to the criteria in [5].

The acceptance in pseudorapidity is kept as large as possible, but the jets will be categorized in three distinct acceptance regions: the central region corresponding to $|\eta| < 2$, the transition region corresponding to $2 < |\eta| < 3$, and the forward region corresponding to $3 < |\eta| < 5$. The jet transverse momenta are required to exceed 20 GeV/c and can be as large as kinematically allowed. The training of the tool will split up the sample in several narrow jet p_T bins, and will be conducted in the central region ($|\eta| < 2$) and the forward region ($3 < |\eta| < 5$). The results will then be extended to cover the transition region, so that the taggers defined in the central region will be commissioned up to $|\eta| = 2.5$, and the forward tagger will be used also in the $2.5 < |\eta| < 3$ pseudorapidity region.

The studies carried on the Monte Carlo simulation will assign a flavor to each reconstructed

jet. In QCD, flavor is not an infrared-safe observable, and is therefore not well defined if not at tree level. We define the jet flavor with an empirical approach, by accessing the generator information (partons and generator jets) in the following way:

- if no generator jet is found with a $\Delta R = 0.3$ cone around the reconstructed jet direction, the jet is considered to be a pile-up (PU) jet;
- if a generator jet is present, the reconstructed jet is matched to the closest (in terms of ΔR) parton in the event (partons are particles in the event with PYTHIA `status = 3` and PDG identification number compatible with either a quark or a gluon). If a parton is found within $\Delta R = 0.3$ of the reconstructed jet direction, the flavor of the parton is assigned to the jet; if not, the jet is considered 'undefined'.

In the above (as in the following), ΔR represents the distance in the $\eta - \phi$ plane, and is defined, for pairs of objects whose rapidities (azimuths) differ by $\Delta\eta$ ($\Delta\phi$), as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

As is shown in more detail in Appendix A, jets initiated by b quarks present hadronization properties which are closer to gluon-initiated jets than to light quark jets. On the other hand, b-jets can be efficiently be identified with the use of b-tagging algorithms, such as the Combined Secondary Vertex (CSV) tagger, which will be used in this note, therefore this tagger will be specialized in the identification of light quark jets. In order to do so, all PFJets considered in this note will be required to be anti-b-tagged, i.e. if a jet is found to have a positive loose Combined Secondary Vertex (CSV) tag, it is discarded.

In addition to being complementary to b-tagging, the quark-gluon discriminator is also meant to be complementary to PU-jet identification. For this reason we require, for jets reconstructed in the tracker-covered region, that they are not identified as PU jets by using the standard requirement on the β^* variable, which is defined as the energy fraction of tracks of the jet which are not compatible with the primary vertex of the event. We therefore require for all jets considered in this study to satisfy:

$$\beta^* < 0.2 * \ln(\mathrm{N_{PV}} - 0.67)$$

where $\mathrm{N_{PV}}$ is the number of reconstructed primary vertexes in the event. Standard PU identification also makes use of another variable, the RMS of the spread of the jet candidates in the $\eta - \phi$ plane. But as this variable is also strongly influenced by hadronization effects, and is discriminating between quarks and gluons, we do not apply any requirement on this variable, in order to avoid to introduce a bias in the sample.

The amount of PU (and underlying event) activity in a given event is measured by the diffuse particle flow energy density (ρ), computed with the median of the transverse momenta of jets clustered with the k_T clustering algorithm with radus parameter $R = 0.6$, after injecting in the event a very large number of infinitely soft particles (ghosts). For the studies presented in this note, two definitions of the ρ variable have been employed:

- the 'standard' ρ, which is computed with all PFCandidates in the event (up to $|\eta| = 5$);
- an energy density variable computed in the same manner, but restricting the median computation to the candidates reconstructed in the tracker acceptance ($|\eta| < 2.5$). This variable will be from here onwards called ρ_{iso}.

The two above measurements of the amount of PU in an event are equivalent, and the two variables have very high degrees of correlation. The use of two different variables is mainly due to historical reasons.

4 Improving the Previous Tagger

Previous studies, conducted in 2011, have already led to the definition of a quark-gluon discriminator [6], which has been used in published results, such as [7]. The tagger was a simple multiplicative likelihood discriminator, based on three variables:

- **Charged Multiplicity**: the number of reconstucted charged hadron candidates in the PFJet;
- **Neutral Multiplicity**: the number of reconstructed photon and neutral hadron candidates in the PFJet;
- $\mathbf{p_T D}$: defined as:

$$p_T D = \frac{\sqrt{\sum p_{\mathrm{T}}^2}}{\sum p_{\mathrm{T}}}$$

  where the sums are extended over all candidates within the PFJet.

This tagger has been developed focusing on the discrimination of hard jets ($p_{\mathrm{T}} \gtrsim 100 GeV$), and was defined only for jets reconstructed within the tracker acceptance. Of the three used variables, the discrimination power was almost exclusively left to $p_T D$up to transverse momenta of 100-200 GeV, and from then on the role of the multiplicities became increasingly important.

The studies reported in this note describe how we have improved the performance of quark-gluon discrimination at CMS. The improvements to the discriminator have been lead by two main directives:

- the extension of the discriminator to the forward rapidity region;
- the optimization of the discrimination performance for soft jets ($p_{\mathrm{T}} < 50$ GeV).

We have furthermore decided to define two distinct taggers: one will be a simple likelihood product, similar to what was done in 2011; the other will be a multivariate discriminant (a multi-layer perceptron, or MLP), implemented within the TMVA [8] framework.

The re-optimization of the tagger performance starts from an improved set of discriminating variables. We will now show how the choice of the new variable set was made.

4.1 The Considered Variables

A large number of variables can be considered in order to discriminate between quark and gluon hadronization. As we have seen, we are looking for variables which are able to distinguish quark and gluon hadronization based on the number, energy and spread of the stable particles produced during such a process. We have therefore tried to define the largest set of variables which could be able of doing so. The complete list of variables we have considered is the following:

- charged multiplicity ($nChg$);
- neutral multiplicity ($nNeutral$);
- total multiplicity ($nPFCand$);
- candidate RMS of in the $\eta - \phi$ plane ($RMSCand$);
- candidate second moment minor axis in the $\eta - \phi$ plane ($axis_2$);
- candidate second moment major axis in the $\eta - \phi$ plane ($axis_1$);
- candidate asymmetry (or $pull$);

- fractional leading candidate transverse momentum (R);
- candidate p_{T} distribution ($p_T D$).

Each variable has been computed in two varieties: the default computation uses all PFCandidates within the jet as input. An additional value of the variable is obtained when restricting the input candidate list: out of all charged constituents, only those whose tracks satisfy quality control (QC) criteria are considered. These criteria require the tracks to be assigned to the hardest reconstructed primary vertex of the event, in a similar way as what is done by the charged hadron subtraction in PFNoPU. Only 'highPurity' tracks assigned to the hardest PV are kept, and they are further required to satisfy $|d_z/\sigma(d_z)| < 5$ and $|d_0/\sigma(d_0)| < 3$. Quality control criteria are intended to mitigate the effect of pile up in the tracker covered region of the detector.

One additional variation was considered, only for the multiplicity and R variables: the set of PFCandidates was restricted to the candidates with a p_{T} greater than 1 GeV. This requirement is also intended to limit the effects of PU, especially in the forward, tracker-less, region of the detector.

What follows is the complete list of variables we have considered, with their definition.

Multiplicities

The simpliest and most studied variable that we can construct is the multiplicity, i.e. the total number of candidates reconstructed within the jet. Its properties are summarized in different articles (e.g. [9]), where it is shown that the ratio of the multiplicity from quark and gluon jets should converge to the color factor ratio $\frac{C_A}{C_F} = \frac{9}{4}$ and provide a good discrimination, expecially at high p_{T}. We construct three multiplicity variables:

- **nPFCand**: the total PFCandidate multiplicity of the PFJet;
- **nChg**: the multiplicity of charged PFCandidates of the PFJet (charged hadrons, electrons, muons);
- **nNeutral**: the neutral PFCandidate multiplicity of the PFJet (photons, neutral hadrons, HF hadrons, HF EM particles).

Quark jets are expected to have, on average, lower values of multiplicities with respect to gluon jets, for a given transverse momentum. Figures 4.1(a,b,c) show the expected normalized distributions for nPFCand, nChg and nNeutral for quark (blue) and gluon (red) jets with $80 < p_{\mathrm{T}} < 120$ GeV reconstructed in the central part of the detector ($|\eta| < 2$).

RMScand

Jets have a conical structure that can be projected in the $\eta - \phi$ plane. From first principles we expect gluon hadronization to produce jets which are 'wider' than jets induced by quark hadronization. The jet width may be quantified by looking at the second moments of the constituent distribution in the $\eta - \phi$ plane.

The simplest way of doing this is to treat the η and ϕ directions democratically, and define a global RMS of the candidate spread:

$$\mathrm{RMScand} = \sqrt{\frac{\sum_i w_i \Delta R_i^2}{\sum_i w_i}}$$

where the sums run on all PFCandidates within the jet, the ΔR_i are distances in the $\eta - \phi$ plane

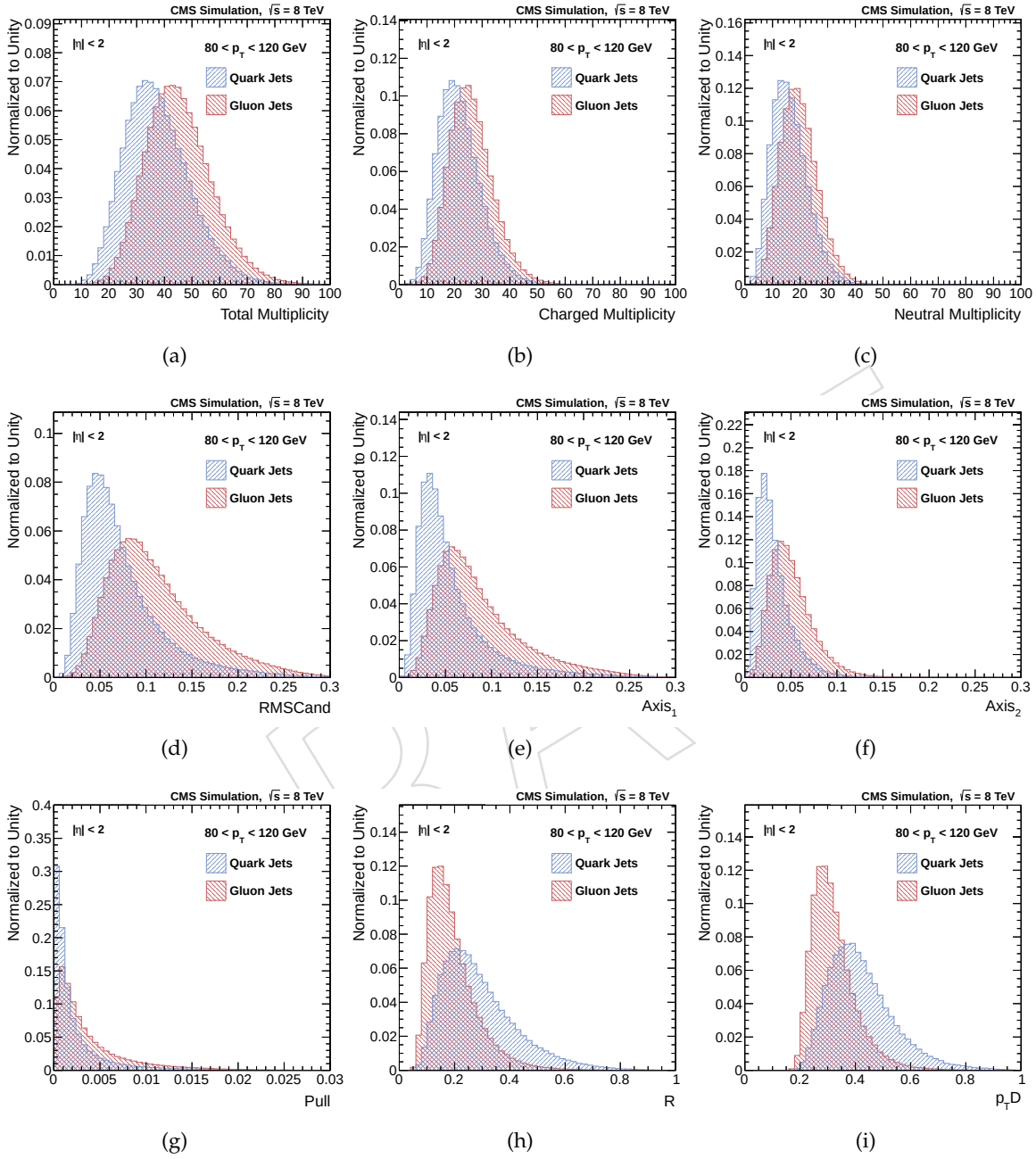

Figure 4.1: Normalized distributions of the considered discriminating variables, for quark (blue) and gluon (red) jets with $80 < p_{\mathrm{T}} < 120$ GeV reconstructed in the central part of the detector ($|\eta| < 2$). (a): total multiplicity (nPFCand); (b): charged multiplicity (nChg); (c): neutral multiplicity (nNeutral); (d): RMScand; (e): axis$_1$; (f): axis$_2$; (g): pull; (h): R; (i): $p_T D$.

($\Delta R_i = \sqrt{\Delta\eta_i^2 + \Delta\phi_i^2}$, where $\Delta\eta_i$ and $\Delta\phi_i$ are, respectively, the rapidity and azimuthal differences between the given candidate and the mean of the candidates), and the w_i are weights. Since jets consist of particles with a power law p_t spectrum, it is natural to compute the jet axes by means of p_t based weights. It was found in [**?**] that the weights, w_i give an optimal separation between gluon and quark jets when they depend on the squared transverse momentum of each constituent particle: $w_i = p_{\mathrm{T},i}^2$. This definition of weights will be used in all variables.

Quark jets are expected to produce, on average, narrower jets and therefore lower values of RMScand with respect to gluon jets, for a given transverse momentum. Figure 4.1(d) shows the expected normalized distribution for RMScand for quark (blue) and gluon (red) jets with $80 < p_{\mathrm{T}} < 120$ GeV reconstructed in the central part of the detector ($|\eta| < 2$).

Axes

A slightly more sophisticated approach is two compute, instead of the simple RMS of the candidate $\eta - \phi$ spread, the two principal components of the second moments. The shape of a jet can be approximated by an ellipse which is characterized by its two principle axes, the major and the minor axis, and the orientation of the major axis in the plane. The shape of the jet can thus be specified in terms of a 2x2 symmetric matrix, M, with the following elements:

$$M_{11} = \sum_i w_i \Delta\eta_i^2 \tag{4.1}$$

$$M_{22} = \sum_i w_i \Delta\phi_i^2 \tag{4.2}$$

$$M_{12} = M_{21} = \sum_i w_i \Delta\eta_i^2 \Delta\phi_i, \tag{4.3}$$

where $\Delta\eta_i$ and $\Delta\phi_i$ are taken as the pseudorapidity and azimuthal distance between each constituent particle, i, and the jet direction.

The major and minor axes of the jet can then be computed as the eigenvalues $\lambda_{1,2}$ of the matrix M by:

$$axis_1 = \left(\lambda_1 / \sum_i w_i\right)^{1/2} \tag{4.4}$$

$$axis_2 = \left(\lambda_2 / \sum_i w_i\right)^{1/2}. \tag{4.5}$$

By defining them in this way, the two axes are statistically independent.

Quark jets are expected to produce, on average, narrower jets and therefore lower values of the axes variables with respect to gluon jets, for a given transverse momentum. Figures 4.1(e,f) show the expected normalized distribution for axis_1 and axis_2 for quark (blue) and gluon (red) jets with $80 < p_{\mathrm{T}} < 120$ GeV reconstructed in the central part of the detector ($|\eta| < 2$).

Pull

Another discriminator is the jet asymmetry or pull, that expresses how much the jet direction is determined by one (or more) particularly high p_{T} particle that are not symmetrically distributed around the jet direction. It is computed as the vector sum of p_{T}^2 weighted distances of each jet constituent with respect to the jet direction (r_i):

$$\mathrm{pull} = \left| \frac{\sum_i w_i |r_i| \vec{r}_i}{\sum_i w_i} \right| \tag{4.6}$$

Quark jets are expected to have, on average, larger values of the pull variable with respect to gluon jets, for a given transverse momentum. Figure 4.1(g) shows the expected normalized distribution for the pull variable for quark (blue) and gluon (red) jets with $80 < p_{\mathrm{T}} < 120$ GeV reconstructed in the central part of the detector ($|\eta| < 2$).

R

The R variable is defined as the fractional transverse momentum of the leading particle inside the jet:

$$R = \frac{\max(p_{\mathrm{T},i})}{\sum_i p_{\mathrm{T},i}}. \tag{4.7}$$

Quark jets are expected to hadronize more asymmetrically with respect to gluon jets, and therefore have R values closer to unity. Figure 4.1(h) shows the expected normalized distribution for the R variable for quark (blue) and gluon (red) jets with $80 < p_{\mathrm{T}} < 120$ GeV reconstructed in the central part of the detector ($|\eta| < 2$).

$p_T D$

A slightly more complex way of assessing how asymmetrically the transverse momentum of a given jet is distributed throughout its constituents is given by the $p_T D$ variable, defined by:

$$p_T D = \frac{\sqrt{\sum_i p_{\mathrm{T},i}^2}}{\sum_i p_{\mathrm{T},i}}$$

From its definition, it stems that $p_T D \to 1$ for jets made of only one particle, which carries all of its momentum, and $p_T D \to 0$ for a jet made of an infinite number of particles, similarly to the R variable.

It is interesting to notice how $p_T D$ and the axes are linked together by an elegant description of the candidates in terms of the introduction of momenta of the jets, and so they can be se as term of the momenta expansion as shown in [10].

Quark jets are expected to hadronize more asymmetrically with respect to gluon jets, and therefore have $p_T D$ values closer to unity. Figure 4.1(i) shows the expected normalized distribution for the $p_T D$ variable for quark (blue) and gluon (red) jets with $80 < p_{\mathrm{T}} < 120$ GeV reconstructed in the central part of the detector ($|\eta| < 2$).

4.2 The Choice of the Variables

The choice of the set of variables on which to base the new discriminator has been done by studying their discriminating performance on the simulation. The discriminating power was assessed by studying their receiver operating characteristics (ROC) in terms of quark efficiency and gluon rejection. These studies have been performed in a specific ρ_{iso} interval ($12 < \rho_{iso} < 14$ GeV), which roughly corresponds to the position of the mode of the data. No dramatic difference in terms of variable performance is expected as a function of ρ_{iso}, nor any has been observed.

The ROC plots are shown in Figures 4.2 and 4.3, respectively for jet reconstructed in the central part of the detector ($|\eta| < 2$) and in the forward ($3 < |\eta| < 5$). In each figure, different transverse momentum intervals are considered: for central jets $30 < p_{\mathrm{T}} < 40$ GeV (top left), $80 < p_{\mathrm{T}} < 120$ GeV (top right) and $300 < p_{\mathrm{T}} < 400$ GeV (bottom); for forward jets $20 < p_{\mathrm{T}} < 40$ GeV (left) and $50 < p_{\mathrm{T}} < 100$ GeV (right).

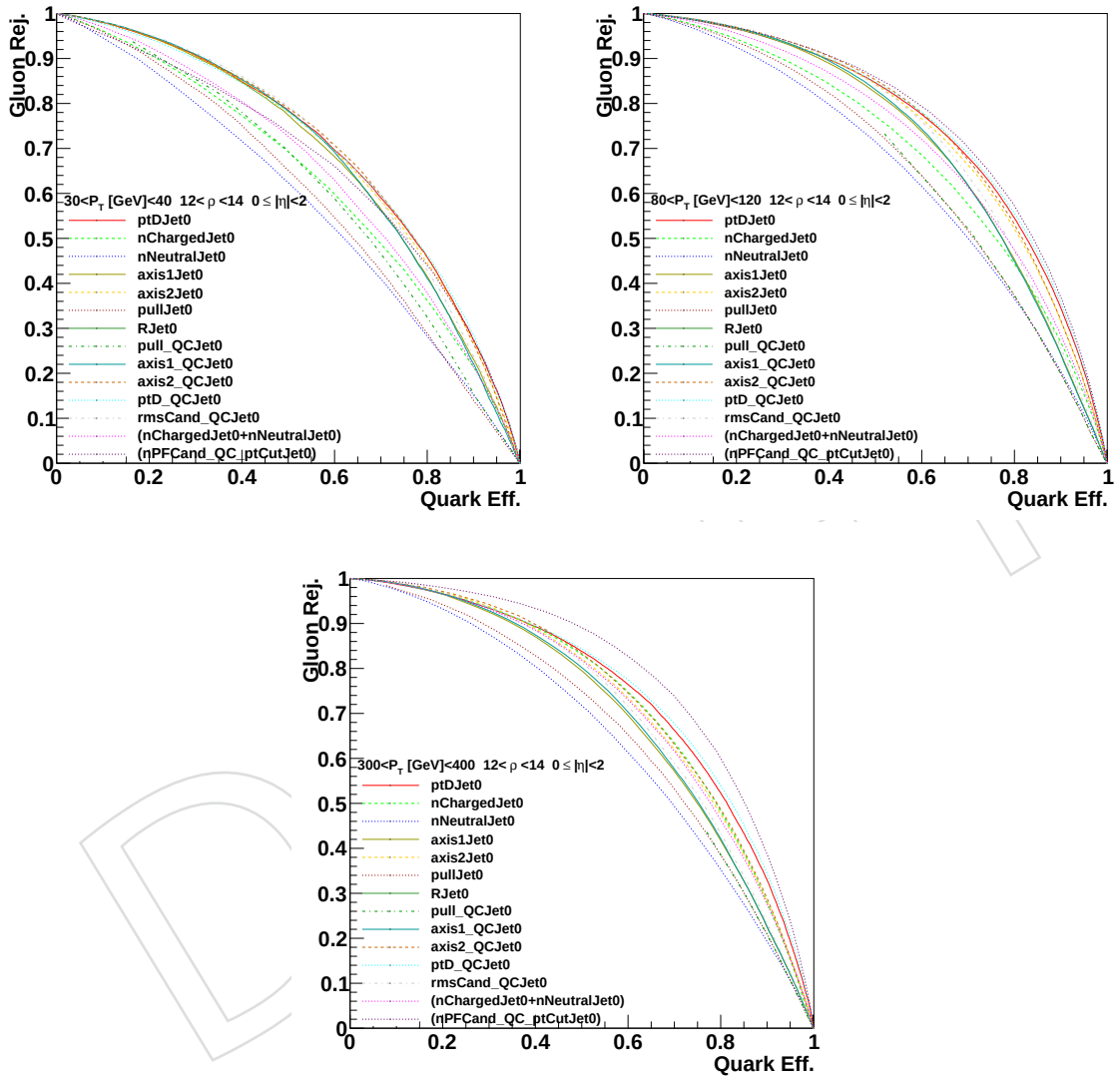

Figure 4.2: Single variable ROC curves for central jets ($|\eta| < 2$) in three transverse momentum bins: $30 < p_{\mathrm{T}} < 40\,\mathrm{GeV}$ (top left), $80 < p_{\mathrm{T}} < 120\,\mathrm{GeV}$ (top right) and $300 < p_{\mathrm{T}} < 400\,\mathrm{GeV}$ (bottom). The figures show the single variable performance in the quark efficiency (x axis) - gluon rejection (y axis) plane.

We start by analyzing the central jet ROC figures. The figures are very busy, but (with the help of some zooming) some results may be assessed:

- in general, quality control (QC) variables perform (slightly) better than their standard counterparts;
- of the multiplicity variables, the charged multiplicity significantly outperforms the neutral multiplicity, but the total multiplicity is found to have even better discrimination; the total multiplicity with quality control on tracks and minimal 1 GeV p_{T} requirement on the neutral (nPFCand_QC_ptCut) shows very good discrimination starting from jets of about 100 GeV onwards;
- out of the 'spread' variables, the second moment minor axis (axis_2) has the best performance, the second moment major axis (axis_1) has the worst, and the averaged RMS (RMScand) lies approximately in between;
- the pull variable does not provide much discrimination power;
- out of the two asymmetrical hadronization variables (R and $p_T D$), $p_T D$ has the best performance, and actually seems to be the variable with the best discrimination power across all the p_{T} range.

Most of these observations are valid also in the forward region (see Figure 4.3). Here no tracker multplicity is present, so the corresponding entries in the legend for the total multiplicity correspond to the total neutral multiplicity. As you can see, the introduction of the minimal 1 GeV transverse momentum threshold increases significantly the discrimination.

We therefore proceed to choose the set of variables on which to build the disciminator. The choice criteria may be summarized as follows: we identify the minimal set of uncorrelated variables which present the highest single-variable discrimination power. From what has been observed above, it is clear that such a set is constituted by the following variables:

- total multiplicity
- axis_2
- $p_T D$

To these three variables, a fourth one can be added:

- axis_1

Although this variable doesn't have very high discrimination power, its definition makes it independent from axis_2, therefore it brings additional, uncorrelated, information, which will improve the discrimination power of a multivariate discriminator.

All these variables are taken in their quality control (QC) variation, and only the neutral candidates which have transverse momentum greater than 1 GeV are used in the computation of the total multiplicity.

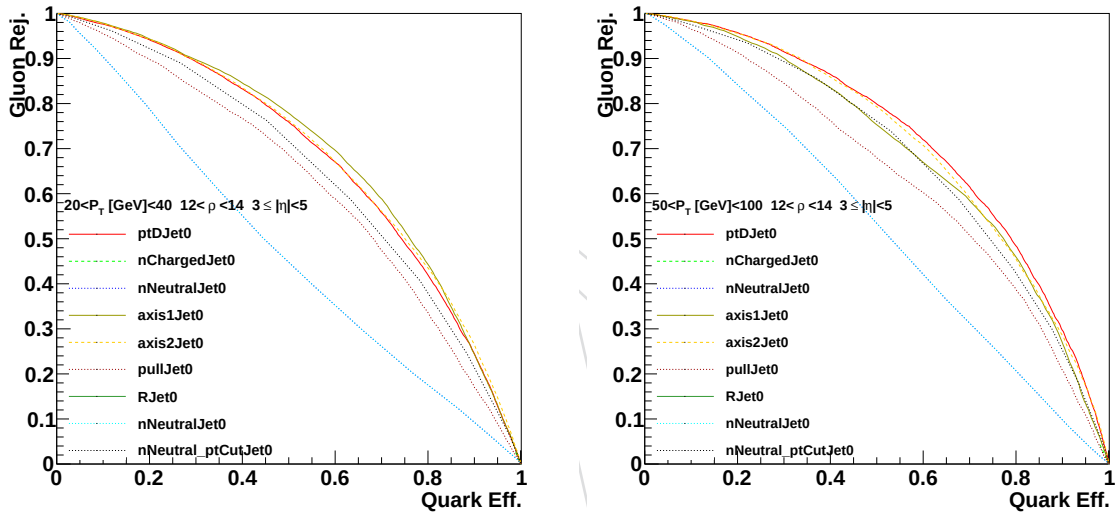

Figure 4.3: Single variable ROC curves for forward jets ($3 < |\eta| < 5$) in two transverse momentum bins: $20 < p_{\mathrm{T}} < 40$ GeV (left) and $50 < p_{\mathrm{T}} < 100$ GeV (right). The figures show the single variable performance in the quark efficiency (x axis) - gluon rejection (y axis) plane.

4.3 The New Taggers

Once the set of variables is found, we have to choose the discriminator. We define two discriminators:

- a simple 3-variable likelihood discriminator, built on the product of nPFCand, axis_2 and $p_T D$;
- a multivariate 4-variable discriminator, based on nPFCand, axis_2, $p_T D$, and axis_1, which uses the Multi-Layer Perceptron (MLP) Artificial Neural Network provided within the TMVA framework.

We will now detail how these two discriminators are built, as they differ not only in the algorithmical computation of the output, but also in the treatment of the input variables' dependance on the jet transverse momentum and on the amount of PU activity.

Likelihood Discriminant

The likelihood discriminant is built as a simple product of the three variables nPFCand, axis_2 and $p_T D$. This approach ensures simplicity, transparency and robustness. As it explicitly neglects the treatment of variable correlations, though, it might be suboptimal in phase space regions in which these are complex, as for instance the high p_T limit.

The probability density functions (PDFs) on which the likelihood is based are built from jets in simulated events, which have been successfully tagged as light quark (u, d, s) or gluon jets. The PDFs are computed separately in two rapidity regions: a central region ($|\eta| < 2$) and a forward region ($3 < |\eta| < 5$). The PDFs obtained in the central region are then used for jets up to $|\eta| = 2.5$, and jets with $2.5 < |\eta| < 3$ use the forward region PDFs.

In order to take into account the strong dependance of the means and shapes of the variables both as a function of the jet transverse momentum and the amount of PU activity in the event, the PDFs are computed double-differentially in bins of jet transverse momentum and ρ_iso. In the central region, a total of 21 bin in transverse momentum are defined in the central region, which span in logarithmic spacing between 20 and 4000 GeV: 20, 26, 32, 40, 51, 64, 80, 101, 127, 159, 201, 252, 317, 400, 503, 633, 797, 1003, 1262, 1589, 2000, 4000 GeV. In the forward region, all jets with transverse momentum greater than 127 GeV are put in one big 127-4000 GeV bin, in order to take into account the obvious energetic limitations are high rapidity.

As for the binning in ρ_iso, it is linearly spaced in 1 GeV stpdf, between 0 and 45 GeV. Then a final bin integrates all values between 45 and 100 GeV.

Once the three variables' probability density functions (in the form of histograms derived from the simulation) are obtained in all of the $p_\text{T} \times \rho$ bins, a likelihood discriminant can be defined. This has been accomplished as a simple product of the three variables' distributions, in each $p_\text{T} \times \rho$ bin. A given PFJet identifies a vector $\vec{x}$ in the three-dimensional space of the structure variables. Probability density functions for gluons (G) and quarks (Q) can then be defined as the product of each variable's probability density function (f^i) computed at the given variable's value ($x[i]$):

$$G(\vec{x}) = \prod_i f_G^i\left(x[i]\right) \qquad Q(\vec{x}) = \prod_i f_Q^i\left(x[i]\right)$$

so that a likelihood estimator can be defined as:

$$L(\vec{x}) = \frac{Q(\vec{x})}{Q(\vec{x}) + G(\vec{x})}$$

and interpreted as the probability of a given PFJet to be originated from a quark parton.

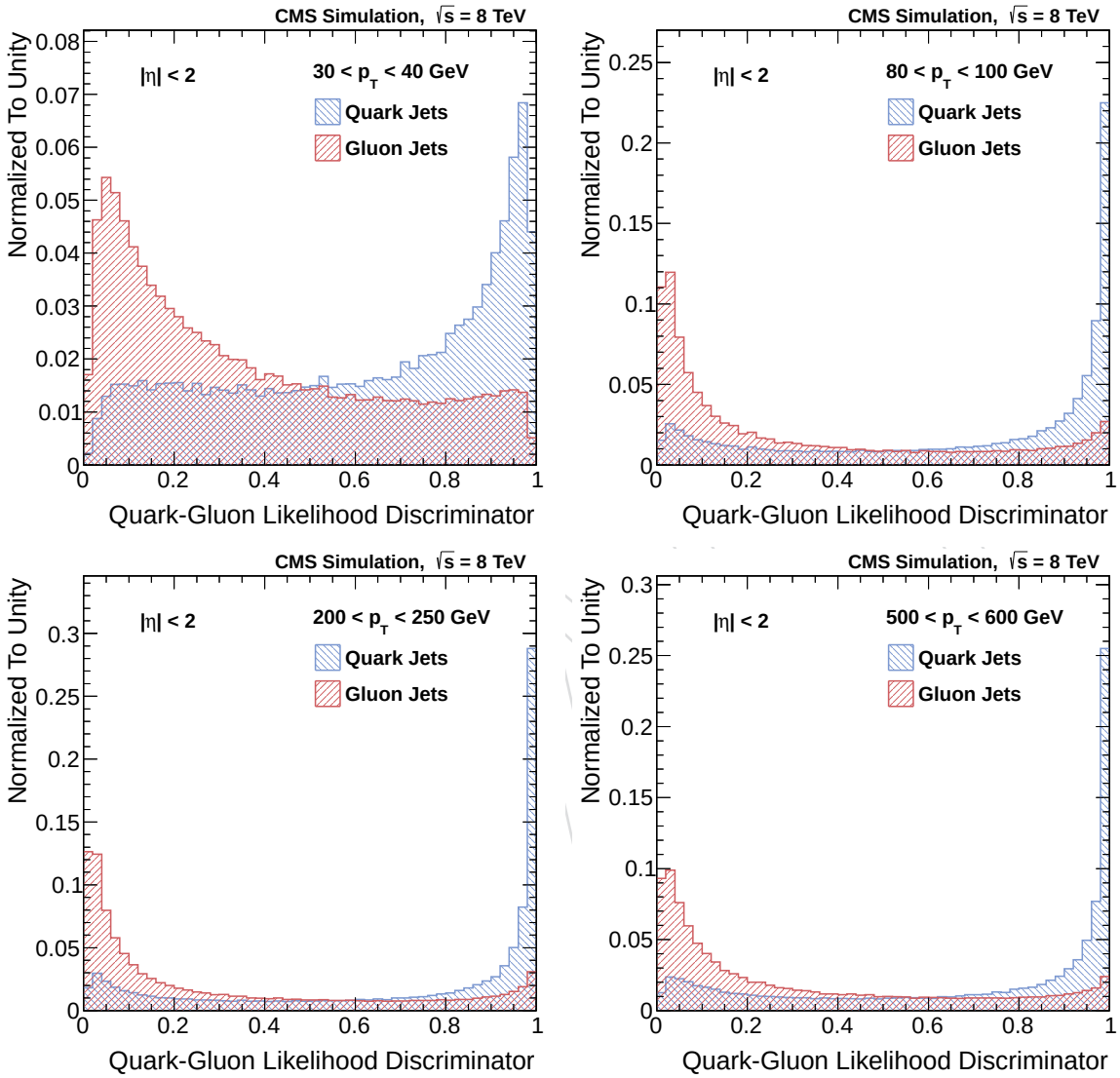

Figure 4.4: Normalized distributions of the Likelihood Quark-Gluon discriminant for quark (blue) and gluon (red) jets reconstructed in the central part of the detector ($|\eta| < 2$). Four representative transverse momentum bins are shown: $30 < p_{\mathrm{T}} < 40$ GeV (top left), $80 < p_{\mathrm{T}} < 100$ GeV (top right), $200 < p_{\mathrm{T}} < 250$ GeV (bottom left) and $500 < p_{\mathrm{T}} < 600$ GeV (bottom right).

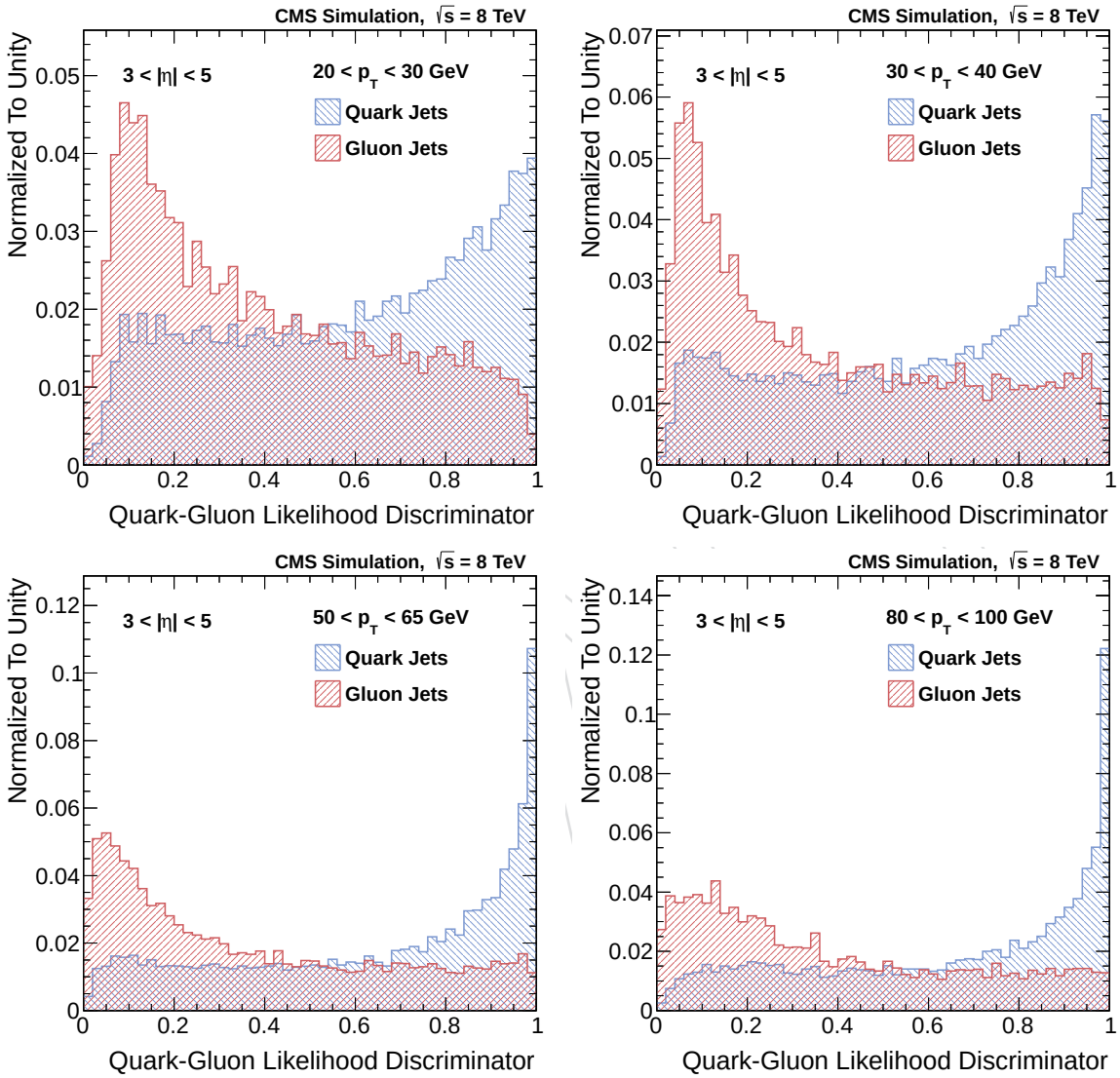

Figure 4.5: Normalized distributions of the Likelihood Quark-Gluon discriminant for quark (blue) and gluon (red) jets reconstructed in the forward part of the detector ($3 < |\eta| < 5$). Four representative transverse momentum bins are shown: $20 < p_{\mathrm{T}} < 30$ GeV (top left), $30 < p_{\mathrm{T}} < 40$ GeV (top right), $50 < p_{\mathrm{T}} < 65$ GeV (bottom left) and $80 < p_{\mathrm{T}} < 100$ GeV (bottom right).

Figures 4.4 and 4.5 show, respectively for the central and forward rapidity regions, normalized Likelihood distributions for quark and gluon jets in the simulation, in four representative transverse momentum bins: $30 < p_{\mathrm{T}} < 40$ GeV, $80 < p_{\mathrm{T}} < 100$ GeV, $200 < p_{\mathrm{T}} < 250$ GeV and $500 < p_{\mathrm{T}} < 600$ GeV in the central region, $20 < p_{\mathrm{T}} < 30$ GeV, $30 < p_{\mathrm{T}} < 40$ GeV, $50 < p_{\mathrm{T}} < 65$ GeV and $80 < p_{\mathrm{T}} < 80$ GeV in the forward region.

MLP Discriminant

The MLP discriminant is built from the MLP method, the recommended neural network in the TMVA framework, using four variables: nPFCand, axis_1, axis_2 and $p_T D$. This method is able to treat correlations and has good robustness against overtraining, but it has little transparency.

The neural network is trained on generated dijet events, which were divided in two rapidity regions: a central region ($|\eta| < 2.5$)) and a forward region ($2.5 < |\eta| < 4.7$). To account for the strong dependence of the variables as a function of the jet transverse momentum, a separate MLP training was done for each of 20 p_{T} bins, logarithmically spaced between 20 GeV and 600 GeV. In the forward region, only 18 bins were used covering a range between 20 GeV and 427 Gev. For every p_{T} bin, the dependence of the variables as a function of the amount of PU activity in the event is measured by a linear fit of the 4 variables versus ρ. This fit is done in the 2 GeV $< \rho <$ 30 GeV range, as shown in figure 4.6. The four variables are then corrected for this dependence before passing them on to the MLP training phase.

A drawback with the MLP approach are the different shapes and ranges of the response distributions in the various p_{T} bins and rapidity regions. Therefore the MLP Quark-Gluon discriminant uses the classification probability, also provided within the TMVA framework, in order to have a discriminating value which can be interpreted in the same way for all p_{T} bins.

In the application phase, the four variables are again corrected for the ρ dependence. The MLP classification probability is retrieved from the training bins which have their average jet p_{T} just above and below the tagging jet p_{T}. The final MLP Quark-Gluon discriminating value is then the linear interpolation of these two outputs.

Figures 4.7 and 4.8 show, respectively for the central and forward rapidity regions, normalized distributions of the MLP Quark-Gluon discriminant for quark and gluon jets in the simulation, in four representative transverse momentum bins: $30 < p_{\mathrm{T}} < 40$ GeV, $80 < p_{\mathrm{T}} < 100$ GeV, $200 < p_{\mathrm{T}} < 250$ GeV and $500 < p_{\mathrm{T}} < 600$ GeV in the central region, $20 < p_{\mathrm{T}} < 30$ GeV, $30 < p_{\mathrm{T}} < 40$ GeV, $50 < p_{\mathrm{T}} < 65$ GeV and $80 < p_{\mathrm{T}} < 80$ GeV in the forward region.

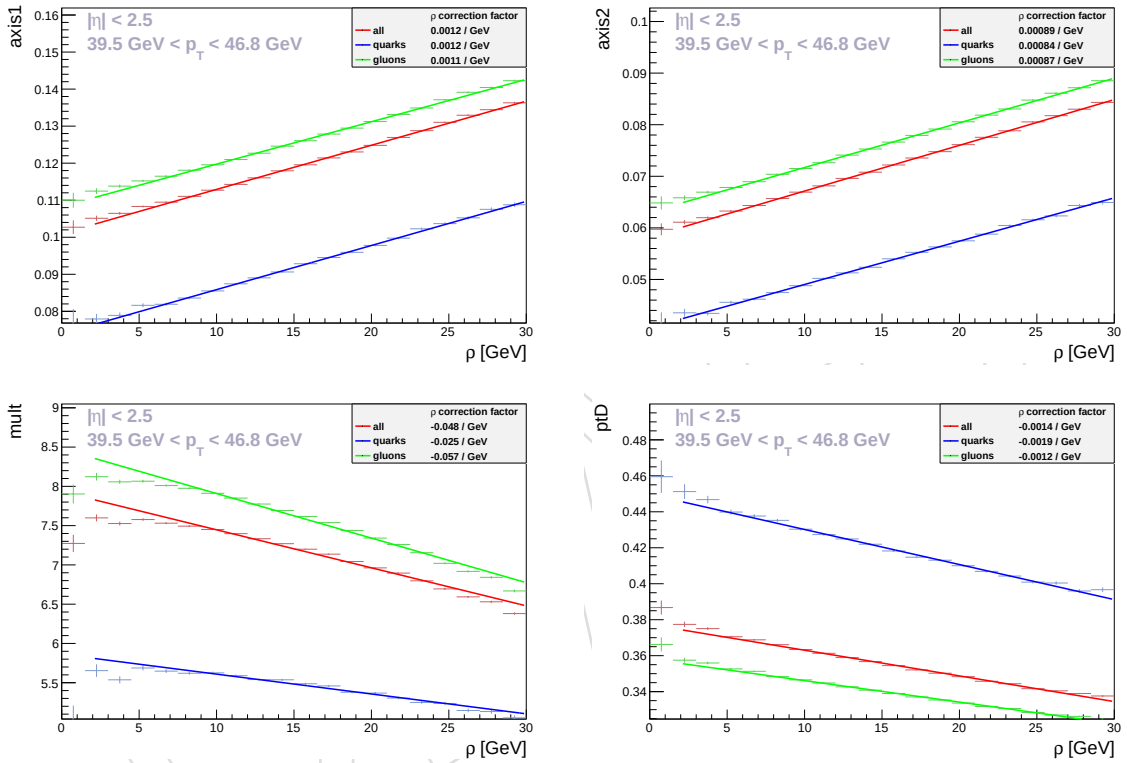

Figure 4.6: Linear fit of the ρ dependence for the axis_1 (top left), axis_2 (top right), nPFCand (bottom left) and $p_T D$ variables (bottom right) in one selected p_T bin. The correction factor is extracted from the fit to the complete sample, the separate gluon and quark fits are shown as reference.

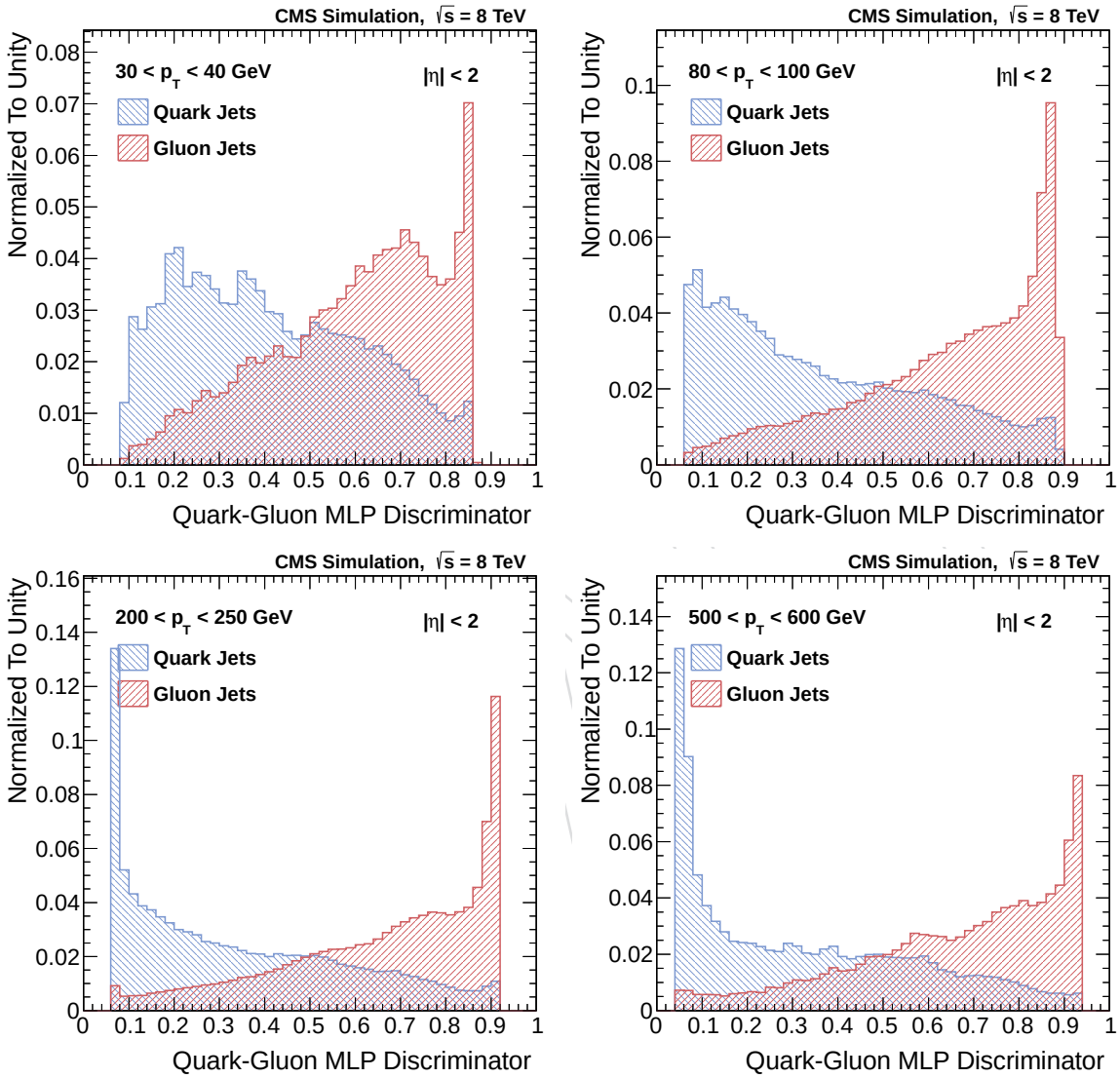

Figure 4.7: Normalized distributions of the MLP Quark-Gluon discriminant for quark (blue) and gluon (red) jets reconstructed in the central part of the detector ($|\eta| < 2$). Four representative transverse momentum bins are shown: $30 < p_{\mathrm{T}} < 40$ GeV (top left), $80 < p_{\mathrm{T}} < 100$ GeV (top right), $200 < p_{\mathrm{T}} < 250$ GeV (bottom left) and $500 < p_{\mathrm{T}} < 600$ GeV (bottom right).

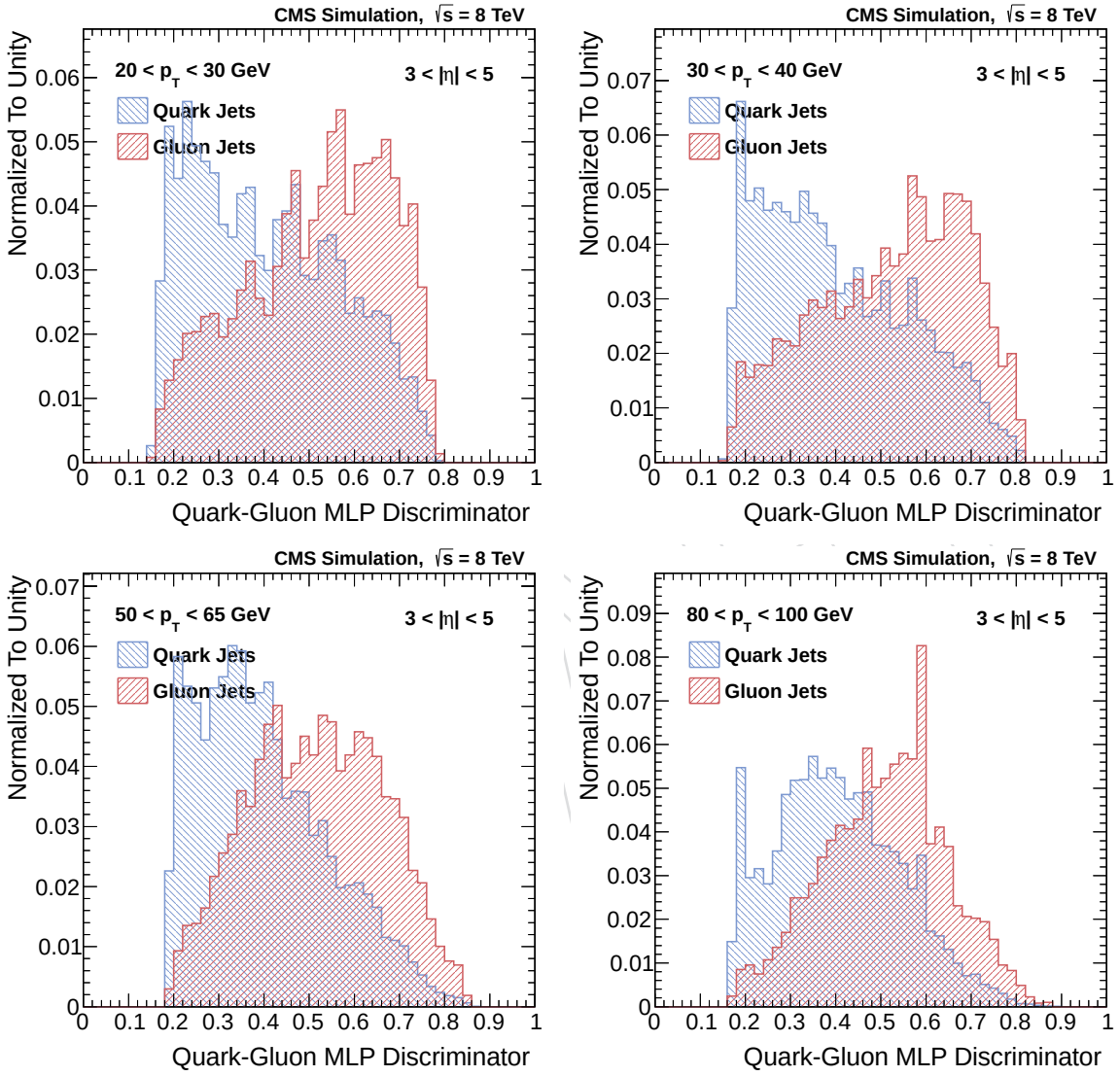

Figure 4.8: Normalized distributions of the MLP Quark-Gluon discriminant for quark (blue) and gluon (red) jets reconstructed in the forward part of the detector ($3 < |\eta| < 5$). Four representative transverse momentum bins are shown: $20 < p_{\mathrm{T}} < 30$ GeV (top left), $30 < p_{\mathrm{T}} < 40$ GeV (top right), $50 < p_{\mathrm{T}} < 65$ GeV (bottom left) and $80 < p_{\mathrm{T}} < 100$ GeV (bottom right).

4.3.1 Performance Comparison

A comparison of the perforance of the two taggers is presented, in terms of ROC curves built on the simulation, in Figures 4.9, 4.10 and 4.11, respectively for central rapidity ($|\eta| < 2$), transition ($2 < |\eta| < 2.5$) and the forward ($3 < |\eta| < 5$), in four representative transverse momentum intervals. The new likelihood discriminant is shown with hollow brown markers and it is compared to the MLP, which is shown with solid green markers. In the tracker covered region, the two are further compared to the old likelihood discriminant, shown in yellow.

The following conclusions can be made:

- both the new likelihood and the MLP increase the discrimination performance of the old tagger, across the phase space;
- in the central part of the detector, the two new taggers have similar performance at low p_T, but the MLP's ability to handle correlations gives it an edge for high p_T (> 200 GeV) jets;
- discrimination has been successfully extended to the forward region of the detector;
- in the forward, the two taggers again have similar performance, but this time the likelihood discriminant seems to lead to better discrimination.

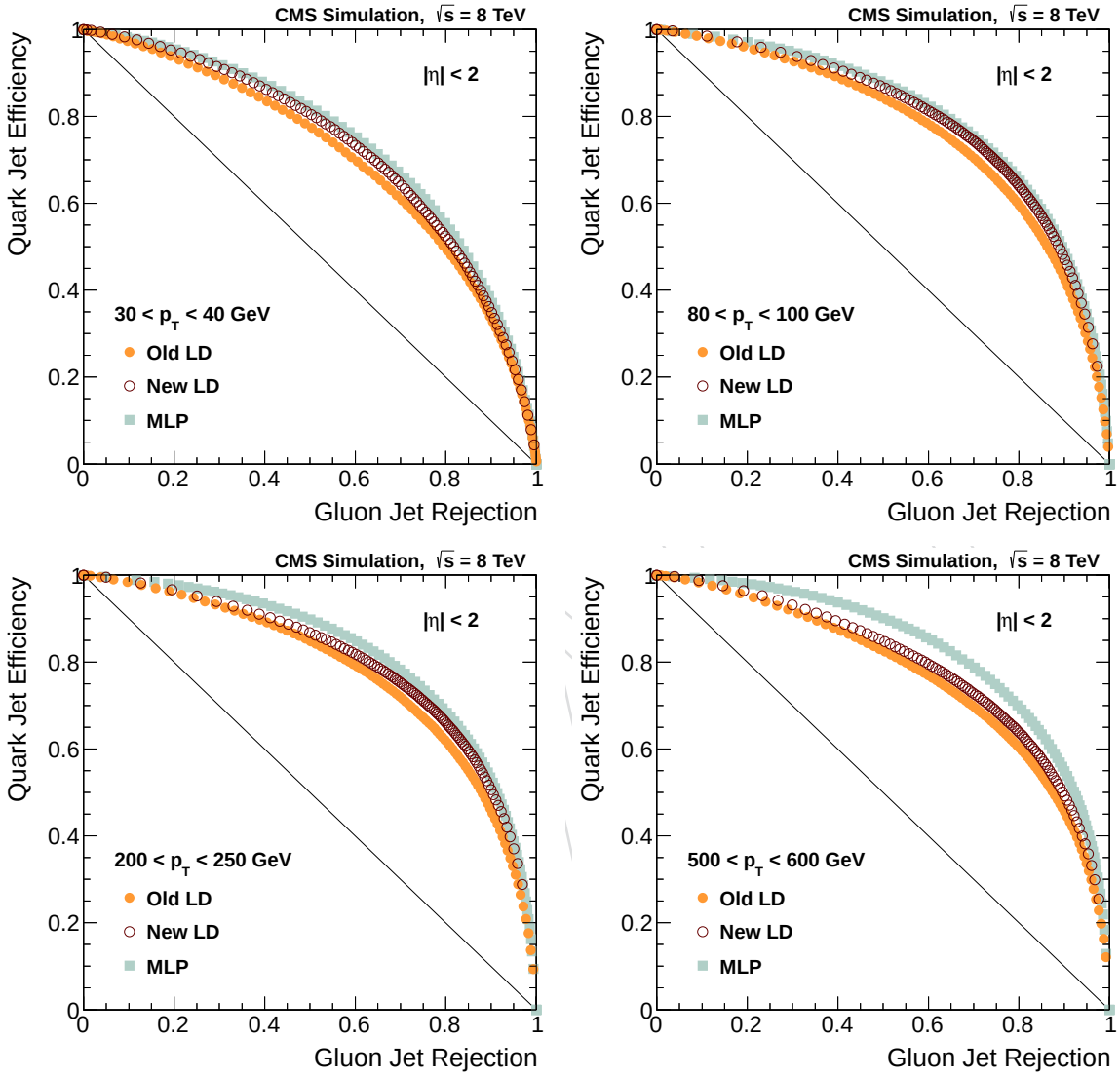

Figure 4.9: Receiver operating characteristic (ROC) curves for jets reconstructed in the central part of the detector ($|\eta| < 2$), in four representative transverse momentum intervals: $30 < p_{\mathrm{T}} < 40$ GeV (top left), $80 < p_{\mathrm{T}} < 100$ GeV (top right), $200 < p_{\mathrm{T}} < 250$ GeV (bottom left) and $500 < p_{\mathrm{T}} < 600$ GeV (bottom right). The likelihood discriminator (open brown markers) is compared to the MLP discriminant (green markers). The two are further compared to the old (2011) likelihood discriminant (yellow markers).

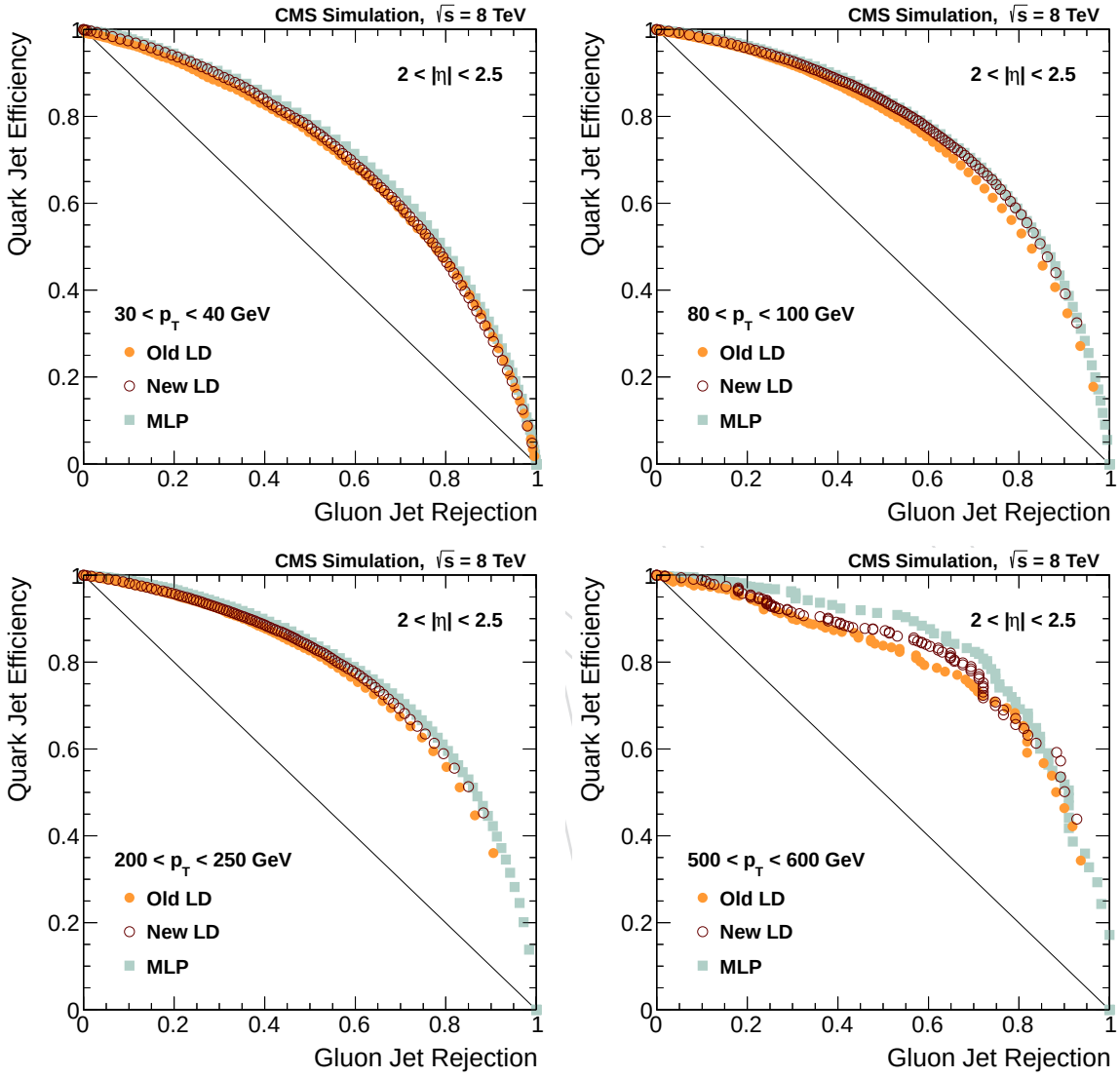

Figure 4.10: Receiver operating characteristic (ROC) curves for jets reconstructed in the detector transition region ($2 < |\eta| < 2.5$), in four representative transverse momentum intervals: $30 < p_{\mathrm{T}} < 40$ GeV (top left), $80 < p_{\mathrm{T}} < 100$ GeV (top right), $200 < p_{\mathrm{T}} < 250$ GeV (bottom left) and $500 < p_{\mathrm{T}} < 600$ GeV (bottom right). The likelihood discriminator (open brown markers) is compared to the MLP discriminant (green markers). The two are further compared to the old (2011) likelihood discriminant (yellow markers).

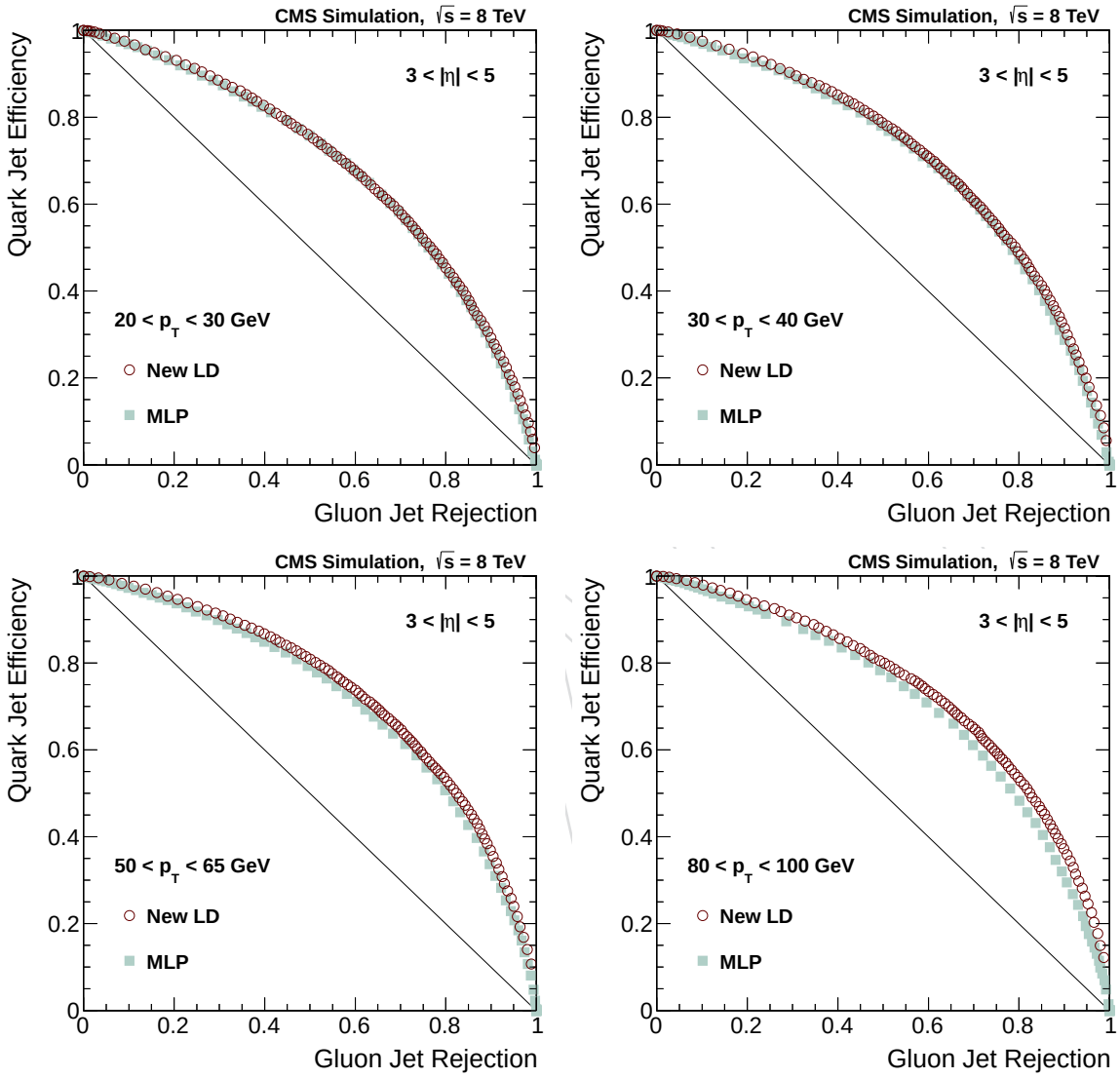

Figure 4.11: Receiver operating characteristic (ROC) curves for jets reconstructed in the detector forward ($3 < |\eta| < 5$), in four representative transverse momentum intervals: $20 < p_{\mathrm{T}} < 30$ GeV (top left), $30 < p_{\mathrm{T}} < 40$ GeV (top right), $50 < p_{\mathrm{T}} < 65$ GeV (bottom left) and $80 < p_{\mathrm{T}} < 100$ GeV (bottom right). The likelihood discriminator (open brown markers) is compared to the MLP discriminant (green markers).

5 Pythia6/Herwig++ Comparisons

As most of the results presented in this note are based on events simulated in Pythia 6, we here want to investigate how different things would look like by using a different parton shower/hadronization model. In order to do so we compare Pythia6 with Herwig++. This will become even more relevant in the following validation Sections, as it is observed than under some kinematical regimes Herwig++ seems to describe the data in a more satisfactory manner.

All of the results presented in this Section are based on the QCD 'flat' samples, namely:

Pythia6: `/QCD_Pt-15to3000_TuneZ2star_Flat_8TeV_pythia6`

Herwig++: `/QCD_Pt-15to3000_TuneEE3C_Flat_8TeV_herwigpp`

Both samples have been reconstructed in the same release (`CMSSW_5_3_X`) with the same pileup scenario (`S10`).

Comparisons between the two generators, for the four input variables, in three representative generator jet transverse momentum / rapidity bins, can be found in Figures 5.1, 5.2, 5.3 and 5.4. As can be seen, for all variables, the observed trend is that quarks have very similar hadronization properties in the two generators (at least for what can be probed by the considered variables), whereas there is a significant difference in gluons. Gluon jets in Herwig++ are much more similar to quarks, compared to Pythia. This is compatible with what has been reported in [11]. This is true for all intervals of transverse momentum and rapidity, we show here some bins just for the sake of conciseness.

This means that, if nature hadronizes like Herwig++, the discriminators presented in this note would have worse discrimination. This is shown in Figure 5.5, where the quark-gluon likelihood receiver operating characteristic obtained on Pythia QCD (open brown markers) is compared to the one obtained on Herwig++ (solid yellow markers), for three transverse momenta in the centerand one in the forward. As can be seen, for a quark efficiency, Herwig++ obtaines 15-20% worse gluon rejection.

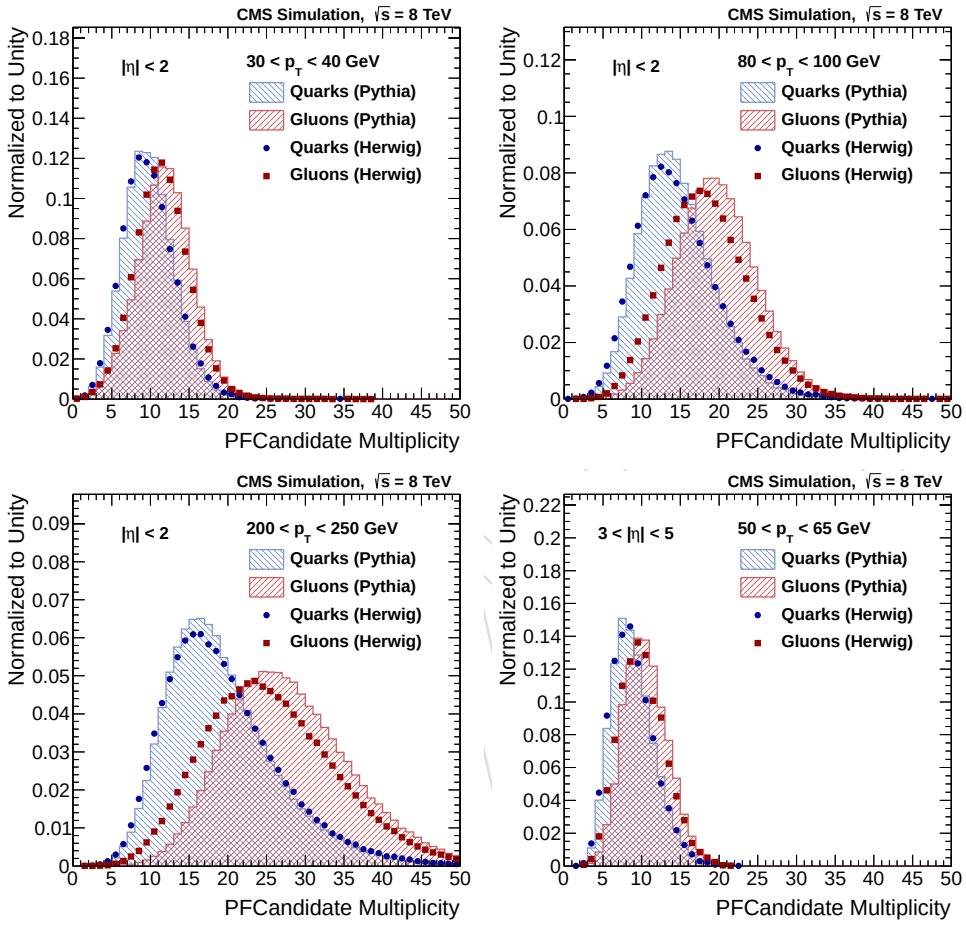

Figure 5.1: Comparison between Pythia and Herwig++ for the total PFCandidate multiplicity: Pythia distributions are shown with hashed histograms, Herwig++ ones with markers. Quark jets are shown in blue, gluon jets in red. Four phase space regions are shown: central jets with transverse momentum between 30 and 40 GeV (top left), 80 and 100 GeV (top right), 200 and 250 GeV (bottom left) and forward jets with transverse momentum between 50 and 65 GeV (bottom right).

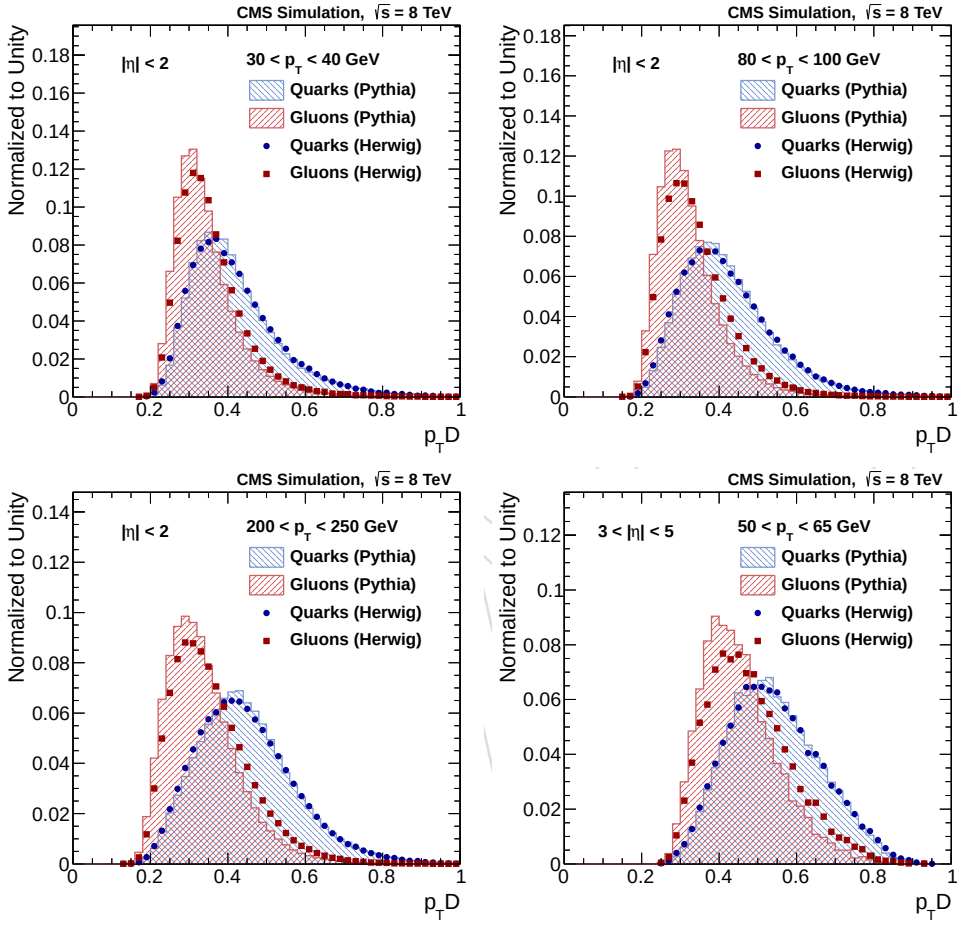

Figure 5.2: Comparison between Pythia and Herwig++ for $p_T D$: Pythia distributions are shown with hashed histograms, Herwig++ ones with markers. Quark jets are shown in blue, gluon jets in red. Four phase space regions are shown: central jets with transverse momentum between 30 and 40 GeV (top left), 80 and 100 GeV (top right), 200 and 250 GeV (bottom left) and forward jets with transverse momentum between 50 and 65 GeV (bottom right).

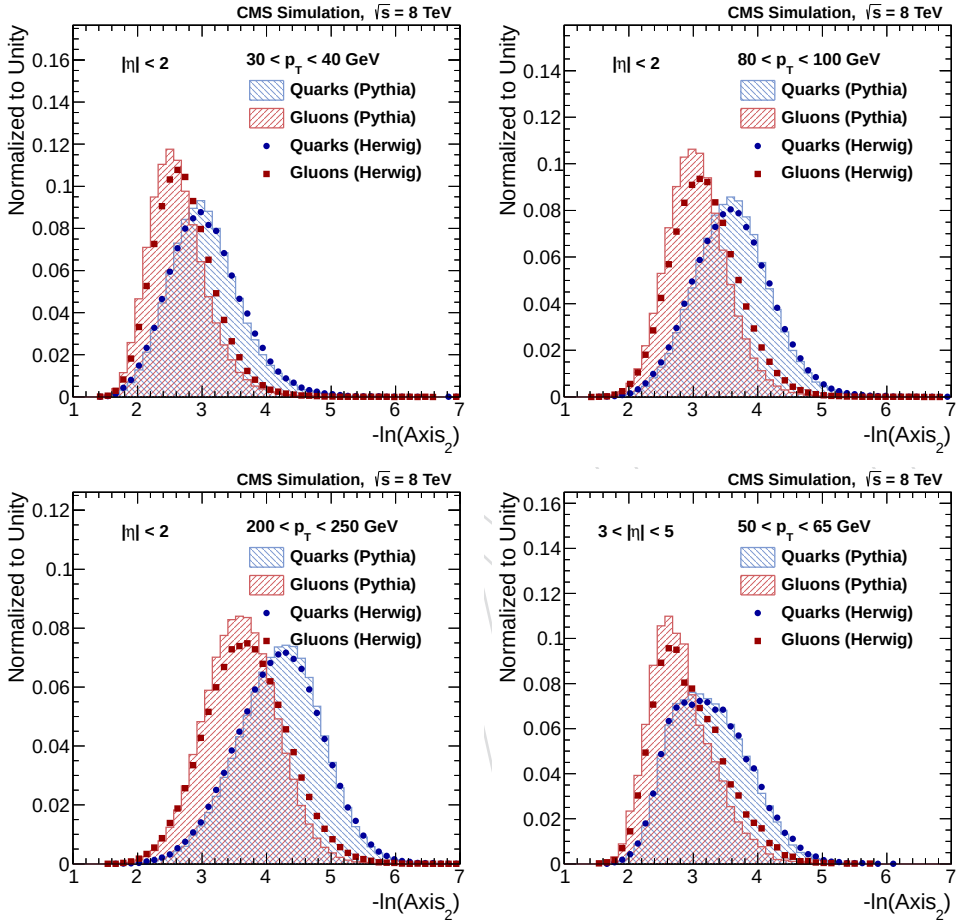

Figure 5.3: Comparison between Pythia and Herwig++ for Axis_2: Pythia distributions are shown with hashed histograms, Herwig++ ones with markers. Quark jets are shown in blue, gluon jets in red. Four phase space regions are shown: central jets with transverse momentum between 30 and 40 GeV (top left), 80 and 100 GeV (top right), 200 and 250 GeV (bottom left) and forward jets with transverse momentum between 50 and 65 GeV (bottom right).

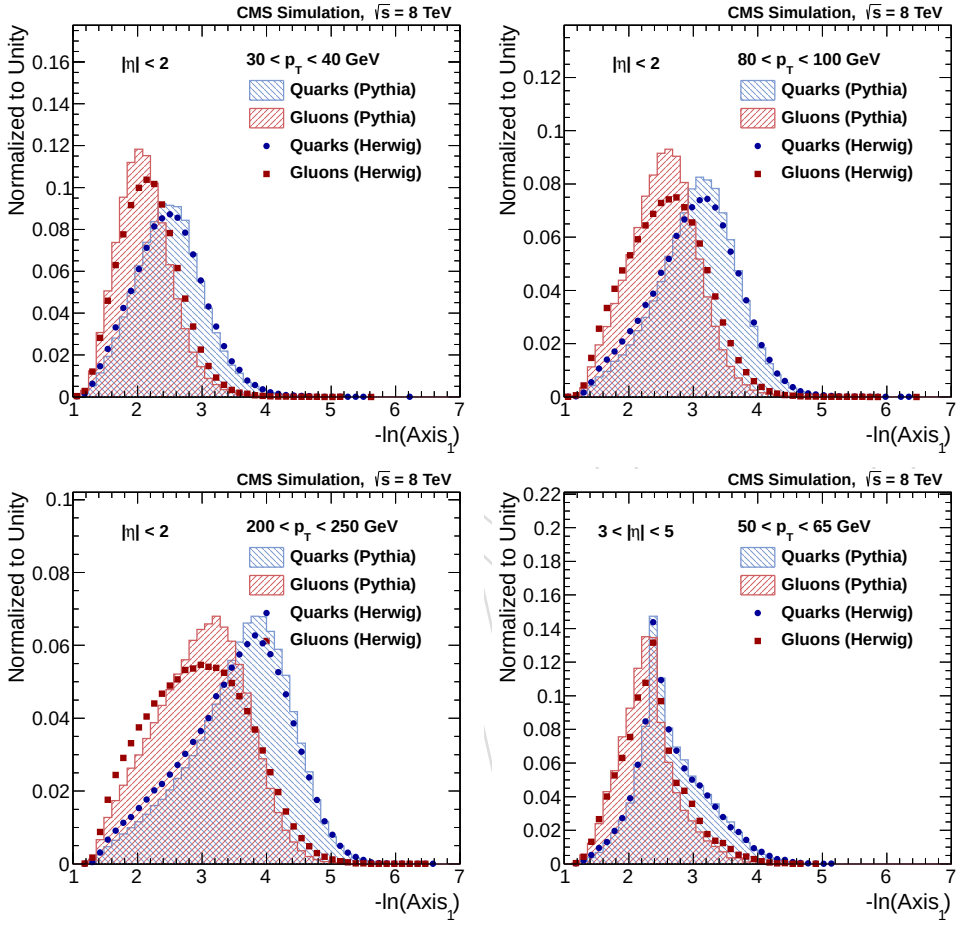

Figure 5.4: Comparison between Pythia and Herwig++ for Axis_1: Pythia distributions are shown with hashed histograms, Herwig++ ones with markers. Quark jets are shown in blue, gluon jets in red. Four phase space regions are shown: central jets with transverse momentum between 30 and 40 GeV (top left), 80 and 100 GeV (top right), 200 and 250 GeV (bottom left) and forward jets with transverse momentum between 50 and 65 GeV (bottom right).

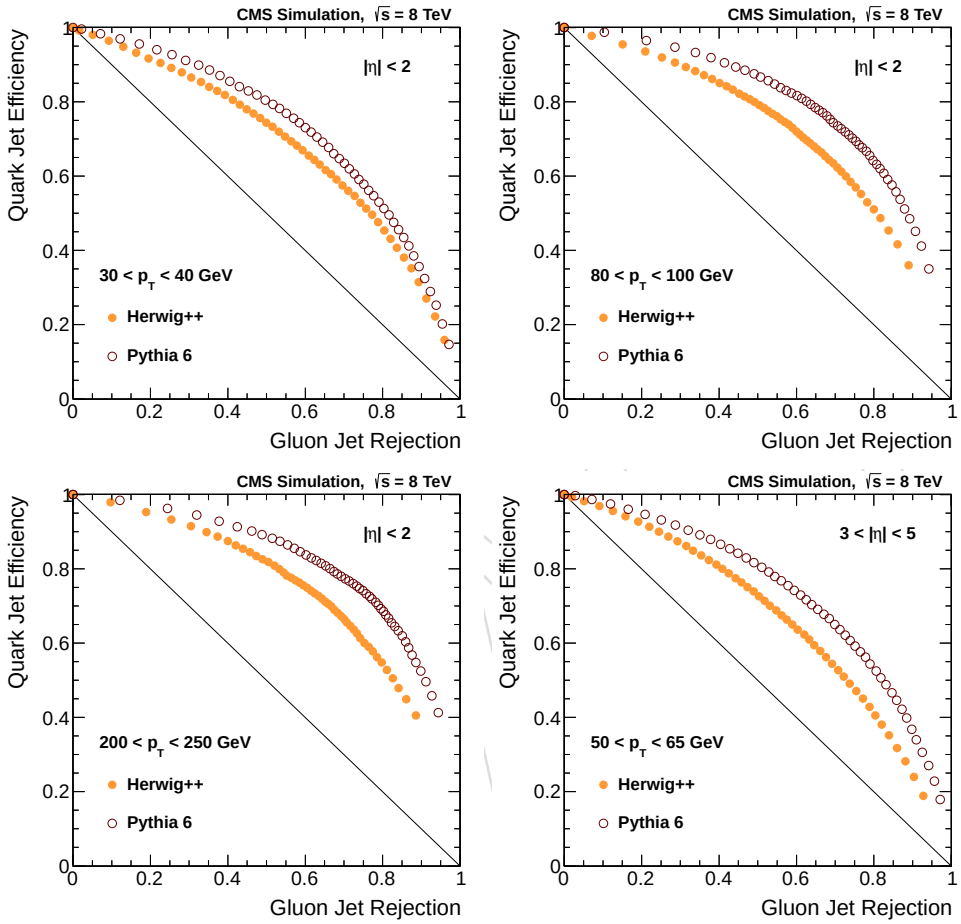

Figure 5.5: Performance, in terms of receiver operating characteristics on the quark efficiency / gluon rejection plane, of the likelihood quark-gluon discriminator for Pythia6 (hollow brown markers) and Herwig++ (yellow markers) jets. Four phase space regions are shown: central jets with transverse momentum between 30 and 40 GeV (top left), 80 and 100 GeV (top right), 200 and 250 GeV (bottom left) and forward jets with transverse momentum between 50 and 65 GeV (bottom right).

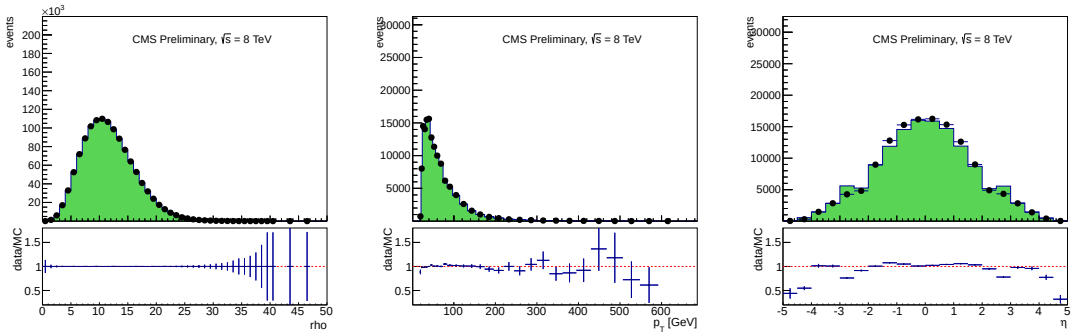

Figure 6.1: Data-MC comparisons for the particle-flow energy density ρ (left), the leading jet transverse momentum (center) and its pseudorapidity (right). The reweighting procedure is defined by the ρ distribution, so it is expected to find perfect agreement there.

6 Validation on data

We will now proceed to validate the discriminators on 8 TeV collision data. In order to do this, two main samples have been identified:

- **Z+jet events**, which are quark-enriched;
- **dijet events**, which are gluon-enriched.

By the simultaneous use of these two control samples, we can verify the functioning of the disciminators on both parton flavors, and across the whole phase space. The following subsections detail the event selection and the obtained results in these two control samples. In what follows, all Monte Carlo distributions are normalized to the integral of the data, as we are interested only in shape comparisons.

6.1 Validation on Z+jet Events

The Z+jet control sample offers a relatively pure quark set of jets, as can be seen in Figure **??**, where the fraction of quark (blue), gluon (red), pile up (yellow) and undefined (grey) leading jets are shown as a function of the hard scattering transverse momentum ($\hat{p}_T$). As can be seen, 80% or more of leading jets in Z+jet events originate from light quark hadronization, across the transverse momentum spectrum.

We choose to use only events in which the Z boson has decayed to muons, as they are less sensitive to pile up effects and therefore provide the cleanest sample. We plan to include also electronic decays of the Z soon. In order to select Z+jet events in data, we run on the full 2012 dataset recorded with the dimuon HLT path `HLT_Mu17_Mu8`, and compare it to the Madgraph Drell-Yan sample (`/DYJetsToLL_M-50_TuneZ2Star_8TeV-madgraph-tarball`) produced in the Summer12 production with pile up profile S10. The event selection further requires:

- the presence of two tight muons of opposite charge;
- the transverse momenta of the muons are required to be greater than 20 and 10 GeV respectively;
- the dimuon invariant mass is required to fall in the 70-110 GeV range;
- the subleading jet in the event is required to have a transverse momentum smaller than 30% that of the dimuon system.

The leading jet is considered, and as previously mentioned it is required to pass jet ID, anti-b-tagging, pile up ID. In order to compare the data and the simulation under similar pile up

conditions, a pile up reweighting procedure is enforced. This is done by simply reweighting the MC distribution of the particle flow energy density ρ to match the one observed in the data. This leads to a satisfactory comparison also for the leading jet's transverse momentum and pseudorapidity distributions, as is shown in Figure 6.1.

Figures 6.2, 6.3, 6.4 and 6.5 show, respectively, the data-MC comparisons of the four input variables (total multiplicity, $p_T D$, Axis_2 and Axis_1), in four representative p_{T} bins, for jets reconstructed in the center of the detector. The corresponding plots for the forward are shown in Figures 6.6, 6.8, 6.9 and 6.10. As can be seen in Figure 6.6, a significant discrepancy between data and the simulation is observed in the multiplicity of forward jets. As will be seen in the next Section, a similar discrepancy is observed in an orthogonal dataset (dijets). We therefore assume it is caused by a mismodeling of pile-up effects in the forward region. The fact that it is seen only in the multiplicity and is not so striking in the other variables is explained by the fact that the multiplicity variables is sensitive to first-order effects, whereas all other variables are sensitive only to higher-order corrections.

We therefore decide to correct the data multiplicity distributions before they are fed to the discriminators. This is similar, conceptually to what is done in L1 jet energy corrections, where different Offset corrections are employed in data and MC, in order to take into account possible discrepancies in pile up modeling. Data multiplicities for forward jets are therefore decreased by one unity. This leads to better compatibilty between the data and the simulation, as can be seen by comparing the distributions before (Figure 6.6) and after (Figure 6.7) the correction.

The corresponding data-MC comparisons for the output discriminants are shown in Figures 6.11 for the likelihood discriminant and in Figure 6.14 for the MLP. The comparisons for the forward discriminators are shown respectively in Figures 6.12 and 6.15 before the multiplicity correction, and in Figures 6.13 and 6.16 after the multiplicity correction. As can be seen significantly better agreement is observed after the application of the correction.

Overall, the level of agreement between the data and the simulation is satisfactory.

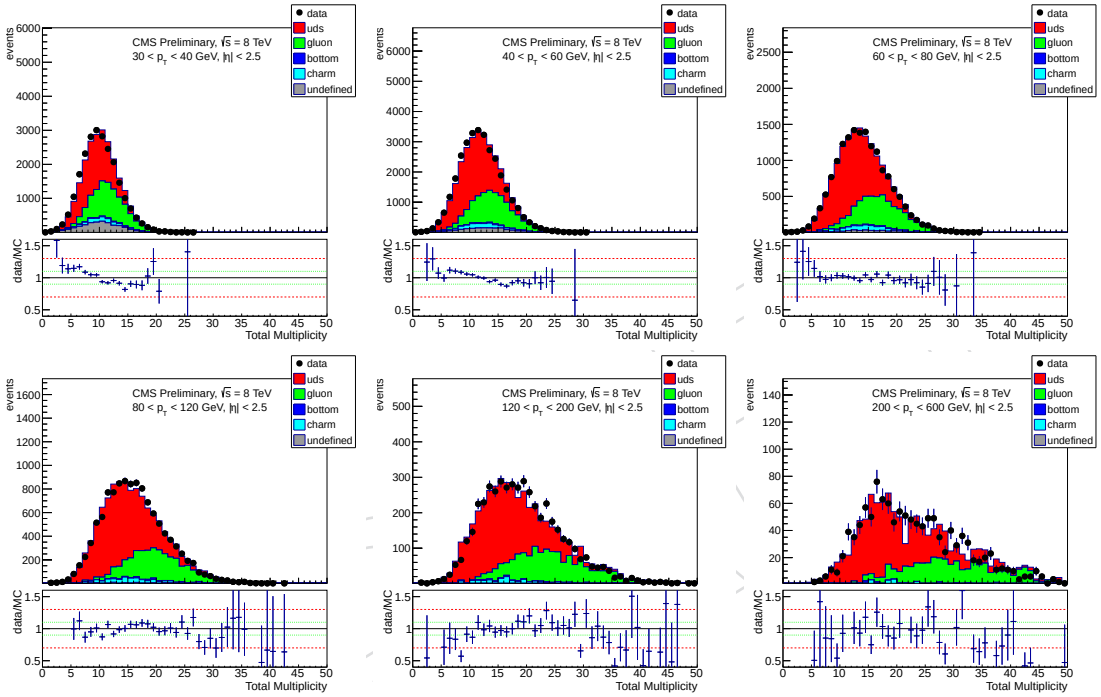

Figure 6.2: Data-MC comparisons for the total PFCandidate multiplicity in central jets in Z+jet events, for six p_{T} bins: 30-40, 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm$10% and $\pm$30%.

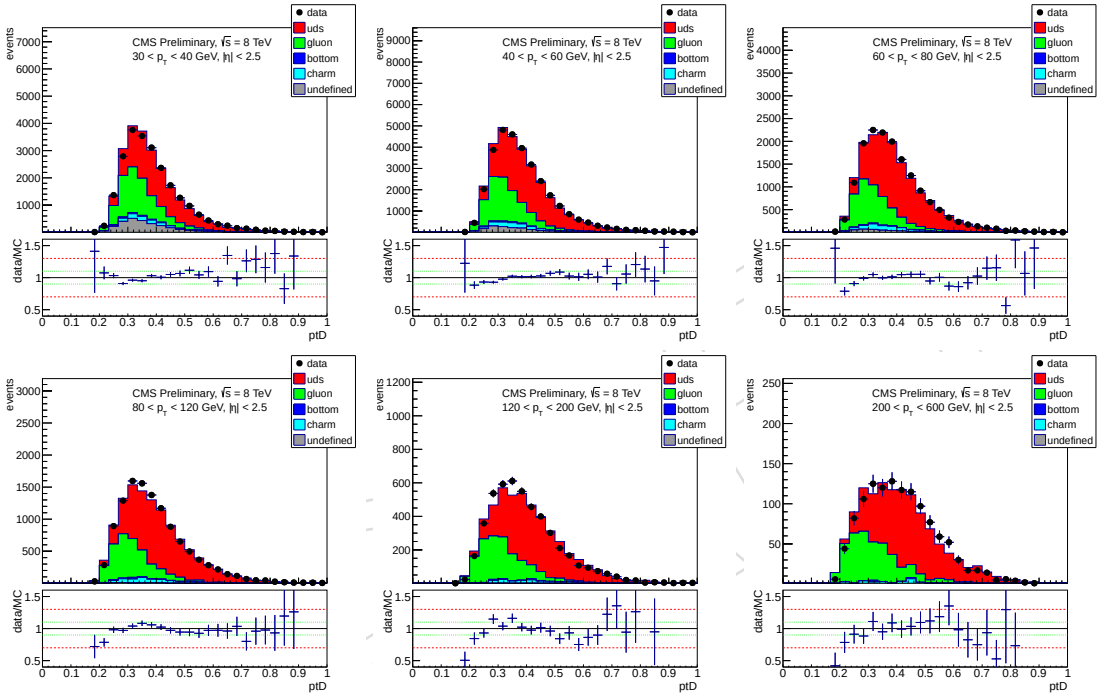

Figure 6.3: Data-MC comparisons for $p_T D$ in central jets in Z+jet events, for six p_{T} bins: 30-40, 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm 10\%$ and $\pm 30\%$.

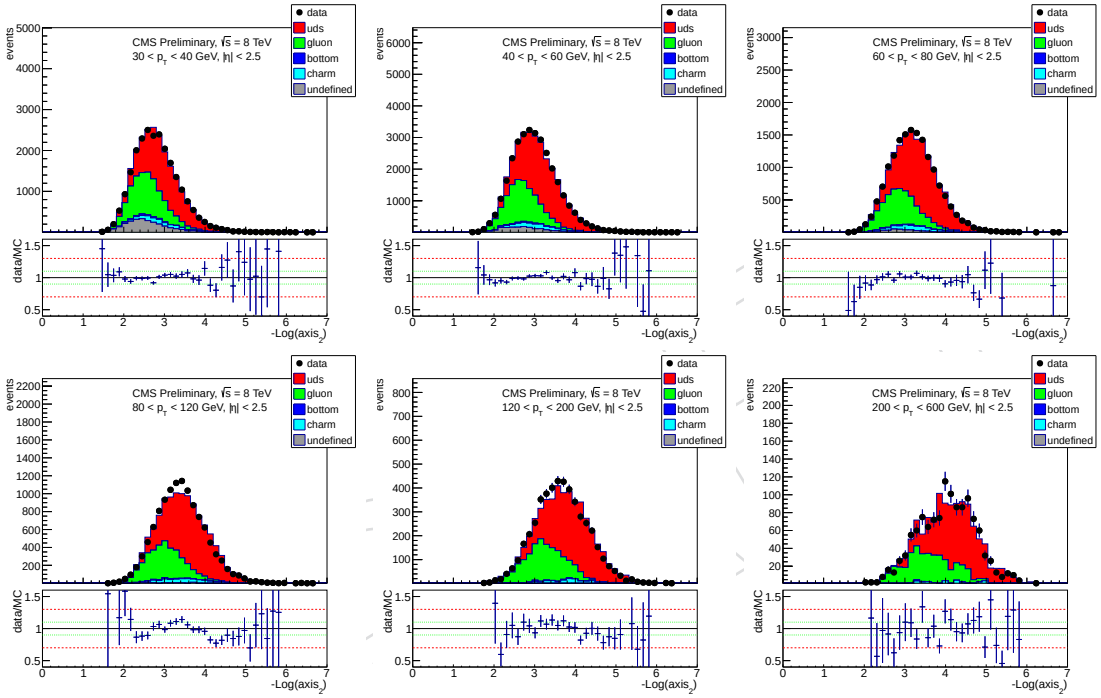

Figure 6.4: Data-MC comparisons for Axis_2 in central jets in Z+jet events, for six p_{T} bins: 30-40, 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm 10\%$ and $\pm 30\%$.

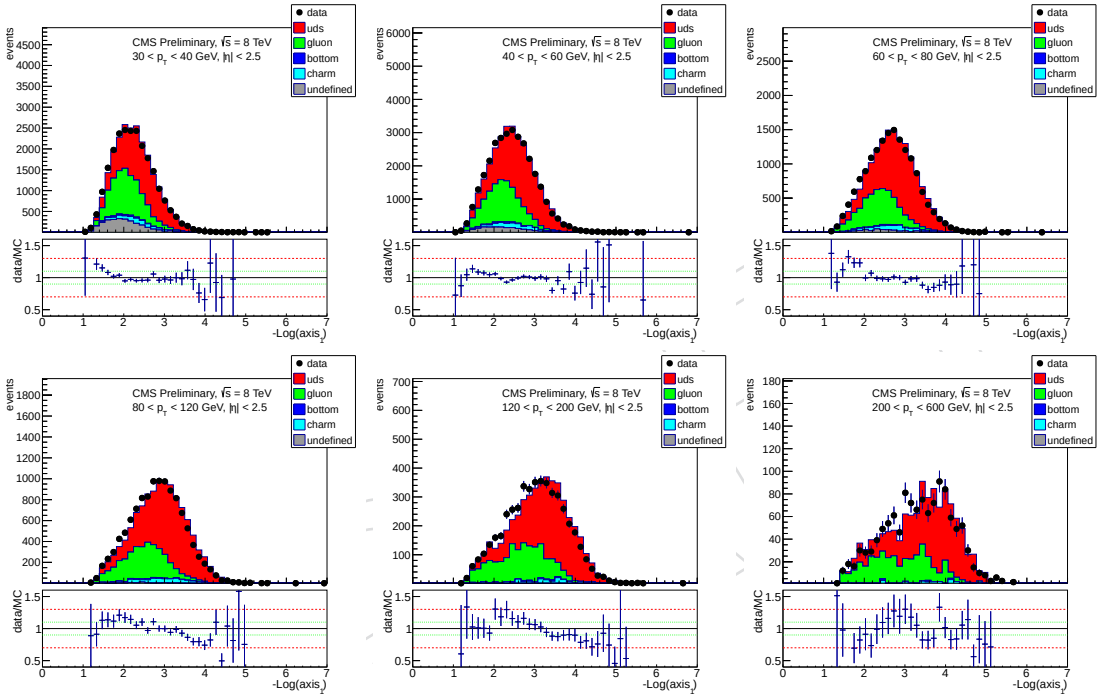

Figure 6.5: Data-MC comparisons for Axis_1 in central jets in Z+jet events, for six p_{T} bins: 30-40, 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm 10\%$ and $\pm 30\%$.

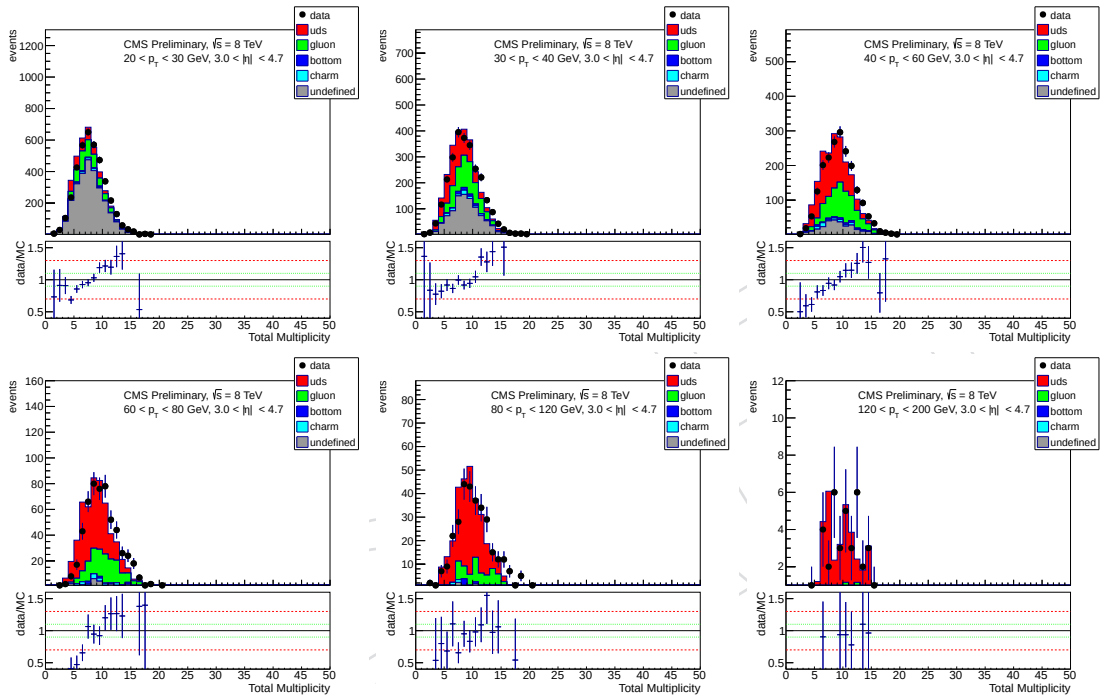

Figure 6.6: Data-MC comparisons for the total PFCandidate multiplicity in forward jets in Z+jet events **before the correction**, for six p_{T} bins: 20-30, 30-40, 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm$10% and $\pm$30%.

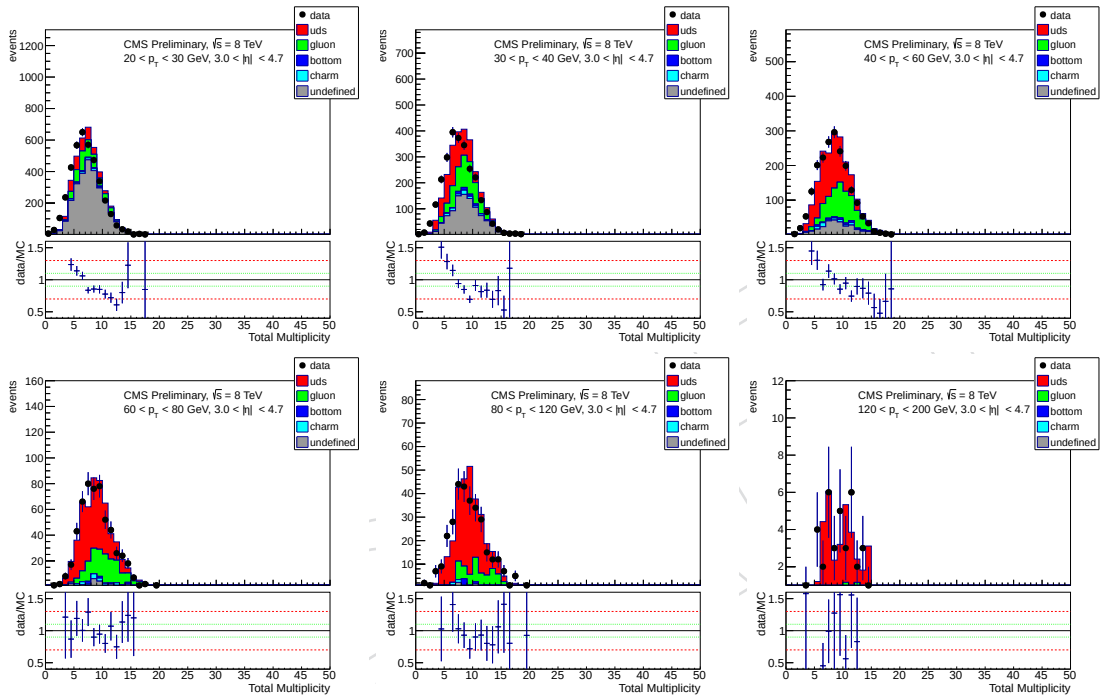

Figure 6.7: Data-MC comparisons for the total PFCandidate multiplicity in forward jets in Z+jet events **after the correction**, for six p_{T} bins: 20-30, 30-40, 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm$10% and $\pm$30%.

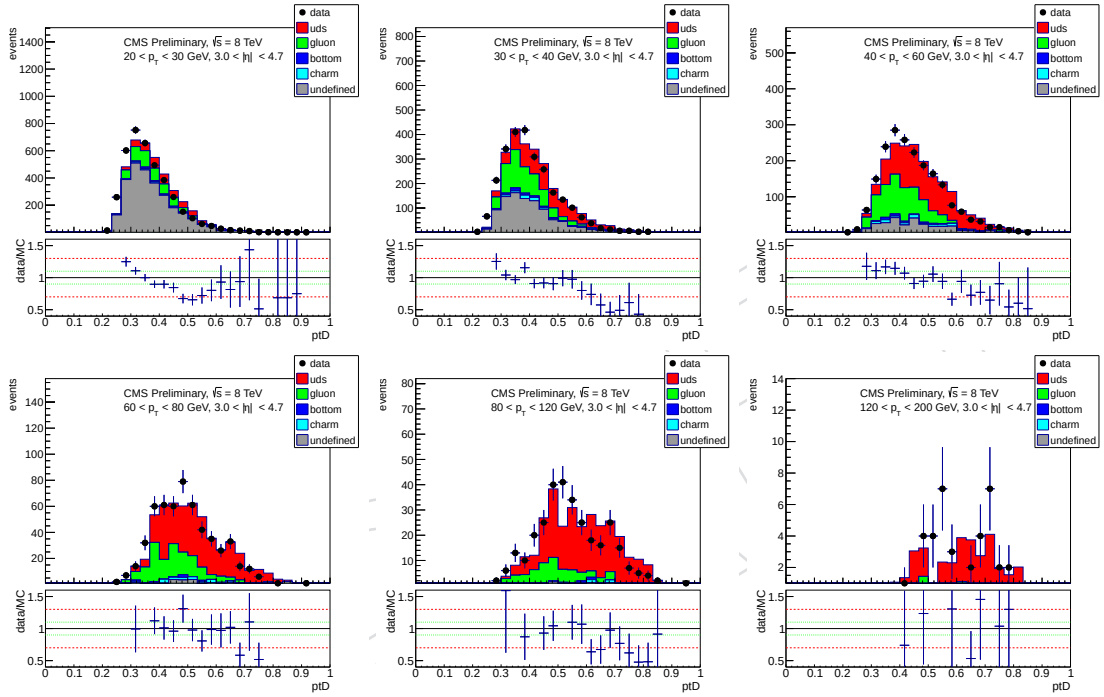

Figure 6.8: Data-MC comparisons for $p_T D$ in forward jets in Z+jet events, for six p_{T} bins: 20-30, 30-40, 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm 10\%$ and $\pm 30\%$.

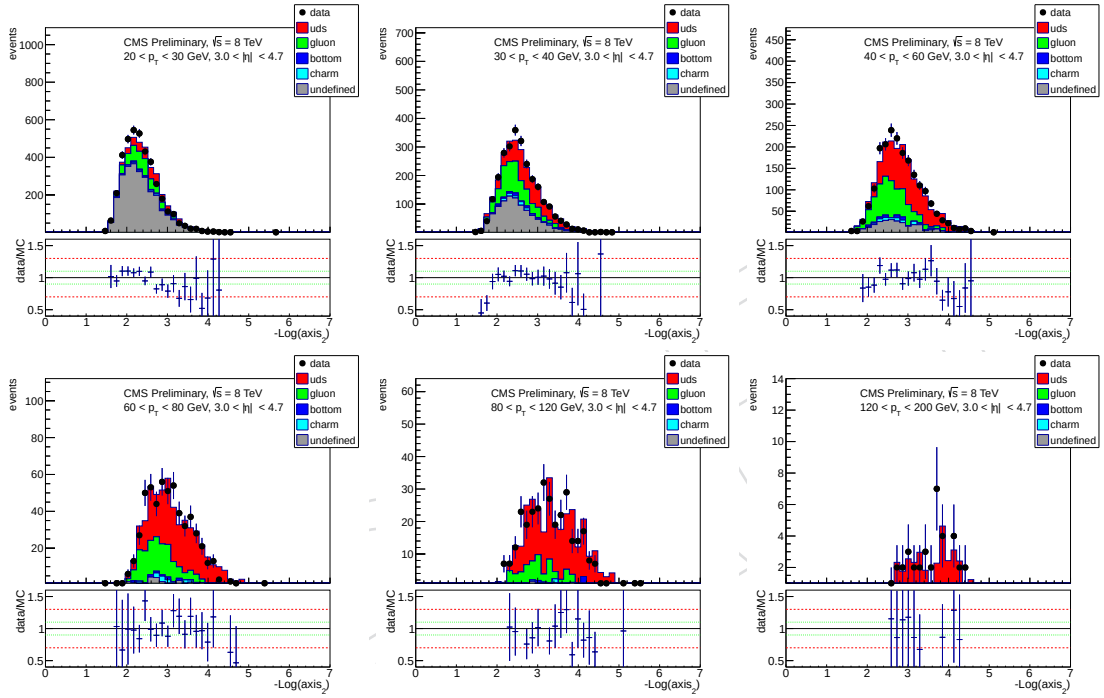

Figure 6.9: Data-MC comparisons for Axis_2 in forward jets in Z+jet events, for six p_{T} bins: 20-30, 30-40, 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm 10\%$ and $\pm 30\%$.

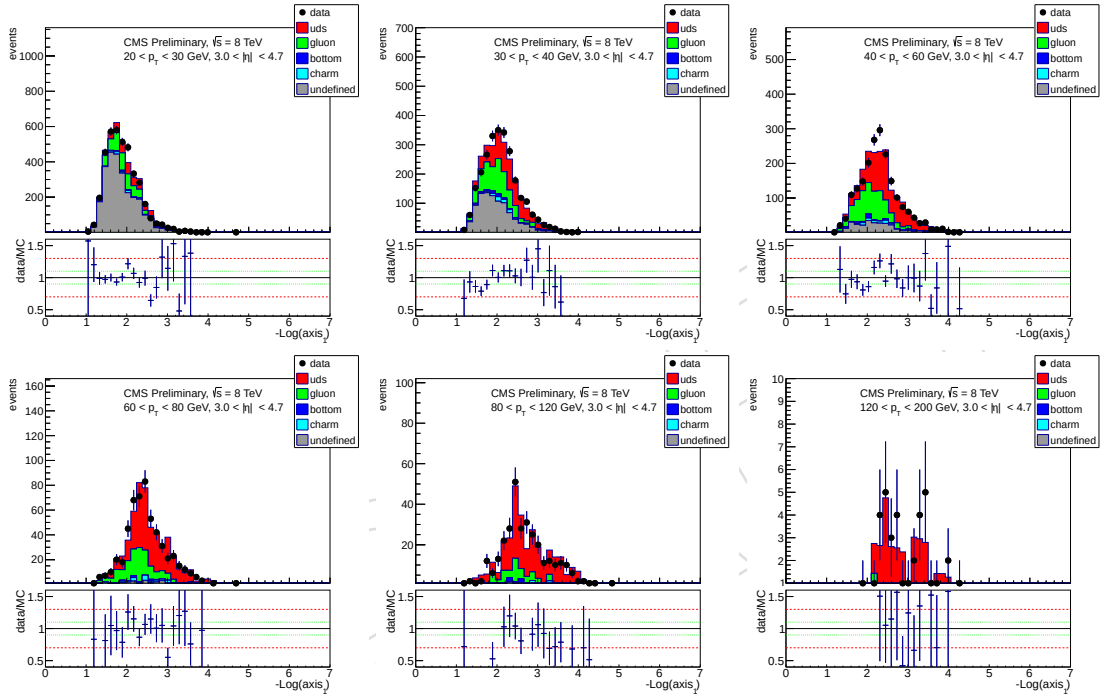

Figure 6.10: Data-MC comparisons for Axis_1 in forward jets in Z+jet events, for six p_T bins: 20-30, 30-40, 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm$10% and $\pm$30%.

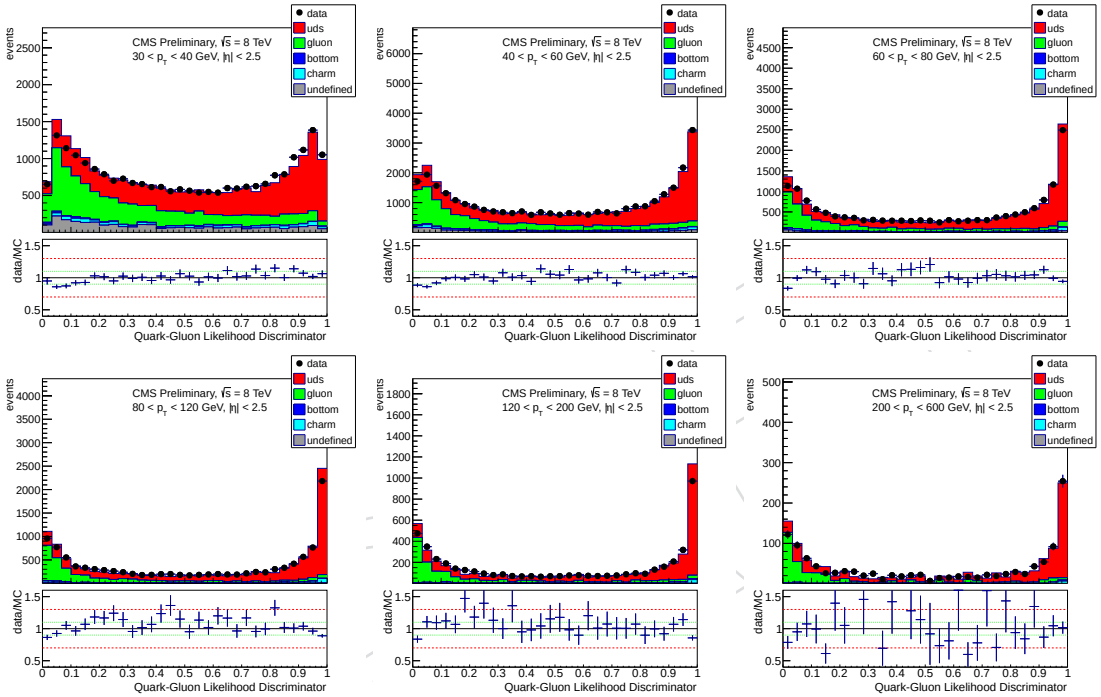

Figure 6.11: Data-MC comparisons for the quark-gluon likelihood discriminant in central jets in Z+jet events, for six p_{T} bins: 30-40, 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm$10% and $\pm$30%.

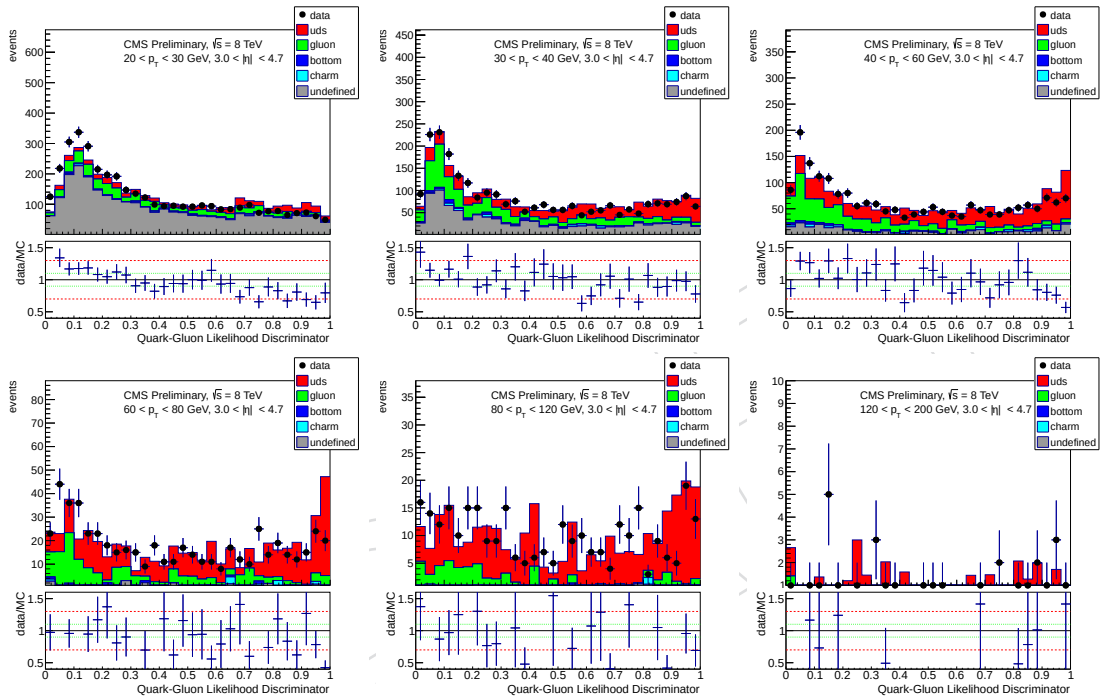

Figure 6.12: Data-MC comparisons for the quark-gluon likelihood discriminant in forward jets in Z+jet events **before the multiplicity correction**, for six p_{T} bins: 20-30, 30-40, 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm 10\%$ and $\pm 30\%$.

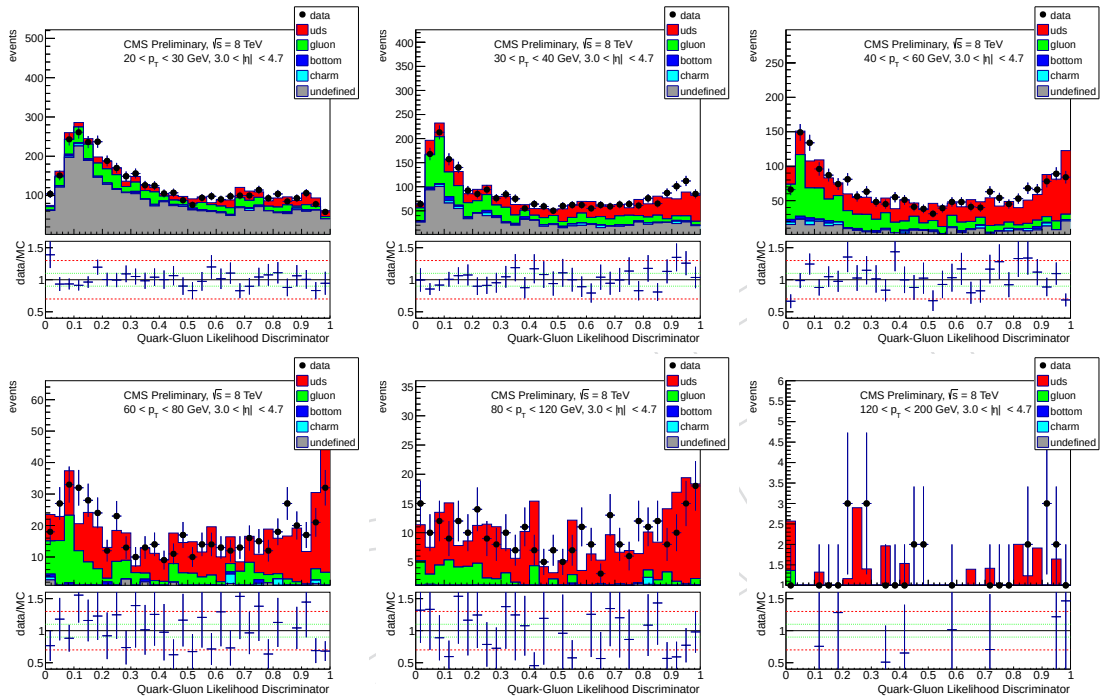

Figure 6.13: Data-MC comparisons for the quark-gluon likelihood discriminant in forward jets in Z+jet events **after the multiplicity correction**, for six p_{T} bins: 20-30, 30-40, 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm$10% and $\pm$30%.

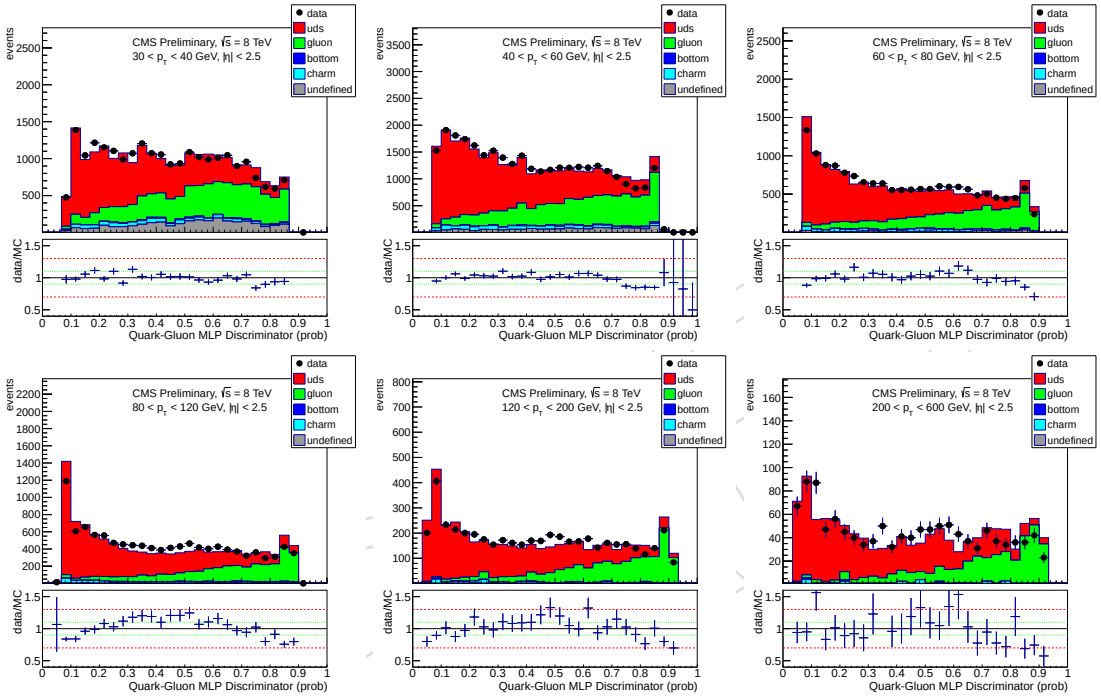

Figure 6.14: Data-MC comparisons for the quark-gluon MLP discriminant in central jets in Z+jet events, for six p_{T} bins: 30-40, 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm$10% and $\pm$30%.

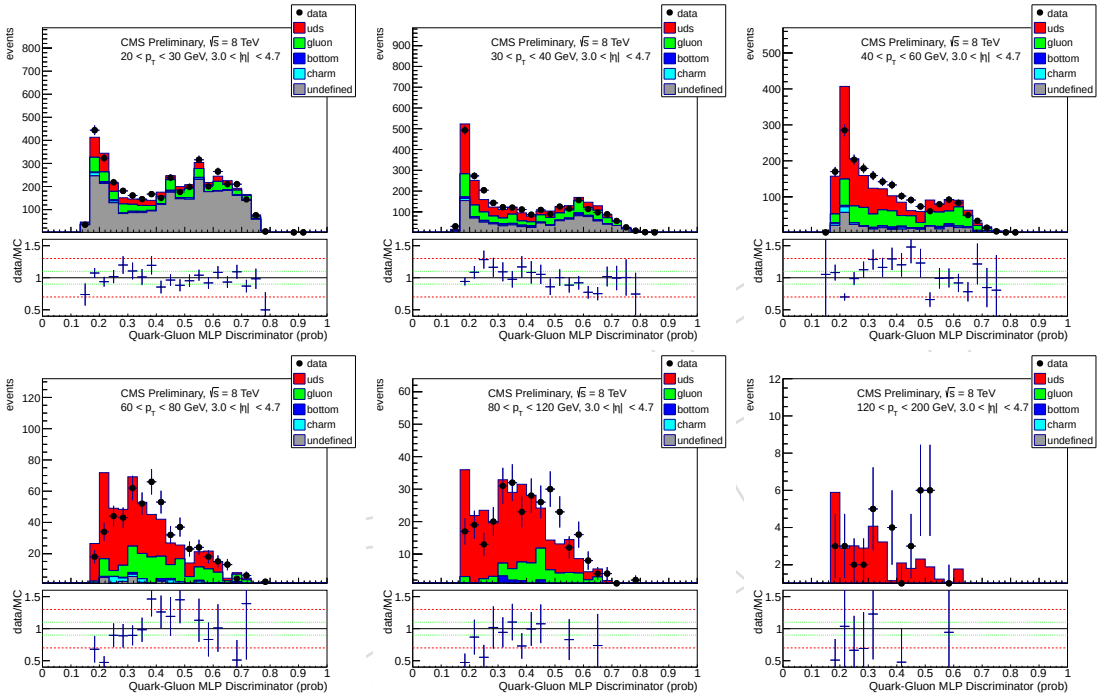

Figure 6.15: Data-MC comparisons for the quark-gluon MLP discriminant in forward jets in Z+jet events **before the multiplicity correction**, for six p_{T} bins: 20-30, 30-40, 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm$10% and $\pm$30%.

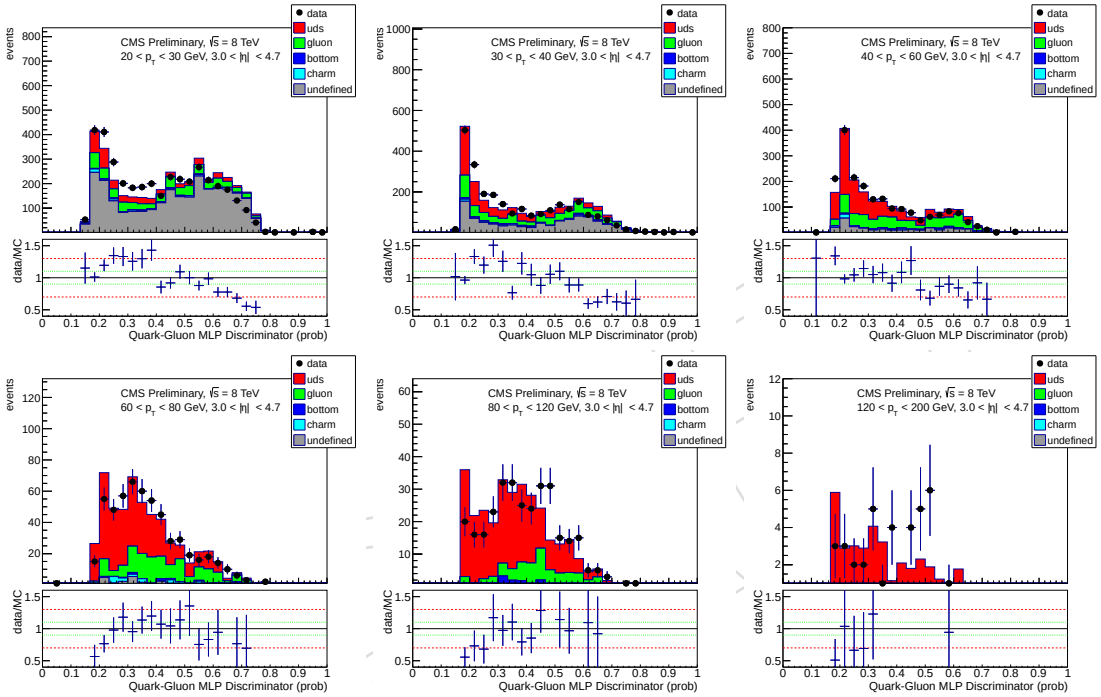

Figure 6.16: Data-MC comparisons for the quark-gluon MLP discriminant in forward jets in Z+jet events **after the multiplicity correction**, for six p_{T} bins: 20-30, 30-40, 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey). The data/MC ratio is showed on the bottom pad, where a solid black line marks unity, and the dotted green and red lines mark, respectively, $\pm$10% and $\pm$30%.

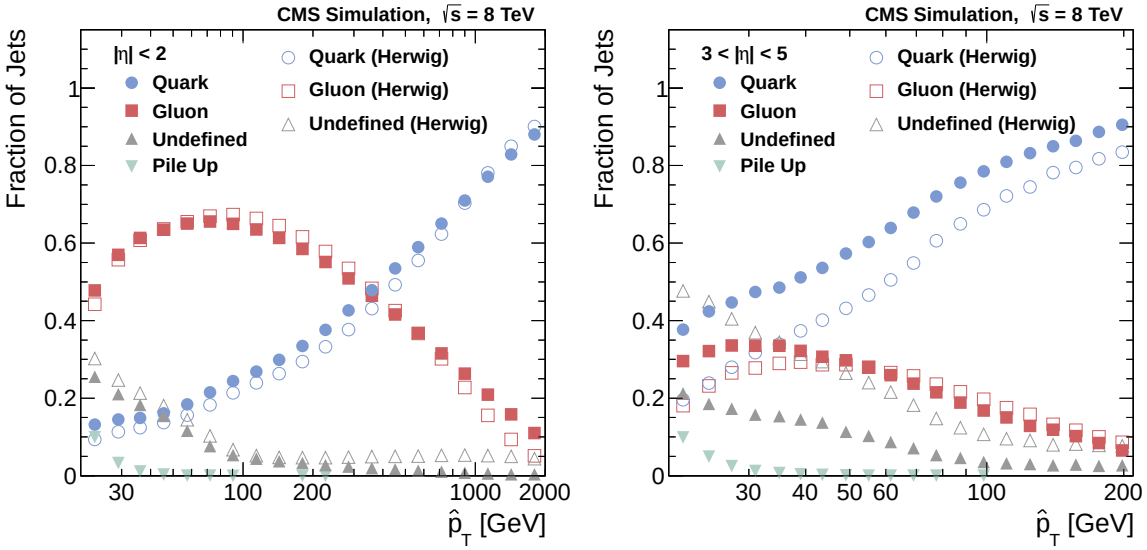

Figure 6.17: Jet flavour composition in dijet events, for central (left) and forward (right) jets, as a function of the event hard scattering transverse momentum ($\hat{p}_T$). Jets originating from light quark partons are shown with blue circles, gluon jets with red squares, undefined jets with grey triangles and pile up jets with yellow triangles. Solid markers indicate events simulated in Pythia 6, whereas hollow ones with Herwig++.

6.2 Validation on Dijet Events

The flavour composition of the dijet dataset is somewhat more complex than the Z+jet case, as can be seen in Figure 6.17, where the flavor composition of jets in dijet events is shown as a function of the hard scattering transverse momentum ($\hat{p}_T$) of the event. The left plot shows central jets, the right forward jets. Jets originating from light quark partons are shown with blue circles, gluon jets with red squares, undefined jets with grey triangles and pile up jets with yellow triangles. Solid markers indicate events simulated in Pythia 6, whereas hollow ones with Herwig++. As can be seen, whereas for $|\eta| < 2$ the majority of jets originate from gluons (at least up to 300-400 GeV), in the forward quark jets are dominant. The dijet sample is therefore a useful gluon enriched sample for central jets, and offers an important cross check to Z+jets for forward quark jets, as it has higher statistics.

Dijet events are selected in data by running on the full 2012 dataset analyzing events recorded with the single jet trigger `HLT_PFJet40`. They are compared to dijet events simulated in Pythia, from the 'Flat' Summer12 dataset reconstructed in CMSSW_5_3_X:

`/QCD_Pt-15to3000_TuneZ2star_Flat_8TeV_pythia6`.

Further comparisons have been carried out on the corresponding Herwig++ sample:

`/QCD_Pt-15to3000_TuneEE3C_Flat_8TeV_herwigpp`.

The event selection further requires:

- two jets with transverse momentum greater than 40 and 20 GeV, respectively;
- the two leading jets are required to be back-to-back in the transverse plane by requiring their azimuthal difference to be greater than 2.5 rad;
- the third jet in the event is required to have a transverse momentum inferior to 30% than the average p_T of the first two jets.

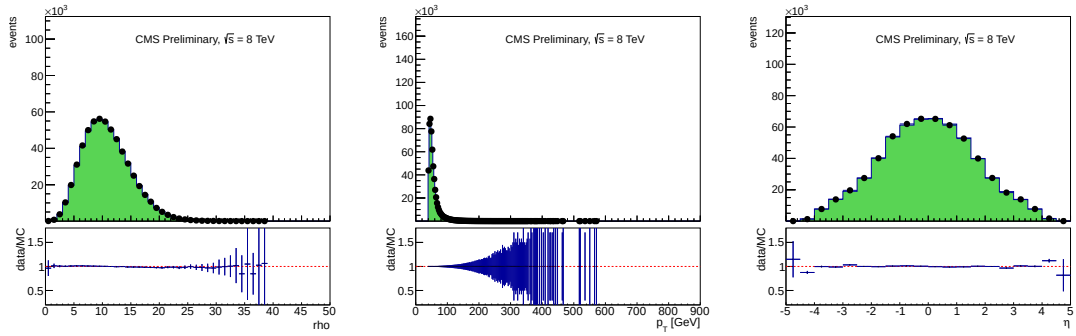

Figure 6.18: Data-MC comparisons for the particle-flow energy density ρ (left), the leading jet transverse momentum (center) and its pseudorapidity (right). The reweighting procedure is defined by the ρ distribution, so it is expected to find perfect agreement there.

Only the leading jet is used to fill the histograms (each in their own p_{T} bin), and it is required to pass the usual jet ID, pile up ID and anti-btagging requirements. In order to match the pile up distribution observed in the data, the Monte Carlo is reweighted based on the *a posteriori* particle flow energy density (ρ) distribution, in the same manner as is done in the Z+jets validation. As the simulated samples have a flat transverse momentum spectrum, though, an additional reweighting is necessary: we reweight simulated events so that the leading jet transverse momentum and pseudorapidity spectra match the ones observed in the data. This is done with a 2-dimensional weight matrix, so that the fact that transverse momentum and pseudorapidity are not completely independent is taken into account. The results of this reweighting is shown in Figure 6.18: the right plot shows the ρ distribution, the center plot shows the leading jet transverse momentum and the right plot its pseudorapidity.

Figures 6.19, 6.20, 6.21 and 6.22 show, respectively, the data-MC comparisons of the four input variables (total multiplicity, $p_T D$, Axis_2 and Axis_1), in four representative p_{T} bins, for jets reconstructed in the center of the detector. The corresponding plots for the forward are shown in Figures 6.23, 6.25, 6.26 and 6.27. Similarly to the validation performed on Z+jet events, also in dijet events we observe a significant discrepancy between the data and the simulation for what concerns the candidate multiplicity forward jets. We apply the same correction to data multiplicities before feeding them to the discriminators, i.e. we sutract one (unity) from the multiplicity. The data-MC comparisons for the forward multiplicity after this correction are shown in Figure 6.24.

The corresponding data-MC comparisons for the output discriminants are shown in Figures 6.28 for the likelihood discriminant and in Figure 6.31 for the MLP. The comparisons for the forward discriminators are shown respectively in Figures 6.29 and 6.32 before the multiplicity correction, and in Figures 6.30 and 6.33 after the multiplicity correction. As can be seen significantly better agreement is observed after the application of the correction.

As was seen in Section 5, Herwig++ predicts gluon jets to be more similar to quark jets than Pythia. This will have significant implications in the dijet sample, which has a large contamination of gluon jets. We have therefore performed a second validation on Herwig++ dijet events. While the single-variable data/MC comparisons are moved to Appendix **??**, we show the data-Herwig++ comparisons for central jets in Figures 6.34 and 6.36, respectively for the likelihood and the MLP discriminants. Similar comparisons for forward jets are shown in Figures 6.35 and **??**. For forward jets the same correction to the jet candidate multiplicity that was used in the Pythia comparisons is used when comparing to Herwig++. It must be noted that whereas for the MLP tagger the same weight files have been used both for Pythia and Herwig++ samples,

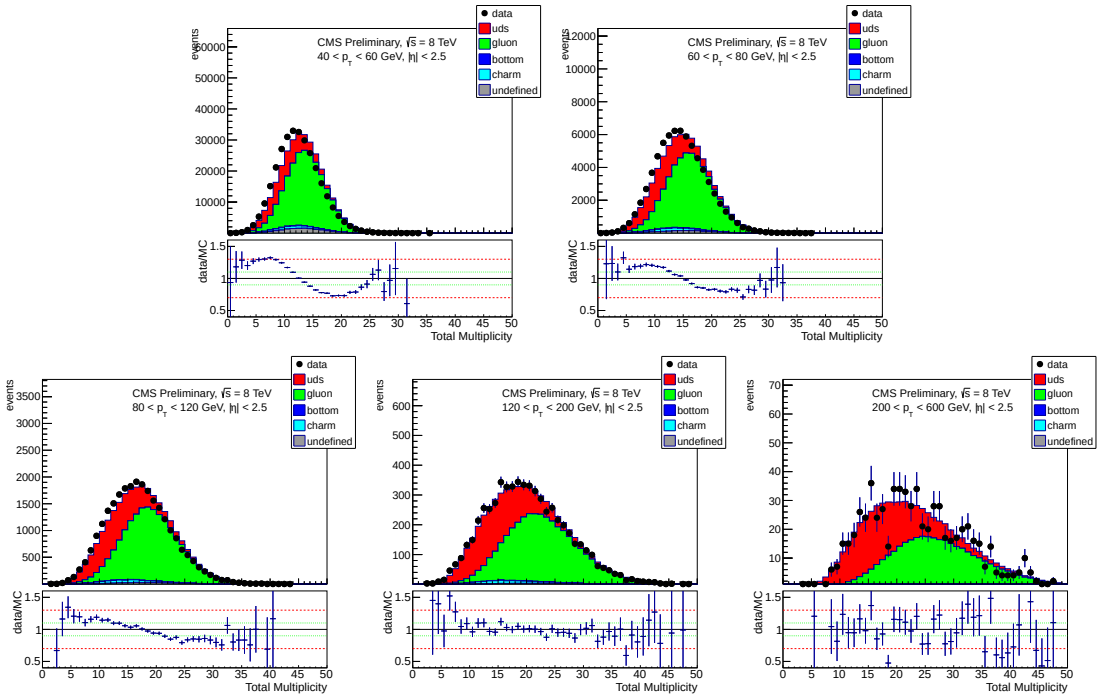

Figure 6.19: Data-MC comparisons for the total PFCandidate multiplicity in central jets in dijet events, for five p_{T} bins: 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

the likelihood has been re-defined for the Herwig++ comparison, by constructing the single variable PDFs on the Herwig++ sample.

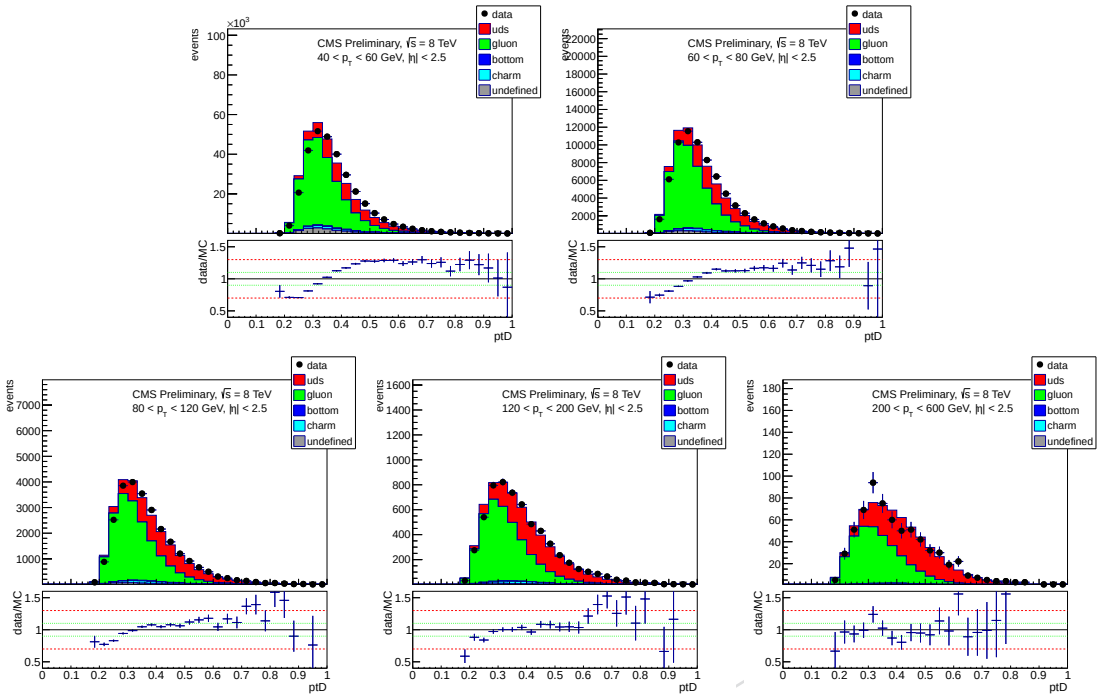

Figure 6.20: Data-MC comparisons for $p_T D$ in central jets in dijet events, for five p_T bins: 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

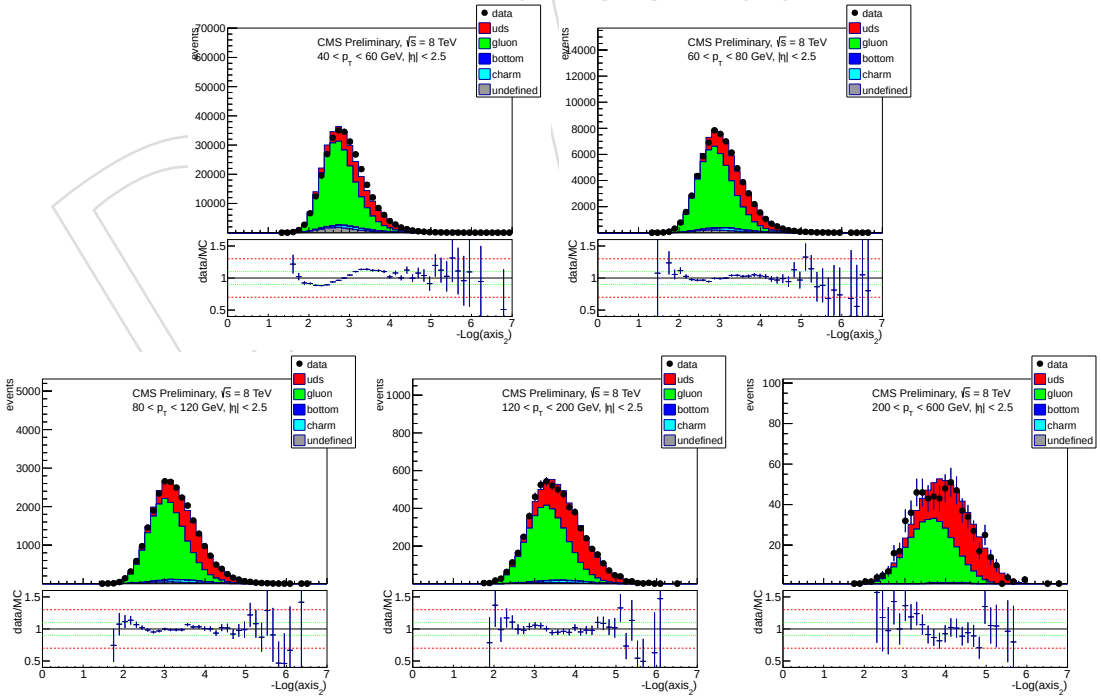

Figure 6.21: Data-MC comparisons for Axis_2 in central jets in dijet events, for five p_T bins: 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

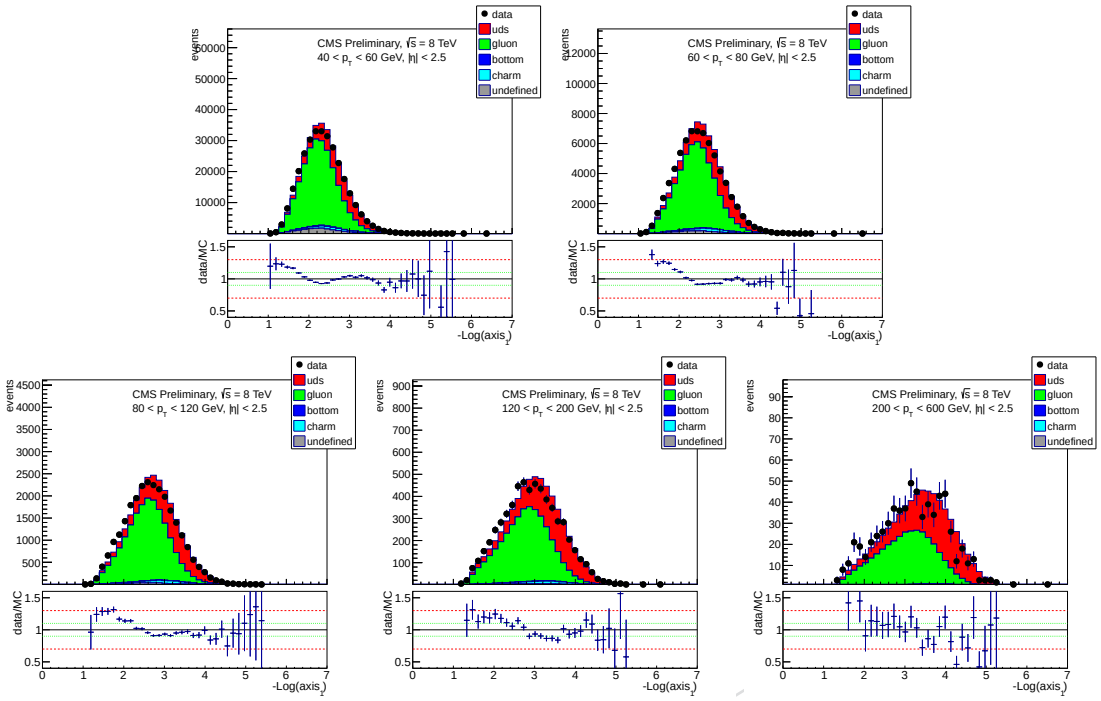

Figure 6.22: Data-MC comparisons for Axis_1 in central jets in dijet events, for five p_T bins: 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

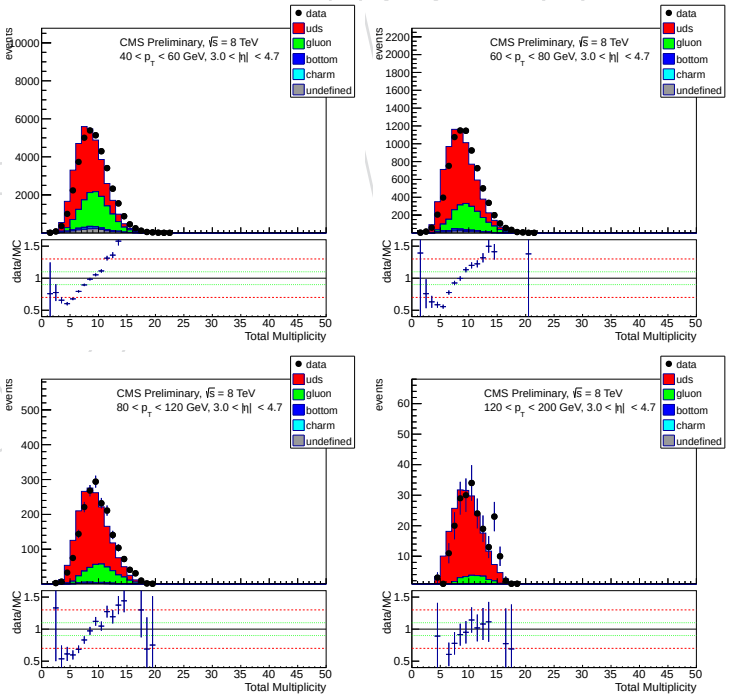

Figure 6.23: Data-MC comparisons for the total PFCandidate multiplicity in forward jets in dijet events **before the correction**, for four p_T bins: 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

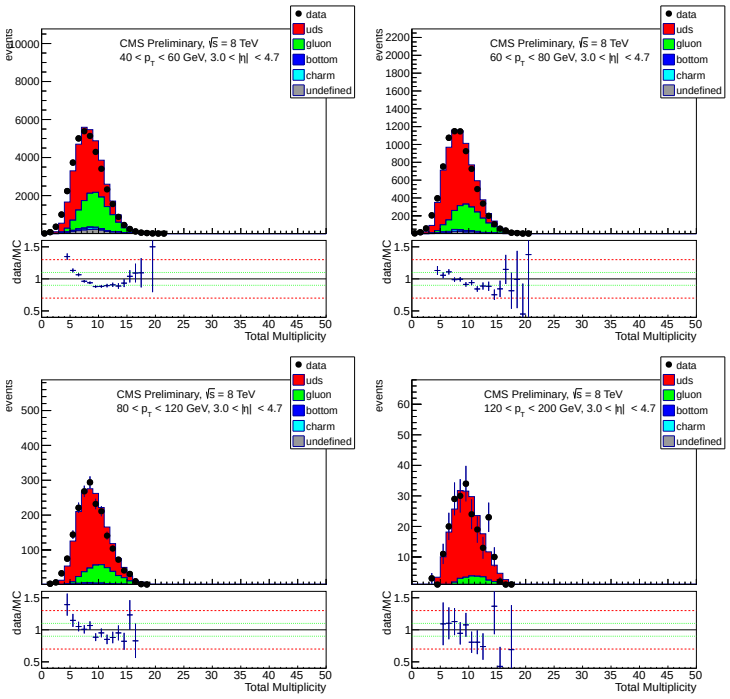

Figure 6.24: Data-MC comparisons for the total PFCandidate multiplicity in forward jets in dijet events **after the correction**, for four p_{T} bins: 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

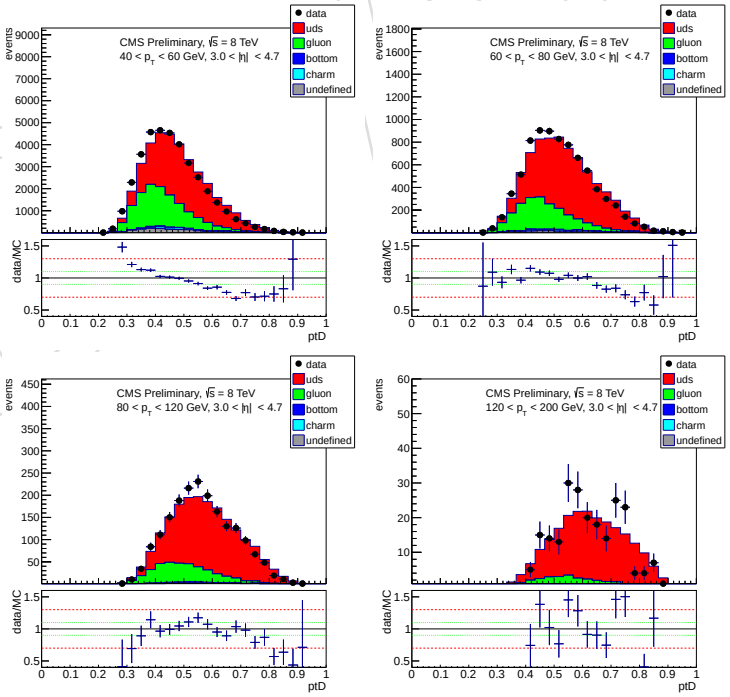

Figure 6.25: Data-MC comparisons for $p_T D$ in forward jets in dijet events, for four p_{T} bins: 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

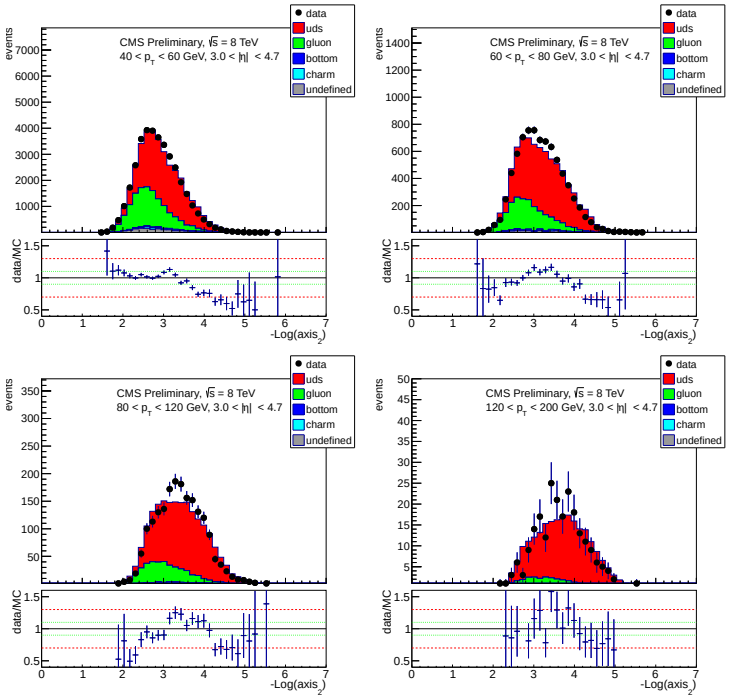

Figure 6.26: Data-MC comparisons for Axis_2 in forward jets in dijet events, for four p_{T} bins: 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

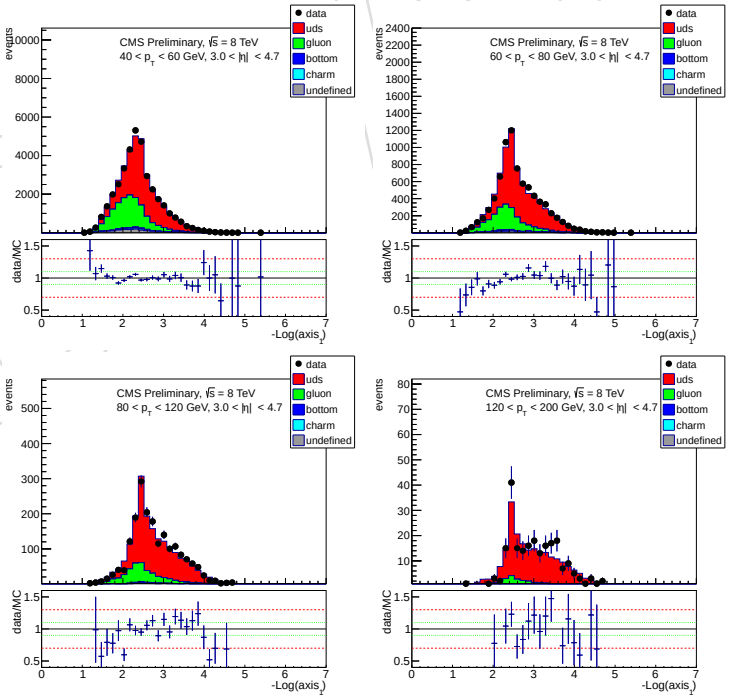

Figure 6.27: Data-MC comparisons for Axis_1 in forward jets in dijet events, for four p_{T} bins: 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

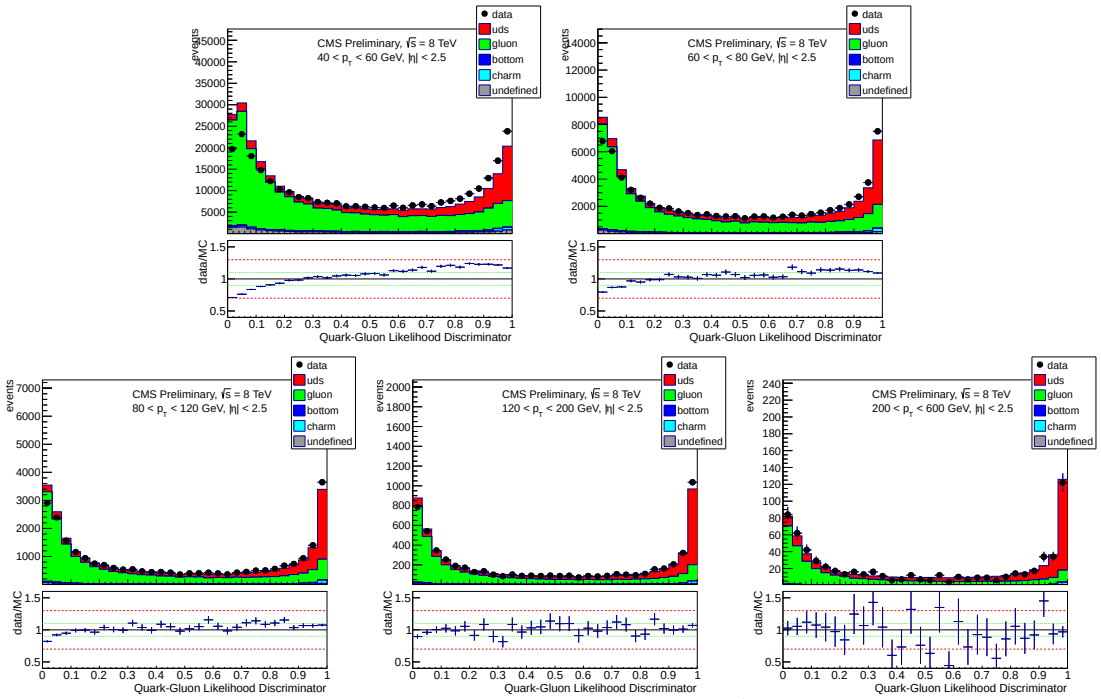

Figure 6.28: Data-MC comparisons for the quark-gluon likelihood discriminant in central jets in dijet events, for five p_{T} bins: 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

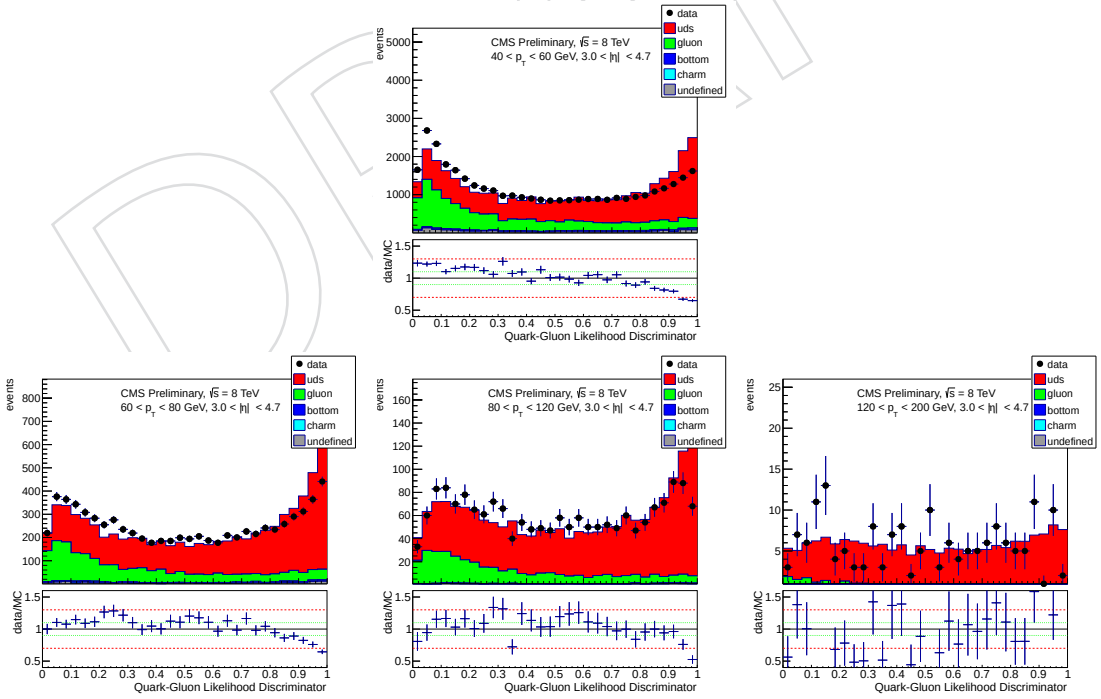

Figure 6.29: Data-MC comparisons for the quark-gluon likelihood discriminant in forward jets in dijet events **before the multiplicity correction**, for four p_{T} bins: 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

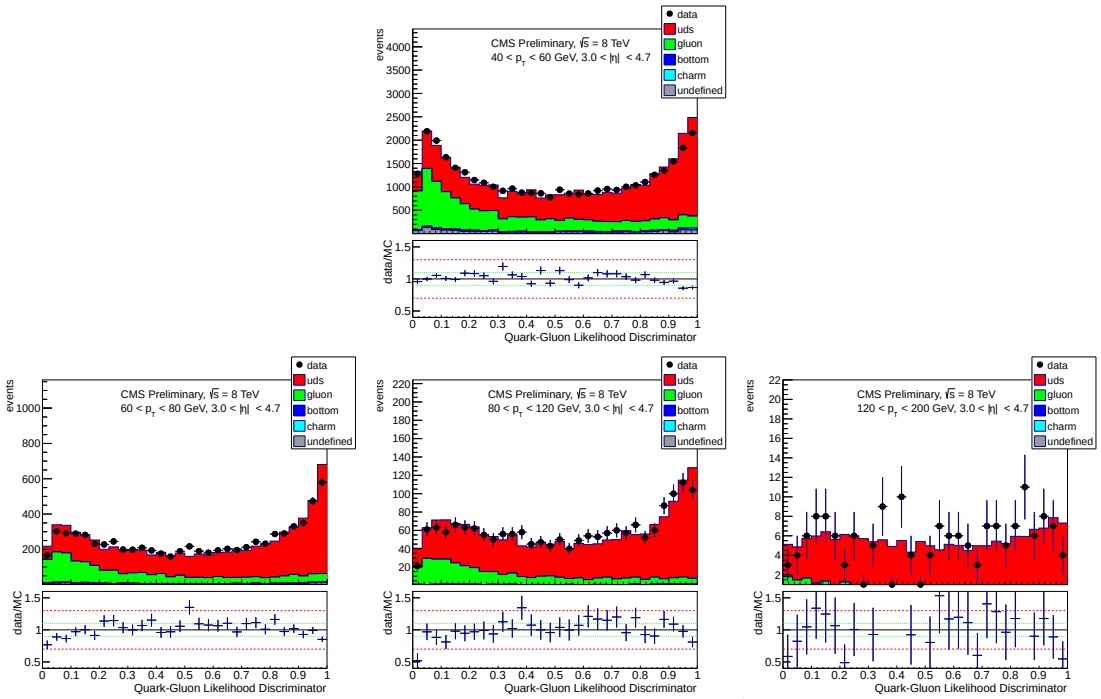

Figure 6.30: Data-MC comparisons for the quark-gluon likelihood discriminant in forward jets in dijet events **after the multiplicity correction**, for four p_{T} bins: 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

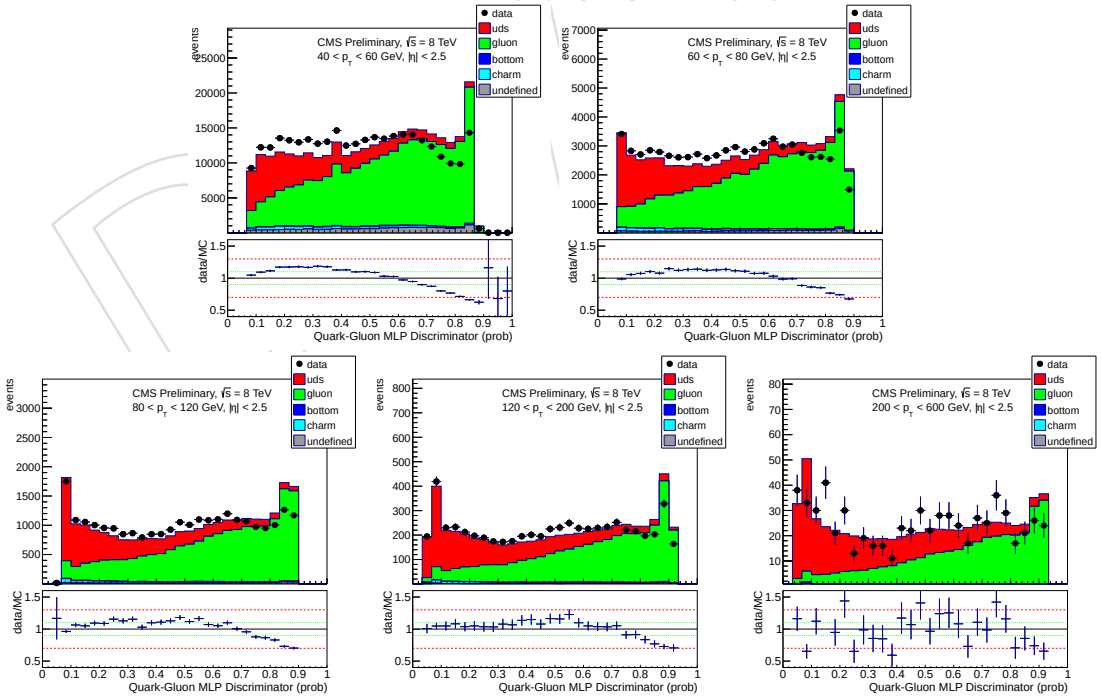

Figure 6.31: Data-MC comparisons for the quark-gluon MLP discriminant in central jets in dijet events, for five p_{T} bins: 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

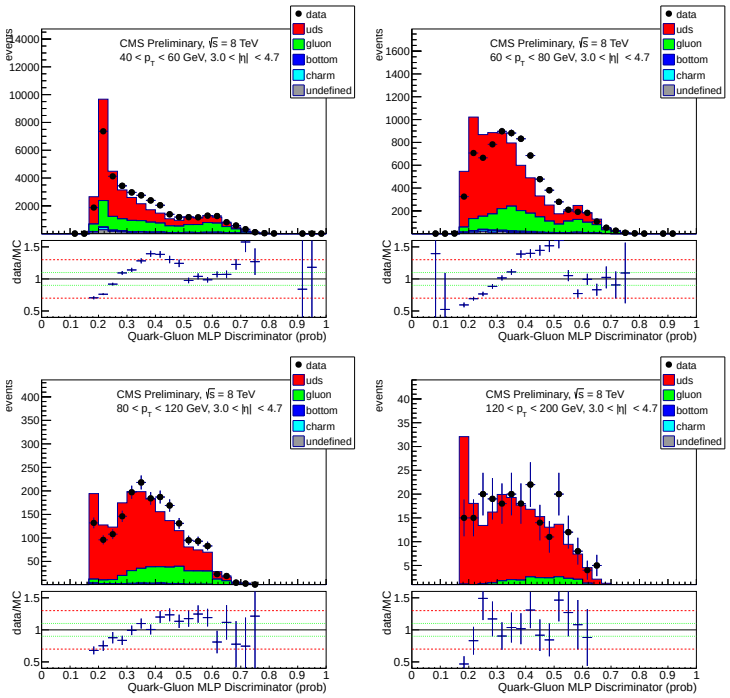

Figure 6.32: Data-MC comparisons for the quark-gluon MLP discriminant in forward jets in dijet events **before the multiplicity correction**, for four p_{T} bins: 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

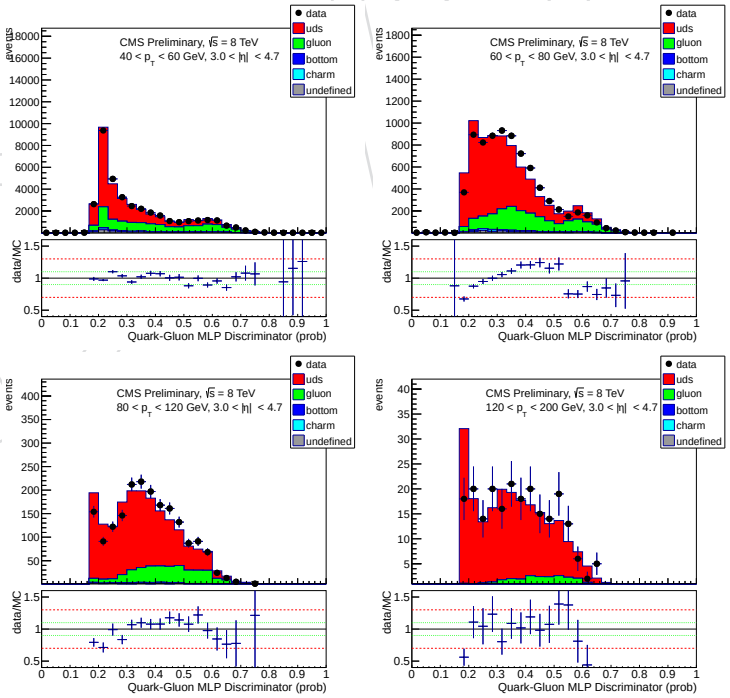

Figure 6.33: Data-MC comparisons for the quark-gluon MLP discriminant in forward jets in dijet events **after the multiplicity correction**, for four p_{T} bins: 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

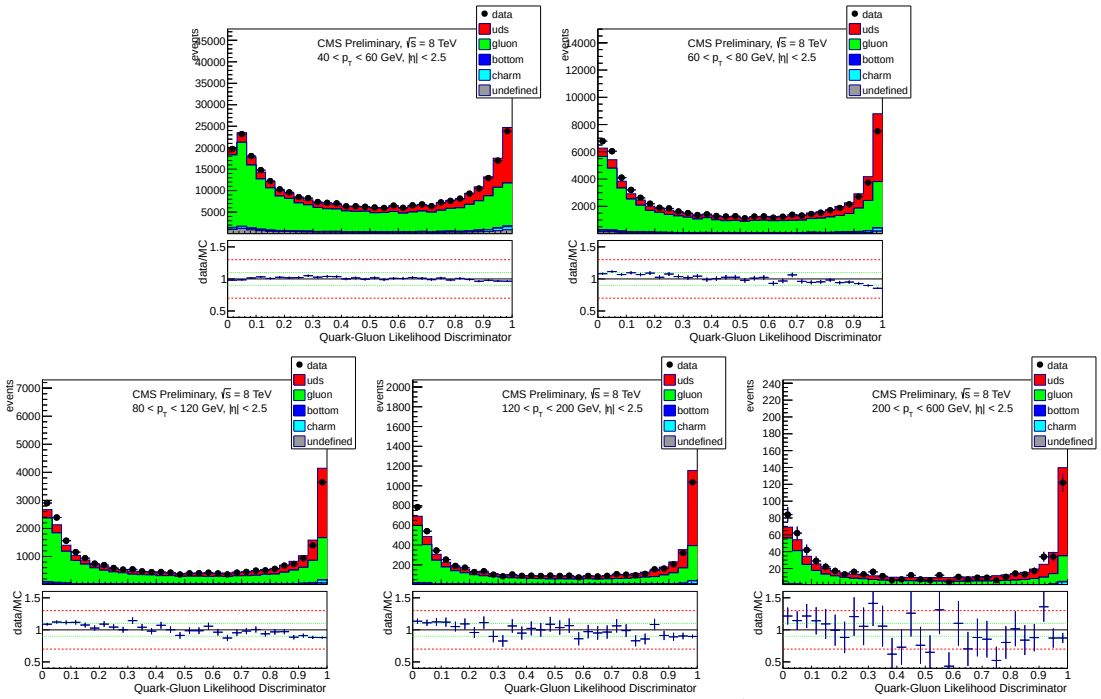

Figure 6.34: Data-Herwig++ comparisons for the quark-gluon likelihood discriminant in central jets in dijet events, for five p_{T} bins: 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

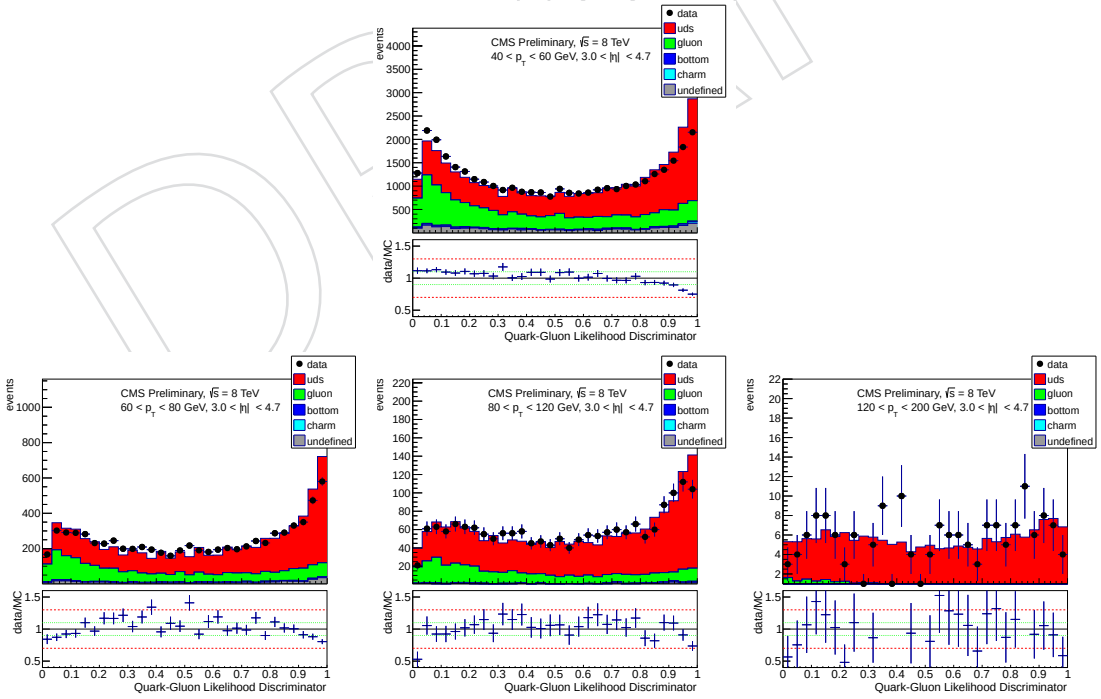

Figure 6.35: Data-Herwig++ comparisons for the quark-gluon likelihood discriminant in forward jets in dijet events **after the multiplicity correction**, for four p_{T} bins: 40-60, 60-80, 80-120 and 120-200 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

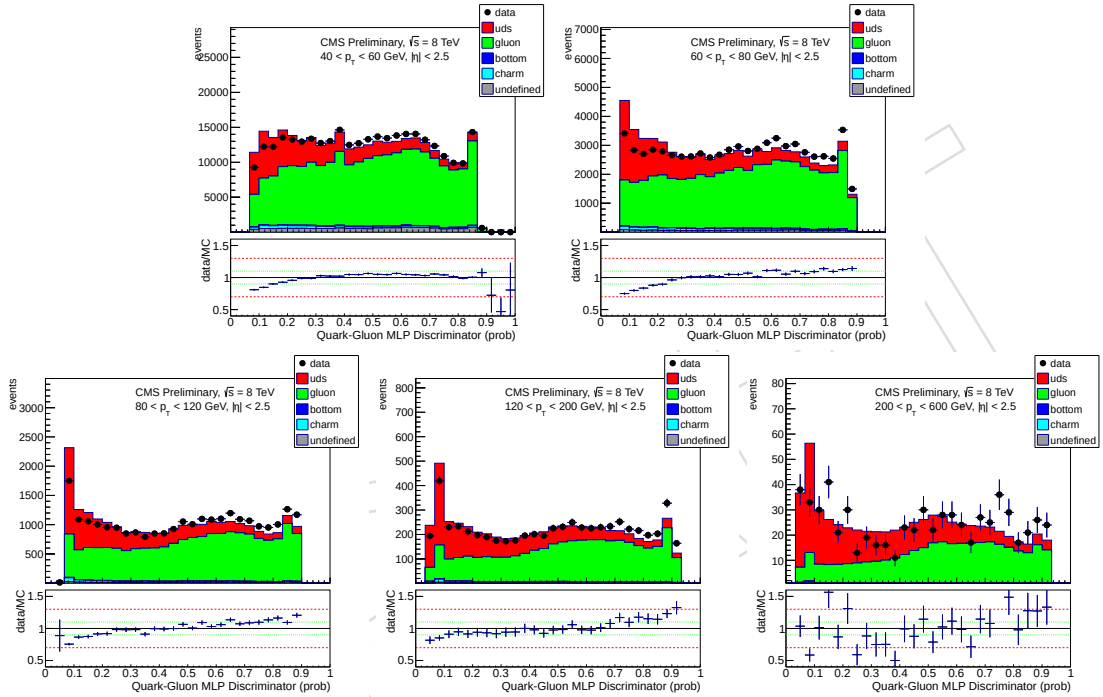

Figure 6.36: Data-Herwig++ comparisons for the quark-gluon MLP discriminant in central jets in dijet events, for five p_{T} bins: 40-60, 60-80, 80-120, 120-200 and 200-600 GeV. The different jet flavours of simulated jets are shown: light quarks (red), gluon (green), bottom (blue), charm (cyan), and undefined (grey).

7 Systematics evaluation

7.1 Usage: data vs data

The comparison between data and data it is far more less sensitive to any model used to derive the selections. The main goal in this case is to show that the analysis is not particularly sensitive to where the cut is placed in the discriminator (or in any other) and it is left to the analysts: a stability under a variation of the cut value of some percent should be expected.

7.2 Usage: data vs mc

This is the first and most delicate systematics that it is need to evaluate, and that allow to make a fair comparison between the results obtained with data and the one with Monte Carlo. A general comparison between data and Monte Carlo can show discrepancies in the discriminator for many reasons. Among them it worth to notice the hadronization model, the detector model, the pile-up model, the hard scattering, the pdf ...

We are therefore interest to assest systematics in order to cover the events not due to phyisics. In order to reduce the impact due to physics, the validation samples used in the validation are used to extract and verify the systematic error.

A genaral usage of the discriminator, being ething the likelihood or the MLP, implies one of the following two scenarios:

- use with cuts
- use with multivariate analysis

In order to allow these two shenario a systematic method which takes into account the whole shape is developed.

Among a set of function that maps the interval $[0,1]$ into itself, the following were considered because have a reasonable number of parameters (≤ 2) that can be used to shift the population through the center or move the population to one of the two estremities.

The functions studied are:

$$f(x,a,b) = \arctan\left(a\tan\left(x\pi - \frac{\pi}{2}\right) + b\right)/\pi + \frac{1}{2} \tag{7.1}$$

$$g(x,a,b) = \tanh\left(a\,\operatorname{arctanh}\left(2x-1\right) + b\right)/2 + \frac{1}{2} \tag{7.2}$$

The behaviour of these two function is shown in Fig. 7.1. It worths to be noticed the difference behaviour of the two set of functions around the extreme points of the interval. This difference allows bigger changes for the function g (Eqn. 7.2) at the extreme with less modification in the middle range. For this reason it perform better on the likelihood. For this reason this function was chosen.

7.2.1 Fit

The Z+Jets data and Monte Carlo are used in order to perform the systematics evaluation. The Z +Jet selection is performed very tight in order to have the less as possible contriubition from background or pile-up with a β^* cut on the jets and a balance in p_{T} and ϕ between the leading jet and the Z, and a jet veto on the second jet.

The minimization is perform w.r.t the χ^2: the Monte Carlo entries of the discriminator are varied according to the chosen function (Eqn. 7.2) and the two parameters are therefore evalueted

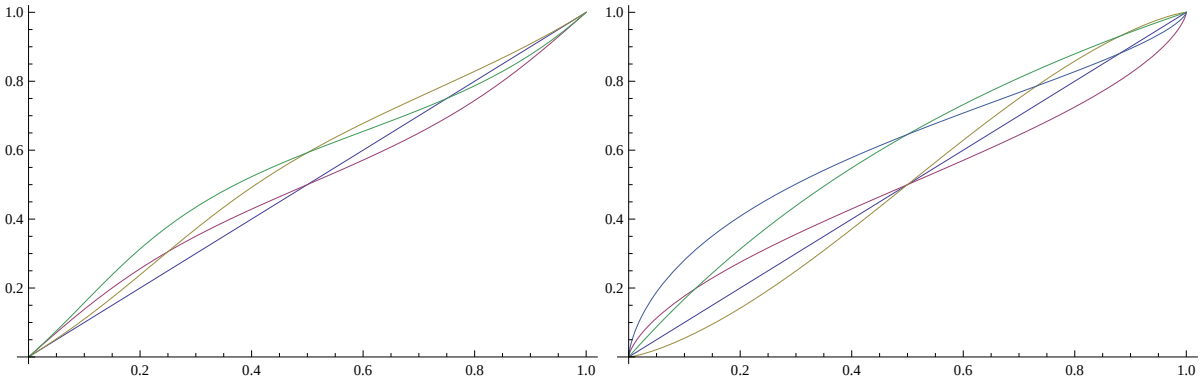

Figure 7.1: Representation of the functions in Equation 7.1 on the left, and in Equation 7.2 on the right for different values of the parameters.

in large bin of the jet p_T,η, and ρ in order to exploit possible differences in any of these directions.

For the MLP an additional step is added to the minimization process: on the given bin and selection the minimum and maximum of the tagger are recorded from the Monte Carlo (t_{min}, t_{max}) and a linear transformation (l) is appended to transform $[t_{min}, t_{max}]$ in $[0, 1]$ and anti-transformed after the function is computed, i.e., the following modified functions are used:

$$\begin{aligned}\tilde{f}(x,a,b,t_{min},t_{max}) &= l^{-1}(x,t_{min},t_{max}) \circ f(x,a,b) \circ l(x,t_{min},t_{max}) \\ &= f\left(\frac{x-t_{min}}{t_{max}-t_{min}},a,b\right)\cdot(t_{max}-t_{min})+t_{min} \qquad (7.3)\\ \tilde{g}(x,a,b,t_{min},t_{max}) &= l^{-1}(x,t_{min},t_{max}) \circ g(x,a,b) \circ l(x,t_{min},t_{max}) \\ &= g\left(\frac{x-t_{min}}{t_{max}-t_{min}},a,b\right)\cdot(t_{max}-t_{min})+t_{min} \qquad (7.4)\end{aligned}$$

In this way the minimization is still on two parameters (a and b), while $t_{min/max}$ are evaluated in advance. If a value will follow out of that interval in the future it will be mapped in the systematic to the closest point in that interval (i.e., aut t_{min} aut t_{max}).

The parameters, result of the minimization process, are reported in Figure 7.2 for the different bins. No trends can be assested from these results.

On Figures 7.3, 7.4 are shown the data with the Monte Carlo before and after the shift for the likelihood tagger and the MLP tagger respectively.

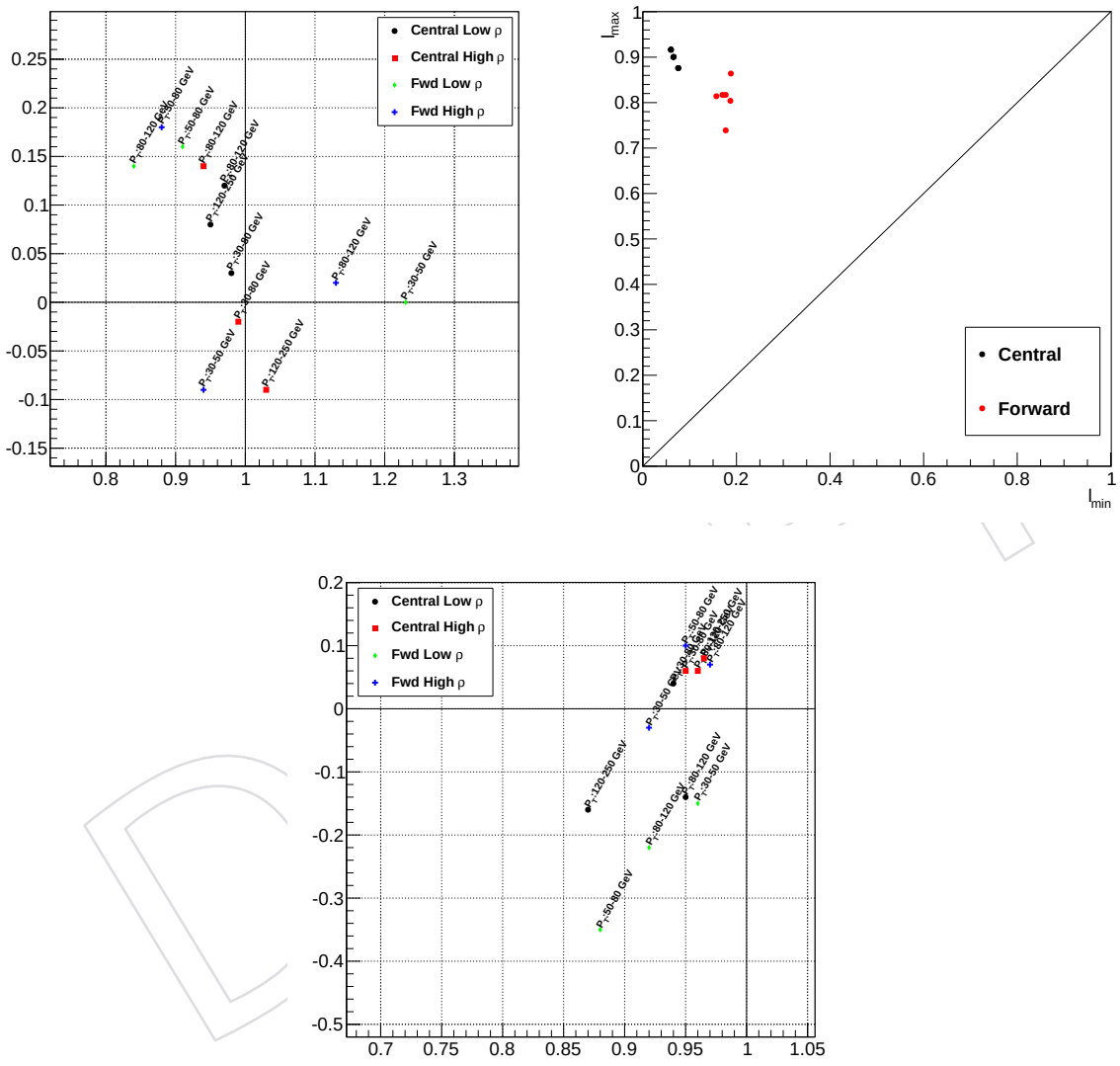

Figure 7.2: Parameters fitted for the chosen function (Eqn. 7.2) with respect to the different bins: on the top the MLP parameters (with the ranges on the right), on the bottom the Likelihood ones.

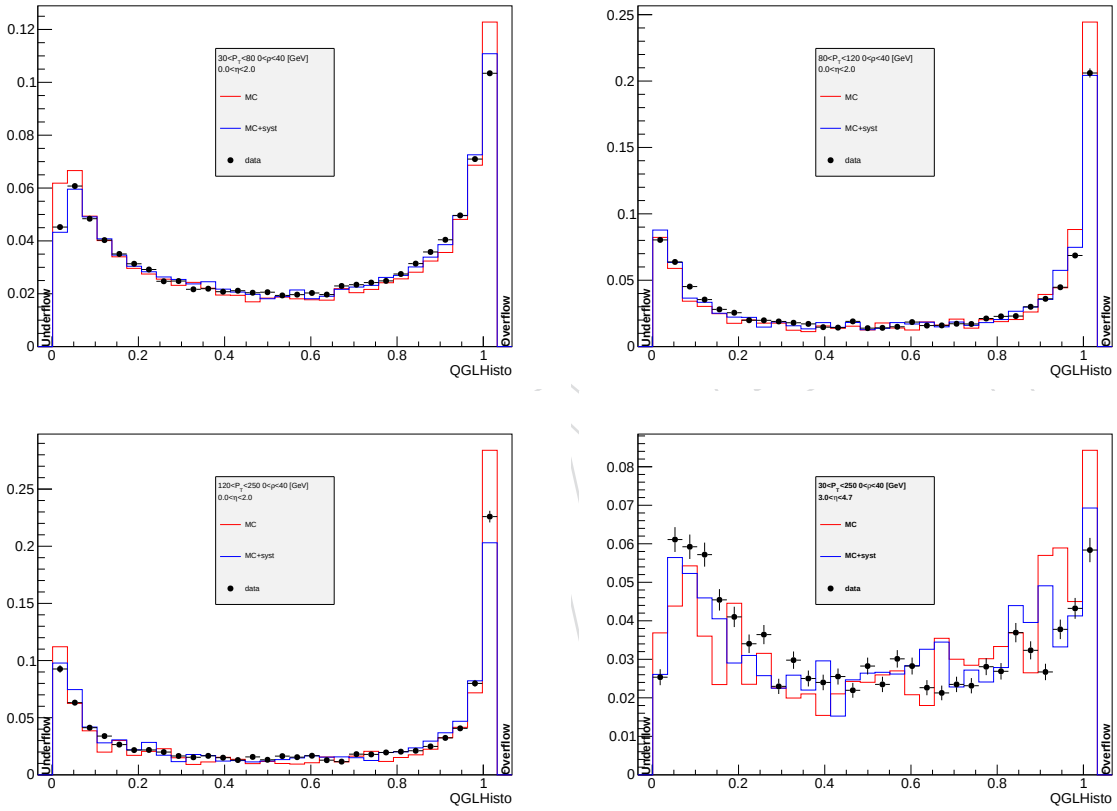

Figure 7.3: Data/MC agreement before and after fit for systematics evaluation in bins of jet p_{T} for the likelihood tagger.

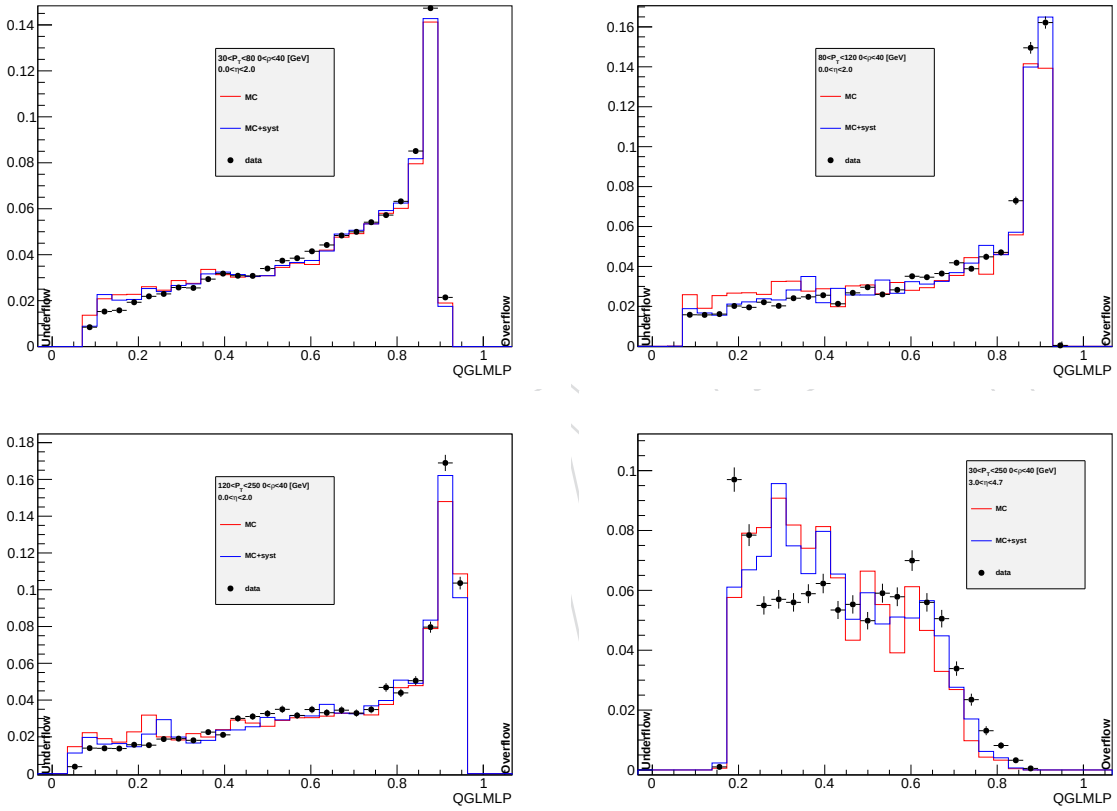

Figure 7.4: Data/MC agreement before and after fit for systematics evaluation in bins of jet p_{T} for the MLP Tagger.

8 Conclusions

A new tool to separate CMS particle flow jets originating form the showering and hadronization of final states gluons or light (uds) quarks has been presented. The new discriminator makes use of internal jet composition properties and exploits known and well established differences in the showering of quarks and gluons. With respect to the previously developed discriminator, the presented ones achieve improved performaces and extend the discrimination power to the CMS forward region, up to $|\eta| = 5$. The new discriminator has been validated using 2012 8 TeV collision data, both with dijet and Z+jet event samples. Finally a recipe to evaluate the systematic uncertainties associated to the use of the new discriminator is given, based on the observed data versus Monte Carlo differences in the validation samples. The recipe makes use of a two-parameter analytical trasformation function that distorts the Monte Carlo output distributions in order to reproduce better the observed data outputs: it can therefore be used to determine uncertainties on Monte Carlo efficiencies and rejections both for a cut-based selection or when using the whole discriminator output for another multivariate analysis step.

References

[1] OPAL Collaboration Collaboration, "A Study of differences between quark and gluon jets using vertex tagging of quark jets", *Z.Phys.* **C58** (1993) 387–404, doi:10.1007/BF01557696.

[2] OPAL Collaboration Collaboration, "A Model independent measurement of quark and gluon jet properties and differences", *Z.Phys.* **C68** (1995) 179–202.

[3] DELPHI Collaboration Collaboration, "Energy dependence of the differences between the quark and gluon jet fragmentation", *Z.Phys.* **C70** (1996) 179–196, doi:10.1007/s002880050095.

[4] ALEPH Collaboration Collaboration, "Quark and gluon jet properties in symmetric three jet events", *Phys.Lett.* **B384** (1996) 353–364, doi:10.1016/0370-2693(96)00849-0.

[5] N. Saoulidou, "Particle Flow Jet Identification Criteria", CMS Note 2010/03, (2010).

[6] A. C. Marini and F. Pandolfi, "Quark-Gluon Discrimination in CMS", CMS Analysis Note AN-2012/074, (2012).

[7] CMS Collaboration, "Search for a Higgs boson in the decay channel $H \to ZZ(*) \to q\bar{q}\ell^-\ell^+$ in pp collisions at $\sqrt{s} = 7$ TeV", *Journal of High Energy Physics* **2012** (2012) 1–36, doi:10.1007/JHEP04(2012)036.

[8] A. Hoecker et al., "TMVA: Toolkit for Multivariate Data Analysis", *PoS* **ACAT** (2007) 040, arXiv:physics/0703039.

[9] E. D. Malaza, "Multiplicity distributions in quark and gluon jets", *Zeitschrift f ur Physik C Particles and Fields* **31** (1986) 143–150. 10.1007/BF01559605.

[10] A. C. Marini et al., "Quark-Gluon Jet Discrimination through Particle Flow Jet Structure", CMS Note 2011/215, (2011).

[11] J. Gallicchio and M. D. Schwartz, "Quark and Gluon Jet Substructure", arXiv:1211.7038.

[12] A. C. Marini and F. Pandolfi, "Quark-Gluon Discrimination", CMS Note 2012/74, (2012).

[13] G. P. Salam, "Towards Jetography", *Eur.Phys.J.* **C67** (2010) 637–686, doi:10.1140/epjc/s10052-010-1314-6, arXiv:0906.1833.

A Bjets

It was already observed that b-jets have similar shape as gluon jets **FIXME:** Ref. We find similar results, in particular we observed that at low p_{T} b-jets are closer to gluon rather than to light quark jets in all variables studied, while at high p_{T} they come closer and closer to the quark jets. The explanation to this behaviour is laid in two opposite facts: from being massive the b-partons are less coupled to the QCD field as measured by **FIXME:** Ref. However, the chain of weak decays which will form the b-jets as we observe them,i.e., after the decay of the heavy B and D mesons produced during hadronization, contribute to enlarge the jet and enrich it in multiplicity. This behaviour is shown in Figure A.1 where the Charged Multiplicity (all) compared to the Charged Multiplicity (QC) show that the extra-number of particles cames from secondary vertexes (and so do not match QC). Also the other variables will be affectted by these weaks decays. The c-jets should have the same but much more less pronunced behaviour. In particular at very high p_{T} c and light (uds) quark are undistinguishable (in our variables).

The goal of this product is not to identify b-jets (or c-jets) rather than to be able to discriminate between light quark jets and gluon jets. Special efforts are already developed in the Collaboration for this kind of jets, and therefore these tools should be used. In particular, since this tool can be used in addition to b-jets tagger, an explicit b-jets vetos was required in the training.

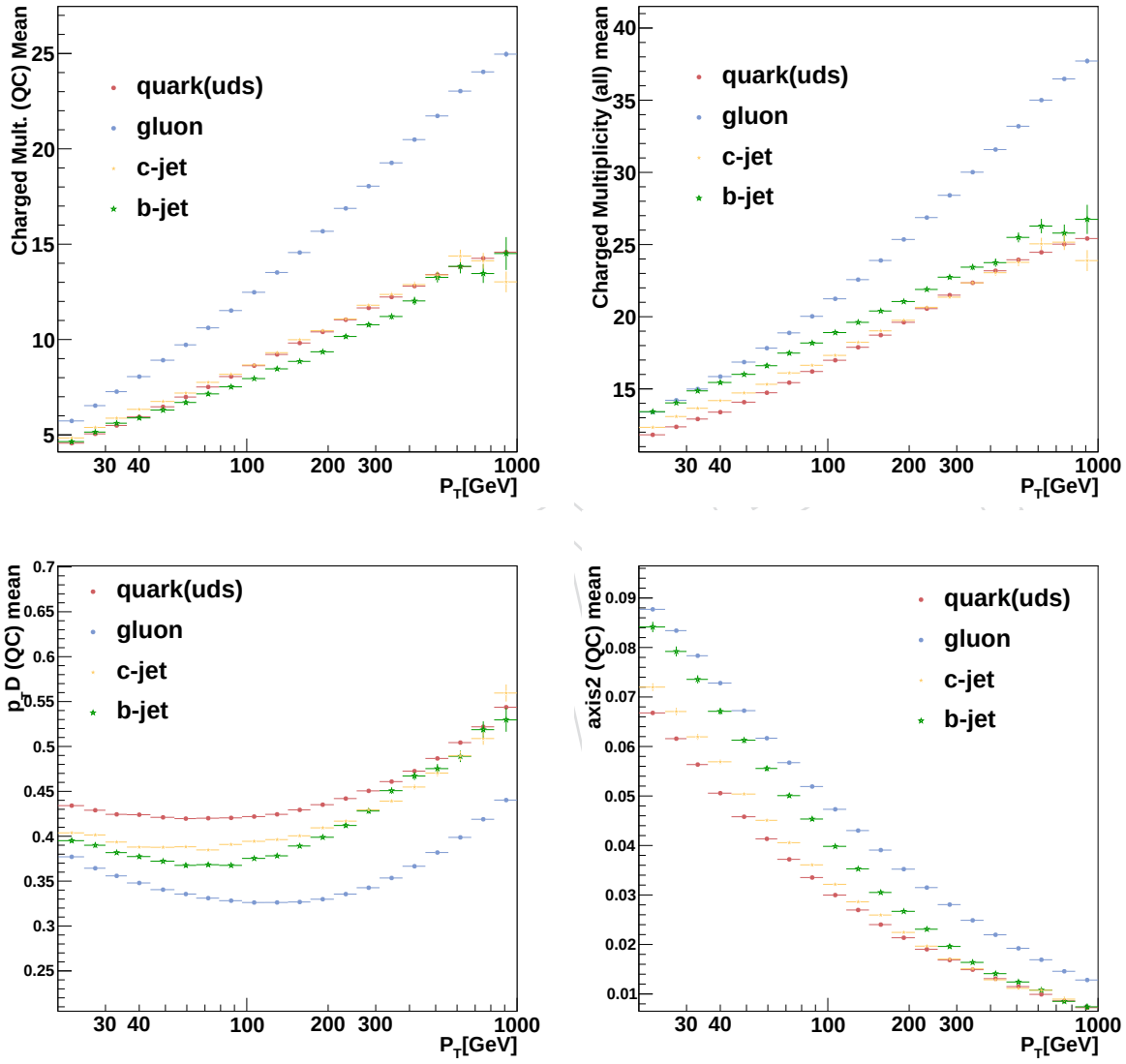

Figure A.1: In the figures is shown the mean of Charged Multiplicity (QC), the Charge Multiplicity, $p_T D$(QC) and axis2 (QC) vs p_T of the jet.