Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector


DOI
10.1140/epjc/s10052-014-3023-z

Publication date
2014

Document Version
Final published version

Published in
European Physical Journal C

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).
Light-quark and gluon jet discrimination in $pp$ collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector

The ATLAS Collaboration
CERN, 1211 Geneva 23, Switzerland

Received: 18 June 2014 / Accepted: 4 August 2014 / Published online: 21 August 2014
© CERN for the benefit of the ATLAS collaboration 2014. This article is published with open access at Springerlink.com

Abstract A likelihood-based discriminant for the identification of quark- and gluon-initiated jets is built and validated using 4.7 fb$^{-1}$ of proton–proton collision data at $\sqrt{s} = 7$ TeV collected with the ATLAS detector at the LHC. Data samples with enriched quark or gluon content are used in the construction and validation of templates of jet properties that are the input to the likelihood-based discriminant. The discriminating power of the jet tagger is established in both data and Monte Carlo samples within a systematic uncertainty of $\approx 10$–$20\%$. In data, light-quark jets can be tagged with an efficiency of $\approx 50\%$ while achieving a gluon-jet mis-tag rate of $\approx 25\%$ in a $p_T$ range between 40 GeV and 360 GeV for jets in the acceptance of the tracker. The rejection of gluon-jets found in the data is significantly below what is attainable using a Pythia 6 Monte Carlo simulation, where gluon-jet mis-tag rates of $10\%$ can be reached for a $50\%$ selection efficiency of light-quark jets using the same jet properties.

1 Introduction

The production of quarks and gluons via strong interactions is the dominant high-momentum-transfer process at the LHC and is a significant background to most new-physics searches. These partons are measured as jets, which are collimated streams of charged and neutral particles, clustered using dedicated algorithms. Corrections to measured quantities are necessary to relate the jets to their parent partons. Many gluons are generated in most common Standard Model processes, such as the inclusive production of jets [1,2]. On the other hand, some processes arising from new-physics models, for example supersymmetry, generate many light quarks [3,4]. The power to discriminate between jets initiated by light quarks and those initiated by gluons would therefore provide a powerful tool in searches for new physics. In case of a discovery of a new particle, such a discriminant could provide valuable information about its nature. Also, some Standard Model measurements rely on the correct identification of the origin of jets, as in the cases of reconstructing a hadronic $W$ decay when measuring the top quark mass, or in the reconstruction of a hadronic $Z$ decay when measuring the Higgs boson mass via $h \rightarrow ZZ \rightarrow \ell\ell q\bar{q}$. These analyses would benefit from such a discriminant. These applications motivate the analysis of the partonic origin of jets that is the focus of this paper.

In perturbative quantum chromodynamics (QCD), the concept of a parton initiating a jet is a fixed-order notion. In the matrix-element calculation of a high-momentum-transfer-process, the outgoing partons appear naïvely much like outgoing particles in the final state. However, only colourless states with two or more partons can form an observable jet. Moreover, in a parton shower, the leading parton is only well defined for a fixed number of splittings. The next step in the shower may change the energy, direction, or flavour of the leading parton. Thus, labelling jets with a specific flavour and interpreting results after such labelling requires a clearly defined procedure [5].

Certain parton branchings can yield an ambiguous jet identity. The labelling of a jet may also depend on the physics goal of the analysis. For example, a jet from the $q\bar{q}'$ decay of a high-momentum $W$ boson produced in a top quark decay can be considered either as a part of a top-quark jet or as a boosted W-boson jet. Nonetheless, many event topologies lend themselves to the identification of a jet as having originated from a specific type of parton in the matrix-element calculation. Such an approach can lead to an unambiguous and meaningful parton labelling for a large majority of jets.

Discrimination between jets of different partonic origin has been attempted previously at several experiments [6–16]. Most work has relied on jet properties that result from the difference in colour charge between the partons. The
colour factors in quantum chromodynamics differ for quarks \((C_F = 4/3)\) and gluons \((C_A = 3)\), and therefore, for example, one expects approximately \(C_A/C_F = 9/4\) times more particles in a gluon-initiated jet than in a jet initiated by a light \((u, d\) or \(s)\) quark. The measured difference in particle multiplicity at OPAL was, in fact, not far from this expectation \([9]\). Because of the showering that produces these additional particles, gluon jets are also expected to be wider and have a softer particle spectrum.

The most successful studies of discrimination between light-quark-initiated and gluon-initiated jets (henceforth, quark-jets and gluon-jets) have taken place at electron-positron colliders \([17,18]\). The selection and identification of “pure” samples of quark- and gluon-jets is considerably more difficult at hadron colliders because of the complication added by beam remnants, initial-state radiation, and multiparton interactions. The presence of multiple soft \(pp\) collisions overlaying the hard-scatter interaction of interest at the LHC further complicates this task. Recently some effort has been devoted to developing kinematic selections that significantly enhance the fraction of quark-jets or gluon-jets in a set of events \([5]\). In addition, discriminants based on jet structure have shown some promise for distinguishing between classes of jets at the LHC \([19]\).

Jets that include, or are initiated by, heavy quarks (bottom and charm) also exhibit properties different from those of quark-jets \([20,21]\). Generally, these jets are wider than quark-jets. They are often identified by long-lived or leptonically decaying hadrons. However, no special discriminant for them is developed here.

This paper is organised as follows. The ATLAS detector is briefly described in Sect. 2. Section 3 describes details of the data and Monte Carlo (MC) samples used, as well as the object reconstruction and event selection. Section 4 introduces the definition of gluon-jets and quark-jets that are used in the remainder of the paper. The jet properties used to build a discriminant from samples with different purities, and the validation of the extraction method using MC event samples, are described in Sect. 5. Section 6 describes the selection of samples based on kinematic variables to enhance quark-jet or gluon-jet fractions and the validation of the extracted properties using those samples. The likelihood-based discriminant is described in Sect. 7, where its performance in MC simulation and in data is discussed. Finally, the conclusions are presented in Sect. 8.

2 ATLAS detector

The ATLAS detector \([22]\) comprises an inner tracking detector, a calorimeter system, and a muon spectrometer. The inner detector (ID) includes a silicon pixel detector, a silicon microstrip detector and a transition radiation tracker. It is immersed in a 2 T axial magnetic field provided by a solenoid and precisely measures the trajectories of charged particles with \(|\eta| < 2.5\).\(^1\) The calorimeter system covers the region \(|\eta| < 4.9\) and is divided into electromagnetic and hadronic compartments. Electromagnetic calorimetry in the region \(|\eta| < 3.2\) is provided by liquid-argon sampling calorimeters with lead absorbers. In the barrel region \((|\eta| < 1.7)\), the hadronic calorimeter comprises scintillator tiles with steel absorbers, and the endcap region \((1.4 < |\eta| < 3.2)\) is covered by a liquid-argon and copper sampling hadronic calorimeter. The calorimetry in the forward region \((3.2 < |\eta| < 4.9)\) is provided by a liquid-argon and copper sampling electromagnetic calorimeter and a liquid-argon and tungsten sampling hadronic calorimeter. The muon spectrometer (MS) covers \(|\eta| < 2.7\) and uses a system of air-core toroidal magnets.

ATLAS has a three-level trigger system to select events. The first-level trigger uses custom-built hardware components and identifies jet, electron and photon candidates using coarse calorimeter information, and muon candidates using coarse tracking information from the muon spectrometer. At the highest level, full event reconstruction, similar to that used in the offline software, is performed to accurately identify and measure objects that determine whether the event is recorded.

3 Data sample and event selection

Several samples are used in the construction and validation of the variables entering the quark/gluon discriminant: dijet events, trijet events, \(\gamma + \text{jet}\) events, \(\gamma + \text{2-jet}\) events, \(\tau \ell\) events and \(W + \text{jet}\) events. After basic data quality requirements are imposed to remove known detector errors and readout problems, the selected dataset corresponds to a total integrated luminosity of \(4.67 \pm 0.08\) fb\(^{-1}\) \([23]\). The data were collected from March to October 2011 at a centre-of-mass energy \(\sqrt{s} = 7\) TeV. The average number of additional \(pp\) collisions per bunch crossing, called “pile-up”, rose during the data-taking period from a few to 15.

\(^1\) ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point (IP) in the centre of the detector and the \(z\)-axis along the beam pipe. The \(x\)-axis points from the IP to the centre of the LHC ring, and the \(y\)-axis points upward. Cylindrical coordinates \((r, \phi)\) are used in the transverse plane, \(\phi\) being the azimuthal angle around the beam pipe. The pseudorapidity is defined in terms of the polar angle \(\theta = -\ln \tan(\theta/2)\), and the rapidity is defined as \(\pm \ln \left(\frac{E + p_T}{E - p_T}\right)\), where \(E\) is the object’s energy and \(p_T\) is its momentum along the \(z\)-axis. The values of \(\eta, \phi\), and \(y\) are determined at the interaction vertex. The variable \(\Delta R = \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2}\) is used to characterise the angular difference between two objects using their \(\eta\) and \(\phi\) directions.
3.1 Monte Carlo simulation

Simulated event samples are generated for comparison with data and for the determination of the systematic uncertainties based on variations in the MC generator settings. For the MC samples, several different generators are used. MADGRAPH [24] is run as a 2 $\rightarrow$ $N$ generator with MLM matching [25], uses the CTEQ 6L1 parton distribution function (PDF) set, and is interfaced to PYTHIA 6 with a version of the ATLAS MC11 Underlying Event Tune 2B (AUET2B) [26] constructed for this PDF set. HERWIG++ [27] is run standalone as a 2 $\rightarrow$ 2 generator and uses the MRST LO** PDF set with the LHC-UE7-2 tune [28]. This tune of HERWIG++ has an improved description of colour reconnection in multiple parton interactions and has been shown to have fair agreement with ATLAS data in minimum-bias observations [28]. PYTHIA 6 is also run standalone as a 2 $\rightarrow$ 2 generator with the MRST LO** PDF set and the AUET2B tune. The AUET2B tune incorporates ATLAS [29] and CDF [30] jet-shape measurements as well as ATLAS fragmentation function measurements at $\sqrt{s} = 7$ TeV [31] and is thus expected to describe inclusive-jet properties well.

Additional pile-up events, which are superimposed on the hard-scattering event, are generated with either PYTHIA 6 [32] with the AUET2B tune using the MRST LO** PDF [33] set, or PYTHIA 8 [34] with the 4C tune [35] using the CTEQ 6L1 PDF set [36]. Choosing between these two pile-up simulations has negligible impact on the analysis. The number of pile-up events in the MC simulation is reweighted to match the conditions found in the data for each trigger selection. The events are passed through the ATLAS detector simulation [37], based on GEANT4 [38], and are reconstructed using the same software as for the data.

3.2 Jet reconstruction, selection and calibration

Jets are constructed from topological clusters of calorimeter cells [39] and calibrated using the EM+JES scheme [1,40]. This scheme is designed to adjust the energy measured in the calorimeter to that of the true particle jets on average. Calorimeter jets are reconstructed using the anti-$k_t$ jet algorithm [41,42] with a four-momentum recombination scheme and studied if calibrated transverse momentum $p_T > 20$ GeV and $|\eta| < 4.5$. Jet-finding radius parameters of both $R = 0.4$ and $R = 0.6$ are studied. Only jets with $|\eta| < 2.1$ are used for building the quark-jet tagger, to guarantee that the jet is well within the tracking acceptance. In the MC simulation, particle jets are reconstructed using the same anti-$k_t$ algorithm with stable, interacting particles as input to the jet algorithm. In all cases, jet finding is done in (rapidity, $\phi$) coordinates and jet calibration is done in $(\eta^{\text{jet}}, \phi^{\text{jet}})$ coordinates.

The reconstructed jets are additionally required to satisfy several data quality and isolation criteria. The data quality cuts are each designed to mitigate the impact of specific non-collision backgrounds [1]. Reconstructed and particle jets are considered isolated if there is no other reconstructed jet (or particle jet) within a cone of size $\Delta R = \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2} < 0.7$ around the jet axis. Only isolated jets are considered in this study. The jet vertex fraction (JVF) is calculated for each jet and used to reject jets originating from pile-up interactions. The JVF is built using information about the origin, along the direction of the beam, of tracks with $\Delta R < 0.4$ ($\Delta R < 0.6$) to the jet axis for $R = 0.4$ ($R = 0.6$) jets and describes the fraction of the jet’s charged particle $p_T$ associated with the primary vertex [40].

3.3 Track selection and associating tracks with jets

Tracks are associated with jets by requiring that the track momentum direction (calculated at the primary vertex) and the jet direction satisfy $\Delta R(\text{jet}, \text{track}) < 0.4$ ($\Delta R(\text{jet}, \text{track}) < 0.6$) for $R = 0.4$ ($R = 0.6$) jets. Track parameters are evaluated at the point of closest approach to the primary hard-scattering vertex, which is the vertex with the highest sum of associated track $p^2_T$. Tracks are required to have $p^4_T > 1$ GeV, at least one pixel hit and at least six hits in the silicon strip tracker, as well as transverse (longitudinal) impact parameters with respect to the hard-scattering vertex $|d_0| < 1$ mm ($|z_0 \cdot \sin(\theta)| < 1$ mm).

The studies in this paper were also performed with a requirement of track $p^4_T > 500$ MeV. No significant changes to the results were found. Requiring $p^4_T > 1$ GeV reduces the sensitivity to pile-up and the underlying event, and this requirement is used for the remainder of the paper. A “ghost association” [43] procedure was also tested in place of $\Delta R$-based matching, and no significant differences are observed. The jet isolation requirement helps to guarantee the similarity of the ghost association procedure and the $\Delta R$ matching.

3.4 Photon selection

Photons with $p^4_T > 25$ GeV are selected with pseudorapidity $|\eta| < 2.37$, excluding the transition region between the barrel and end-cap calorimeters (1.37 $< |\eta| < 1.52$). Only the leading photon in the event is considered. The photons are required to satisfy the preselection and “tight” photon cuts described in Ref. [44]. An additional isolation cut requiring less than 5 GeV of transverse energy in a cone of size $\Delta R = 0.4$ around the photon is imposed to increase the purity of the sample [40]. The photons are additionally required to be well separated from calorimeter defects and to not be within
\[ \Delta R < 0.4 \] of a jet arising from non-collision backgrounds or out-of-time pile-up.

### 3.5 Lepton selection

Isolated electrons and muons are used to select \( W + \text{jet} \) and \( t\bar{t} \) events. Electron candidates are formed by matching clusters found in the electromagnetic calorimeter to tracks reconstructed in the ID in the region \(|\eta| < 2.47\) and are required to have transverse energy \( E_T > 25 \) GeV. To ensure good containment of electromagnetic showers in the calorimeter, the transition region \( 1.37 < |\eta| < 1.52 \) is excluded as for photons. The electron candidates must pass the “tight” selection criteria based on the lateral and transverse shapes of the clusters described in Ref. [45] but updated for 2011 running conditions. Reconstructed tracks in the ID and the MS are combined to form muon candidates, which are selected in the region \(|\eta| < 2.5\) and are required to have \( p_T > 20 \) GeV. The selection efficiency for electrons and muons in simulated events, as well as their energy and momentum scale and resolution, are adjusted to reproduce those observed in \( Z \rightarrow \ell\ell \) events in data [45]. To reduce the contamination from jets identified as leptons, requirements are placed on the total momentum carried by tracks within \( \Delta R = 0.3 \) of the lepton and on calorimeter energy deposits within \( \Delta R = 0.2 \), excluding the track and energy of the lepton itself. For muons, the scalar sum of the \( p_T \) of these neighbouring tracks must be less than 2.5 GeV, while the sum of this close-by calorimeter \( E_T \) must be less than 4 GeV. For electrons, the sum of calorimeter \( E_T \) must be less than 6 GeV. Additionally, leptons are required to be consistent with originating from the primary hard-scattering vertex. They are required to have \( |z_0| < 10 \) mm, and the ratio of \( d_0 \) to its uncertainty (\( d_0 \) significance) must be smaller than 3.0 for muons and 10.0 for electrons, due to the wider distribution found in signal electrons caused by bremsstrahlung.

### 3.6 Trigger and event selection

All events must have a vertex with at least three associated tracks with \( p_T > 150 \) MeV. Other event selection requirements are described below.

#### 3.6.1 Dijet and trijet samples

The dijet sample is selected using single-jet triggers with various thresholds [46], which are fully efficient for jets with \( p_T > 40 \) GeV. Each jet \( p_T \) bin is filled exclusively by a single trigger that is fully efficient for jets in that \( p_T \) range, following Ref. [1]. The trijet sample uses the same trigger selection as the dijet sample. This guarantees that studies using the jet with the third highest \( p_T \) in each event are not biased by the trigger.

### 3.6.2 \( \gamma + \text{jet and } \gamma + 2\text{-jet samples} \)

The \( \gamma + \text{jet} \) sample is selected using single-photon triggers. The lowest threshold single-photon trigger is fully efficient for photons with \( p_T > 25 \) GeV. For this sample, a back-to-back requirement for the photon and the leading jet, \( \Delta \phi > 2.8 \), is imposed. An additional veto on soft radiation is also applied to further reduce background contamination [40]: the uncalibrated \( p_T \) of the sub-leading jet is required to be less than 30% of the photon \( p_T \). Relying on the \( p_T \) balance of the photon and jet, each jet \( p_T \) bin is filled exclusively by a single-photon trigger that provides a fully efficient selection.

The same triggers are used in the \( \gamma + 2\text{-jet} \) sample in each region of jet \( p_T \). Since the sub-leading jet \( p_T \) is lower than that of the leading jet by definition, this selection is also not biased by jet reconstruction effects.

#### 3.6.3 \( W + \text{jet sample} \)

The \( W + \text{jet} \) sample is selected using a single-electron or single-muon trigger. The event selection, following Ref. [47], requires exactly one charged lepton (electron or muon) and that it matches the trigger accepting the event, a transverse mass\(^3\) \( m_T > 40 \) GeV, missing transverse momentum \( E_T^{\text{miss}} > 25 \) GeV, and at most two jets (to reject \( t\bar{t} \) backgrounds). The triggers are fully efficient for electrons and muons satisfying the offline \( p_T \) requirements.

In events in which two jets are reconstructed, only the jet with the highest \( p_T \) is studied.

#### 3.6.4 \( t\bar{t} \) sample

Top quark pair events in which exactly one of the \( W \) bosons produced by the top quarks decays to an electron or a muon are selected as described in Ref. [48]. The event selection requires that exactly one electron or muon is reconstructed and that it matches the trigger accepting the event. Background suppression cuts of \( m_T > 40 \) GeV (\( m_T > 60 \) GeV) and \( E_T^{\text{miss}} > 25 \) GeV (\( E_T^{\text{miss}} > 20 \) GeV) in the electron (muon) channel, and at least four jets with \( p_T^{\text{jet}} > 25 \) GeV, \(|JVF| > 0.75\) and \(|\eta^{\text{jet}}| < 2.5\) are also required. Two of the selected jets must be identified as arising from a \( b \)-quark (\( b \)-tagged) using the MV1 algorithm, which combines several tracking variables into a multi-variate discriminant, with the 60% efficiency working point [49].

After this selection, the background contamination in the \( t\bar{t} \) sample is of the order of 10% and consists mainly of events
from \( W/Z + \text{jets} \) or single top-quark production. The contribution from multi-jet background after the requirement of two \( b \)-tagged jets is about 4% [48]. The background contamination in the selected data sample has no sizable impact in the studies performed. The change in the results when including the background in the analysis is small, and the sample is therefore assumed to be pure \( t\bar{t} \).

4 Jet labelling in Monte Carlo simulation

One natural definition of the partonic flavour of a jet in a Monte Carlo event is given by matching the jet to the closest outgoing parton (in \( \Delta R \)) from the matrix-element calculation, which represents a fixed-order QCD event record. In generators with \( 2 \rightarrow 2 \) matrix elements, such a matching scheme is clear only for the two leading jets at most. To simplify the task for analyses using different MC simulations, jets are matched to the highest-energy parton in the parton shower record within a \( \Delta R \) equal to the radius parameter of the jet algorithm. Using this method, only a small fraction of the jets (<1% around jet \( p_T = 50 \) GeV and fewer above 100 GeV) are not assigned a partonic flavour. Studies with PYTHIA 6 and MADGRAPH indicate that jets with significant energy contributions from more than one distinct parton (e.g. overlap of initial- and final-state radiation) are rare in the samples used. The jet isolation requirement restricts the wide-angle QCD radiation of the jet and further guarantees the accuracy of the labelling based on the parton shower record.

Jets are identified as originating from \( c \)- and \( b \)-quarks by requiring one \( c \)- or \( b \)-hadron with \( p_T > 5 \) GeV in the MC record within a \( \Delta R \) equal to the radius parameter of the jet. Jets with two \( c \)- or \( b \)-hadrons are identified as including a gluon splitting to \( c\bar{c} \) or \( b\bar{b} \). Both classes are considered separately from quark- and gluon-jets. The labelling of \( b \)-jets supersedes that of \( c \)-jets, which itself supersedes the quark and gluon labelling. In the samples used, other than \( t\bar{t} \), the fraction of heavy-flavour jets is relatively small. The variables used for quark- and gluon-jet discrimination are sufficiently different for each of these jet types to require an independent treatment.

In MC event generators with matching schemes [25,50, 51], it is possible to use the outgoing partons from the matrix-element calculation to label jets. Only jets above the matching scale can be identified in this manner, and only in exclusively showered events (i.e. events with the same number of jets at the matrix-element level and after showering). To avoid the need to tag jets originating from partons created in the parton shower, the matching scale must be chosen to be much lower than the minimum \( p_T \) of the jets for which the tagger is designed and commissioned. Labelling of jets based on the highest-energy parton is consistent with labelling based on the matrix-element calculation for isolated jets in the samples used here. The former is therefore used in this paper.

For the construction of templates and the examination of data, only ensembles of jets are considered. The parton record of the MC simulation is notused. Instead, the fractions of quark- and gluon-jets in each sample are calculated using the matrix-element event record, and only these fractions are used to describe the average composition of the jet ensemble.

5 Determination of quark-jet and gluon-jet properties

In previous theoretical [5] and experimental [40] studies, the jet width and the number of tracks associated with the jet were found to be useful for identifying the partonic origin of a jet. As discussed in Sect. 1, the larger colour factor associated with a gluon results in the production of a larger number of particles and a softer hadron \( p_T \) spectrum after the shower. To define the optimal discriminant, several jet properties are examined for their ability to distinguish the partonic origin of a jet and for their stability against various experimental effects, including pile-up. As these jet properties depend on the jet kinematics, the analysis of the properties and the resulting discriminant are separated into bins of jet \( p_T \) and \( \eta \). The \( p_T \) bin width is dictated by a combination of the jet resolution and the number of available events in data, and the \( \eta \) bins coarsely follow the detector features.

5.1 Discriminating variables

Useful discriminating variables, such as the number of particles associated with a jet, may be estimated using either the number of charged-particle tracks in the inner detector or using the number of topological clusters of energy inside the jet [40]. Although they are limited to charged particles, and thus miss almost half of the information in a typical jet, jet properties built from tracks have three practical advantages over calorimeter-based properties. First, they may include particles that have sufficiently low \( p_T \) that they are not measured by the calorimeter, or which are in the regime where the ID momentum measurement is more accurate than the energy measurement of the calorimeter. Second, charged particles bend in the magnetic field of the ID. Additional particles from the underlying event brought into the jet produce a background in the calorimeter, and particles that are sufficiently bent are lost to the calorimeter jet. However, both classes of particles can be correctly assigned using their momenta calculated at the interaction point. Third, tracks can be easily associated with a specific vertex. This association dramatically reduces the pile-up dependence of track-based observables. Similar arguments hold in the calculation of jet shape variables.
The variables surveyed as potential inputs to the quark/gluon tagging discriminant are:

- Number of reconstructed tracks ($n_{\text{trk}}$) in the jet.
- Calorimeter width:
  \[ w = \frac{\sum_i p_{T,i} \times \Delta R(i, \text{jet})}{\sum_i p_{T,i}}, \]
  where the sum runs over the calorimeter energy clusters that are part of the jet.
- Track width, defined similarly to the calorimeter width but with the sum running over associated tracks.
- Track-based energy–energy-correlation (EEC) angularity:
  \[ \text{angEEC} = \frac{\sum_i \sum_j p_{T,i} \times p_{T,j} \times (\Delta R(i, j))^\beta}{(\sum_i p_{T,i})^2}, \]
  where the index $i$ runs over tracks associated with the jet, $j$ runs over tracks associated with the jet while $j > i$, and $\beta$ is a tunable parameter [52,53].

The discriminating power ("separation") of a variable $x$ is calculated as in Ref. [54] to investigate the effectiveness of each variable in a quark/gluon tagger in a sample with equal fractions of quarks and gluons:

\[ s = \frac{1}{2} \int \frac{(p_q(x) - p_g(x))^2}{p_q(x) + p_g(x)} \text{d}x = \frac{1}{2} \sum_i \frac{(p_{q,i} - p_{g,i})^2}{p_{q,i} + p_{g,i}}, \]

where $p_q(x)$ and $p_g(x)$ are the normalised distributions of the variables for quark- and gluon-jets, and where the second expression applies to histograms, with the sum running over the bins of the histogram. This definition corresponds to the square of the statistical uncertainty that one would get in a maximum-likelihood fit when fitting for the fraction of quark- or gluon-jets using the given variable, divided by the square of the uncertainty in the case of perfect separation. While this is not a variable that relates easily to quantities of interest for tagging, its interpretation is independent of the shape of the distributions, allowing for comparisons that are independent of the tagging efficiency. Using this definition, Fig. 1 shows, for different variables, the separation between quark-jets and gluon-jets as a function of jet $p_T$ for jets built with the anti-$k_T$ algorithm with $R = 0.4$ using the PYTHIA 6 dijet MC simulation. In this simulation, the two most powerful variables are the EEC angularity with $\beta = 0.2$ and the number of tracks associated with the jet. The jet width built using the associated tracks is the weakest discriminant and the calorimeter-based width is somewhat stronger, and of comparable power to that of the EEC angularity with $\beta = 1.0$.

Track-based variables show excellent stability against pile-up and significant discrimination power between quark- and gluon-jets. The dependence of the mean calorimeter width on the number of reconstructed vertices is about five times stronger than the dependence of any of the variables considered for the final discriminant and at low jet $p_T$ is up to $\approx 1.5\%$ per primary vertex. At high jet $p_T$, the dependence is negligible for all variables. While it is possible to correct the inputs or to use a pile-up-dependent selection to allow the use of calorimeter-based variables without introducing a pile-up dependence in the tagger, such an approach is not pursued in this paper. Although Fig. 1 suggests using the charged particle multiplicity and the EEC angularity with $\beta = 0.2$ to build the tagger, a larger linear correlation between these two variables makes this tagger perform worse at high $p_T$ than the tagger built using the charged particle multiplicity and the track width. Furthermore, differences between data and MC simulation are reduced when using the latter tagger. For this reason, track width and $n_{\text{trk}}$ are used to build the discriminant used in the rest of this paper. The linear correlations between $n_{\text{trk}}$ and track width are at the 15% level at low $p_T$, increasing to 50% at high $p_T$. Thus, the variables add independent information about the properties of the jet. For EEC angularity with $\beta = 0.2$, the linear correlation with $n_{\text{trk}}$ is about 75% with a weak dependence on $p_T$. Still, the study of the EEC angularities and the evolution of their tagging performance as a function of $\beta$ is interesting for reasons discussed in Ref. [53]. Since this discussion is not relevant for the tagger developed in this paper, it is relegated to 1.

\[ \text{Springer} \]
5.2 Extraction of pure templates in data

To construct a discriminant, the properties of “pure” quark- and gluon-jets must be determined. As these properties depend on the modelling of non-perturbative effects, they are extracted from data to avoid reliance on MC simulations. The extraction can be performed using unbiased samples of pure quark- and gluon-jets or, alternatively, several mixed samples for which the admixture is known well theoretically. The use of pure samples is explored in detail in Sect. 6 as a validation procedure but is not used to determine the performance of the tagger in data, due to the limited number of events available and the difficulties in obtaining samples with negligible gluon and light-quark contaminations. The use of mixed samples is described below in detail, since it is used to create an operational tagger for data.

Distributions of properties of quark-jets or gluon-jets are extracted using the dijet and $\gamma$+jet event samples and the fraction of quark- and gluon-jets predicted by PYTHIA 6 with the AUET2B tune. For each bin of jet $\eta, p_T$, and jet property (track width, number of tracks, or the two-dimensional distribution of the these), a set of linear equations is solved:

$$P_i(\eta, p_T) = f_q(\eta, p_T) \times P_{q,i}(\eta, p_T) + f_g(\eta, p_T) \times P_{g,i}(\eta, p_T) + f_c(\eta, p_T) \times P_{c,i}(\eta, p_T) + f_b(\eta, p_T) \times P_{b,i}(\eta, p_T),$$

(1)

where $P_i$ is the value of the relevant distribution in bin $i$ of the distribution in the dijet or $\gamma$+jet sample, $f_q$ and $f_g$ are the light-quark and gluon fractions predicted by PYTHIA at a given $\eta$ and $p_T$, and $P_{q,i}$ and $P_{g,i}$ are the values of the relevant distribution for quark- and gluon-jets in bin $i$ of the distribution. The fractions $f_c$ and $f_b$ for c-jets and b-jets are relatively small. They are taken from the MC simulation, together with the corresponding distributions $P_c$ and $P_b$. The same is true for the fractions and distributions for $g \to c\bar{c}$ and $g \to b\bar{b}$, not shown in Eq. 1 for brevity. By using the different fractions of light quarks and gluons in dijet and $\gamma$+jet events in each $p_T$ and $\eta$ bin, the expected “pure” jet sample properties ($P_q$ and $P_g$) can be estimated. In these samples, the b-jet and c-jet fractions are typically below 5 – 10%. The studies are performed in three bins of $|\eta|$: $|\eta| < 0.8$, $0.8 < |\eta| < 1.2$ and $1.2 < |\eta| < 2.1$.

An additional term $f_{fake,i}(\eta, p_T) \times P_{fake,i}(\eta, p_T)$ must be added to the distributions in the $\gamma$+jet sample to account for events in which the reconstructed (“fake”) photon arises from a jet with energy deposits mostly within the electromagnetic calorimeter. The term is estimated from data using a sideband counting technique, developed and implemented in Refs. [40, 44]. The method uses regions defined with varying levels of photon isolation and photon identification criteria, estimating the number of background events in the signal region from those in the background regions, after accounting for signal leakage into the background regions.

Knowledge of $P_i$ and $f_i$ for the dijet and $\gamma$+jet samples allows the extraction of pure quark- and gluon-jet $n_{bk}$ and track width distributions from the data. The method can be tested in the MC simulation, comparing the properties of jets labelled in MC as quark- or gluon-jets and the properties extracted using Eq. (1) to demonstrate consistency. Figure 2 (top) shows the mean number of tracks and the mean track width as a function of the jet $p_T$, separated using either the MC flavour labels or the extraction procedure in the same MC events for jets with $|\eta| < 0.8$. Differences are observed between the average of the distributions in the dijet and $\gamma$+jet samples. This biases the extracted distributions for gluon-jets to be more like the gluon-jet properties in the dijet sample. The same is true for quark-jets and the $\gamma$+jet sample. The differences are larger at low $p_T$ and for the track width distributions. The bias demonstrates a sample dependence, which is included as a systematic uncertainty on the performance of the discriminant built from these jet properties. These differences are, however, small compared to the differences between quark- and gluon-jets, demonstrating the sensitivity of the extraction method. Similar results are obtained for jets reconstructed with radius parameter $R = 0.6$ and in other $|\eta|$ regions.

Figure 2 (bottom) shows the same MC simulation points as Fig. 2 (top), but here the data are used in the extraction. Relatively good agreement is found between data and PYTHIA AUET2B for the track width of gluon-jets and for the number of tracks in quark-jets. However, the mean number of associated tracks is significantly smaller for gluon-jets in the data than in the PYTHIA MC simulation. Similarly, the mean track width is larger in data than in the MC simulation for quark-jets.

Both these differences make gluon-jets and quark-jets more similar, reducing the discrimination power of these properties in data. Differences between the PYTHIA MC simulation and the data are also present in some of the other variables originally considered. These differences translate into non-negligible differences in the corresponding discriminants. For this reason, a fully data-driven tagger is built.

5.3 Systematic uncertainties on the extraction procedure

The distributions extracted from data can be used to build a data-driven tagger, and to evaluate its performance in data. Uncertainties on the extracted pure quark- and gluon-jet property templates are thus propagated through as uncertainties on the performance of the tagger. The systematic effects considered can be classified into four categories: uncertainties on the input fractions ($f_{x,i}$), uncertainties on the input shapes ($P_{x,i}$), uncertainties on the fake photon
Fig. 2 Average a, c $n_{\text{trk}}$ and b, d track width for quark- (solid symbols) and gluon-jets (open symbols) as a function of reconstructed jet $p_T$ for isolated jets with $|\eta| < 0.8$. Results are shown for distributions obtained using the in-situ extraction method in Pythia 6 simulation (black circles, a, b)) or data (black circles, c, d), as well as for labelled jets in the dijet sample (triangles) and in the $\gamma$+jet sample (squares). The error bars represent only statistical uncertainties. Isolated jets are reconstructed using the anti-$k_t$ jet algorithm with radius parameter $R = 0.4$.

The bottom panels show the ratio of the results obtained with the in-situ extraction method to the results in the dijet and $\gamma$+jet MC samples background, and sample-dependence effects. This last category includes, for example, differences in quark-jet properties between samples, which result in different quark-jet rejection across the various samples. This effect is the one that causes the inconsistency in the extraction method, illustrated in Fig. 2. Sample-dependent effects are included as a systematic uncertainty rather than deriving a separate tagger for each event selection and MC simulation.

Because jets with different observable properties have different calorimeter response, an additional uncertainty in the jet energy scale arises from the modelling of the response as a function of the discriminant in the MC simulation. The resulting uncertainties on the jet energy response after tagging, in addition to the standard jet energy scale uncertainties, are determined to be below 1 % using a $\gamma$+jet $p_T$-balance study following the procedures described in Ref. [40].
5.3.1 Input fraction uncertainties

The fraction of quark- and gluon-jets can change when going from a leading-order calculation to a next-to-leading-order (NLO) calculation, changing the renormalisation/factorisation scale, or changing the PDF set.

The first two effects are examined by comparing the PYTHIA and MADGRAPH calculations, which have different renormalisation/factorisation scales and different ways of simulating real emissions. Similarly, the potential effect of the real emissions is also probed by comparing the matrix-element labelling and the highest-energy parton labelling. A 5% uncertainty that is anti-correlated between quark- and gluon-jets is applied to cover the maximum variation seen in these comparisons. This uncertainty is uncorrelated amongst samples.

The potential mis-modelling of the fraction of quark- and gluon-jets in the MC simulation due to limitations of the PDFs is estimated using several PDF sets. The PDF sets use different fitting procedures (MRST, CTEQ and NNPDF sets), different orders in the perturbation theory expansion (MSTW2008lo for LO, CT10 for NLO) and different assumptions about the $\alpha_s$-calculation (MRST2007lmod for LO$^*$ and MRSTMCal for LO$^{**}$). A 5% uncertainty, anti-correlated between quark- and gluon-jets, conservatively covers the differences between the various PDF sets. This uncertainty is considered uncorrelated between the dijet and $\gamma+$jet samples because no significant trend is observed between the samples as the PDF set is changed.

5.3.2 Heavy-flavour input uncertainties

The fractions of $b$-jets and $c$-jets are varied by ±20% in the dijet sample, following Ref. [55], and by ±50% in the $\gamma+$jet sample to estimate a conservative uncertainty. As the fractions of $b$-jets and $c$-jets are small, these uncertainties remain sub-leading. The two input fractions are varied independently. The differences in the results obtained after the extraction of the pure quark- and gluon-jet properties are added in quadrature to obtain the total systematic uncertainty from this effect.

Uncertainties on the properties of $b$-jets are determined using a $t\bar{t}$ sample, described in Sect. 3. The purity of this sample is generally better than 95%. An envelope 10% uncertainty is included on the $b$-jet properties as a result of comparisons of $b$-jet properties between data and several MC simulations. The validation is performed using tagged jets. Differences between tagged and inclusive $b$-jets in the MC simulation are found to be within the assigned uncertainty.

For $c$-jets, several templates with 10% increases in the rates of 2-prong, 3-prong, and 4-prong decays are used to estimate the effect of changes to the $c$-hadron decay. These different $c$-jet distributions are propagated through the extraction procedure and the largest difference is used as the systematic uncertainty on the performance of the tagger due to this effect.

5.3.3 Fake photon background uncertainty

Several variations in the background to the $\gamma+$jet sample are considered. The identification requirements used to define the regions for the background estimation method are changed, resulting in purity differences of up to 10% for low-$p_T$ jets. The same procedure is used to estimate an uncertainty on the jet properties in the fake background. An uncertainty of up to 4% covers the changes in the means of the property distributions. These differences are propagated to the discriminant distribution to obtain a systematic uncertainty due to the purity estimate. An additional uncertainty covering the full shape correction to $P_{\text{fake}}$ for signal leakage into the background regions of the sideband counting method is included as well, amounting to less than a 3% change in the means of the property distributions.

5.3.4 Sample-dependence uncertainty

The application to a signal sample of a quark/gluon discriminant derived in a specific set of samples (or sample admixtures) rests upon the assumption that sample dependence is negligible, or that it can at least be parameterised as a function of visible properties of the event. One such property is the degree of isolation of the jet, which requires separate treatment. However, there are other effects, such as colour flow, that are much harder to constrain using the available data and may lead to a sample-dependence of jet properties.

Uncertainties on the jet properties are estimated first from differences between the $\gamma+$jet and dijet samples of the properties of quark- and gluon-jets. These are representative of the differences observed when comparing several different samples. Events generated with PYTHIA 6 and HERWIG++ are also tested for this effect. The envelope of these variations is used to estimate a systematic uncertainty due to the sample dependence of the jet properties. This systematic uncertainty is sensitive to statistical uncertainties in the MC simulation. These statistical uncertainties are estimated and used to smooth the $p_T$ dependence of the uncertainty following the procedure described in Ref. [56]. The sample dependence is consistently the dominant systematic uncertainty for all jet $p_T$ bins. The differences between MC labelled samples derive from differences in observable properties in the dijet and $\gamma+$jet samples. It is thus critical to consider these effects when estimating uncertainties on the tagging efficiency.

The properties of non-isolated jets differ from those of isolated jets, in general. In both the data and the MC simulation, isolated jet properties show no significant dependence on the $\Delta R$ to the nearest reconstructed jet for $\Delta R > 0.7$. As

---

the discriminant constructed here uses only jets satisfying this isolation criterion, no additional uncertainty due to the effect of jet non-isolation is applied.

An additional uncertainty arises from an incorrect description of the $p_T$-dependence of the tagging variables for samples with a significantly different jet $p_T$ spectrum from that of the dijet and $\gamma+\text{jet}$ samples with which the discriminant was constructed. This accounts for the differences in bin-to-bin migrations in the various samples. As this uncertainty is dependent entirely on the sample to which the discriminant is applied, it is not explicitly included here.

6 Validation with event-level kinematic cuts

The jet property templates extracted in the previous section can be further validated using high-purity quark- and gluon-jet samples. Largely following the work in Ref. [19], events are selected using basic kinematic cuts and event-level selection criteria to study purified samples of quark-jets and gluon-jets. These event selections are independent of the properties of individual jets and thus do not bias them. By including several different selections, the importance of colour flow and other sample-dependent effects can be evaluated using data.

The jets that are not tagged as $b$-jets in the $t\bar{t}$ sample, particularly in the case of events with exactly four jets, are mostly light-flavour jets. However, because of impurities introduced by gluon contamination and $W \rightarrow c\bar{s}$ decays, they are not sufficiently pure to be of use in this study.

6.1 Validation of gluon-jet properties

As protons have a large gluon component at low $x$, inclusive low-$p_T$ jet production at the LHC has a high rate of gluon-jet production. However, the fractions drop rapidly as jet $p_T$ increases. Particularly at moderate- and high-$|\eta|$, the relative rate of gluon-jet production exceeds 50% only below 150 GeV in jet $p_T$.

Multi-jet events from QCD contain relatively more gluon radiation than the inclusive jet sample. The radiation is typically soft, implying that the third-leading jet will often be a gluon-jet. A useful kinematic discriminant that can further purify a multi-jet sample, discussed in Ref. [19], is:

$$\xi = |\eta_3| - |\eta_1 - \eta_2|, \quad \text{(2)}$$

where $\eta_i$ is the pseudorapidity of the $i$th leading jet. A selection based on this variable can provide gluon-jet purity over 90%, at the price of significantly reduced efficiency.

To evaluate the modelling of gluon-jet properties, events in data with $\xi < 1$ are compared to those extracted using the template technique described in Sect. 5. The track multiplicity and jet width are shown in Fig. 3a, b. The mean values of properties obtained using the purified and (regular) mixed samples generally agree within statistical and systematic uncertainties. Systematic uncertainties in this figure are calculated as detailed in Sect. 5.3, and symmetrised around the central value.

6.2 Validation of quark-jet properties

Events containing photons are widely used as an enriched sample of quark-jets. By selecting events with photons produced in association with exactly one jet, a sample of quark-jets that is up to 80% pure for jets with $p_T > 150$ GeV can be constructed. Although the further enrichment of quark-jets in this sample is difficult, it is possible to obtain higher purities using events with a photon and two jets [19]. If no other selection cuts are applied, these events have a lower quark-jet fraction than inclusive $\gamma$-jet production. However, a kinematic selection can help to identify jets seeded by the parton that is most likely to have radiated the photon. As that parton must have had electric charge, selecting these jets enhances the purity of quark-jets and rejects gluon-jets.

Following Ref. [19], a variable is defined that allows the kinematic separation of quark-jets and gluon-jets:

$$\xi = \eta_{\text{jet 1}} \times \eta_\gamma + \Delta R_{\text{(jet 2, }\gamma)},$$

where $\eta_\gamma$ ($\eta_{\text{jet 1}}$) is the $\eta$ of the photon (leading jet), and $\Delta R_{\text{(jet 2, }\gamma)}$ gives the difference in $\eta$-$\phi$ space between the sub-leading jet and the photon. By imposing a requirement on this variable, purities over 90% can be achieved, although with a significant loss of events.

To evaluate the modelling of quark-jet properties, events with $\xi < 1$ are compared in data with those extracted using the template technique described in Sect. 5. The track multiplicity and jet width are shown in Fig. 3c, d. The two sets of data agree within statistical and systematic uncertainties. These results also hold in higher $|\eta|$ bins and for jets reconstructed with the anti-$k_t$ algorithm with $R = 0.6$.

Additionally, the production of a $W$ boson in association with a jet can be used to provide a relatively pure sample of quark-jets. A useful variable in constructing the sample is the jet “charge”, defined as

$$c_j = \frac{\sum q_i \times |\mathbf{p}_i \cdot \hat{j}|^{1/2}}{\sum |\mathbf{p}_i \cdot \hat{j}|^{1/2}},$$

where the sums run over all tracks associated with the jet, $\hat{j}$ is a unit three-vector pointing in the direction of the jet momentum, $\mathbf{p}_i$ is the track momentum three-vector, and $q_i$ is the track charge. This variable has been found to be useful in discriminating jets originating from positively charged quarks from those originating from negatively charged quarks [57–59].
The leading contribution to $W$ production results in a jet with charge opposite to that of the $W$ boson. The main backgrounds are from gluon-jets, including those in events with jets misidentified as leptons, which should have a charge distribution that is approximately Gaussian and centred at zero.\footnote{This is not quite the case, as the initial state at the LHC is more often positively charged than negatively charged.}

A pure sample of $W$ events, selected as described in Sect. 3, is divided into events in which the leading jet has a charge with the same sign as the identified lepton (SS) and those in which the charge is opposite (OS). Templates are then constructed for jet properties in the SS and OS samples, and the SS sample is used to subtract the gluon-jet contribution from the OS template.

Comparisons between the mean of the OS minus SS distributions in data and MC simulation are shown in Fig. 4.
7 Light-quark/gluon tagger construction and performance

The discriminant for quark- and gluon-jets is based on a simple likelihood ratio that uses the two-dimensional extracted distributions of $n_{\text{trk}}$ and track width for quark- and gluon-jets:

$$L = \frac{q}{q + g},$$

where $q$ ($g$) represents the normalised two-dimensional distribution for quark-jets (gluon-jets). A selection on $L$ is used in each bin to discriminate quark- and gluon-jets. This discriminant is built in bins of jet $p_T$ and $\eta$. The two-dimensional distributions are first smoothed using a Gaussian kernel and then appropriately rebinned to build the discriminant distribution in such a way that all bins are populated sufficiently.

The performance of the tagger is determined using the two-dimensional extracted distributions of $n_{\text{trk}}$ and track width in data and those obtained for labelled jets in MC simulations. Systematic uncertainties on the evaluated performance are estimated using alternative templates as described in Sect. 5.3. Table 1 summarises this performance for jets with $|\eta| < 0.8$. The efficiencies for gluon-jets and quark-jets are evaluated only at certain operating points with fixed light-quark efficiency. Statistical uncertainties are evaluated using pseudoexperiments. Systematic uncertainties are combined in quadrature and affect both the quark- and gluon-jet efficiency in data. Large differences between MC simulation and data in the variables used translate into large scale factors in the gluon-jet efficiency. Practically, analyses using this tagger would apply the appropriate MC tagger to MC simulation and the data tagger to data. These scale factors are needed for each MC tagger to create event weights for the MC simulation, so that the efficiency in the MC simulation matches the measured efficiency in such analyses. Three representative $p_T$ bins are shown in the table.

The difference in efficiency between data and MC simulation is particularly large for the tightest operating point at high $p_T$. It improves for the loosest operating points and is generally better for the lowest $p_T$ bins. The efficiencies extracted from data show a much weaker dependence on $p_T$ than is suggested by PYTHIA 6. No strong dependence on $\eta$ is observed in any sample. The performance obtained here in PYTHIA 6 compares well with the generator-level studies presented in Ref. [5]. The systematic uncertainties are dominated by the uncertainty due to the sample dependence.

The efficiencies of the tagger in MC simulation and in data are summarised in Fig. 5, where the performance estimated from labelled jets in dijet MC simulations and extracted data are shown. Two MC simulation-based taggers were used to produce this figure, one developed using distributions extracted in PYTHIA 6, which is applied to the PYTHIA 6 samples, and another derived from HERWIG++, used for the

The data show reasonable agreement with the MC simulation, generally within the statistical uncertainties. The points on these curves disagree at the 10% level with extracted or purified quark-jet results shown in previous figures due to a non-closure effect in the method observed at low $p_T$ in the MC simulation. Results from the $W+1$-jet MC simulation using generator-based labelling are in agreement with the quark-jet results from the dijet samples shown in Fig. 2.
Table 1 Summary of the performance of the quark-jet tagger on quark- and gluon-jets in data and PYTHIA6 MC simulation for jets built with the anti-$$k_t$$ algorithm with $$R = 0.4$$ and with $$|\eta| < 0.8$$

<table>
<thead>
<tr>
<th>Monte Carlo</th>
<th>Data</th>
<th>Scale factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$$\epsilon_{\text{quark}}$$ (%)</td>
<td>$$\epsilon_{\text{quark}}$$ (%)</td>
</tr>
<tr>
<td></td>
<td>$$\epsilon_{\text{quark}}$$ (%)</td>
<td>$$\epsilon_{\text{quark}}$$ (%)</td>
</tr>
<tr>
<td>$$p_T = 60–80$$ GeV</td>
<td>30</td>
<td>8.4</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>21.0</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>41.5</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>69.9</td>
</tr>
<tr>
<td>$$p_T = 110–160$$ GeV</td>
<td>30</td>
<td>5.7</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>13.9</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>29.7</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>64.8</td>
</tr>
<tr>
<td>$$p_T = 310–360$$ GeV</td>
<td>30</td>
<td>3.9</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>23.5</td>
</tr>
<tr>
<td></td>
<td>90</td>
<td>58.9</td>
</tr>
</tbody>
</table>

The first error corresponds to the statistical uncertainty, while the second corresponds to the systematic uncertainty. The scale factor is the ratio of data to MC simulation.

HERWIG++ samples. As expected from Sect. 5.2, the data do not agree well with either PYTHIA 6 or HERWIG++. Differences between data and PYTHIA 6 are within systematic uncertainties at low $$p_T$$, but are more significant at high $$p_T$$ for those points for which a large sample is available in the data. The tagger performs worse in HERWIG++ than on data at low $$p_T$$ (Fig. 5a), but there is fair agreement in its performance for high $$p_T$$ jets (Fig. 5b). Comparable results are observed for higher $$|\eta|$$ ranges, but with larger statistical uncertainties.

The performance can also be calculated using the relatively pure samples obtained in trijet and $$\gamma$$+jet events (see Sect. 6). The efficiencies obtained using purified samples are compared in Fig. 6 to those obtained using the extracted discriminant distribution. The agreement within systematic uncertainties, particularly in Fig. 5a, further validates the extraction method. Some small differences, like those in Fig. 5b, should be expected from impurities in the quark and gluon purified samples. A comparison of performance in jets with radius parameters of $$R = 0.4$$ and $$R = 0.6$$ in data and MC simulation is shown in Fig. 7. The performance is comparable with the two jet sizes.

8 Conclusions

Several variables that are sensitive to differences between quark- and gluon-jets were studied in various MC simulations and 4.7 fb$$^{-1}$$ of $$\sqrt{s} = 7$$ TeV pp collision data collected with the ATLAS detector at the LHC during the year 2011. Two of these variables, chosen to be relatively weakly correlated and stable against pile-up, were used to build a likelihood-based discriminant to select quark-jets and reject gluon-jets. Because of non-negligible differences in these variables between data and MC simulations, a data-driven technique was developed to extract the discriminant from the data and the MC simulations independently. This technique exploits significant, $$p_T$$ dependent differences in the quark- and gluon-jet content between dijet and $$\gamma$$+jet samples.

A detailed study of the jet properties reveals that quark- and gluon-jets look more similar to each other in the data than in the PYTHIA 6 simulation and less similar than in the HERWIG++ simulation. As a result, the ability of the tagger to reject gluons at a fixed quark efficiency is up to a factor of two better in PYTHIA 6 and up to 50% worse in HERWIG++ than in data. Gluon-jet efficiencies in data of $$\approx 11\%$$ and 80% are achieved for quark-jet efficiencies of $$\approx 30\%$$ and 90%, respectively. Relative uncertainties of $$\approx 5–50\%$$ ($$\approx 3–20\%$$) were evaluated for the estimate of these gluon-jet (quark-jet) efficiencies, with the uncertainties increasing for operating points with lower quark-jet efficiency. These uncertainties are dominated by differences in the properties of quark- and gluon-jets in the calibration samples (dijet and $$\gamma$$+jet) and are potentially caused by effects such as colour flow, which can make radiation around jets different for jets in different samples, even if they share the same partonic origin. These differences are predicted to be of different magnitude by the two generators studied and, without further insight, prevent final-state-dependent tags to be developed. The differences between the properties in the two samples are typical of the variations of the properties observed in other samples studied.
The likelihood-based discriminants were studied independently in kinematically purified gluon-jet and quark-jet samples in data. Agreement is found within systematic uncertainty between the properties that are used to build the discriminant for the pure samples and the mixed samples. The same is true for the tagger efficiencies.

Because their properties differ, the same likelihood-ratio discriminant cannot be applied to non-isolated jets. However,
using the methodology described in this paper, a discriminant for non-isolated jets with typical rejections and efficiencies comparable to those of the isolated-jet discriminant can be derived.

**Acknowledgments** We thank CERN for the very successful operation of the LHC, as well as the support staff from our institutions without whom ATLAS could not be operated efficiently. We acknowledge the support of ANPCyT, Argentina; YerPhI, Armenia; ARC, Australia; BMWF and FWF, Austria; ANAS, Azerbaijan; SSTC, Belarus; CNPq and FAPESP, Brazil; NSERC, NRC and CFI, Canada; CERN; CONICYT, Chile; CAS, MOST and NSFC, China; COLCIENCIAS, Colombia; MSMT CR, MPO CR and VSC CR, Czech Republic; DNRF, DNSRC and Lundbeck Foundation, Denmark; EPLANET, ERC and NSRF, European Union; IN2P3-CNRS, CEA-DSM/IRFU, France; GNSF, Georgia; BMBF, DFG, HGF, MPG and AvH Foundation, Germany; GSRT and NSRF, Greece; ISF, MINERVA, GIF, I-CORE and Benoziyo Center, Israel; INFN, Italy; MEXT and JSPS, Japan; CNRST, Morocco; FOM and NWO, Netherlands; BRF and RCN, Norway; MNiSW and NCN, Poland; GRICES and FCT, Portugal; MNE/IFA, Romania; MES of Russia and ROSATOM, Russian Federation; JINR; MSTD, Serbia; MSSR, Slovakia; ARRS and MIZŠ, Slovenia; DST/NRF, South Africa; MNI SW and NCN, Poland; GRICES and FCT, Portugal; MNE/IFA, Romania; MES of Russia and ROSATOM, Russian Federation; JINR; MSTD, Serbia; MSSR, Slovakia; ARRS and MIZŠ, Slovenia; DST/NRF, South Africa; MINECO, Spain; SRC and Wallenberg Foundation, Sweden; SER, SNSF and Cantons of Bern and Geneva, Switzerland; NSC, Taiwan; TAEK, Turkey; STFC, the Royal Society and Leverhulme Trust, United Kingdom; DOE and NSF, United States of America. The crucial computing support from all WLCG partners is acknowledged gratefully, in particular from CERN and the ATLAS Tier-1 facilities at TRIUMF (Canada), NDGF (Denmark, Norway, Sweden), CC-IN2P3 (France), KIT/GridKA (Germany), INFN-CNAF (Italy), NSC, Taiwan; TAEK, Turkey; STFC, the Royal Society and Leverhulme Trust, United Kingdom; DOE and NSF, United States of America. The crucial computing support from all WLCG partners is acknowledged gratefully, in particular from CERN and the ATLAS Tier-1 facilities at TRIUMF (Canada), NDGF (Denmark, Norway, Sweden), CC-IN2P3 (France), KIT/GridKA (Germany), INFN-CNAF (Italy), NL-T1 (Netherlands), PIC (Spain), ASGC (Taiwan), RAL (UK) and BNL (USA) and in the Tier-2 facilities worldwide.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

Funded by SCOAP3 / License Version CC BY 4.0.

**Appendix A: Performance of EEC angularities**

The EEC angularities described in Sect. 5.1 include a free parameter $\beta$ that affects the performance of the variables for quark/gluon discrimination. Recent studies [53] suggest that smaller values of the exponent $\beta$ provide stronger gluon-jet rejection for the same quark-jet efficiency. Figure 8 shows one minus the gluon-jet efficiency (for comparison with Ref. [53]) as a function of $\beta$, for a fixed quark-jet efficiency of 50%, in data and Pythia MC simulation. The MC simulation shows a weak dependence on $\beta$, with optimal performance for a $\beta$ value between 0.2 and 0.4. A similar trend is observable in data at high jet $p_T$. At low jet $p_T$, however, the performance falls off with lower $\beta$, with the highest few $\beta$ points showing comparable performance. The worst performance at all $p_T$ is given by $\beta = 0$, which uses the $p_T$ of the tracks without angular information, emphasising the high-$p_T$ tracks at the core of the jet for which the tracking momentum resolution is worse and inefficiencies and fakes due to shared hits between tracks in the detector become more common. At high jet $p_T$, the point at $\beta = 0$ shows reduced systematic uncertainties with respect to many of the other points, though the sample dependence is still significant.
from the two samples used by the method show differences in dependence of these variables. Since quark- and gluon-jets plate method developed in this paper is sensitive to the sample in building and testing the tagger in MC simulation. The template observed in data, as well as the importance of correct MC event generator tunes that describe the jet properties and their potential sample dependence. This serves to emphasise the importance of data-based validation of quark-jet/gluon-jet discriminants, as MC simulation may not correctly describe the jet properties observed in data, as well as the importance of correct MC event generator tunes that describe the jet properties and their potential sample dependence.

Fig. 8 One minus the gluon-jet efficiency (for comparison with Ref. [53]) as a function of $\beta$ (see Sect. 5.1) for EEC angularities, calculated using extracted jet properties from data (solid circles) and dijet PYTHIA 6 MC simulation (solid squares) for jets with $p_T < 80$ GeV and $210 < p_T < 260$ GeV and $|\eta| < 0.8$ reconstructed with the anti-$k_t$ algorithm with $R = 0.4$. The shaded bands represent the total systematic uncertainty on the data points.

It should be noted that the dependence on $\beta$ resembles more closely that found in Ref. [53] when using quark- and gluon-jets from either the dijet or $\gamma +$jet samples exclusively in building and testing the tagger in MC simulation. The template method developed in this paper is sensitive to the sample dependence of these variables. Since quark- and gluon-jets from the two samples used by the method show differences in these variables, the method is not capable of distinguishing this trend. The significant uncertainties on the data points in Fig. 8 are mostly from the uncertainties associated with this sample dependence.
105 Institute for Mathematics, Astrophysics and Particle Physics, Radboud University Nijmegen/Nikhef, Nijmegen, Netherlands
106 Nikhef National Institute for Subatomic Physics and University of Amsterdam, Amsterdam, The Netherlands
107 Department of Physics, Northern Illinois University, DeKalb, IL, USA
108 Budker Institute of Nuclear Physics, SB RAS, Novosibirsk, Russia
109 Department of Physics, New York University, New York, NY, USA
110 Ohio State University, Columbus, OH, USA
111 Faculty of Science, Okayama University, Okayama, Japan
112 Homer L. Dodge Department of Physics and Astronomy, University of Oklahoma, Norman, OK, USA
113 Department of Physics, Oklahoma State University, Stillwater, OK, USA
114 Palacký University, RCPTM, Olomouc, Czech Republic
115 Center for High Energy Physics, University of Oregon, Eugene, OR, USA
116 LAL, Université Paris-Sud and CNRS/IN2P3, Orsay, France
117 Graduate School of Science, Osaka University, Osaka, Japan
118 Department of Physics, University of Oslo, Oslo, Norway
119 Department of Physics, Oxford University, Oxford, UK
120 (a) INFN Sezione di Pavia, Pavia, Italy; (b) Dipartimento di Fisica, Università di Pavia, Pavia, Italy
121 Department of Physics, University of Pennsylvania, Philadelphia, PA, USA
122 Petersburg Nuclear Physics Institute, Gatchina, Russia
123 (a) INFN Sezione di Pisa, Pisa, Italy; (b) Dipartimento di Fisica E. Fermi, Università di Pisa, Pisa, Italy
124 Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA, USA
125 (a) Laboratorio de Instrumentacao e Fisica Experimental de Particulas-LIP, Lisbon, Portugal; (b) Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal; (c) Department of Physics, University of Coimbra, Coimbra, Portugal; (d) Centro de Física Nuclear da Universidade de Lisboa, Lisbon, Portugal; (e) Departamento de Fisica, Universidade do Minho, Braga, Portugal; (f) Departamento de Fisica Teorica y del Cosmos and CAFPE, Universidad de Granada, Granada, Spain; (g) Dep Fisica and CEFITEC of Faculdade de Ciencias e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
126 Institute of Physics, Academy of Sciences of the Czech Republic, Prague, Czech Republic
127 Czech Technical University in Prague, Prague, Czech Republic
128 Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic
129 State Research Center Institute for High Energy Physics, Protvino, Russia
130 Particle Physics Department, Rutherford Appleton Laboratory, Didcot, UK
131 Physics Department, University of Regina, Regina, SK, Canada
132 Ritsumeikan University, Kusatsu, Shiga, Japan
133 (a) INFN Sezione di Roma, Rome, Italy; (b) Dipartimento di Fisica, Sapienza Università di Roma, Rome, Italy
134 (a) INFN Sezione di Roma Tor Vergata, Rome, Italy; (b) Dipartimento di Fisica, Università di Roma Tor Vergata, Rome, Italy
135 (a) INFN Sezione di Roma Tre, Rome, Italy; (b) Dipartimento di Matematica e Fisica, Università Roma Tre, Rome, Italy
136 (a) Faculté des Sciences Ain Chock, Réseau Universitaire de Physique des Hautes Energies-Université Hassan II, Casablanca, Morocco; (b) Centre National de l’Energie des Sciences Techniques Nucleaires, Rabat, Morocco; (c) Faculté des Sciences Semlalia, Université Cadi Ayyad, LPHEA-Marrakech, Marrakech, Morocco; (d) Faculté des Sciences, Université Mohamed Premier and LPTPM, Oujda, Morocco; (e) Faculté des Sciences, Université Mohammed V-Agdal, Rabat, Morocco
137 DSM/IRFU (Institut de Recherches sur les Lois Fondamentales de l’Univers), CEA Saclay (Commissariat à l’Energie Atomique et aux Energies Alternatives), Gif-sur-Yvette, France
138 Santa Cruz Institute for Particle Physics, University of California Santa Cruz, Santa Cruz, CA, USA
139 Department of Physics, University of Washington, Seattle, WA, USA
140 Department of Physics and Astronomy, University of Sheffield, Sheffield, UK
141 Department of Physics, Shinshu University, Nagano, Japan
142 Fachbereich Physik, Universität Siegen, Siegen, Germany
143 Department of Physics, Simon Fraser University, Burnaby, BC, Canada
144 SLAC National Accelerator Laboratory, Stanford, CA, USA
Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovak Republic; (b) Department of Subnuclear Physics, Institute of Experimental Physics of the Slovak Academy of Sciences, Kosice, Slovak Republic

Department of Physics, University of Cape Town, Cape Town, South Africa; (b) Department of Physics, University of Johannesburg, Johannesburg, South Africa; (c) School of Physics, University of the Witwatersrand, Johannesburg, South Africa

Department of Physics, Stockholm University, Stockholm, Sweden; (b) The Oskar Klein Centre, Stockholm, Sweden

Physics Department, Royal Institute of Technology, Stockholm, Sweden

Departments of Physics and Astronomy and Chemistry, Stony Brook University, Stony Brook, NY, USA

Department of Physics and Astronomy, University of Sussex, Brighton, UK

School of Physics, University of Sydney, Sydney, Australia

Institute of Physics, Academia Sinica, Taipei, Taiwan

Department of Physics, Technion: Israel Institute of Technology, Haifa, Israel

Department of Physics, Aristotle University of Thessaloniki, Thessaloniki, Greece

International Center for Elementary Particle Physics and Department of Physics, The University of Tokyo, Tokyo, Japan

Graduate School of Science and Technology, Tokyo Metropolitan University, Tokyo, Japan

Department of Physics, Tokyo Institute of Technology, Tokyo, Japan

Department of Physics, University of Toronto, Toronto, ON, Canada

(a) TRIUMF, Vancouver, BC, Canada; (b) Department of Physics and Astronomy, York University, Toronto, ON, Canada

Faculty of Pure and Applied Sciences, University of Tsukuba, Tsukuba, Japan

Department of Physics and Astronomy, Tufts University, Medford, MA, USA

Centro de Investigaciones, Universidad Antonio Narino, Bogota, Colombia

Department of Physics and Astronomy, University of California Irvine, Irvine, CA, USA

(a) INFN Gruppo Collegato di Udine, Sezione di Trieste, Udine, Italy; (b) ICTP, Trieste, Italy; (c) Dipartimento di Chimica, Fisica e Ambiente, Università di Udine, Udine, Italy

Department of Physics, University of Illinois, Urbana, IL, USA

Department of Physics and Astronomy, University of Uppsala, Uppsala, Sweden

Instituto de Física Corpuscular (IFIC) and Departamento de Física Atómica, Molecular y Nuclear and Departamento de Ingeniería Electrónica and Instituto de Microelectrónica de Barcelona (IMB-CNM), University of Valencia and CSIC, Valencia, Spain

Department of Physics, University of British Columbia, Vancouver, BC, Canada

Department of Physics and Astronomy, University of Victoria, Victoria, BC, Canada

Department of Physics, University of Warwick, Coventry, UK

Waseda University, Tokyo, Japan

Department of Particle Physics, The Weizmann Institute of Science, Rehovot, Israel

Department of Physics, University of Wisconsin, Madison, WI, USA

Fakultät für Physik und Astronomie, Julius-Maximilians-Universität, Würzburg, Germany

Fachbereich C Physik, Bergische Universität Wuppertal, Wuppertal, Germany

Department of Physics, Yale University, New Haven, CT, USA

Yerevan Physics Institute, Yerevan, Armenia

Centre de Calcul de l’Institut National de Physique Nucléaire et de Physique des Particules (IN2P3), Villeurbanne, France

a Also at Department of Physics, King’s College London, London, UK
b Also at Institute of Physics, Azerbaijan Academy of Sciences, Baku, Azerbaijan
c Also at Particle Physics Department, Rutherford Appleton Laboratory, Didcot, UK
d Also at TRIUMF, Vancouver, BC, Canada
e Also at Department of Physics, California State University, Fresno, CA, USA
f Also at CPPM, Aix-Marseille Université and CNRS/IN2P3, Marseille, France
g Also at Università di Napoli Parthenope, Naples, Italy
h Also at Institute of Particle Physics (IPP), Canada
i Also at Department of Physics, St. Petersburg State Polytechnical University, St. Petersburg, Russia
j Also at Department of Financial and Management Engineering, University of the Aegean, Chios, Greece