Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC

The ATLAS Collaboration

DOI
10.1140/epjc/s10052-019-6847-8

Publication date
2019

Document Version
Final published version

Published in
European Physical Journal C

License
CC BY

Citation for published version (APA):
https://doi.org/10.1140/epjc/s10052-019-6847-8
Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC

ATLAS Collaboration
CERN, 1211 Geneva 23, Switzerland

Received: 24 August 2018 / Accepted: 3 April 2019 / Published online: 30 April 2019
© CERN for the benefit of the ATLAS collaboration 2019

Abstract The performance of identification algorithms (“taggers”) for hadronically decaying top quarks and W bosons in pp collisions at \( \sqrt{s} = 13 \) TeV recorded by the ATLAS experiment at the Large Hadron Collider is presented. A set of techniques based on jet shape observables are studied to determine a set of optimal cut-based taggers for use in physics analyses. The studies are extended to assess the utility of combinations of substructure observables as a multivariate tagger using boosted decision trees or deep neural networks in comparison with taggers based on two-variable combinations. In addition, for highly boosted top-quark tagging, a deep neural network based on jet constituent inputs as well as a re-optimisation of the shower deconstruction technique is presented. The performance of these taggers is studied in data collected during 2015 and 2016 corresponding to 36.1 fb\(^{-1}\) for the \( t \bar{t} \) and \( \gamma + \text{jet} \) and 36.7 fb\(^{-1}\) for the dijet event topologies.

Contents

1 Introduction ........................................... 1
2 ATLAS detector ...................................... 2
3 Data and simulated samples ......................... 2
4 Jet substructure techniques ....................... 3
  4.1 Jet reconstruction ............................... 4
  4.2 Jet labelling .................................. 4
  4.3 Tagging techniques ............................. 5
    4.3.1 Jet moments .............................. 5
    4.3.2 Topocluster-based Tagger .............. 6
    4.3.3 Shower deconstruction .................. 6
    4.3.4 HEPTopTagger ........................... 6
5 Tagger optimisation .................................. 6
  5.1 Cut-based optimisation ....................... 7
  5.2 Jet-moment-based multivariate taggers .......... 7
  5.3 Topocluster-based deep neural network tagger 11
  5.4 Shower deconstruction tagger ............... 12
  5.5 Summary of tagger performance studies in simulation 13
6 Performance in data ................................. 14
  6.1 Signal efficiency in boosted \( t \bar{t} \) events ........ 14
    6.1.1 Analysis and selection .................. 14
    6.1.2 Signal efficiencies ..................... 20
  6.2 Background rejection from multijet and \( \gamma + \text{jet} \) events ............. 25
    6.2.1 Analysis and selection ............... 28
    6.2.2 Background rejection measurements .... 29
  6.3 Systematic uncertainties ...................... 32
7 Conclusion .......................................... 35
Appendix .............................................. 38
A BDT and DNN hyper-parameters .................... 38
References ............................................. 38

1 Introduction

With the increase of the Large Hadron Collider (LHC) [1] centre-of-mass energy to 13 TeV in Run 2, it is important for searches for physics phenomena beyond the Standard Model to probe processes involving highly boosted massive particles, such as W and Z bosons and top quarks [2–4], as well as Standard Model measurements using these techniques [5–7]. To fully exploit these final states, it is important to reconstruct and accurately identify the hadronic decay modes of these massive particles which serve as an effective tool to reject events produced by background processes and improve the sensitivity in searches for physics beyond the Standard Model. Techniques to achieve this aim were studied by both the ATLAS and CMS collaborations during the course of Run 1 of the LHC [8–11]. In this paper, these studies are performed with Run 2 data with particular attention to the investigation of multivariate techniques based on both jet shape observables and an approach using the jet constituents as input observables in addition to the optimisation of the shower deconstruction technique for highly boosted top-quark tagging.

In Sect. 2 the ATLAS detector is briefly described, followed by a description of the Monte Carlo and data samples...
used in the analysis in Sect. 3. The set of jet reconstruction and tagging techniques investigated in this work is described in Sect. 4. The optimisation procedure for each tagger, as well as a comprehensive comparison of the tagging techniques using Monte Carlo simulation are presented in Sect. 5. In Sect. 6, the pp collision data recorded in 2015 and 2016 are used to evaluate the performance of these tagging techniques, with the measurement of signal and background efficiencies using boosted lepton+jet \( \ell \ell \), dijet and \( \gamma + \text{jet} \) topologies and the robustness of the various techniques when confronted with varying levels of event pile-up. Finally, concluding remarks are given in Sect. 7.

2 ATLAS detector

The ATLAS detector [12,13] at the LHC covers nearly the entire solid angle around the collision point.\(^1\) It consists of an inner tracking detector (ID) surrounded by a thin superconducting solenoid, electromagnetic and hadronic calorimeters, and a muon spectrometer composed of three large superconducting toroid magnets and precision tracking chambers. For this study, the most important subsystems are the calorimeters, which cover the pseudorapidity range \( |\eta| < 4.9 \). Within the region \( |\eta| < 3.2 \), electromagnetic calorimetry is provided by barrel and endcap high-granularity lead/liquid-argon (LAr) sampling calorimeters, with an additional thin LAr presampler covering \( |\eta| < 1.8 \) to correct for energy loss in material upstream of the calorimeters. Hadronic calorimetry is provided by a steel/scintillator-tile calorimeter, segmented into three barrel structures within \( |\eta| < 1.7 \), and two copper/LAr hadronic endcap calorimeters which instrument the region \( 1.5 < |\eta| < 3.2 \). The forward region \( 3.1 < |\eta| < 4.9 \) is instrumented with copper/LAr and tungsten/LAr calorimeter modules.

Inside the calorimeters, the inner tracking detector measures charged-particle trajectories in a 2 T axial magnetic field produced by the superconducting solenoid. It covers a pseudorapidity range \( |\eta| < 2.5 \) with pixel and silicon microstrip detectors, and the region \( |\eta| < 2.0 \) with a straw-tube transition radiation tracker.

The muon spectrometer (MS) comprises separate trigger and high-precision tracking chambers measuring the deflection of muons in a magnetic field generated by superconducting air-core toroid magnets. The precision chamber system covers the region \( |\eta| < 2.7 \) with three layers of monitored drift tubes, complemented by cathode strip chambers in the forward region where the background is highest. The muon trigger system covers the range \( |\eta| < 2.4 \) with resistive plate chambers in the barrel and thin gap chambers in the endcap regions.

A two-level trigger system is used to select events for offline analysis [14]. The first step, named the level-1 trigger, is implemented in hardware and uses a subset of detector information to reduce the event rate from 40 MHz to 100 kHz. This is followed by a software-based high-level trigger which reduces the final event rate to an average of 1 kHz.

3 Data and simulated samples

The taggers described in this article were initially designed, as described in Sect. 5, using Monte Carlo (MC) simulated samples for two signal processes (i.e. events containing the decay of heavy resonances) and one background process (i.e. light quark and gluon jets). The dijet process was used to simulate jets from gluons and non-top quarks. It was modelled using the leading-order Pythia8 (v8.186) [15] generator with the NNPDF2.3LO [16] parton distribution function (PDF) set and a set of tuned parameters called the A14 tune [17]. Events were generated in slices of leading jet transverse momentum \( (p_T) \) to sufficiently populate the kinematic region of interest (between 200 and 2500 GeV). Event-by-event weights were applied to correct for this generation methodology and to produce the expected smoothly falling \( p_T \) distribution of the multijet background. The signal samples containing either high-\( p_T \) top-quark or W-boson jets were obtained from two physics processes modelling phenomena beyond the Standard Model. For the W-boson sample, high-mass sequential standard model [18] \( W \rightarrow WZ \rightarrow q\bar{q}q\bar{q} \) events were used. For the top-quark sample, high-mass sequential standard model \( Z' \rightarrow t\bar{t} \) events were used as a source of signal jets. Both the W bosons and top quarks were required to decay hadronically. The two signal processes were simulated using the Pythia8 [15] generator with the NNPDF2.3LO PDF set and A14 tune for multiple values of the resonance (\( W \) or \( Z' \) boson) mass between 400 and 5000 GeV in order to populate the entire \( p_T \) range\(^2\) from 200 to 2500 GeV and to reduce the impact of MC statistical uncertainties on the calculated signal efficiencies.

For the study of W-boson and top-quark jets in data, described in Sect. 6, a number of MC samples are needed

---

\(^1\) 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 upwards. Cylindrical coordinates \((r, \phi)\) are used in the transverse plane, \(\phi\) being the azimuthal angle around the z-axis. The pseudorapidity is defined in terms of the polar angle \( \theta \) as \( \eta = -\ln \tan(\theta/2) \). Angular distance is measured in units of \( \Delta R \equiv \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2} \).

\(^2\) As the combination of these signal samples with different generated heavy resonance masses results in irregular top-quark and W-boson \( p_T \) distributions, the events are reweighted at the generator level to either a constant or a falling jet \( p_T \) distribution, as is typical for light jets. This procedure is described in Sect. 5.2.
to model both the $t\bar{t}$ signal and backgrounds. The POWHEG-Box v2 generator [19–21] was used to simulate $t\bar{t}$ and single-top-quark production in the $W$- and $s$-channels at next-to-leading order (NLO), while for the single-top-quark $t$-channel process, the NLO POWHEG-Box v1 generator and the CT10 [22] NLO PDF set was used. For all processes involving top quarks, the parton shower, fragmentation, and the underlying event were simulated using PYTHIA6 (v6.428) [23] with the CTEQ6L1 [24] PDF set and the corresponding Perugia 2012 tune (P2012) [25]. The top-quark mass was set to 172.5 GeV. The $h_{\text{damp}}$ parameter, which controls the matching of the matrix element to the parton shower, was set to the mass of the top quark. The $t\bar{t}$ process is normalised to the cross-sections predicted to next-to-next-to-leading order (NNLO) in $\alpha_S$ and next-to-next-to-leading logarithm (NNLL) in soft-gluon terms while the single-top-quark processes are normalised to the NNLO cross-section predictions [26].

Several additional variations of the $t\bar{t}$ generator are used for the estimation of modelling uncertainties. Estimates of the parton showering, hadronisation modelling and underlying-event uncertainty are derived by comparing results obtained with the POWHEG-Box v2 generator interfaced to HERWIG++ (v2.7.1) [27] instead of PYTHIA6. To estimate the hard-scattering modelling uncertainty, the NLO MADGRAPH5_AMC@NLO (v2.2.1) generator [28] (hereafter referred to as MC@NLO) is used with PYTHIA6. To estimate the uncertainty in the modelling of additional radiation, the POWHEG-Box v2 generator with PYTHIA6 is used with modified renormalisation and factorisation scales ($\times 2$ or $\times 0.5$) and a simultaneously modified $h_{\text{damp}}$ parameter value ($h_{\text{damp}} = m_{\text{top}}$ or $h_{\text{damp}} = 2 \times m_{\text{top}}$) as described in Ref. [29].

Samples of $W/Z+$jets and Standard Model diboson ($W/W/Z/ZZ$) production were generated with final states that include either one or two charged leptons. The SHERPA [30] generator version 2.1.1 and version 2.2.1 were used to simulate these processes at NLO with the CT10 PDF set to simulate the diboson and $W/Z+$jets production processes, respectively. The $W/Z+$jets events are normalised to the NNLO cross-sections [31].

For the study of $\gamma +$jet events in data, events containing a photon with associated jets were simulated using the SHERPA 2.1.1 generator, requiring a photon transverse momentum above 140 GeV. Matrix elements were calculated with up to four partons at LO and merged with the SHERPA parton shower [32] using the ME+PS@LO prescription [33]. The CT10 PDF set was used in conjunction with the dedicated parton shower tune developed by the SHERPA authors.

The MC samples were processed through the full ATLAS detector simulation [34] based on GEANT4 [35]. Additional simulated proton–proton collisions generated using PYTHIA8 (v8.186) with the A2M [17] tune and MSTW2008LO PDF set [36] were overlaid to simulate the effects of additional collisions from the same and nearby bunch crossings (pile-up), with a mean number of 24 collisions per bunch crossing. All simulated events were then processed using the same reconstruction algorithms and analysis chain as is used for the data.

Data were collected in three broad categories to study the signal efficiency and background rejection. For the signal, a set of observed top-quark and $W$-boson jet candidates is obtained from a sample of $t\bar{t}$ candidate events in which one top quark decays semileptonically and the other decays hadronically, the lepton-plus-jets decay signature. The background is studied using data samples enriched in dijet events and $\gamma +$jet events. In addition to covering different $p_T$ regions, the dijet and $\gamma +$jet samples differ in what partons initiated the jets under study. In the $\gamma +$jet topology the jets are mostly initiated by quarks over the full $p_T$ range studied, while for the dijet topology the fraction of quarks initiating jets is slightly smaller than the gluon fraction at low $p_T$ and becomes large at high $p_T$. The data for the $t\bar{t}$ and $\gamma +$jet studies were collected during normal operations of the detector and correspond to an integrated luminosity of 36.1 fb$^{-1}$. For the dijet analysis, additional data where the toroid magnet was turned off are used. This adds an additional 0.6 fb$^{-1}$. For both datasets, only data collected while all relevant detector subsystems were fully functional and in which at least one primary vertex was reconstructed with at least five associated ID tracks consistent with the LHC beam spot are used [37].

The lepton-plus-jets events were collected with a set of single-electron and single-muon triggers that became fully efficient for $p_T$ of the reconstructed lepton greater than 28 GeV. The dijet events were collected with a single large-$R$ jet trigger, where the jet was reconstructed using the same algorithm described in Sect. 4.1 and with a radius parameter of $R = 1.0$. This trigger became fully efficient for an offline jet $p_T$ of approximately 450 GeV. The $\gamma +$jet jet events were collected with a single-photon trigger that became fully efficient for an offline photon $p_T$ of approximately 155 GeV.

4 Jet substructure techniques

The identification of hadronic jets originating from the decay of boosted $W$ and $Z$ bosons and top quarks can broadly, and somewhat arbitrarily, be divided into two stages: jet reconstruction and jet tagging. In the first, the hadronic energy flow of the event is exclusively divided into a number of jets, composed of constituents, with the primary goal being to most accurately reconstruct the interesting energy flows in the case of true signal jets while suppressing contributions from the underlying event and event pile-up. In the second, the information about the jet constituents is distilled into a single observable by different means to obtain a criterion
by which to identify a jet as originating from a hadronically decaying massive particle, such as a $W$ boson or a top quark. A number of techniques and observables pertaining to these two categories have been described and investigated extensively in previous work [8,9] with only a short summary of the relevant techniques presented here. In the case of the identification of $W$ bosons, the techniques and conclusions are more broadly applicable to both $W$ and $Z$ bosons, with dedicated studies concerning the separation of $W$-boson jets from $Z$-boson jets performed in Ref. [38].

4.1 Jet reconstruction

In this work, jets are reconstructed with the intention of capturing the full energy flow resulting from the decay of a massive particle. This reconstruction primarily uses inputs in the form of noise-suppressed topological clusters of calorimeter cells [39] that are individually calibrated to correct for effects such as the non-compensating response of the calorimeter and inactive material, and which are assumed to be massless [40]. These topoclusters are then used as inputs to build two different types of jets. The first uses the anti-$k_T$ algorithm [41] with a radius parameter $R = 1.0$ to form jets which are further trimmed to remove the effects of pile-up and the underlying event. Trimming [44] is a grooming technique in which the original constituents of the jets are reclustered using the $k_T$ algorithm [45] with a radius parameter $R_{\text{sub}}$ to produce a collection of subjets. These subjets are then discarded if they have less than a specific fraction ($f_{\text{cut}}$) of the $p_T$ of the original jet. The trimming parameters used here are $R_{\text{sub}} = 0.2$ and $f_{\text{cut}} = 0.05$. These large-$R$ jets are then calibrated in a two-step procedure that first corrects the jet energy scale and then the jet mass scale [40,46]. The resulting set of constituents forms the basis from which further observables are calculated. The second type of jet clustering, needed for the HEPTopTagger algorithm [47,48], makes use of the Cambridge/Aachen (C/A) jet algorithm [49,50] with a radius parameter of $R = 1.5$ which aims to identify top-quark jets across a broad $p_T$ range, in particular reaching low $p_T$. These jets, used in conjunction with the HEPTopTagger algorithm described in Sect. 4.3.4, are also groomed to mitigate the effects of pile-up. Trimming with subjet radius parameter of $R_{\text{sub}} = 0.2$ and momentum fraction $f_{\text{cut}} = 0.05$, the same as those used in the trimming of the anti-$k_T$ $R = 1.0$ jet collection, is found to produce jet reconstruction and identification performance independent of the average number of interactions per bunch crossing.

In simulation, in addition to jets reconstructed from detector-level observables, a set of jets based on generator-level information is also used to characterise the performance of a given tagging algorithm. These jets are reconstructed with the anti-$k_T$ algorithm with a radius parameter $R = 1.0$, using stable particles from the hard scatter with lifetimes greater than 10 ps, excluding muons and neutrinos, as constituents. These jets, to which no trimming algorithm is applied, are referred to as truth jets, and the related observables are denoted by the superscript “true”.

4.2 Jet labelling

As the aim of this study is the evaluation of the performance of jet tagging algorithms, the labelling of the particle that initiated the jet is of particular importance. For signal jets, this labelling is based on the partonic decay products of the particle of interest ($W$ boson or top quark) in a three-step process. First, reconstructed jets are matched to truth jets with a matching criterion of $\Delta R_j^{\text{true}} < 0.75$. Next, those truth jets are matched to truth $W$ bosons and top quarks ($W$, $t$) with a matching criterion of $\Delta R_j^{\text{true, particle}} < 0.75$. Finally, the partonic decay products of the parent $W$ boson or top quark (two quarks for hadronically decaying $W$ bosons and an additional $b$-quark) are matched to the reconstructed jet. A reconstructed jet is labelled as a $W$-boson or top-quark jet if the parent particle and all of its direct decay products are contained within a region in $(\eta, \phi)$ with $\Delta R < 0.75 \times R_j^{\text{true}}$, where $R_j^{\text{true}}$ is the jet radius parameter. In the case of $W$ bosons, this means that both of the daughter partons from the $W \rightarrow q\bar{q}'$ decay are contained within the jet. For jets matched to the parent $W$ boson, at $p_T \sim 200$ GeV only 50% of the jets are fully contained when using this criterion while for $p_T > 500$ GeV the containment rises to nearly 100%. In the case of top-quark jets, the possible final-state topologies for the jet are more complex, including the possibility of the large-$R$ jet containing only the $b$-quark from the top decay, only the two quarks from the $W$-boson decay, or a pairing of a $b$-quark and one of the daughter $W$-boson quarks within $\Delta R < 0.75 \times R_j^{\text{true}}$ around the jet axis. As seen in Fig. 1, the fraction of large-$R$ jets falling into each category depends strongly on the $p_T$ of the parent particle with only 60% of jets being fully contained at 600 GeV and with 100% containment not being reached even at 1500 GeV. The value $0.75 \times R_j^{\text{true}}$ for the jet labelling criteria is chosen as a compromise between the resulting labelling efficiency and the resolution of the top-quark and $W$-boson jet mass peak. The jet $p_T$ dependence of the variation in containment, particularly in the case of top-quark tagging in which a top-quark jet is labelled as such only when the top parton, the $b$-quark from its decay as well as the two light quarks from the subsequent $W$-boson decay are contained within the region $\Delta R < 0.75 \times R_j^{\text{true}}$ around

---

3 When using topoclusters [39] as jet inputs, the trimming algorithm, as opposed to pruning [42] or split-filtering [43] was found to be optimal in terms of accurately reconstructing the important aspects of the energy flow as shown in Ref. [8].
the jet axis, serves as a strong motivation for the various optimisation strategies described in Sect. 5.

4.3 Tagging techniques

After reconstructing the jet as a collection of constituents, a number of methods can be used to classify a jet as originating from a heavy particle (W boson or top quark) decay as opposed to a light jet originating from gluons and quarks of all flavours other than top quarks. The motivation behind the various techniques differs, but they all attempt to form a decision criterion by which to identify a jet as originating from a W boson or top quark.

4.3.1 Jet moments

The first broad class of observables studied for classification are directly based on the constituents of the trimmed jet and attempt to quantify a particular feature of the jet in an analytic way. Of these features, the most powerful is the jet mass, which for a jet formed from the decay of a heavy particle has a scale associated with the mass of the particle, whereas for light jets high masses are less likely as they need to be generated through QCD emissions. Traditionally, the jet mass was calculated as the invariant mass of the collection of topclusters of the trimmed jet (m_{calo}) [8]. However, at very high p_T, the resolution of this observable decreases when energy depositions from individual particles begin to merge in clusters. To mitigate this effect, the fine spatial granularity of the inner detector is used to calculate the jet mass as the invariant mass of the ghost-associated [51] charged-particle tracks scaled by the ratio of the transverse momenta of the trimmed jet and the associated tracks to form the track-assisted mass (m^{TA}). To achieve good performance across a broad range of jet transverse momenta, an average of m_{calo} and m^{TA}, weighted by the inverse of their resolutions is calculated to form the combined mass (m^{comb}) [46].

In addition to the jet mass, a number of other observables quantify the extent to which the jet constituents are clustered or uniformly dispersed and can be used to augment the discrimination power from the jet mass alone. This can be done by explicitly using a set of axes (e.g. N-subjettiness, τ_{21} and τ_{32}), declustering the jet (e.g. splitting measures, $\sqrt{d_{12}}$ and $\sqrt{d_{23}}$), or using all jet constituents to quantify the dispersion of the jet constituents in an axis-independent way (e.g. planar flow or energy correlation functions). In previous ATLAS studies [8,9], it was found that for W boson tagging, energy correlation variables, in particular $D_2$, were the best-performing tagging observables while for top-quark tagging the N-subjettiness ratio, τ_{32}, was found to be optimal among the techniques considered. This can be understood from an analytical point of view in the context of W-boson tagging [52] and is attributed to additional wide-angle radiation present in parton jets originating from W-boson decays, which is more fully exploited in the energy correlation functions than in the N-subjettiness moments.

The full set of jet moments studied in this work is summarised in Table 1 while a more complete description of the observables under study can be found in Ref. [8]. These moments are studied individually when paired with the jet mass (m^{comb}) as well as in multivariate combinations, similar to those studied in Refs. [10,53,54], with the intention of exploiting correlations between the observables and creating a more powerful single discriminant across a broad p_T range from 200 to 2000 GeV, the range commonly probed in searches.

![Fig. 1 Containment of the W-boson (a) and top-quark (b) decay products in a single truth-level anti-k_t R = 1.0 jet as a function of the particle’s transverse momentum](image-url)
4.3.2 Topocluster-based Tagger

All of the jet moments presented in Sect. 4.3.1 and summarized in Table 1 make use of a specific physical motivation to distil the individual jet constituent measurements into a single observable. However, recent simulation-based studies have found that the more direct use of the jet constituents as inputs to a machine-learning algorithm can lead to significant improvements in discriminating power as compared to more traditional, jet-moment-based discriminants. Therefore, in this work, a classifier that makes use of lower-level input observables is investigated which focuses specifically on the identification of high-\(p_T\) top quarks with \(p_T > 450\) GeV. This classifier is referred to as “TopoDNN” throughout the work.

4.3.3 Shower deconstruction

Shower deconstruction (SD) [70] is an approach which attempts to classify jets according to the compatibility of the radiation pattern of the jet with a predefined set of parton shower hypotheses in a manner similar to the matrix element method [71]. For a set of input subjets, intended to be representative of the partonic decay products of the top quark, loose compatibility with the decay of a top quark is ensured by requiring that the jet has at least three subjets, that two or more subjets have a mass in a window centred around the \(W\)-boson mass (\(\Delta m_W\)), and that at least one more subjet can be added to obtain a total mass in a window centred around the top-quark mass (\(\Delta m_{\text{top}}\)). If the jet passes these requirements, then a set of potential shower histories is constructed for the signal and background models. Each shower history represents a possible means by which the chosen model could have resulted in the given subject configuration. A probability is assigned to each shower history based on the parton shower model from which the \(\chi\) variable is defined as the likelihood ratio of the signal and background hypotheses. The logarithm of this likelihood ratio \(\log \chi\) is used as the final discriminant. The precise values of the parameters in this algorithms are described in Sect. 5.4.

4.3.4 HEPTopTagger

An alternative approach to top-quark tagging is the HEPTopTagger (HTT) algorithm [47,48]. Unlike the previous observables that are calculated from the constituents of the \(R = 1.0\) trimmed jets, this technique relies on reconstructing jets using the C/A algorithm with \(R = 1.5\) to allow the tagging of fully contained boosted top quarks to be effective at lower values of \(p_T\) (> 200 GeV) and to take advantage of the C/A clustering sequence which attempts to reverse the decay structure of the top-quark decay. The constituents of the ungroomed uncalibrated C/A jet are analysed with the HEPTopTagger algorithm, which identifies the hard jet substructure and tests it for compatibility with the 3-prong pattern of hadronic top-quark decays using an algorithm which is designed to mitigate the effects of pile-up by removing low-\(p_T\) portions of the jet. The HEPTopTagger studied in this paper is the original algorithm, from Ref. [47], not the extended HEPTOPTAGGER2 algorithm [72] and is executed with \(m_{\text{cut}} = 50\) GeV, \(R_{\text{filt}}^{\text{max}} = 0.25\), \(N_{\text{filt}} = 5\), \(f_W = 15\%\), settings found to be optimal in Ref. [9]. The result of the algorithm is a top-quark-candidate four-vector. The jet is considered to be tagged if the mass of this resultant top-quark-candidate four-vector is between 140 and 210 GeV and its \(p_T\) is larger than 200 GeV.

5 Tagger optimisation

A wide variety of techniques, described in Sect. 4, exist for identifying \(W\)-boson and top-quark jets. In this section, each of these techniques is explored and optimised and an inclusive comparison of the performance of each technique is made based on the \(W\)-boson or top-quark (signal) efficiency and light-jet (background) rejection, defined as the inverse of the background efficiency. This performance is quantified in exclusive kinematic regimes based on the \(p_T\) of the associated anti-\(k_T\), \(R = 1.0\) truth jet (\(p_T^{\text{true}}\)) to more closely resemble the kinematics of the parent particle and allow comparison of taggers employing different jet clustering algorithms. Finally, to mitigate any bias in the tagging performance due to differences between the \(p_T\) spectra of the signal and background,

<table>
<thead>
<tr>
<th>Observable</th>
<th>Variable</th>
<th>Used for</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibrated jet kinematics</td>
<td>(p_T, m_{\text{comb}})</td>
<td>Top, (W)</td>
<td>[46]</td>
</tr>
<tr>
<td>Energy correlation ratios</td>
<td>(e_3, C_2, D_2)</td>
<td>Top, (W)</td>
<td>[52,56]</td>
</tr>
<tr>
<td>N-subjettiness</td>
<td>(\tau_{1,2,21})</td>
<td>Top, (W)</td>
<td>[57,58]</td>
</tr>
<tr>
<td>Fox–Wolfram moment</td>
<td>(R_{\text{FW}}^2)</td>
<td>Top</td>
<td>[59,60]</td>
</tr>
<tr>
<td>Splitting measures</td>
<td>(z_{\text{cut}})</td>
<td>(W)</td>
<td>[61,62]</td>
</tr>
<tr>
<td>Planar flow</td>
<td>(P)</td>
<td>(W)</td>
<td>[63]</td>
</tr>
<tr>
<td>Angularity</td>
<td>(a_3)</td>
<td>(W)</td>
<td>[64]</td>
</tr>
<tr>
<td>Aplanarity</td>
<td>(A)</td>
<td>(W)</td>
<td>[60]</td>
</tr>
<tr>
<td>KtDR</td>
<td>(K_{\text{DR}})</td>
<td>(W)</td>
<td>[65]</td>
</tr>
<tr>
<td>Qw</td>
<td>(Q_w)</td>
<td>Top</td>
<td>[61]</td>
</tr>
</tbody>
</table>
jet samples, the simulated signal samples described in Sect. 3 are combined and weighted (separately for W bosons and top quarks) such that the truth $p_T$ distribution of the ensemble of signal jets matches that of the light-jet background.

5.1 Cut-based optimisation

The first approach to tagging is based on selection cuts on jet shape observables. This approach was studied in preparation for Run 2 [73,74] to provide a set of guiding techniques that were used extensively in searches. The primary goal of these taggers is to provide a simple set of selections on jet moments that yield a constant signal efficiency as a function of the transverse momentum of the jet across a broad $p_T$ range, thus being widely applicable. In the case of W-boson tagging, one of these observables is taken to be $m^{\text{comb}}$ and the discrimination power is augmented by a selection on another jet moment defined in Table 1, while in the case of top tagging, a more inclusive strategy is explored where all pairwise combinations of jet moments are investigated. This optimisation is performed as a function of the $p_T$ of the associated anti-$k_T$, $R = 1.0$ truth jet for both W-boson and top-quark tagging. The tagging strategy resulting from this optimisation provides a benchmark in terms of tagging performance to which other tagging strategies can be compared.

This simple tagger is optimised using a sample of signal W-boson or top-quark jets as well as background light jets extracted from the samples described in Sect. 3. In each event the two reco jets matched to the two highest-$p_T$ truth jets within $|\eta| < 2$ are studied. In the case of signal, W-boson (top-quark) jets are retained if they are truth labelled as such according to the procedure in Sect. 4.2 and have a transverse momentum greater than 200 GeV (350 GeV). In the case of background, no labelling procedure is applied and the two highest-$p_T$ jets from the dijet sample are retained.

For this study, the general optimisation procedure to determine the two-variable selection criteria is the same for both W-boson and top-quark jet tagging. For each pair of observables, the selection criteria which give the chosen signal efficiency and the largest background rejection are considered optimal and taken as the selection criteria in that region of jet $p_T$. In the case of $m^{\text{comb}}$, the selection region is two-sided for W-boson tagging, selecting a region near $m_W$, and one-sided in the case of top-quark tagging, selecting an inclusive region of high jet mass. In the case of the other jet moments, the selection criteria are always one-sided, the direction of which depends on the particular observable in question. This procedure is repeated for exclusive bins of jet $p_T$ and a sequence of selection criteria for each of the jet moment observables is derived. Finally, this sequence of selection criteria is parameterised by a smooth function dependent on the jet $p_T$. All single-sided cuts are parameterised as a function of $p_T$ with a polynomial function to describe features which occur due to correlation of the combined-tagger variable. In the case of the W-boson tagging, the $m^{\text{comb}}$ selection is fit using a four-parameter ($p_1$) function of the form $\sqrt{(p_0 / p_T + p_1)}^2 + (p_2 \cdot p_T + p_3)^2$ chosen to encapsulate the dominant effects on the jet mass resolution.

Throughout this work, the targeted signal efficiencies are taken to be constant with respect to jet $p_T$ with values of 50% for W-boson tagging and 80% for top-quark tagging. These signal efficiency working points are largely based on those commonly used in searches for physics beyond the Standard Model. In the case of top-quark tagging, a working point with higher efficiency is commonly used because the dominant backgrounds involve processes including real top quarks [2,75] while in the case of searches for signals involving W-boson jets, the backgrounds are largely dominated by processes involving light-quark jets [76,77] thereby requiring a selection that more effectively rejects background at the expense of signal efficiency.

In Fig. 2, the resulting background rejections as a function of the jet $p_T^{\text{true}}$ are shown for a selection of the most powerful two-variable combinations. Based on this study, in the case of W tagging, the combination of $m^{\text{comb}}$ and $D_2$ is most powerful in the kinematic range of interest and is taken as the baseline pairing for W tagging. However, at higher jet $p_T^{\text{true}}$, where the power of $D_2$ decreases, $\sqrt{D_{12}}$ retains constant discrimination power. In the case of top-quark jet tagging, the behaviour of the most powerful taggers provide a large background rejection at low $p_T^{\text{true}}$, plateauning at a lower value for high jet $p_T^{\text{true}}$ mostly due to the migration of the light-jet mass distribution to higher values and a looser $\tau_{32}$ cut to maintain the constant signal efficiency. The two-variable combinations that do not involve mass perform marginally better than those with mass across the entire kinematic range studied. As a consequence, the specific cut-based top-quark jet tagger used in an analysis may depend on the context of the analysis and not on the performance alone. Therefore, the baseline two-variable cut-based top-quark jet tagger is selected to be the one composed of one-sided selections on $m^{\text{comb}}$ and $\tau_{32}$, as it has been commonly used in ATLAS.

5.2 Jet-moment-based multivariate taggers

Some of the moments presented in Sect. 4.3.1 contain complementary information and it has been shown that combining these observables by creating a multivariate W-boson or top-quark classifier provides higher discrimination, albeit to differing degrees [10,78,79]. In this work boosted decision tree (BDT) and deep neural network (DNN) algorithms are investigated following a procedure similar to the one in Ref. [79]. The goal is to discriminate W-boson and top-quark jets from light jets and to provide a single jet-tagging discriminant that is widely applicable in place of the single jet moment, described in Sect. 5.1, to augment the discrimi-
nation of $m_{\text{comb}}$ alone across a broad $p_T$ range, providing another widely applicable and more powerful tagger.

The two algorithmic classes used here, BDTs and DNNs, are explored in parallel to determine if one of the architectures is better suited to exploit differences between the input observables and their correlations among high-level variables in signal and background. The DNN used here is a fully-connected feed-forward network. Given that both algorithms have access to the same set of input features, of which there are approximately ten, it is expected that the discrimination power will be approximately the same. The internal settings, so called hyper-parameters, used for the BDTs and DNNs are summarized in Appendix A. For the design of all multivariate discriminants, exclusive subsamples of signal and background jets are derived from the more inclusive sample selected as in Sect. 5.1 to be used separately for the training and testing of the discriminant. To ensure that all jet substructure features are well-defined for the training, two additional selection criteria are applied to the jet mass ($m_{\text{comb}} > 40 \text{ GeV}$) and number of constituents ($N_{\text{const}} \geq 3$). The jets which fail to meet these criteria are not used in the training. However, in the evaluation of the performance of the tagger, such jets are classified as background jets only if they fail the $m_{\text{comb}}$ requirement, taking this auxiliary selection into account in the calculation of the signal efficiency and background rejection. The chosen input observables used for either W-boson or top-quark tagging are the full set of observables summarised in Table 1, noting that both the jet mass ($m_{\text{comb}}$) and transverse momentum are directly used as inputs. Therefore, when defining a final working point for this tagger, unlike in the case of the cut-based taggers in Sect. 5.1, no additional direct selection beyond the $m_{\text{comb}} > 40 \text{ GeV}$ requirement is imposed on the mass. Finally, in the design of the classifiers, all studies are performed in a wide $p_T$ range, providing another widely applicable and more powerful tagger.

The two algorithmic classes used here, BDTs and DNNs, are explored in parallel to determine if one of the architectures is better suited to exploit differences between the input observables and their correlations among high-level variables in signal and background. The DNN used here is a fully-connected feed-forward network. Given that both algorithms have access to the same set of input features, of which there are approximately ten, it is expected that the discrimination power will be approximately the same. The internal settings, so called hyper-parameters, used for the BDTs and DNNs are summarized in Appendix A. For the design of all multivariate discriminants, exclusive subsamples of signal and background jets are derived from the more inclusive sample selected as in Sect. 5.1 to be used separately for the training and testing of the discriminant. To ensure that all jet substructure features are well-defined for the training, two additional selection criteria are applied to the jet mass ($m_{\text{comb}} > 40 \text{ GeV}$) and number of constituents ($N_{\text{const}} \geq 3$). The jets which fail to meet these criteria are not used in the training. However, in the evaluation of the performance of the tagger, such jets are classified as background jets only if they fail the $m_{\text{comb}}$ requirement, taking this auxiliary selection into account in the calculation of the signal efficiency and background rejection. The chosen input observables used for either W-boson or top-quark tagging are the full set of observables summarised in Table 1, noting that both the jet mass ($m_{\text{comb}}$) and transverse momentum are directly used as inputs. Therefore, when defining a final working point for this tagger, unlike in the case of the cut-based taggers in Sect. 5.1, no additional direct selection beyond the $m_{\text{comb}} > 40 \text{ GeV}$ requirement is imposed on the mass. Finally, in the design of the classifiers, all studies are performed in a wide $p_T$ range, providing another widely applicable and more powerful tagger.

### Fig. 2
The W-boson (a) and top-quark tagging (b) background rejection as a function of jet $p_T^{\text{true}}$ for the best performing two-variable combinations at fixed signal efficiency.

### Background Rejection

Background rejection ($1 / \epsilon_{\text{bkg}}$) as a function of jet $p_T^{\text{true}}$ for the best performing two-variable combinations at fixed signal efficiency.

### Table 1

<table>
<thead>
<tr>
<th>Observable</th>
<th>Signal Efficiency</th>
<th>Background Rejection</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_{\text{comb}}$</td>
<td>50%</td>
<td>80%</td>
</tr>
<tr>
<td>$N_{\text{const}}$</td>
<td>&gt; 3</td>
<td>&gt; 3</td>
</tr>
<tr>
<td>$N_{\text{comb}}$</td>
<td>&gt; 40 GeV</td>
<td>&gt; 40 GeV</td>
</tr>
</tbody>
</table>

Relative signal efficiency and relative background rejection take into account only the jets that satisfy the training criteria, where relative signal efficiency is defined as

$$\epsilon_{\text{sig}}^{\text{rel}} = \frac{N_{\text{tagged}}}{N_{\text{tagged}} + N_{\text{untagged}}}$$

and in a similar manner, relative background rejection is defined as $1 / \epsilon_{\text{bkg}}^{\text{rel}}$. The smallest set of variables which reaches the highest relative background rejection within statistical uncertainties is selected. The minimum number of selected

---

the jets in a testing set. The relative background rejection set of jets and the relative performance is evaluated using group, the DNN classifier is constructed using the training variables tested for the DNN is shown in Table 2. For each evidence on other substructure variables. A summary of all the features of the substructure they describe and their dependence on the momentum scale of the jet substructure objects, what is chosen by selecting variables according to their dependence in background rejection of roughly a factor of two over the entire kinematic range studied. This is presumably due to the merging of calorimeter energy depositions and subsequent loss of granularity in discerning substructure information. However, in the case of top-quark tagging, the improvements in performance are more sizeable, showing increases in background rejection of roughly a factor of two over the entire kinematic range studied. This is presumably due to the greater complexity of the top-quark decay in contrast to that of the isolated \( W \) boson, indicating that among the observables studied here, excluding the multivariate classifiers, no single observable adequately captures the full set of features that provide ability to discriminate signal from background. There are richer correlations between the observables that can be further exploited by the multivariate classification algo-

![ATLAS Simulation](attachment:image.png)

**Fig. 3** The relative background rejection of the jet-shape-based BDT discriminant for different sets of variables, with more variables added successively at the 50% (W-boson tagging) and 80% (top-quark tagging) relative signal efficiency working point for W-boson (a) and top-quark (b) tagging. Only jets which satisfy the training criteria are considered when calculating the relative signal efficiency and relative background rejection. The performance is evaluated with constant \( \rho_T^{\text{true}} \) spectra. Uncertainties are not presented. The horizontal dashed lines indicate the level of performance saturation, while the vertical dashed lines and solid arrow represent the set of jet moments used in the final construction of the discriminant.

variables is 11 for W-boson tagging and 10 for top-quark tagging. The relative background rejection achieved at each stage for both classifiers is shown in Fig. 3.

In a similar manner, the observables used in the DNN classifier are chosen by comparing the performance when using different sets of input variables to find the set of observables which gives the largest relative background rejection at a fixed relative signal efficiency. In this case, variables are not added in succession due to the time requirements to train the large number of networks. Instead, groups of observables are chosen by selecting variables according to their dependence on the momentum scale of the jet substructure objects, what features of the substructure they describe and their dependence on other substructure variables. A summary of all the variables tested for the DNN is shown in Table 2. For each group, the DNN classifier is constructed using the training set of jets and the relative performance is evaluated using the jets in a testing set. The relative background rejection achieved inclusively in jet \( \rho_T^{\text{true}} \) is shown in Fig. 4. The performance of the DNN tagger depends on both the number of variables and the information content in the group. The chosen groups of inputs for W-boson tagging and top-quark tagging are listed in Table 2. Within statistical uncertainties, the number of variables necessary for maximum rejection at a fixed relative signal efficiency of 50% (W-boson tagging) and 80% (top-quark tagging) is found to be 12 variables for W-boson tagging (Group 8 in Table 2) and 13 variables for top-quark tagging (Group 9 in Table 2).

Similarly to the cut-based two-variable optimised taggers, for the chosen BDT and DNN taggers the working points are defined as a function of the reconstructed jet \( p_T \) so that they yield constant signal efficiencies versus \( p_T \). In both cases, the target signal efficiency working point is obtained by the fixed jet mass requirement of \( m_{\text{comb}}> 40 \text{ GeV} \), relevant \( \Lambda^\text{const} \) criteria and a single-sided selection on the relevant discriminant. The performance of the resulting BDT and DNN discriminants is characterised by the background rejection, evaluated as a function of jet \( \rho_T^{\text{true}} \), for a fixed signal efficiency of 50% (W-boson tagging) and 80% (top-quark tagging), where the relative variation of the signal efficiency for the fixed-efficiency taggers is less than 5%. It can be seen in Fig. 5 that in the case of W-boson tagging, the performance improvements beyond the cut-based taggers are highest at low jet \( p_T \) and decrease at higher \( p_T^{\text{true}} \), presumably due to the merging of calorimeter energy depositions and subsequent loss of granularity in discerning substructure information.
Table 2 A summary of the set of observables that were tested for W-boson and top-quark tagging for the various DNN input observable groups as well as the final set of DNN and BDT input observables as chosen using Figs. 3 and 4

<table>
<thead>
<tr>
<th>W boson tagging</th>
<th>Chosen inputs</th>
<th>Top quark tagging</th>
<th>Chosen inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observable</strong></td>
<td><strong>Test groups</strong></td>
<td><strong>Chosen inputs</strong></td>
<td><strong>Test groups</strong></td>
</tr>
<tr>
<td>$m_{comb}$</td>
<td>1   2   3   4   5   6   7   8   9</td>
<td>BDT</td>
<td>DNN</td>
</tr>
<tr>
<td>$p_T$</td>
<td>1   2   3   4   5</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$e_3$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$C_2$</td>
<td>0   0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$D_2$</td>
<td>0   0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$\tau_1$</td>
<td>0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$\tau_2$</td>
<td>0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$\tau_3$</td>
<td>0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$\tau_{21}$</td>
<td>0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$\tau_{32}$</td>
<td>0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$R_F^{W}$</td>
<td>0   0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$q$</td>
<td>0   0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$a_3$</td>
<td>0   0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$A$</td>
<td>0   0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$\varepsilon_{\text{cut}}$</td>
<td>0   0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$\sqrt{d_{22}}$</td>
<td>0   0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$\sqrt{d_{23}}$</td>
<td>0   0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$K_t DR$</td>
<td>0   0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
<tr>
<td>$Q_w$</td>
<td>0   0   0   0   0</td>
<td>6   7   8   9</td>
<td>BDT</td>
</tr>
</tbody>
</table>

Fig. 4 Distributions showing the training with different set of variables and relative improvement in performance for the DNN W-boson (a) and top-quark (b) taggers at the 50% and 80% relative signal efficiency working point, respectively. The grouping of observables was decided prior to training and discriminator performance evaluation. Only jets which satisfy the training criteria are considered when calculating the relative signal efficiency and relative background rejection. The performance is evaluated with constant $p_T^{\text{true}}$ spectra. Uncertainties are not presented.
5.3 Topocluster-based deep neural network tagger

Recently, a number of jet phenomenology studies have found that using lower-level information more directly pertaining to the jet energy flow can lead to further improvements in the ability to distinguish signal $W$-boson and top-quark jets from light jets [66–69, 80–84]. Furthermore, it was seen in Fig. 5 that the performance gains for the high-level variables BDT and DNN combination are significantly larger for top-quark tagging than for $W$-boson tagging. Consequently, a top-quark jet tagger based directly on the jet constituents, focusing on the high-$p_T$ top quarks with $p_T > 450$ GeV, is designed.

The jet tagger based on low-level jet input information studied in this work closely follows that described in Ref. [68] and the reader is referred there for a more in-depth review of the optimisation of the techniques used; only a brief summary is provided in the following. The first aspect of note which sets this tagger apart from those studied in Refs. [66, 67, 69] is that there is no use of pixelation in this tagger, similar to the taggers studied in Ref. [80]. Compared to the taggers studied in Ref. [80], the architecture of this tagger does not employ sequenced, variable-length inputs and the input features used in this tagger are the four-vectors of fixed-number of topoclusters in the individual large-$R$ anti-$k_t$ trimmed jet in the $(p_T, \eta, \phi)$ representation, noting that topoclusters are taken as massless by convention. As a preprocessing step, the $p_T$ of each constituent four-vector is normalised by 1/1700 to bring the scale of the input network features within the same magnitude between approximately 0 and 1. The $(\eta, \phi)$ location of the set of constituents is then transformed by a process that involves a translation, a rotation, and a flip based on the assumed three-subjet topology of a top-quark decay. Of the full set of constituents, only the 10 highest-$p_T$ constituents are used as input to the neural network. This was found to provide optimal background rejection for this network architecture as compared to using more or fewer clusters and can be qualitatively understood by examining the fraction of the jet $p_T$ carried by each of the clusters, shown in Fig. 6 where the distribution of the $p_T$-fraction for a subset of the 10 highest-$p_T$ clusters is shown along with the mean value of each of the 20 highest-$p_T$ cluster distributions. It is seen that the first 10 clusters, on average carry more than 99% of the $p_T$ of the jet. Therefore, including further clusters saturates the information for the network to disentangle when discriminating signal from background. If a jet has fewer than 10 constituents, the remaining inputs to the neural network are taken to be null vectors. The three components of each four-vector are used as input to a fully connected neural network with four hidden layers composed of 300, 102,
12 and 6 nodes, respectively. This network architecture was determined through manual hyper-parameter tuning, exploring configurations with between 4–6 layers and 40–1000 nodes per layer, and where the used architecture and hyper-parameters are exactly the same as the one used in [68]. The network is trained on jets where only the initial top parton is required to be matched to the reconstructed jet obtained from the Z (signal) and light jets (background) in the high-p_T region from 450 to 2400 GeV in p_T. To remove bias in the training due to the difference in kinematics between the signal and background samples, a subset of the background ensemble of jets is selected in a random fashion such that the jet p_T distribution is the same in both signal and background, as opposed to the BDT and DNN taggers described in Sect. 4.3.3, which use event-by-event reweighting. However, in Run 2, shower deconstruction was recommissioned in the context of the search for a heavy W’ boson decaying to a top quark and a bottom quark where the mass-splitting between the W’ and the top quark was large enough to produce top quarks with momenta of roughly 1 TeV and above [85]. The approach taken in Run 1 to reconstruct the subjet inputs to the shower deconstruction algorithm was found to have a low signal efficiency, largely due to the subjet multiplicity falling below three and therefore producing a set of subjets that are unable to fulfill the initial consistency checks between the subjet pairings and triplets with W-boson and top-quark masses, respectively. This drop in efficiency was recovered by altering the manner in which subjets are constructed to instead use the exclusive-k_t jet clustering algorithm, run on the constituents of the trimmed large-R jet. Since splitting scales are less dependent on the large-R jet p_T than the geometric distance between the jet and its constituents, a stopping criterion is imposed to halt clustering if k_t splitting scales larger than 15 GeV are found. At that stage, the resulting set of subjets are used as subjet inputs to shower deconstruction. Because the computation time of shower deconstruction scales exponentially with the number of input subjets, the total number of subjets is limited to at most the six highest-p_T subjets, compared to a limit of nine in Run 1, with no loss in performance. Finally, the parameters controlling the top-quark topology check using subjet pairings and triplets, m_W and m_top respectively, were fixed to 20 GeV and 40 GeV, the same as in Run 1.
5.5 Summary of tagger performance studies in simulation

A direct comparison of the performance of all of the tagging techniques, described in Sect. 4 and individually optimised in Sect. 5, is important in providing guidance as to which technique can be most beneficial when applied in an analysis. The primary metric used to assess the performance of the taggers is the background rejection as a function of the signal efficiency, characterised in the form of a receiver operating characteristic (ROC) curve, shown in Figs. 7 and 8 for $W$-boson and top-quark tagging, respectively, for both a low- and high-$p_T$ kinematic region. For comparison, two relatively simple cut-based taggers composed of selections on $m_{\text{comb}}$ and a single substructure observable are shown. In the case of $W$-boson tagging, a fixed mass window requirement of $60 < m_{\text{comb}} < 100$ GeV is applied and a cut on the $D_2$ observable is used for the ROC curve. In the case of top-quark tagging, the mass selection is one-sided, requiring $m_{\text{comb}} > 60$ GeV, and a requirement on $\tau_{32}$ is varied to obtain the ROC curve. These simple taggers, along with the specific working points tuned

Fig. 7 The performance comparison of the $W$-boson taggers in a low-$p_T^{\text{true}}$ (a) and high-$p_T^{\text{true}}$ (b) bin. The performance is evaluated with the $p_T^{\text{true}}$ distribution of the signal jets weighted to match that of the dijet background samples.

Fig. 8 The performance comparison of the top-quark taggers in a low-$p_T^{\text{true}}$ (a) and high-$p_T^{\text{true}}$ (b) bin. The performance is evaluated with the $p_T^{\text{true}}$ distribution of the signal jets weighted to match that of the dijet background samples.
to give constant signal efficiency and maximal background rejection, are provided as a point of reference for subsequent optimisations that were performed for studies of the more advanced techniques.

When examining Figs. 7 and 8, it can be seen that a careful tuning of the simple two-variable cut-based taggers can lead to sizeable gains due to taking into consideration the correlation between \(m^{\text{comb}}\) and the auxiliary jet moment observable. The gains are significantly larger in the case of the BDT- and DNN-based high-level observable discriminants, and lead to larger improvements for top-quark tagging than for \(W\)-boson tagging. However, the BDT and DNN algorithms perform similarly to each other for all signal efficiencies, indicating that they are leveraging the correlations of the input jet moment observables equally well. Therefore, when studying the performance of these tagging techniques in data in Sect. 6, only the DNN-based taggers are included. The performance of the BDT-based taggers was studied and found to be similar. Finally, in the case of top-quark tagging, where more dedicated tagging techniques are studied, the conclusion is similar. Dedicated approaches, including shower deconstruction and HEPTopTagger, are more performant than a simple cut-based approach on \(m^{\text{comb}}\) and \(R_{32}\), but the combination of many jet-moment observables in a BDT or DNN yields the best overall performance out of the techniques tested in this study. Of particular note, however, is the comparison of the BDT and the fully-connected feed-forward DNN taggers using high-level observables and those using lower-level inputs, namely the jet constituents, here taken to be topoclusters. The performance of these two approaches is similar, with the TopoDNN tagger having slightly higher background rejection at high jet \(p_T\), resulting in conclusions qualitatively similar to those found in Ref. [69], particularly at high jet \(p_T\) where the details of the signal sample used for training are less relevant.

6 Performance in data

The taggers studied in the previous sections are validated using signal and background-enriched data samples collected during 2015 and 2016 at a centre-of-mass energy of \(\sqrt{s} = 13\) TeV and corresponding to an integrated luminosity of 36.1 fb\(^{-1}\). In the case of \(W\)-boson and top-quark jets, the lepton-plus-jets \(\tilde{t}\) signature is used, which provides a sample of signal jets in a \(p_T\) range of approximately 200–1000 GeV. In the case of background light jets, two topologies are studied: a \(\gamma +\) jet sample enriched in light-quark jets and spanning a \(p_T\) range of approximately 200–2000 GeV and a multijet sample which probes a mixture of light-quark and gluon jets in a \(p_T\) range of approximately 500–3500 GeV. The primary aim of these studies is to validate the modelling of the Monte Carlo simulation in data for the techniques studied in Sect. 4. This is achieved by directly studying the full spectrum of a subset of important observables used in the tagging as well as directly measuring both the signal efficiency and background efficiency of the various techniques in the phase space accessible in this data sample. In the case of the measured signal efficiency and background rejection, the performance is evaluated differentially as a function of the jet transverse momentum as well as the average number of interactions per bunch crossing (\(\mu\)).

6.1 Signal efficiency in boosted \(\tilde{t}\tilde{t}\) events

To study the modelling of signal \(W\)-boson and top-quark large-\(R\) jet tagging, a sample of data enriched in \(\tilde{t}\tilde{t}\) events where one top quark decays hadronically and the other semileptonically in both the electron and the muon decay channel is selected in a similar manner to Refs. [8,9]. The inclusive sample of events is decomposed into two exclusive subsamples, enriched in \(W\)-boson jets and top-quark jets, based on the proximity of a \(b\)-jet to the large-\(R\) jet. The inclusive distributions of the key observables used in each tagging method are examined and the signal efficiency is measured for a set of fixed signal efficiency working points, for which systematic uncertainties can be derived and associated with a particular tagging method.

6.1.1 Analysis and selection

To select the inclusive set of lepton-plus-jets \(\tilde{t}\tilde{t}\) events, both the data and Monte Carlo simulated events are required to pass either an inclusive electron trigger or an inclusive muon trigger, where the thresholds were varied between the 2015 and 2016 datasets due to increases in instantaneous luminosity. In the electron channel, events from the 2015 data-taking period are required to pass at least one of three triggers: one isolated electron with \(p_T > 24\) GeV, one electron with \(p_T > 60\) GeV without any isolation requirement, or one electron with \(p_T > 120\) GeV without any isolation requirement and relaxed identification criteria. In the 2016 data-taking period, the thresholds of these electron triggers required \(p_T > 26\) GeV, \(p_T > 60\) GeV and \(p_T > 140\) GeV, respectively. In the muon channel, events from the 2015 data-taking period are required to pass at least one of two muon triggers: one isolated muon with \(p_T > 20\) GeV or one muon with \(p_T > 50\) GeV and no isolation requirement. In the 2016 data-taking period, the thresholds of these triggers required \(p_T > 26\) GeV and \(p_T > 50\) GeV, respectively.

Events are then required to contain exactly one electron or muon candidate with \(p_T > 30\) GeV that is matched to the trigger-level counterpart associated with the appropriate trigger. Electron candidates are reconstructed as ID tracks that are matched to a cluster of energy in the electromagnetic calorimeter. Electron candidates are required to be within...
| $\eta$ | < 2.47, excluding the calorimeter transition region from 1.37 < |$\eta$| < 1.52, and satisfy the “tight” likelihood-based identification criterion based on shower shape and track selection requirements [86, 87]. Muons are reconstructed as tracks found in the ID that are matched to tracks reconstructed in the muon spectrometer. They are required to be within |$\eta$| < 2.5 and are required to satisfy the “medium” muon identification quality criteria defined in Ref. [88]. For both electrons and muons, the reconstructed lepton candidate is required to be isolated from additional activity in the event by imposing isolation criterion defined by a sum of $p_T$ of tracks in an isolation cone with variable radius depending on the lepton $p_T$ [88, 89].

In addition to identified leptons, small-radius jets are used to reconstruct the missing transverse momentum and identify the signal topology. These jets are reconstructed from topoclusters calibrated to the electromagnetic scale using the anti-$k_T$ algorithm with a radius parameter of $R = 0.4$. The energy of these jets is corrected for the effects of pile-up by using a technique based on jet area [51] and the jet energy is further corrected using a jet energy scale calibration based on both Monte Carlo simulation and data [40]. To ensure that the reconstructed jets are well-measured, they are required to have $p_T > 20$ GeV, |$\eta$| < 2.5 and to satisfy “loose” quality criteria to prevent mismeasurements due to calorimeter noise spikes and non-collision backgrounds [90]. For jets with $p_T < 60$ GeV and |$\eta$| < 2.4, a requirement that the jets described in Sect. 4.2. The statistical uncertainty of the background prediction (Stat. uncert.) results from limited Monte Carlo statistics as well as the limited size of the data sample used in the data-driven estimation of the multijet background.

Fig. 9 A comparison of the observed data and predicted MC distributions of the mass of the leading $p_T$ anti-$k_T$ trimmed jet in the event for the $W$ boson (a) and top quark (b) selections in a sample enriched in lepton+jets $t\bar{t}$ events. Simulated distributions are normalised to data. The $t\bar{t}$ sample is divided into a set of subsamples (e.g. $t\bar{t}$ (top)) based on criteria.

Fig. 10 A comparison of the observed data and predicted MC distributions of the mass of the leading $p_T$ C/A $R = 1.5$ trimmed jet in events passing the top-quark selection in a sample enriched in lepton+jets $t\bar{t}$ events. Simulated distributions are normalised to data. The $t\bar{t}$ sample is divided into a set of subsamples (e.g. $t\bar{t}$ (top)) based on criteria described in Sect. 4.2. The statistical uncertainty of the background prediction (Stat. uncert.) results from limited Monte Carlo statistics as well as the limited size of the data sample used in the data-driven estimation of the multijet background.
Fig. 11 A comparison of the observed data and predicted MC distributions of the anti-kt, $R = 1.0$ trimmed jet $D_2$ (a) and $\tau_{32}$ (b) for the W-boson and top-quark selections, respectively, in a sample enriched in lepton+jets $t\bar{t}$ events. Simulated distributions are normalised to data. The $t\bar{t}$ sample is divided into a set of subsamples (e.g. $t\bar{t}$ (top)) based on criteria described in Sect. 4.2. The statistical uncertainty of the background prediction (Stat. uncert.) results from limited Monte Carlo statistics as well as the limited size of the data sample used in the data-driven estimation of the multijet background.

Fig. 12 A comparison of the observed data and predicted MC distributions of the anti-kt, $R = 1.0$ trimmed jet DNN W discriminant for W boson (a) and top quark (b) tagging for the respective event selections in a sample enriched in lepton+jets $t\bar{t}$ events. Simulated distributions are normalised to data. The $t\bar{t}$ sample is divided into a set of subsamples (e.g. $t\bar{t}$ (top)) based on criteria described in Sect. 4.2. The statistical uncertainty of the background prediction (Stat. uncert.) results from limited Monte Carlo statistics as well as the limited size of the data sample used in the data-driven estimation of the multijet background.
arise from the primary vertex, using the ID tracks associated with the jet, is imposed to suppress pile-up jets [91].

For the identification of $b$-quark candidate jets, jets reconstructed from ID tracks with the anti-$k_t$ algorithm with radius parameter $R = 0.2$ are used. These jets are $b$-tagged using a multivariate discriminant based on impact parameter and secondary vertex information [92]. The 70% signal efficiency point selection is used. Event-by-event scale factors, evaluated in $t\bar{t}$ events [93], are applied to account for mismodelling of the selection efficiency.

The missing transverse momentum is reconstructed as the negative vectorial sum of the momenta of all reconstructed physics objects in the plane transverse to the beamline [94]. In this case, the sum consists of the single identified lepton and the full set of reconstructed and fully calibrated small-$R$ calorimeter jets as well as ID tracks not associated with the lepton or jets. These ID tracks are included to account for the soft hadronic energy flow in the event. In the following the magnitude of the missing transverse momentum vector is denoted by $E_T^{\text{miss}}$.

First, events containing a leptonically decaying $W$ boson are preselected by requiring one electron or muon candidate with $p_T > 30$ GeV and rejecting events that contain additional electrons or muons with $p_T > 25$ GeV. The missing transverse momentum is required to be greater than 20 GeV and the scalar sum of $E_T^{\text{miss}}$ and the transverse mass of the leptonically decaying $W$ boson candidate must satisfy $E_T^{\text{miss}} + m_W^T > 60$ GeV. To ensure the topology is consistent with a $t\bar{t}$ event, at least one small-$R$ jet is required to have $p_T > 25$ GeV and to be close to the lepton ($\Delta R(\text{lepton, jet}) < 1.5$). To study $W$-boson and top-quark tagging, the highest-$p_T$ large-$R$ jet is studied, which is either a trimmed anti-$k_t$, $R = 1.0$ jet or a C/A $R = 1.5$ jet in the case of HEPTopTagger, with $p_T > 200$ GeV and $|\eta| < 2.0$. The C/A jets are also trimmed using the same trimming parameters as for the anti-$k_t$ jets, such that their kinematics are robust against pile-up. Since HEPTopTagger is designed to tag ungroomed jets, the constituents of the C/A jet before trimming are used as inputs to the tagging algorithm. The signal top-quark jet candidate is required to be well-separated from the semileptonic top-quark decay by requiring $\Delta R > 1.5$ between the large-$R$ jet and the small-
$R$ jet close to lepton. Additionally, the angular separation in the transverse plane between the lepton and the large-$R$ jet is required to be $\Delta\phi > 2.3$.

Finally, the sample preselected as above is divided into two subsamples, intended to be representative of a fully contained top-quark decay or an isolated and fully contained $W$-boson decay based on the proximity of a $b$-tagged track jet to the highest-$p_T$ large-$R$ jet. The track jets are clustered from at least two tracks using the anti-$k_t$ algorithm with a radius parameter of $R = 0.2$. All tracks must fulfil $|\eta| < 2.5$ and $p_T > 10$ GeV. The sample enriched in top quarks (“top-quark selection”) is defined by requiring a $b$-tagged track jet to have an angular separation of $\Delta R (b$-jet, large-$R$ jet) < 1.0 ($\Delta R (b$-jet, large-$R$ jet) < 1.5) from the large-$R$ anti-$k_t$ trimmed jet (C/A jet). In order to enhance the fraction of fully contained top quarks, an additional requirement of $p_T > 350$ GeV ($p_T > 200$ GeV) is also applied. The sample enriched in $W$-boson jets (“$W$-boson selection”) is defined by requiring a $b$-tagged track jet to have angular separation $\Delta R (b$-jet, large-$R$ jet) > 1.0 from the large-$R$ anti-$k_t$ trimmed jet. Because the geometrical separation of the daughter $b$-quark and the top parton decreases with increasing $p_T$, this requirement limits the efficiency of the $W$-boson selection at high jet $p_T$, which limits the kinematic reach to approximately 600 GeV. These requirements result in relatively pure samples of $W$-boson and top-quark jets as shown in Fig. 9 for the anti-$k_t$ $R = 1.0$ trimmed jet mass, including the full set of systematic uncertainties summarised in Sect. 6.3, while Fig. 10 shows the C/A $R = 1.5$ trimmed jet mass. The disagreement between the peak positions in Monte Carlo simulation and data observed near $m_W$ and $m_{top}$ is attributed to a mismodelling of the jet mass scale as studied in Ref. [95]. In this paper, the $t\bar{t}$ and single-top Monte Carlo samples are divided into three subsamples based on the jet labelling criteria outlined in Sect. 4.2 to highlight the fraction of events in each sample of interest (“$t\bar{t}$ (top)”) and “$t\bar{t}$ ($W$)”), with all other events in these samples being grouped together in a single subsample (“$t\bar{t}$ (other)”). The backgrounds are derived from the Monte Carlo simulations

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Fig15}
\caption{A comparison of the observed data and predicted MC distributions of the HEPTopTagger mass for the top-quark event selection in a sample enriched in lepton-plus-jets $t\bar{t}$ events. Simulated distributions are normalised to data. The $t\bar{t}$ sample is divided into a set of subsamples (e.g. $t\bar{t}$ (top)) based on criteria described in Sect. 4.2. The statistical uncertainty of the background prediction (Stat. uncert.) results from limited Monte Carlo statistics as well as the limited size of the data sample used in the data-driven estimation of the multijet background.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Fig16}
\caption{The anti-$k_t$ trimmed jet mass distribution in the pass (a) and fail (b) categories for the $m^{comb} + \tau_2$ top-quark tagger working point after the chi-square fit has been performed. The templates shown here are those used in the chi-square fit for the extraction of the three normalisation factors. The first, $t\bar{t}$ signal, includes only the $t\bar{t}$ (top) contribution, while $t\bar{t}$ background includes contributions from $t\bar{t}$ ($W$) and $t\bar{t}$ (other) and the non-$t\bar{t}$ background component includes all other backgrounds. Only statistical uncertainties are shown.}
\end{figure}
Fig. 17 The signal efficiency on contained W-boson jets for the two-variable $m_{\text{comb}} + D_2$ W-boson tagger as a function of the large-$R$ jet $p_T$ (a) and the average number of interactions per bunch crossing $\mu$ (b) in data and simulation. Statistical uncertainties of the signal efficiency measurement in data and simulation are shown as error bars in the top panel. In the bottom panel, the ratio of the measured signal efficiency in data to that estimated in Monte Carlo simulation is shown with statistical uncertainties as error bars on the data points and the sum in quadrature of statistical and systematic uncertainties as a shaded band. When considering experimental uncertainties arising from the large-$R$ jet, only those coming from the jet energy scale and resolution are considered.

Fig. 18 The signal efficiency on contained W-boson jets for the jet shape-based DNN W-boson tagger as a function of the large-$R$ jet $p_T$ (a) and the average number of interactions per bunch crossing $\mu$ (b) in data and simulation. Statistical uncertainties of the signal efficiency measurement in data and simulation are shown as error bars in the top panel. In the bottom panel, the ratio of the measured signal efficiency in data to that estimated in Monte Carlo is shown with statistical uncertainties as error bars on the data points and the sum in quadrature of statistical and systematic uncertainties as a shaded band. When considering experimental uncertainties arising from the large-$R$ jet, only those coming from the jet energy scale and resolution are considered.

Described in Sect. 3, with the exception of the multijet background, which is estimated using a data-driven method based on looser lepton selection criteria with a dedicated evaluation of the probability of prompt lepton reconstruction and the probability of fake/non-prompt lepton reconstruction, as was performed in Ref. [2]. The event yield in the simulation is normalised to that in the data at this stage of the selection throughout Sect. 6.1.1.

The primary tagging observables used by the other tagging techniques described in Sect. 4 are examined in Figs. 11, 12, 13, 14 and 15. For these spectra, the full set of systematic uncertainties described in Sect. 6.3 are included for the $D_2$ and $\tau_{32}$ observables, whereas for the other spectra, no dedicated experimental systematic uncertainty in the scale or resolution of the observable itself is included. Instead, the mismodelling of the simulation relative to data is taken into account as a derived uncertainty in the in situ measurement of the signal efficiency of the tagger itself, in a manner similar to that commonly used to evaluate mismodelling in the detector response in the context of the identification of heavy-flavour
Signal efficiency ($\epsilon_{\text{sig}}$) in data to that estimated in Monte Carlo is shown with statistical uncertainties as error bars on the data points and the sum in quadrature of statistical and systematic uncertainties as a shaded band. When considering experimental uncertainties arising from the large-$R$ jet, only those coming from the jet energy scale and resolution are considered.

jets [96]. However, for nearly all regions of phase space, the overall relative yield of data is well-described by the Monte Carlo prediction within the theoretical uncertainties, derived from the comparison of various $t\bar{t}$ Monte Carlo generators.

6.1.2 Signal efficiencies

Due to the relatively high purity of the samples of $W$-boson and top-quark jets that result from the selection described in Sect. 6.1.1, it is possible to measure the signal efficiency in data. This measurement, when compared with the Monte Carlo prediction, can be used to estimate the systematic uncertainty of a particular tagging method when applied in the context of an independent analysis. It can also be used to provide an in situ correction in the form of a jet-by-jet efficiency scale factor [93,96]. Because the aim is to provide an efficiency measurement for a particular tagging method, it is necessary to define selection criteria based on the particular tagging discriminants described in Sect. 4 for which the comparison of the Monte Carlo prediction to data was
Fig. 21 The signal efficiency on contained top-quark jets for the TopoDNN top-quark tagger as a function of the large-$R$ jet $p_T$ (a) and the average number of interactions per bunch crossing $\mu$ (b) in data and simulation. Statistical uncertainties of the signal efficiency measurement in data and simulation are shown as error bars in the top panel. In the bottom panel, the ratio of the measured signal efficiency in data to that estimated in Monte Carlo is shown with statistical uncertainties as error bars on the data points and the sum in quadrature of statistical and systematic uncertainties as a shaded band. When considering experimental uncertainties arising from the large-$R$ jet, only those coming from the jet energy scale and resolution are considered.

Fig. 22 The signal efficiency on contained top-quark jets for the Shower Deconstruction top-quark tagger as a function of the large-$R$ jet $p_T$ (a) and the average interactions per bunch crossing $\mu$ (b) in data and simulation. Statistical uncertainties of the signal efficiency measurement in data and simulation are shown as error bars in the top panel. In the bottom panel, the ratio of the measured signal efficiency in data to that estimated in Monte Carlo is shown with statistical uncertainties as error bars on the data points and the sum in quadrature of statistical and systematic uncertainties as a shaded band. When considering experimental uncertainties arising from the large-$R$ jet, only those coming from the jet energy scale and resolution are considered.

shown in a rather inclusive selection of signal-like events in Sect. 6.1.1. In particular, the seven tagger working points for which the signal efficiency is measured here are:

- $D_2 + m_{\text{comb}}$ ($W$ boson): A pair of selections on $m_{\text{comb}}$ and $D_2$, tuned as a function of $p_T$, that give the largest background rejection for a fixed 50% signal efficiency for fully contained $W$-boson jets;
- $m_{\text{comb}} + \tau_{32}$ (top quark): A pair of selections on $m_{\text{comb}}$ and $\tau_{32}$, tuned as a function of $p_T$, that give the largest background rejection for a fixed 80% signal efficiency for fully contained top-quark jets;
- DNN ($W$ boson): A single-sided selection of $m_{\text{comb}} > 40$ GeV and a selection on the DNN discriminant, tuned to give a fixed 50% signal efficiency as a function of $p_T$ for fully contained $W$-boson jets;
- DNN (top quark): A single-sided selection of $m_{\text{comb}} > 40$ GeV and a selection on the DNN discriminant, tuned to give a fixed 80% signal efficiency as a function of $p_T$ for fully contained top-quark jets;
The results are shown in Figs. 17 and 18 for the Run-1 analysis [9].

The signal efficiency is measured in data and obtained in simulation, which is based on the numbers of predicted signal and background events, as demonstrated in Fig. 16 in the case of the simple $m_{\text{comb}} + \tau_3$ top-quark tagger. For the measurement of the HEPTopTagger signal efficiency, the fit is performed using distributions of the mass of the leading C/A trimmed jet instead. Distributions of events that either pass or fail the tagger under study are fit simultaneously. The total normalisation of each grouped background component is allowed to float and is extracted in the fit, while the efficiency of the tagger on background events is fixed to the value in Monte Carlo simulation. Normalisations of signal distributions in the pass and fail categories ($N_{\text{tagged}}^{\text{fitted signal}}$ and $N_{\text{not tagged}}^{\text{fitted signal}}$) are extracted from the fit. Therefore, the tagger efficiency for signal events in data can be extracted as

$$\epsilon_{\text{data}} = \frac{N_{\text{tagged}}^{\text{fitted signal}}}{N_{\text{tagged}}^{\text{fitted signal}} + N_{\text{not tagged}}^{\text{fitted signal}}}.$$ 

This can be compared to the tagger efficiency in Monte Carlo simulation, which is based on the numbers of predicted signal events that pass, $N_{\text{signal}}^{\text{tagged}}$, and fail, $N_{\text{signal}}^{\text{not tagged}}$, the tagger under study:

$$\epsilon_{\text{MC}} = \frac{N_{\text{tagged}}^{\text{signal}}}{N_{\text{signal}}^{\text{tagged}} + N_{\text{signal}}^{\text{not tagged}}}.$$ 

The signal efficiency is measured in data and obtained in simulations as a function of the $p_T$ of the large-$R$ jet as well as the average number of interactions per bunch crossing ($\mu$). The results are shown in Figs. 17 and 18 for the $W$-boson
Fig. 24 A comparison of the observed data and predicted MC distributions of the mass of the leading $p_T$ anti-$k_t$ trimmed jet in events for the multijet (a) and $\gamma + \text{jet}$ (b) selections. The data-driven normalisation correction, described in Sect. 6.2.1, is shown in the legend beside the specific sample to which it applies. Systematic uncertainties are indicated as a band in the lower panel and include all experimental uncertainties related to the selection of events, as well as the reconstruction and calibration of the large-$R$ jet.

Fig. 25 A comparison of the observed data and predicted MC distributions of the mass of the leading $p_T$ C/A $R = 1.5$ trimmed jet in events for the multijet (a) and $\gamma + \text{jet}$ (b) selections. The data-driven normalisation correction, described in Sect. 6.2.1, is shown in the legend beside the specific sample to which it applies. Systematic uncertainties are indicated as a band in the lower panel and include all experimental uncertainties related to the selection of events, as well as the reconstruction and calibration of the large-$R$ jet.
Fig. 26 A comparison of the observed data and MC predictions in the multijet and $\gamma + \text{jet}$ event samples for the anti-$k_t$ $R = 1.0$ trimmed jet $D_2$ (a, c) and $r_{32}$ (b, d) spectra. The data-driven normalisation correction, described in Sect. 6.2.1, is shown in the legend beside the specific sample to which it applies. Systematic uncertainties are indicated as a band in the lower panel and include all experimental uncertainties related to the selection of events, as well as the reconstruction and calibration of the large-$R$ jet taggers and in Figs. 19, 20, 21, 22 and 23 for the top-quark taggers.

The signal efficiency for the $W$-boson and top-quark taggers in Monte Carlo simulation is compatible with the measured efficiency in data within uncertainties. In the case of the $W$-boson tagger working points, there is a systematic difference between the target 50% signal efficiency and that measured in data due to event topology differences between $W$-boson jets from these two samples, as was investigated in Ref. [8]. The total uncertainty of the measured signal efficiency is typically about 50% and 15% for the $W$-boson and top-quark tagger efficiencies, respectively, and is largely dominated by the subtraction of the non-contained top-quark contribution. In most of the kinematic phase space, these uncertainties are dominated by systematic uncertainties, described in Sect. 6.3, specifically by the theoretical
Fig. 27 A comparison of the observed data and MC predictions in the multijet and $\gamma + \text{jet}$ event samples for the anti-$k_t$ $R=1.0$ trimmed jet spectra of the $W$-boson (a, c) and top-quark (b, d) DNN discriminants. The data-driven normalisation correction, described in Sect. 6.2.1, is shown in the legend beside the specific sample to which it applies. Systematic uncertainties are indicated as a band in the lower panel and include all experimental uncertainties related to the selection of events, as well as the reconstruction and calibration of the large-$R$ jet uncertainties in $t\bar{t}$ modelling, largely coming from the subtraction of the component of the $t\bar{t}$ Monte Carlo prediction that consists of either non-$W$-boson jets or non-contained top-quark jets.

When examining the measured signal efficiency as a function of the average number of interactions per bunch crossing, it is found to be quite robust against increasing levels of event pile-up, even when considering only the statistical uncertainties due to the size of the data sample, noting that the systematic uncertainties are correlated between bins.

6.2 Background rejection from multijet and $\gamma + \text{jet}$ events

In addition to studying the modelling of signal $W$-boson and top-quark jets using a sample of $t\bar{t}$ events, the behaviour of background light jets is studied in two sets of events (enriched in multijet and $\gamma + \text{jet}$ processes) to cover a broad kine-
Fig. 28 A comparison of the observed data and MC predictions in the multijet (a) and $\gamma$ + jet (b) event samples for the anti-$k_t$ $R=1.0$ trimmed jet spectra of the TopoDNN top tagger discriminant. The data-driven normalisation correction, described in Sect. 6.2.1, is shown in the legend beside the specific sample to which it applies. Systematic uncertainties are indicated as a band in the lower panel and include all experimental uncertainties related to the selection of events, as well as the reconstruction and calibration of the large-$R$ jet.

Fig. 29 A comparison of the observed data and MC predictions in the multijet (a) and $\gamma$ + jet (b) event samples for the anti-$k_t$ $R=1.0$ trimmed jet spectra of the log $\chi$ shower deconstruction discriminant. The data-driven normalisation correction, described in Sect. 6.2.1, is shown in the legend beside the specific sample to which it applies. Systematic uncertainties are indicated as a band in the lower panel and include all experimental uncertainties related to the selection of events, as well as the reconstruction and calibration of the large-$R$ jet.
Fig. 30 A comparison of the observed data and predicted MC distributions in the multijet (a) and $\gamma +$ jet (b) event samples for the HEPTopTagger mass. The data-driven normalisation correction, described in Sect. 6.2.1, is shown in the legend beside the specific sample to which it applies. Systematic uncertainties are indicated as a band in the lower panel and include all experimental uncertainties related to the selection of events, as well as the reconstruction and calibration of the large-$R$ jet. The difference in shape of the HEPTopTagger mass distribution between the multijet and the $\gamma +$ jet selections, in particular the absence of a pronounced top-mass peak in the $\gamma +$ jet selection, is caused by the difference in the jet $p_T$ thresholds.

Fig. 31 A comparison of the observed data and predicted MC distributions of the anti-$k_t$, $R = 1.0$ trimmed jet $m^{\text{comb}}$ observable for events from the multijet (a) or $\gamma +$ jet (b) selections that pass the selection on the jet-shape-based $W$-boson DNN tagger. The data-driven normalisation correction, described in Sect. 6.2.1, is shown in the legend beside the specific sample to which it applies. Systematic uncertainties are indicated as a band in the lower panel and include all experimental uncertainties related to the selection of events, as well as the reconstruction and calibration of the large-$R$ jet.
matic range and probe the behaviour of quark- and gluon-enriched regions of phase space separately [97]. The first sample, multijet events, provides a means to study a mixture of light-quark and gluon jets in the kinematic range from $p_T$ of approximately 450–3000 GeV while the $\gamma +$ jet sample is greatly enhanced in the fraction of quark jets produced and provides a means to study jets with $p_T$ from $\sim 200$ to 2000 GeV. As in the case of the study of signal W-boson and top-quark jets in Sect. 6.1, the distributions of important tagging observables are examined and the background rejection is quantified in both data and Monte Carlo simulation.

**6.2.1 Analysis and selection**

To select the multijet sample, events are selected in both data and Monte Carlo simulation using a single-jet trigger based on a single large-$R$ anti-$k_t$ trimmed jet with $R = 1.0$ with an online requirement of $E_T > 360$ GeV during 2015 data taking and 420 GeV in 2016. Events are then required to have at least one fully-calibrated large-$R$ anti-$k_t$ trimmed jet with radius 1.0 with $p_T > 450$ GeV so that the trigger is fully efficient. After this selection, the modelling of the highest-$p_T$ large-$R$ jet (both anti-$k_t$, $R = 1.0$ trimmed and C/A $R = 1.5$) in the event is examined with respect to both the PYTHIA and HERWIG++ generators described in Sect. 3.

In the case of the $\gamma +$ jet sample, events are selected in both data and Monte Carlo simulation with a single-photon trigger which selects photons satisfying “loose” quality criteria and which pass an online requirement of $E_T > 120$ GeV in 2015 and 140 GeV in 2016. Photon candidates are required to be within $|\eta| < 2.5$ and satisfy a likelihood-based identification criterion based on shower shape observables in the electromagnetic calorimeter as well as the relative amount of energy in the hadronic and electromagnetic calorimeters, and are required to be isolated from other activity in the event. Both the identification and isolation criteria are required to satisfy the “tight” working point described in Ref. [98]. In addition, large-$R$ jets are required to have $p_T > 200$ GeV, $|\eta| < 2.0$ and to be well-separated from the reconstructed photon with $\Delta\phi(\text{jet}, \gamma) > \frac{\pi}{2}$. Finally, events with at least one photon with $E_T > 155$ GeV are selected to ensure that the trigger is fully efficient.

In both selections, the normalisation of the simulated multijet and $\gamma +$ jet predictions is derived directly from data after the initial inclusive selection, taking into account the small contribution from hadronically decaying W-boson, Z-boson and $t\bar{t}$ events. First the predicted contribution from processes containing real hadronically decaying W bosons and top quarks is subtracted from data. The remaining Monte Carlo samples are then normalised to reproduce the same yield as the background-subtracted data.

Figures 24 and 25 show a comparison of the distributions of the leading anti-$k_t$, $R = 1.0$ and C/A $R = 1.5$ jet mass in the inclusive multijet and $\gamma +$ jet selections. In addition, the
Fig. 33 The estimated light-jet rejection $1/\epsilon_{\text{bkg}}$ as a function of the leading jet $p_T$ and the average number of interactions per bunch crossing $\mu$ for the two-variable $W$-boson tagger in the multijet (a, c) and $\gamma +$ jet (b, d) selection.

6.2.2 Background rejection measurements

In a similar manner to the measurement of the signal efficiency in Sect. 6.1.2, the background rejection $1/\epsilon_{\text{bkg}}$ is measured for the $W$-boson and top-quark tagging working points described in Sect. 6.1.2. This measurement is performed in both the multijet and $\gamma +$ jet topologies as a function of the transverse momentum of the highest-$p_T$ jet in the event, taken to be the leading jet studied in Sect. 6.2.1, as well as $\mu$.

The approach in this measurement is simpler than the chi-square fit approach used in Sect. 6.1.2 due to the purity of
Fig. 34 The estimated light-jet rejection $1/\epsilon_{\text{bkg}}$ as a function of the leading jet $p_T$ and the average number of interactions per bunch crossing $\mu$ for the DNN $W$-boson tagger in the multijet (a, c) and $\gamma + \text{jet}$ (b, d) selection

these samples. In particular, after subtracting the signal contamination from data and performing the normalisation of the multijet and $\gamma + \text{jet}$ samples in the inclusive selection described in Sect. 6.2.1, the background efficiency is calculated directly as the fraction of events that satisfy the full set of tagging criteria in data and in Monte Carlo simulation. The results are shown in Figs. 33, 34, 35, 36, 37, 38 and 39, for the full set of tagging techniques. In the case of $W$-boson tagging (Figs. 33, 34), the dependence of the background rejection on jet $p_T$ arises from the requirement of a fixed signal efficiency. At low jet $p_T$, there is a non-negligible fraction of signal $W$-boson jets which are not sufficiently collimated due to radiation from parton shower outside of the jet area despite the signal labelling requirement on the $\Delta R$ between the quarks from $W$-boson decay and the jet axis. As a result, a broader jet-mass selection is required to maintain the 50% signal efficiency. As the jet $p_T$ increases, the sample of signal jets becomes better contained within a radius of 1.0, thereby allowing a stricter mass requirement, and the background rejection increases. However, the $W$-boson signal jets become fully contained at $p_T \sim 800$ GeV, and with increasing jet $p_T$ the experimental resolution worsens and the Sudakov peak of the light-jet mass migrates into the signal region, thereby leading to a degradation of the background rejection.
Good agreement is generally observed between the predicted and measured rejections. For the multijet topology, the \textsc{Pythia8} prediction of the background rejection describes the observed one, while the \textsc{Herwig++} prediction is lower than the rejection in data. Although the rejections for the two topologies are similar, there are relatively large uncertainties at higher jet $p_T$, with clear differences observed between the generators examined for the dominant samples in each topology. In particular, in the case of \textit{W}-boson tagging, it is observed that these generator differences are larger for the more complex jet-shape DNN tagger, shown in Fig. 34, than for the cut-based tagger, shown in Fig. 33. In the case of top-quark tagging, in addition to the trend between the jet-shape DNN and cut-based taggers, a similar trend can be seen in which a more algorithmically involved classifier, namely the TopoDNN tagger, shown in Fig. 37, shows larger differences between generators than the jet-shape DNN tagger, shown in Fig. 36.

When examining the background rejection with respect to $\mu$, in the case of \textit{W}-boson tagging, a trend of increasing background rejection for higher $\mu$ exists. This is observed in both the multijet and $\gamma + \text{jet}$ topologies and found to be the same size for both the $m_{\text{comb}} + D_2$ \textit{W}-boson tagger and the jet shape-based DNN \textit{W}-boson tagger. In the case of top-quark tagging, the $m_{\text{comb}} + \tau_3$ top-quark tagger, the jet shape-based DNN top-quark tagger, and the TopoDNN tagger show no
clear trend as a function of pile-up, likely due to the high-
$p_T$ regime selected by the top-quark taggers. However, the
Shower Deconstruction top-quark tagger shows minor trends
with the background rejection decreasing as the level of pile-
up increases. The background rejection of the HEPTopTagger
shows little dependence on $\mu$. In all cases, this trend is well-
described by the Monte Carlo simulation.

6.3 Systematic uncertainties

A number of sources of systematic uncertainty enter into
the evaluation of the modelling of data by the Monte Carlo
simulation. These uncertainties derive both from theoretical
assumptions within the Monte Carlo predictions and from the
reconstruction and calibration of the detector response to the
physics objects and therefore affect the three topologies to
varying degrees. These sources of uncertainty, their effect in
this analysis, and the manner in which they are estimated are
summarised in Tables 3 and 4. Systematic uncertainties are
propagated to the signal efficiency measurement by repeating
the fit for varied templates that correspond to each systematic
uncertainty source and comparing the extracted efficiency for
the varied and nominal templates.

From this set of uncertainties, those originating from the
measurement of leptons, photons, anti-$k_t$ $R = 0.4$ calorim-
eter jets, and the $E_T^{\text{miss}}$ soft term are found to be negligible in
Fig. 37 The estimated light-jet rejection $1/\epsilon_{bkg}$ as a function of the leading jet $p_T$ and the average number of interactions per bunch crossing $\mu$ for the TopoDNN top-quark tagger in the multijet (a, c) and $\gamma +$ jet (b, d) selection.

all cases. Additionally, the uncertainty related to the estimation and subsequent subtraction of the multijet background in the $tt$ analysis in Sect. 6.1 and the background with a real hadronic $W/Z$-boson or top-quark decay in the multijet and $\gamma +$ jet analysis in Sect. 6.2 are found to be negligible. The uncertainties due to the application of flavour tagging in the $tt$ analysis of signal jets are subdominant and affect the yield results with an impact of the order of 20% in the region of $m_{\text{comb}}$ below 100 GeV. Similarly, the component of the flavour tagging uncertainties pertaining to the misidentification of light-flavour jets as $b$-jets tend to have a larger effect at low values of the multivariate classifier score, in Figs. 12 and 13 where non-top-quark jet contributions are more dominant. However, due to the localization of these effects, they have a negligible impact on the measurement of the signal efficiency. The uncertainties in both the scale and resolution of the observable of interest (e.g. $m_{\text{calo}}$, $D_2$ and $\tau_{32}$) are evaluated by comparing the large-$R$ jets formed from calorimeter cell topoclusters to those formed from ID tracks [95]. These sources of uncertainty generally cause small (10%) changes in the yield of events near the most highly populated regions of the distributions of observables but are generally the dominant uncertainties when examining both the tails of these distributions and the regions near $m_W$ and $m_{\text{top}}$ in Sect. 6.1. Likewise, in the case of the HEPTopTagger, the subjet energy scale uncertainty, which itself is based on Run 1 studies, is
Fig. 38 The estimated light-jet rejection $1/\epsilon_{\text{bkg}}$ as a function of the leading jet $p_T$ and the average number of interactions per bunch crossing $\mu$ for the shower deconstruction top-quark tagger in the multijet (a, c) and $\gamma+\text{jet}$ (b, d) selection.

A dominant source of systematic uncertainty in the shape of the HEPTopTagger mass near $m_{\text{top}}$ but this uncertainty does not propagate strongly into the final evaluation of the signal efficiency due to the broad mass window selection described in Sect. 4.3.4.

The dominant systematic uncertainties of these techniques are those related to the theoretical modelling of the Monte Carlo predictions. In particular, in Sect. 6.1, the contribution of the uncertainty in the modelling of parton shower and hadronisation is dominant in all cases, leading to variations in the yield of the Monte Carlo when examining the distributions of $m_{\text{comb}}$, $D_2$, and $r_{32}$ of up to 30%. This is also true when examining the modelling of the multivariate classifiers, shown in Figs. 12 and 13. In the tails of these distributions, the uncertainty in the modelling of additional radiation in $t\bar{t}$ events yields variations that are comparable in size. The same behaviour can be observed in the study of the modelling of light jets, particularly in Sect. 6.2.1, where predictions from both PYTHIA8 and HERWIG++ show shape differences ranging up to approximately 25% for certain jet moments as well as for the DNN top tagger. As seen in Sects. 6.1.2 and 6.2.2, these uncertainties manifest themselves as large variations in the measured signal efficiency and background rejection. In the case of the tagging efficiency measurement of top quarks in particular, the measured signal efficiency is found to be
susceptible to both the truth-level labelling of the top quark and the particular working point chosen for the tagger.

7 Conclusion

Various methods to tag boosted, hadronically decaying $W$ bosons and top quarks are studied in data and simulation. A number of techniques, including the use of physically motivated jet moments, shower deconstruction and the HEPTopTagger which were studied in Run 1 are re-optimised for use in LHC Run 2 conditions. Additionally, the multivariate combination of high-level jet moments using boosted decision trees and neural networks as well as the combination of low-level energy flow information in the form of topoclusters using a deep neural network is studied both in data and Monte Carlo simulation. The performance of these techniques is evaluated using Monte Carlo simulation for jets in the $p_T$ range from 500 to 2000 GeV and compared in terms of the central value of the background rejection at fixed signal efficiency. This study indicates that a multivariate combination of information can enhance performance to exceed that of techniques based on more physically motivated individual
<table>
<thead>
<tr>
<th>Source</th>
<th>Affected topologies</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event generator choice</td>
<td>$t\bar{t}$</td>
<td>Hard-scattering modelling uncertainty estimated as the difference between POWHEG+HERWIG and MC@NLO+HERWIG [29]</td>
</tr>
<tr>
<td>Showering choice</td>
<td>$t\bar{t}$</td>
<td>Parton shower and hadronisation modelling uncertainty estimated from the difference between POWHEG+PYTHIA6 and POWHEG+HERWIG [29]</td>
</tr>
<tr>
<td>Modelling of extra QCD radiation</td>
<td>$t\bar{t}$</td>
<td>Uncertainty in amount of initial/final-state radiation estimated as the difference between the nominal POWHEG+PYTHIA6 generator and radLo and radHi tunes of POWHEG+PYTHIA6. The radiation variations include variation of renormalisation and factorisation scales and the $h_{\text{damp}}$ Parameters [29]</td>
</tr>
<tr>
<td>$t\bar{t}$ total cross-section</td>
<td>$t\bar{t}$</td>
<td>Uncertainty in normalisation of $t\bar{t}$ MC contribution of magnitude $\pm 5.5%$ [26]</td>
</tr>
<tr>
<td>uncertainty</td>
<td>$t\bar{t}$</td>
<td>Uncertainty in normalisation of single-top MC contribution of magnitude $\pm 5.3%$ (a conservative estimate enveloping the uncertainties on $t$-channel, $s$-channel and $Wt$-channel)</td>
</tr>
<tr>
<td>Single-top total cross-section</td>
<td>$t\bar{t}$</td>
<td>Uncertainty in normalisation of $W$+jets MC contribution of magnitude $\pm 5.0%$ [99]</td>
</tr>
<tr>
<td>uncertainty</td>
<td>$t\bar{t}$</td>
<td>Uncertainty arising from the choice of renormalisation and factorisation scale, CKKW matching scale and QSF scale [100]. For the renormalisation and factorisation scale, $\times 0.5$ and $\times 2$ variations are considered and the renormalisation and factorisation scales are varied independently as well as in correlated and anti-correlated ways. The envelope of the variations is considered as the final renormalisation+factorisation scale uncertainty</td>
</tr>
<tr>
<td>$W$+jets total cross-section</td>
<td>$t\bar{t}$</td>
<td>Uncertainty in normalisation of $W$+jets MC contribution of magnitude $\pm 5.0%$ [99]</td>
</tr>
<tr>
<td>uncertainty</td>
<td>$t\bar{t}$</td>
<td>Uncertainty in subtraction of processes containing a hadronically decaying top quark or vector boson, conservatively taken to be $25%$</td>
</tr>
<tr>
<td>$W$+jets theory scale uncertainties</td>
<td>$t\bar{t}$</td>
<td>The uncertainty in subtraction of processes containing a hadronically decaying top quark or vector boson, conservatively taken to be $25%$</td>
</tr>
<tr>
<td>Signal normalisation</td>
<td>Multijet, $\gamma$+jet</td>
<td></td>
</tr>
</tbody>
</table>

The performance of the various tagging techniques is studied using a sample of 36.1 fb$^{-1}$ of 13 TeV proton–proton collision data collected by the ATLAS detector at the LHC in 2015 and 2016. A sample of lepton-plus-jets $t\bar{t}$ events is used to study the signal $W$-boson and top-quark jet tagging efficiency and compare the predicted efficiency in Monte Carlo simulation to that in data for a set of working points for the tagging strategies from which in situ calibrations and systematic uncertainties can be derived. Likewise, background light-jet-enriched event topologies are studied using multijet and $\gamma$+jet samples. We have demonstrated that tagging efficiencies and the relevant uncertainties for both signal and background can be extracted from data. This opens opportunities for complex $W$-boson and top-quark taggers utilizing with ATLAS data in the future. In general, it is found that the inputs to and the performance of the studied $W$-boson and top-quark taggers currently in use in physics analyses are well-modelled by Monte Carlo simulations. However, in all studies, it is found that the primary limiting factor in the description of the tagging efficiency by the Monte Carlo prediction derives from the theoretical modelling of the Monte Carlo processes studied, particularly the parton shower and hadronisation model of the $t\bar{t}$ process. Finally, the small pile-up dependence of each tagger working point is characterised to understand the relative susceptibility of each strategy to pile-up contamination within the jet. In general, the signal efficiency is found to be quite robust against increased levels of event pile-up whereas the background rejection shows residual pile-up dependence, particularly in the case of the $W$ taggers. In all cases, however, the dependence is well-described by the Monte Carlo simulation.
Table 4 Summary of experimental systematic uncertainties considered in the performance measurements in data

<table>
<thead>
<tr>
<th>Source</th>
<th>Affected topologies</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-$k_t$ $R=1.0$ trimmed jet</td>
<td>$t\bar{t}$, multijet, $\gamma$ + jet</td>
<td>The uncertainty in the scale of the detector response for all jet moments derived by comparing the calorimeter quantity with the reference track jet [8]</td>
</tr>
<tr>
<td>Anti-$k_t$ $R=1.0$ trimmed jet</td>
<td>$t\bar{t}$, multijet, $\gamma$ + jet</td>
<td>The uncertainty in the resolution of the detector response conservatively estimated as a 2% absolute uncertainty in $p_t$, a 20% relative uncertainty in jet mass, and a 15% relative uncertainty in all other jet moments [95]</td>
</tr>
<tr>
<td>C/A $R=1.5$ subjett energy scale</td>
<td>$t\bar{t}$, multijet, $\gamma$ + jet</td>
<td>The uncertainty in the scale of the detector energy response for subjets used in the HEPTopTagger algorithm conservatively estimated to be 3% based on Run 1 studies [9]</td>
</tr>
<tr>
<td>Anti-$k_t$, $R=0.4$ jet energy scale and resolution</td>
<td>$t\bar{t}$</td>
<td>The uncertainty in the scale and resolution of the detector response for the jet $p_T$ derived from simulation and in situ calibration measurements [40]</td>
</tr>
<tr>
<td>$E_T^{miss}$ track soft term</td>
<td>$t\bar{t}$</td>
<td>The uncertainty on the component of the $E_T^{miss}$ calculation due to energy flow that is unassigned to a calibrated physics object, estimated in-situ in Z+jet events [94]</td>
</tr>
<tr>
<td>Flavour tagging</td>
<td>$t\bar{t}$</td>
<td>The uncertainty in the scale factor correcting the efficiency response of the detector to identify heavy-flavour $b$- and $c$-jets as well as light-flavour jets derived in situ using $t\bar{t}$ events [92, 96]</td>
</tr>
<tr>
<td>Lepton reconstruction and calibration</td>
<td>$t\bar{t}$</td>
<td>The uncertainty in the scale factor correcting the efficiency to trigger on, reconstruct, and identify leptons as well as uncertainties in their energy and $p_T$ scale and resolution [88, 89, 101]</td>
</tr>
<tr>
<td>Photon reconstruction and calibration</td>
<td>$\gamma$ + jet</td>
<td>The uncertainty in the scale factor correcting the efficiency to trigger on, reconstruct, and identify photons [102] as well as uncertainties in their energy scale and resolution [103]</td>
</tr>
<tr>
<td>Multijet background normalisation</td>
<td>$t\bar{t}$</td>
<td>The uncertainty in the data-driven prediction of the yield of multijet events, conservatively taken to be 50% based on the estimate in Ref. [2]</td>
</tr>
<tr>
<td>Multijet lepton misreconstruction efficiencies</td>
<td>$t\bar{t}$</td>
<td>The statistical uncertainty of the real and fake/non-prompt lepton reconstruction efficiencies estimated in Ref. [2] is propagated through the matrix method</td>
</tr>
<tr>
<td>Luminosity uncertainty</td>
<td>$t\bar{t}$, multijet, $\gamma$ + jet</td>
<td>A 2.1% relative uncertainty in the MC yield, based on the luminosity uncertainty of the combined 2015+2016 dataset based on [104]</td>
</tr>
<tr>
<td>Pile-up uncertainty</td>
<td>$t\bar{t}$, multijet, $\gamma$ + jet</td>
<td>Uncertainty in the reweighting of MC pile-up profile to the measured pile-up profile in data based on disagreement between instantaneous luminosity in data and in simulation [104]</td>
</tr>
</tbody>
</table>

Acknowledgements 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; BMWFW 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 and DNSRC, Denmark; IN2P3-CNRS, CEA-DRF/IRFU, France; SRNSFG, Georgia; BMBF, HGF, and MPG, Germany; GSRT, Greece; RGC, Hong Kong SAR, China; ISF and Benoziyo Center, Israel; INFN, Italy; MEXT and JSPS, Japan; CNRST, Morocco; NWO, Netherlands; RCN, Norway; MNiSW and NCN, Poland; FCT, Portugal; MINErFA, Romania; MES of Russia and NRC KI, Russian Federation; JINR; MEStD, Serbia; MSSR, Slovakia; ARRS and MIZŠ, Slovenia; DST/NRF, South Africa; MINECO, Spain; SRC and Wallenberg Foundation, Sweden; SERI, SNSF and Cantons of Bern and Geneva, Switzerland; MOST, Taiwan; TAEK, Turkey; STFC, United Kingdom; DOE and NSF, United States of America. In addition, individual groups and members have received support from BCKDF, CANARC, CRC and Compute Canada, Canada; COST, ERC, ERDF, Horizon 2020, and Marie Skłodowska-Curie Actions, European Union; Investissements d’Avenir Labex and Idex, ANR, France; DFG and AvH Foundation, Germany; Herakleitos, Thales and Aristeia programmes co-financed by EU-ESF and the Greek NSRF, Greece; BSF-NSF and GIF, Israel; CERCA Programme Generalitat de Catalunya, Spain; The Royal Society and Leverhulme Trust, United Kingdom. The crucial computing support from all WLCG partners is acknowledged gratefully, in particular from CERN, 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), the Tier-2 facilities worldwide and large non-WLCG resource providers. Major contributors of computing resources are listed in Ref. [113].

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors’ comment: “All ATLAS scientific output is published in journals, and preliminary results are made available in Conference Notes. All are openly available, without restriction on use by external parties beyond copyright law and the standard conditions agreed by CERN. Data associated with journal publications are also made available: tables and data from plots (e.g. cross section values, likelihood profiles, selection efficiencies, cross section limits, ...) are stored in appropriate repositories such as HEPDATA (http://hepdata.cedar.ac.uk/). ATLAS also strives to make additional material related to the paper available that allows a reinterpretation of the data.
in the context of new theoretical models. For example, an extended encapsulation of the analysis is often provided for measurements in the framework of RIVET (http://rivet.hepforge.org/). This information is taken from the ATLAS Data Access Policy, which is a public document that can be downloaded from http://opendata.cern.ch/record/413 [opendata.cern.ch].]

Open Access  This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. Funded by SCOAP3.

Appendix

A BDT and DNN hyper-parameters

In this section, a description of the tuned hyper-parameters of the BDT and DNN are presented in Tables 5 and 6.

### Table 5  Brief description of the BDT parameters and the chosen parameters

<table>
<thead>
<tr>
<th>Setting name</th>
<th>Description</th>
<th>Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software package</td>
<td>Package used for training</td>
<td>TMVA 4.2.1 [105]</td>
</tr>
<tr>
<td>BoostType</td>
<td>Type of boosting technique</td>
<td>GradientBoost</td>
</tr>
<tr>
<td>NTrees</td>
<td>Number of trees in the forest</td>
<td>500</td>
</tr>
<tr>
<td>MaxDepth</td>
<td>Max depth of the decision tree allowed</td>
<td>20</td>
</tr>
<tr>
<td>MinimumNodeSize</td>
<td>Minimum fraction of training events required in a leaf node</td>
<td>1.0%</td>
</tr>
<tr>
<td>Shrinkage</td>
<td>Learning rate for GradientBoost algorithm</td>
<td>0.5</td>
</tr>
<tr>
<td>UseBaggedBoost</td>
<td>Use only a random (bagged) subsample of all events for growing the trees in each iteration</td>
<td>True</td>
</tr>
<tr>
<td>BaggedSampleFraction</td>
<td>Relative size of bagged event sample to original size of the data sample</td>
<td>0.5</td>
</tr>
<tr>
<td>SeparationType</td>
<td>Separation criterion for node splitting</td>
<td>GiniIndex</td>
</tr>
<tr>
<td>nCuts</td>
<td>Number of grid points in variable range used in finding optimal cut in node splitting</td>
<td>500</td>
</tr>
</tbody>
</table>

### Table 6  Chosen DNN parameters and architecture for shape-based $W^-$boson and top-quark tagging

<table>
<thead>
<tr>
<th></th>
<th>$W^-$boson tagging</th>
<th>Top-quark tagging</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software package</td>
<td>Keras 1.0.8 with Theano backend, lwtnn 2.0</td>
<td></td>
<td>[106–108]</td>
</tr>
<tr>
<td>Layer type</td>
<td>Dense</td>
<td>Dense</td>
<td>[106]</td>
</tr>
<tr>
<td>Number of hidden layers</td>
<td>4</td>
<td>5</td>
<td>[106]</td>
</tr>
<tr>
<td>Architecture</td>
<td>16, 14, 9, 6</td>
<td>18, 16, 14, 10, 5</td>
<td>–</td>
</tr>
<tr>
<td>Activation function</td>
<td>Rectified linear unit (relu)</td>
<td>Rectified linear unit (relu)</td>
<td>[109]</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam</td>
<td>Adam</td>
<td>[110]</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.0001</td>
<td>0.00005</td>
<td>[110]</td>
</tr>
<tr>
<td>L1 regulariser</td>
<td>0.001</td>
<td>0.001</td>
<td>[109]</td>
</tr>
<tr>
<td>NN weight initialisation</td>
<td>Glorot uniform</td>
<td>Glorot uniform</td>
<td>[111]</td>
</tr>
<tr>
<td>Batch size</td>
<td>200</td>
<td>200</td>
<td>[109]</td>
</tr>
<tr>
<td>Batch normalisation</td>
<td>Yes</td>
<td>Yes</td>
<td>[112]</td>
</tr>
<tr>
<td>Number of epochs</td>
<td>100 with early stopping</td>
<td>100 with early stopping</td>
<td>[106]</td>
</tr>
<tr>
<td>Training input group</td>
<td>Group 8</td>
<td>Group 9</td>
<td>–</td>
</tr>
</tbody>
</table>

References

8. ATLAS Collaboration, Identification of Boosted, Hadronically Decaying W Bosons and Comparisons with ATLAS Data Taken...
156 (a) E. Andronikashvili Institute of Physics, Iv. Javakhishvili Tbilisi State University, Tbilisi, Georgia; (b) High Energy Physics Institute, Tbilisi State University, Tbilisi, Georgia
157 Department of Physics, Technion, Israel Institute of Technology, Haifa, Israel
158 Raymond and Beverly Sackler School of Physics and Astronomy, Tel Aviv University, Tel Aviv, Israel
159 Department of Physics, Aristotle University of Thessaloniki, Thessaloniki, Greece
160 International Center for Elementary Particle Physics and Department of Physics, University of Tokyo, Tokyo, Japan
161 Graduate School of Science and Technology, Tokyo Metropolitan University, Tokyo, Japan
162 Department of Physics, Tokyo Institute of Technology, Tokyo, Japan
163 Tomsk State University, Tomsk, Russia
164 Department of Physics, University of Toronto, Toronto, ON, Canada
165 (a) TRIUMF, Vancouver, BC, Canada; (b) Department of Physics and Astronomy, York University, Toronto, ON, Canada
166 Division of Physics and Tomonaga Center for the History of the Universe, Faculty of Pure and Applied Sciences, University of Tsukuba, Tsukuba, Japan
167 Department of Physics and Astronomy, Tufts University, Medford, MA, USA
168 Department of Physics and Astronomy, University of California Irvine, Irvine, CA, USA
169 Department of Physics and Astronomy, University of Uppsala, Uppsala, Sweden
170 Department of Physics, University of Illinois, Urbana, IL, USA
171 Instituto de Física Corpuscular (IFIC), Centro Mixto Universidad de Valencia - CSIC, Valencia, Spain
172 Department of Physics, University of British Columbia, Vancouver, BC, Canada
173 Department of Physics and Astronomy, University of Victoria, Victoria, BC, Canada
174 Fakultät für Physik und Astronomie, Julius-Maximilians-Universität Würzburg, Würzburg, Germany
175 Department of Physics, University of Warwick, Coventry, UK
176 Waseda University, Tokyo, Japan
177 Department of Particle Physics, Weizmann Institute of Science, Rehovot, Israel
178 Department of Physics, University of Wisconsin, Madison, WI, USA
179 Fakultät für Mathematik und Naturwissenschaften, Fachgruppe Physik, Bergische Universität Wuppertal, Wuppertal, Germany
180 Department of Physics, Yale University, New Haven, CT, USA
181 Yerevan Physics Institute, Yerevan, Armenia

a Also at Borough of Manhattan Community College, City University of New York, NY, USA
b Also at California State University, East Bay, United States of America
c Also at Centre for High Performance Computing, CSIR Campus, Rosebank, Cape Town, South Africa
d Also at CERN, Geneva, Switzerland
e Also at CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille, France
f Also at Département de Physique Nucléaire et Corpusculaire, Université de Genève, Genève, Switzerland
g Also at Departament de Física de la Universitat Autònoma de Barcelona, Barcelona, Spain
h Also at Departamento de Física Teorica y del Cosmos, Universidad de Granada, Granada (Spain), Spain
i Also at Department of Applied Physics and Astronomy, University of Sharjah, Sharjah, United Arab Emirates
j Also at Department of Financial and Management Engineering, University of the Aegean, Chios, Greece
k Also at Department of Physics and Astronomy, University of Louisville, Louisville, KY, USA
l Also at Department of Physics and Astronomy, University of Sheffield, Sheffield, UK
m Also at Department of Physics, California State University, Fresno, CA, USA
n Also at Department of Physics, California State University, Sacramento, CA, USA
o Also at Department of Physics, King’s College London, London, UK
p Also at Department of Physics, St. Petersburg State Polytechnical University, St. Petersburg, Russia
q Also at Department of Physics, Stanford University, United States of America
r Also at Department of Physics, University of Fribourg, Fribourg, Switzerland
s Also at Department of Physics, University of Michigan, Ann Arbor, MI, USA
t Also at Dipartimento di Fisica E. Fermi, Università di Pisa, Pisa, Italy
u Also at Giresun University, Faculty of Engineering, Giresun, Turkey.
v Also at Graduate School of Science, Osaka University, Osaka, Japan
w Also at Hellenic Open University, Patras, Greece