Search for neutral long-lived particles in pp collisions at √s = 13 TeV that decay into displaced hadronic jets in the ATLAS calorimeter

The ATLAS Collaboration

DOI
10.1007/JHEP06(2022)005

Publication date
2022

Document Version
Final published version

Published in
Journal of High Energy Physics

License
CC BY

Citation for published version (APA):
Search for neutral long-lived particles in $pp$ collisions at $\sqrt{s} = 13$ TeV that decay into displaced hadronic jets in the ATLAS calorimeter

The ATLAS collaboration

E-mail: atlas.publications@cern.ch

ABSTRACT: A search for decays of pair-produced neutral long-lived particles (LLPs) is presented using 139 fb$^{-1}$ of proton-proton collision data collected by the ATLAS detector at the LHC in 2015–2018 at a centre-of-mass energy of 13 TeV. Dedicated techniques were developed for the reconstruction of displaced jets produced by LLPs decaying hadronically in the ATLAS hadronic calorimeter. Two search regions are defined for different LLP kinematic regimes. The observed numbers of events are consistent with the expected background, and limits for several benchmark signals are determined. For a SM Higgs boson with a mass of 125 GeV, branching ratios above 10% are excluded at 95% confidence level for values of $c$ times LLP mean proper lifetime in the range between 20 mm and 10 m depending on the model. Upper limits are also set on the cross-section times branching ratio for scalars with a mass of 60 GeV and for masses between 200 GeV and 1 TeV.

KEYWORDS: Hadron-Hadron Scattering

ArXiv ePrint: 2203.01009
1 Introduction

Many of the scenarios proposed to address some of the open questions of the Standard Model (SM) predict the existence of new non-SM particles whose lifetimes can be long enough for their decays to be significantly displaced from the interaction point but still within the ATLAS detector. Examples of theories involving such suitably long-lived particles (LLPs) are various supersymmetric (SUSY) models [1–7]; neutral naturalness models [8–11] which feature a hidden sector (HS) [12–14] that addresses the hierarchy problem; models that seek to incorporate dark matter [15–18], or explain the matter-antimatter asymmetry of the universe [19]; and models that lead to massive neutrinos [20, 21] which provide an explanation for the origin of light neutrino masses and mixings.

The decay of an LLP created in proton-proton collisions in the ATLAS detector could produce one of a variety of highly unconventional signatures, depending on the detector subsystem in which the LLP decays. Searches for promptly decaying particles often have very low sensitivity to such signatures and hence the study of LLPs requires dedicated analyses. This paper presents a search sensitive to neutral LLPs decaying chiefly in the calorimeters of the ATLAS detector. This allows the analysis to probe values of $c\tau$ (where $c$ is the speed of light and $\tau$ is the mean proper lifetime of the LLP) ranging between a few centimetres and a few tens of metres.
A HS benchmark model [12–14, 22, 23] is studied, in which the SM and HS are connected via a heavy neutral boson (Φ), which decays into two long-lived neutral scalar bosons (s). While Φ could be the SM Higgs boson, this analysis considers mediators with masses (mΦ) ranging from 60 GeV to 1000 GeV, and scalars with masses between 5 GeV and 475 GeV. The decay Φ → ss → f ¯f′f′ is considered (see figure 1), where f refers to fermions and ¯f to anti-fermions. The branching ratio (BR, with value B) of each decay mode depends on the mass of the scalar, with the dominant decay being to the heaviest quark that is kinematically accessible. In the hypothesised physics process with the largest scalar mass ms (475 GeV), decays to top quarks dominate (B > 99%). Conversely, in the hypothesis with the lightest ms (5 GeV), decays to charm quarks dominate (B ∼ 75%), followed by decays to τ-leptons (B ∼ 25%). In the other cases, the relation among the BRs is approximately constant and typically 85:8:5 for ¯b ¯b, ¯cc, and τ+τ−, respectively. A single analysis strategy, described in section 4, is used for all mass hypotheses regardless of the scalar LLP’s decay mode. The cτ of LLPs in HS models are usually not constrained by theory, apart from a rough upper limit of cτ < 10^8 m given by the cosmological constraint of Big Bang Nucleosynthesis [24], and could be short enough for the LLPs to decay inside the ATLAS detector volume.

The SM fermions from the LLP decay result in jets whose origins may be far from the interaction point (IP) of the colliding protons, leading to so-called displaced vertices or displaced jets. If the LLP decay occurs in the calorimeters, the decay products are collimated enough to be reconstructed as a single jet which is narrow, trackless and with an unusually high proportion of its energy in the hadronic calorimeter. Since pair-produced LLPs are considered, this analysis requires two such non-standard jets, with two selections targeting different LLP kinematic regimes. One is optimised for models with mΦ ≤ 200 GeV (referred to as low-ET models), and the other for models with mΦ > 200 GeV (high-ET models), where ET denotes transverse energy. The dominant background process that mimics this signal is SM multijet production, in cases where the jets are composed mainly of neutral hadrons or where some of the tracks are misreconstructed. Despite the low probability for a prompt jet to produce a signal-like jet, the SM multijet cross-section is high enough for this to be the dominant background in this search. Other contributions come from jets reconstructed from non-collision background consisting of cosmic rays and beam-induced background (BIB) [25]. The latter is composed of LHC beam-gas interactions and beam-halo interactions with the collimators upstream of the ATLAS detector, resulting
in muons travelling parallel to the beam-pipe. This analysis makes use of a new per-jet neural network to discriminate signal-like jets from non-displaced jets or BIB-like jets, and a boosted decision tree to separate signal from background events. The per-jet neural network replaces a boosted decision tree used in the previous version of the analysis, using an adversarial training scheme to minimise the effect of mis-modelling in the input variables. This is the first time such a technique has been used in an ATLAS analysis. The background estimation is performed using a data-driven method.

Previous searches for similar signatures (pair-produced neutral LLPs decaying hadronically) at hadron colliders have been performed at the Tevatron [26, 27] and at the LHC. The search for displaced jets at LHCb in ref. [28] is sensitive to \(c\tau\) values from \(\sim 1\) mm to \(\sim 0.1\) m. The most recent CMS searches at 13 TeV [29–32] involve jets with displaced vertices in the tracking system, and are sensitive to \(c\tau\) values from \(\sim 1\) mm to \(\sim 1\) m. Previous ATLAS searches at 13 TeV looked for displaced vertices in the tracking system [33, 34], pairs of reconstructed vertices in the muon spectrometer [35], or the combination of one displaced vertex in the muon spectrometer and one in the inner tracking detector [36]. These ATLAS searches are complementary, and together provide coverage of \(c\tau\) values extending from effectively prompt to \(\sim 200\) m. This analysis is an update of a search for pairs of displaced hadronic jets in the ATLAS calorimeters [37], with significant improvements to the displaced-jet identification, and using the full LHC Run 2 dataset with 139 fb\(^{-1}\) of 13 TeV data instead of only data from 2016. It is sensitive to long-lived particle \(c\tau\) values between approximately 20 mm and 10 m, depending on the model.

The paper is structured as follows. The ATLAS detector is described in section 2. The collection of the data and generation of samples of simulated events are then discussed in section 3. The trigger and event selection are detailed in section 4, followed by a discussion of the estimation of the background yield in the search regions in section 5. The systematic uncertainties are summarised in section 6. The statistical interpretation of the data is described in section 7, and the conclusions are given in section 8.

2 ATLAS detector

The ATLAS detector [38] at the LHC covers nearly the entire solid angle around the collision point.\(^1\) It is a multipurpose detector consisting of an inner tracking detector surrounded by a thin superconducting solenoid, electromagnetic and hadronic calorimeters, and a muon spectrometer incorporating three large superconducting toroidal magnets. The inner-detector system is immersed in a 2 T axial magnetic field and provides charged-particle tracking in the range \(|\eta| < 2.5\).

The high-granularity silicon pixel detector covers the vertex region and typically provides four measurements per track. The layer closest to the interaction point is known as the

---

\(^1\)ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point 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 \((\tau, \phi)\) 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 \(\Delta R \equiv \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2}\).
insertable B-Layer [39–41]. It was added in 2014 and provides high-resolution hits at small radius to improve the tracking performance. The pixel detector is surrounded by the silicon microstrip tracker, which usually provides four three-dimensional measurement points per track. These silicon detectors are complemented by the transition radiation tracker, with coverage up to $|\eta| = 2.0$.

The calorimeter system covers 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) electromagnetic calorimeters (together referred to as the ECal), with an additional thin LAr presampler covering $|\eta| < 1.8$ to correct for energy loss in material upstream of the calorimeters. The ECal extends from 1.5 m to 2.0 m in radial distance $r$ in the barrel and from 3.6 m to 4.25 m in $|z|$ in the endcaps. Hadronic calorimetry is provided by a steel/scintillator-tile calorimeter (HCal), segmented into three barrel structures within $|\eta| < 1.7$, and two copper/LAr hadronic endcap calorimeters covering $|\eta| > 1.5$. The HCal covers the region from 2.25 m to 4.25 m in $r$ in the barrel (although the HCal active material extends only up to 3.9 m) and from 4.3 m to 6.05 m in $|z|$ in the endcaps. The solid angle coverage is completed with forward copper/LAr and tungsten/LAr calorimeter modules optimised for electromagnetic and hadronic measurements, respectively.

The calorimeters have a highly granular lateral and longitudinal segmentation. Including the presamplers, there are seven sampling layers in the combined central calorimeters (the LAr presampler, three in the ECal barrel and three in the HCal barrel), and the endcap regions provide up to eight sampling layers (the presampler, three in ECal endcaps and four in HCal endcaps). The forward calorimeter modules provide three sampling layers in the forward region. The total amount of material in the ECal corresponds to 24–35 radiation lengths in the barrel and 35–40 radiation lengths in the endcaps. The combined depth of the calorimeters for hadronic energy measurements is more than 9 hadronic interaction lengths nearly everywhere across the full detector acceptance.

The muon spectrometer comprises separate trigger and high-precision tracking chambers measuring the deflection of muons in the magnetic field generated by the superconducting air-core toroids. The field integral of the toroids ranges between 2.0 and 6.0 T m across most of the detector.

The ATLAS detector selects events using a tiered trigger system [42]. The level-1 (L1) trigger is implemented in custom electronics and reduces the event rate from the LHC crossing frequency of 40 MHz to a design value of 100 kHz. The second level, known as the high-level trigger (HLT), is implemented in software running on a commodity PC farm that processes the events and reduces the rate of recorded events to 1 kHz. An extensive software suite [43] is used in the reconstruction and analysis of real and simulated data, in detector operations, and in the trigger and data acquisition systems of the experiment.

3 Data and simulation samples

Data collected by the ATLAS detector during the period 2015–2018 from proton-proton ($pp$) collisions at $\sqrt{s} = 13$ TeV are used in this search. Only data collected during stable beam conditions in which all detector subsystems were operational are considered [44]. These
<table>
<thead>
<tr>
<th>Trigger</th>
<th>Collected integrated luminosity [fb$^{-1}$]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2015</td>
</tr>
<tr>
<td>High-$E_T$ CalRatio trigger with $E_T &gt; 60$ GeV</td>
<td>3</td>
</tr>
<tr>
<td>High-$E_T$ CalRatio trigger with $E_T &gt; 100$ GeV</td>
<td>—</td>
</tr>
<tr>
<td>Low-$E_T$ CalRatio trigger (2016 version)</td>
<td>—</td>
</tr>
<tr>
<td>Low-$E_T$ CalRatio trigger (2018 version)</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 1. CalRatio triggers which were available during the LHC Run 2 data-taking, and corresponding integrated luminosity collected in each period. The high-$E_T$ CalRatio trigger with $E_T > 60$ GeV was disabled in 2017 for instantaneous luminosities higher than $1.4 \times 10^{34}$ cm$^{-2}$s$^{-1}$. Two versions of the low-$E_T$ CalRatio trigger were used, with slight differences in their algorithms. The details are reported in section 4.

Data were collected with a set of dedicated LLP signature-driven triggers and separated into four datasets, defined according to the triggers used to collect them. The search is performed on the main dataset, composed of all data events passing at least one of the two types of CalRatio triggers [45] running on bunch crossings where protons were present in both beams. The name “CalRatio” refers to the ratio of energies deposited in the hadronic and electromagnetic calorimeters. As described in detail in section 4, these include the low-$E_T$ CalRatio triggers and high-$E_T$ CalRatio triggers. Different versions of these triggers ran during the full data-taking period. The amount of data collected in each case is summarised in table 1, with at least one of these triggers running at any given time during the entire data-taking. Two additional datasets were collected for the study of non-collision backgrounds. The BIB dataset, used for the study of BIB events faking signal events, was collected from events failing the CalRatio trigger BIB-removal algorithm. The cosmics dataset, used for the estimation of cosmic-ray events passing the analysis selection, was collected from events recorded during empty bunch crossings, as described in section 4. Finally, a dijet dataset is selected using a single-jet-based trigger and vetoing on the CalRatio triggers to make it orthogonal to the main dataset. This dataset is used in the neural network training described in section 4.2 and in the calculation of some of the systematic uncertainties involved in the analysis.

The HS $\Phi \rightarrow ss$ signal samples were generated using the MadGraph5_aMC@NLO v2.6.2 [46] generator at leading order (LO) with the NNPDF2.3LO parton distribution function (PDF) set [47]. In these samples, $\Phi$ is produced via gluon-gluon fusion. The $\Phi$ transverse momentum distribution for the samples is reweighted to match that obtained for corresponding next-to-leading-order (NLO) predictions using the same event generator. A production cross-section of $48.6\,\text{pb}$, taken from a next-to-next-to leading-order calculation [48], is assumed when normalising results for the case where the mediator is the SM Higgs boson (the mass of which was set to 125 GeV). Parton showering and hadronisation was modelled using Pythia 8.230 [49] with the A14 set of tuned parameters (tune) [50]. Several sets of samples were generated, with different assumptions for the masses of the
mediator \((m_\Phi \in [60, 1000] \text{ GeV})\) and LLPs \((m_s \in [5, 475] \text{ GeV})\). These mass ranges ensure an extensive LLP boost spectrum, allowing a wide variety of topologies to be studied. The LLP mean proper lifetime at which each of the signal samples was generated \((\tau_{\text{gen}})\) was determined so as to maximise the fraction of decays in the ATLAS hadronic calorimeter and muon system, and is quoted in relevant tables and figures throughout this document. Some of the samples were generated for two assumptions about the LLP generated lifetime: one sample is used to study the signal throughout the analysis, while the other sample (with the alternative lifetime assumption but exactly the same mass choices) is used to validate the procedure for extrapolating limits to different mean proper lifetimes of the long-lived scalar.

The dominant SM background in this analysis is SM multijet production. Although a data-driven method is used to perform the background estimation, Monte Carlo (MC) simulated multijet events are needed to train the per-jet neural network to discriminate between signal jets and the multijet background, and to evaluate some of the systematic uncertainties. The samples were generated at LO with Pythia 8.186 \([51]\) using the A14 tune for parton showering and hadronisation. The NNPDF2.3lo PDF set \([47]\) was used.

The effect of multiple interactions in the same and neighbouring bunch crossings (pile-up) was modelled by overlaying the simulated hard-scattering event with inelastic \(pp\) collision events generated with Pythia 8.186 \([51]\) using the NNPDF2.3lo PDF set and the A3 tune \([52]\).

The detector geometry and response were simulated with Geant4 \([53, 54]\). The standard ATLAS reconstruction software is used for both simulation and collision data.

4 Trigger and event selection

This section describes how the data were collected, processed and selected for the analysis. The main dataset was collected with dedicated CalRatio triggers, and auxiliary datasets are also defined so as to be enriched in background events such as BIB, cosmic rays, or SM dijets. Using a combination of data and simulated samples, per-jet neural networks are trained to distinguish three classes of objects: signal-like jets, BIB-like jets, and non-displaced jets. The neural network training architecture involves an adversarial network, which is used to ensure that differences between data and simulation in the training samples are not exploited. Details are provided in section 4.2. Per-event boosted decision trees (one each for low-\(E_T\) and high-\(E_T\) signals) make use of the per-jet neural network scores as well as several other event-level variables to construct discriminants which separate signal from all sources of background. Some additional requirements are applied to maximise the sensitivity of the final selection. Finally, the data-driven ABCD method, explained in detail in section 5, is used to estimate the remaining background in the final selection. In this method the boosted decision tree score is used as one of two axes which define the ABCD plane.

4.1 Triggers and event preselection

Events are selected by the CalRatio triggers \([45]\), which are designed to identify jets that result from neutral LLPs decaying near the outer radius of the ECal or within the HCal. The combined L1 and HLT selections make use of the main characteristics of jets resulting
Figure 2. Trigger efficiency for simulated signal events as a function of the LLP $p_T$ for (a) one of the low-$E_T$ signal samples and (b) one of the high-$E_T$ signal samples for HLT CalRatio triggers seeded by the high-$E_T$ L1 triggers with $E_T$ thresholds of 60 GeV and 100 GeV and by the two versions of the low-$E_T$ L1 triggers. Only statistical uncertainties are shown.

from the decay of a neutral LLP in the calorimeters: due to the late development of these jets, they appear narrower than standard jets, they have a high fraction of their energy deposited in the HCal, and they are isolated from tracks. Two types of such triggers are used, differing only in the L1 trigger selection.

At L1, two different triggers are used. The high-$E_T$ L1 triggers select narrow jets with energy deposits above certain thresholds (either $E_T > 60$ GeV or $E_T > 100$ GeV) in a $0.2 \times 0.2 (\Delta \eta \times \Delta \phi)$ region of the ECal and HCal combined [55]. The low-$E_T$ L1 trigger makes use of the L1 topological trigger system [56] by accepting events where the largest energy deposit, with total-$E_T$ threshold at 30 GeV, does not geometrically overlap in the $\eta-\phi$ plane with any energy deposits with $E_T > 3$ GeV in the ECal. The veto on ECal deposits ensures a high value of the ratio at L1 of energy deposited in the HCal to energy deposited in the ECal, $E_H/E_{EM} > 9$, rejecting a large portion of the background with typical values of $E_H/E_{EM} < 1$, and allowing the $E_T$ threshold to be kept low. This looser $E_T$ requirement increases the efficiency for lower-$p_T$ LLPs. In its 2016 version, the trigger requires that if there is more than one HCal deposit, the second-most energetic one is also isolated from energy deposits in the ECal. This condition was removed in 2018, which helped to increase the trigger efficiency as seen in figures 2 and 3.

At the HLT, a two-step selection algorithm is applied to CalRatio triggers, regardless of the L1 selection. In the first step, calorimeter deposits are clustered into jets using the anti-$k_t$ algorithm [57, 58] with radius parameter $R = 0.4$. The standard jet-cleaning requirements [59] applied offline in most ATLAS analyses reject jets with high values of $E_H/E_{EM}$, one of the key characteristics of the displaced hadronic jets, and are therefore not included in these triggers. A dedicated cleaning algorithm (referred to as CalRatio jet
cleaning and applied starting in 2016) is applied instead. This algorithm is based on the standard jet cleaning described in ref. [59] but with the requirements on the jet $E_H/E_{EM}$ substituted by a condition to reject any jet with more than 85% of its energy associated with a single calorimeter sampling layer and with an absolute value of its negative energy higher than 4 GeV, calculated as the sum of the energy in all cells with negative energy, caused by electronic and pile-up noise. At least one of the HLT jets passing the CalRatio jet cleaning is required to satisfy $E_T > 30$ GeV, $|\eta| < 2.5$ and $\log_{10}(E_H/E_{EM}) > 1.2$. Jets satisfying these requirements are used to determine $0.8 \times 0.8$ regions in $\Delta \eta \times \Delta \phi$ centred on the jet axis in which to perform charged-particle reconstruction (tracking). Triggering jets are required to have no tracks with $p_T > 2$ GeV within $\Delta R = 0.2$ of the jet axis.

In the second step, events passing the CalRatio triggers are required to pass a BIB-removal algorithm. Events which pass the first step but fail this algorithm are collected to form the BIB dataset. Muons from BIB enter the HCal horizontally and may radiate a photon via bremsstrahlung, generating an energy deposit that may be reconstructed as a signal-like jet. BIB deposits are more likely to be aligned in $\phi$ since they travel parallel to the LHC beam pipe. Furthermore, muons from BIB leave these energy deposits in the calorimeter earlier than particles resulting from a collision in the same bunch crossing. The BIB-removal algorithm applied at the HLT therefore identifies events as containing BIB if the triggering jet has at least four HCal-barrel cells close together in $\phi$ and in the same calorimeter layer but not belonging to the same jet and requiring that every cell’s timing be consistent with that of a BIB deposit [60]. This algorithm has been estimated to have an efficiency of approximately 70% in BIB identification.

**Figure 3.** Trigger efficiency for simulated signal events as a function of the LLP decay position in the $x$–$y$ plane ($L_{xy}$) for LLPs decaying in the barrel ($|\eta| < 1.4$) for (a) one of the low-$E_T$ signal samples and (b) one of the high-$E_T$ signal samples for HLT CalRatio triggers seeded by the high-$E_T$ L1 triggers with $E_T$ thresholds of 60 GeV and 100 GeV and by the two versions of the low-$E_T$ L1 triggers. Only statistical uncertainties are shown.
All events in the main dataset are required to pass at least one of the two CalRatio triggers described in this section. The same HLT selection is also active in empty bunch crossings. These are crossings where protons are absent in both beams and isolated from filled bunches by at least five unfilled bunches on either side. Events in empty bunch crossings that have at least one $0.2 \times 0.2 (\Delta \eta \times \Delta \phi)$ calorimeter energy deposit with $E_T > 30$ GeV at L1, and which pass the two steps of the HLT selection algorithm, are stored in the cosmics dataset.

The dijet dataset is collected using the single-jet L1 trigger with the lowest energy threshold in any given data-taking period, and at the HLT, events are required to have at least one jet with $p_T > 400$ GeV or $p_T > 420$ GeV depending on the data-taking period. Events passing any of the CalRatio triggers described above are vetoed from this selection.

The CalRatio trigger efficiency for simulated signal LLPs is defined as the fraction of jets spatially matched ($\Delta R < 0.2$) to one of the generated LLPs that fire the trigger. The trigger efficiency as a function of triggering LLP particle-level $p_T$ is shown in figure 2 for two signal samples. Only generated LLPs decaying in the calorimeters are considered in this plot. The 60-GeV-high-$E_T$ CalRatio trigger, which is seeded by the high-$E_T$ L1 trigger with an $E_T$ threshold of 60 GeV, provides the highest efficiency (about 80%) to select LLPs with $p_T > 100$ GeV. The 2018 version of the low-$E_T$ CalRatio trigger (seeded by the low-$E_T$ L1 trigger) has a similar efficiency for high-$p_T$ LLPs, although it is slightly lower in some $p_T$ ranges. For LLPs with $p_T < 100$ GeV, both versions of the low-$E_T$ CalRatio trigger are able to recover part of the inefficiency that is present in the high-$E_T$ CalRatio trigger. This is especially relevant in the low-$E_T$ samples, which are mostly populated by low-$p_T$ LLPs.

The trigger efficiency also depends strongly on the LLP decay position, as shown in figure 3 for two samples of simulated signal events. The efficiency as a function of LLP decay length in the $x$–$y$ plane is shown for LLPs decaying in the calorimeter’s barrel ($|\eta| < 1.4$). The selection is most efficient in the HCal for both triggers. The fact that the low-$E_T$ samples contain a large fraction of LLPs with low $p_T$ makes these samples have a lower trigger efficiency. The total per-event trigger efficiency in signal is obtained by applying a logical OR of the CalRatio triggers, according to the conditions and luminosities listed in table 1 on a per-event basis, to the complete signal sample integrated over all LLP transverse momenta and for any decay position. Given that the CalRatio triggers accept events where at least one of the jets fulfils the two-step HLT conditions, this efficiency accounts for events where one or both LLPs decay in the calorimeters. In the HS signal, the total trigger efficiency rises from 1.6% in the sample with $m_\Phi = 125$ GeV and $m_s = 55$ GeV to 28.8% for the $m_\Phi = 1000$ GeV and $m_s = 50$ GeV sample.

For every event, primary vertices (PV) with at least two tracks with $p_T > 500$ MeV are reconstructed. If more than one PV candidate is found, the one with the largest sum of the squares of the transverse momenta of all its associated tracks is chosen as the PV. Clean jets are used at several steps of the event selection. They are required to have $p_T > 40$ GeV, $|\eta| < 2.5$ and to pass the CalRatio jet cleaning. To select events with trackless jets, an additional event-level variable, $\sum \Delta R_{\text{min}}(\text{jet}, \text{tracks})$, is used. The quantity $\Delta R_{\text{min}}(\text{jet}, \text{tracks})$ is defined as the angular distance between the jet axis and the closest track with $p_T > 2$ GeV that is associated with the PV, and $\sum \Delta R_{\text{min}}(\text{jet}, \text{tracks})$
is calculated by summing this distance over all the clean jets with $p_T > 50$ GeV. Events with no displaced decays have a very small value of this variable. Every displaced jet contributing to the sum causes a considerable increase in its value, making this variable a good discriminator between signal and multijet background.

For an event to pass the analysis preselection, it is required to have passed at least one of the CalRatio triggers, to contain a PV and at least two clean jets, and to have $\sum \Delta R_{\text{min}}(\text{jet, tracks}) > 0.5$. After preselection, $\sum \Delta R_{\text{min}}(\text{jet, tracks})$ still has good discrimination power and it is used in the data-driven background estimation described in section 5.

4.2 Displaced-jet-tagging neural network

A neural network (NN) is trained on a per-jet basis to discriminate displaced jets in signal from non-displaced jets and from jets generated by BIB. The architecture chosen for the network is a set of convolutional layers fed into a long short-term memory (LSTM) layer using Keras [61] (version 2.2.4) with TensorFlow [62] (version 1.12.0) as the backend. The displaced-jet-tagging NN is trained on SM multijet MC samples, the BIB dataset and a combination of the simulated HS signal samples. Each signal sample is split into two orthogonal subsets, one used for training and the other for application in the rest of the analysis chain. In the signal samples, only jets matched to MC generated LLPs decaying after passing through the ATLAS inner detector are used in the training. In the SM multijet samples, the two highest-$p_T$ clean jets in the event are used in the training. In the BIB dataset, only the triggering jet is used. About 1 million jets of each class are obtained, with an 80:10:10 training, testing, validation split. Signal jets with small boosts are in general harder to discriminate from BIB and SM multijets than those with large boosts. In order to obtain the optimal training for all boost regimes, two separate NNs are trained. The low-$E_T$ NN is trained using the HS signal samples with $m_{\Phi} = 60, 125, 200$ GeV, which typically contain low-$p_T$ signal jets with small boosts. The high-$E_T$ NN is trained on the HS samples with $m_{\Phi} = 400, 600, 1000$ GeV, dominated by high-$p_T$ jets.

The input variables are low-level features associated with each jet, using information from the entirety of the ATLAS detector: positions, momenta, impact parameter and quality variables of inner-detector tracks which are within $\Delta R = 0.2$ of jets; momenta, timing information and positions of calorimeter topoclusters (collections of calorimeter cells used in jet reconstruction [63]) associated with each jet, as well as the fraction of the jet energy deposited in each layer of the ECal and HCal; and spatial and timing information for muon track segments within $\Delta \phi = 0.2$ of a jet. In addition, the momenta and positions of the jets themselves are also used. Overall, the input to the neural network consists of 10 track variables for up to 20 tracks per jet, 12 topocluster variables for up to 30 topoclusters per jet, 6 muon-segment variables for up to 30 muon segments per jet, and 3 jet variables. To check the modelling of variables prior to training, simulated distributions of the per-jet NN inputs are compared with data and fair agreement is generally observed, although some variables display some mismodelling; this is addressed during training by using the adversarial technique described below.
The displaced-jet-tagger uses a combination of two specialised network architectures. The first one is a one-dimensional (1D) convolutional neural network (CNN) which acts as a feature extractor. A convolutional layer performs a series of convolution operations on a 1D array of inputs, corresponding to the variables per subdetector listed above, separately for the tracker, calorimeter and muon-segment variables. A kernel size of 1 is used for the 1D convolutions, with a series of filters being applied which up-sample the features per object in the first filter, then down-sample as the filter goes down in size, working as a feature extractor. In this way the network can exploit the correlations between the input variables.

The second step is the LSTM [64] network, where the memory uses relations between subsequent inputs (for example, between subsequent topoclusters) which allows the network to exploit the correlations between the different topoclusters, tracks and muon segments separately. The topocluster and track inputs to the LSTM are ordered in descending $p_T$, which was shown to enhance the performance of the LSTM. No ordering of the muon segments is applied, but up to 30 muon segments per jet are included which is enough to contain all muon segments for the vast majority of input jets.

The outputs of the individual networks learning from the different subdetectors are concatenated with the jet variables as input to one network which has a number of fully connected layers, named ‘dense layers’. The output of this final, concatenated network is three predictions: signal NN score, SM multijet NN score and BIB NN score. Each prediction is a score between 0 and 1 and it is closer to 1 the more confident the NN is that a jet belongs to this particular class; the three scores add up to 1. Parameters such as learning rate, regularisation and learning nodes per layer are optimised by scanning a multidimensional grid.

Finally, given that the NN is trained on simulation samples as well as on data samples, an adversarial network is added to the output of the jet-tagging NN with the objective of mitigating the dependence of the displaced-jet-tagger outputs on potential mismodelling of any of the input variables in simulation. The adversarial network concept is based on ref. [65] but built specifically for the input variable mismodelling circumstances of this search. An illustration of the architecture is shown in figure 4. The adversarial network is trained on a selection of SM multijet MC and collision data events in a dijet control region separate from the signal selection used to train the principal NN. This dijet control region is constructed from events in the dijet dataset which pass the following requirements: the leading and subleading jets must pass standard quality selections and have $p_T > 400 \text{ GeV}$ and $p_T > 60 \text{ GeV}$ respectively; they must have high angular separation ($|\Delta \phi| > 1.6$); and the magnitude of the vectorial sum of transverse momenta of these jets ($H_T^{\text{miss}}$) is required to be below $120 \text{ GeV}$.

The NN learns the differences between simulation and real data and feeds this information back to the jet-tagging neural network such that the latter network avoids taking advantage of the MC mismodelling. The whole network tries to minimise the loss function. This is achieved by simultaneously minimising the principal network’s loss function (learning to discriminate Signal from BIB and SM multijets) and maximising the adversary’s loss function (learning not to discriminate between simulation and data). During training the cross-entropy function $H(\hat{x})$ is used, combining the cross-entropy of the main network...
Figure 4. A simplified diagram of how the adversarial network is structured. The principal network which discriminates between signal, SM multijets and BIB is shown with the inputs moving through the network as black arrows. Dense layers refer to fully-connected layers, while LSTM refers to Long-Short Term Memory layers described in the text. The adversary which classifies jets as being MC simulation or data is shown with the inputs moving through the network as orange arrows. The input variables are labelled $n \times m$ where $n$ is the maximum of the number of tracks, topoclusters or muon segments per jet and $m$ is the number of distinct variables per input object.

where $\vec{x}$ is the signal, multijet and BIB jets and the cross-entropy of the adversarial network where $\vec{x}$ is the input variables of the control region MC simulation and data jets. The adversary part is modulated by a scale factor $\lambda$, with a value of $\lambda = 10$ chosen to minimize the impact of mismodelling in the final NN output while maintaining good separation. The adversarial cross-entropy term is subtracted from the main cross-entropy term to give the total loss function. The adversary greatly improves the level of agreement between data and SM multijet simulation in the jet-tagging NN outputs, reducing the impact of any potential mismodeling of the NN input variables along the analysis chain. Figure 5 shows the BIB NN score in the dijet control region in data and in simulation, before and after including the adversarial network in the training. The remaining small discrepancies in the modelling of the NN output variables are covered by a dedicated uncertainty, which is described in section 6 and applied to the signal efficiency only, as the background estimate in this analysis is data-driven.

Over-fitting during training is monitored by checking the loss in the training and testing sets periodically. No over-training is seen, which would be characterised as a large deviation between training and testing loss.

The performance of the final trained algorithm can be seen in figure 6. The training took between 50 and 100 run-throughs of the input data to converge, depending on the training. In the event selection, both NNs (low- and high-$E_T$) are evaluated in every clean
Figure 5. The NN output scores in the dijet control region for (a) the low-$E_T$ training with no adversary network and (b) with an adversary network included, (c) for the high-$E_T$ training with no adversary network and (d) with an adversary network included. Statistical uncertainties are shown in all plots. In cases where training with adversary networks is considered, the systematic uncertainty related to modelling discrepancies is included as well.

jet of every signal sample. This provides a better selection than using only the low-$E_T$ (high-$E_T$) NN in the low-mass (high-mass) signal samples because of the residual presence of high-$p_T$ jets in the low-mass samples and low-$p_T$ jets in the high-mass samples.

4.3 Event selection

A per-event boosted decision tree (BDT) is designed with the foremost objective of discriminating BIB events from signal events where at least two jets have a considerable displacement. As explained at the beginning of this section, the CalRatio triggers contain an algorithm to reject BIB. However, this algorithm is not 100% effective, and hence some BIB contamination is expected in the main dataset. The BIB data sample contains SM multijet events on top of the BIB event that caused them to be selected by the trigger. Consequently, even if no multijet sample is explicitly used in the training, the per-event BDT is able to discriminate signal from BIB as well as from multijet background. A combination of signal
Figure 6. The NN output scores for (a,b) signal, (c,d) BIB and (e,f) SM multijets shown for low-$E_T$ and high-$E_T$ training. The jets used here are part of the validation set, and follow the selection used when choosing jets to train the NN in signal, multijets and BIB.
samples in the HS model is used to train the per-event BDT against BIB data. As in the NN training, only part of each signal sample is used for training, leaving the other part for interpretation.

Four CalRatio jet candidates are defined as the two clean jets with the highest low-\(E_T\) signal-NN-score and the two clean jets with the highest high-\(E_T\) signal-NN-score in the event (\(\text{jet}^{\text{sig}1}, \text{jet}^{\text{sig}2}, \text{jet}^{\text{sig}1b}, \text{jet}^{\text{sig}2b}\)). Analogously, four BIB jet candidates are defined as the two clean jets with the highest low-\(E_T\) BIB-NN-score and the two clean jets with the highest high-\(E_T\) BIB-NN-score in the event (\(\text{jet}^{\text{BIB}1}, \text{jet}^{\text{BIB}2}, \text{jet}^{\text{BIB}1b}, \text{jet}^{\text{BIB}2b}\)). The NN outputs of these jets are used as input variables for the per-event BDTs. Other event-level variables are used in the training: examples include the distance \(\Delta R\) between the two high-\(E_T\) CalRatio jet candidates (\(\text{jet}^{\text{sig}1}, \text{jet}^{\text{sig}2}\)) and \(H_{\text{miss}}/H_T\), where \(H_T\) is the scalar sum of the transverse momenta of the jets.

Following the same strategy as used for the per-jet NN and to obtain optimal signal-to-background discrimination at any jet \(p_T\), two versions of the per-event BDT are trained: the high-\(E_T\) BDT is trained for the analysis of the high-mass (\(m_\Phi = 400\) to 1000 GeV) signal samples, with a high jet-\(p_T\) distribution, and the low-\(E_T\) BDT is trained for low-mass (\(m_\Phi = 125\) to 200 GeV) signal samples, with a softer jet-\(p_T\) distribution. The two versions use the same set of input variables and the same BIB background sample. They differ only in the signal samples used during training. Only signal events where at least one of the LLPs decays after passing through the inner detector are considered in the training. Figure 7 shows the distribution of the per-event BDT outputs from five signal samples, as well as from the main data and BIB data. The BDT output in the BIB dataset peaks at slightly lower values than in the main data. Using the jet’s time, measured as the energy-weighted average of the timing for each calorimeter cell related to the jet and corrected by the corresponding time-of-flight from the interaction point, and \(z\)-coordinate measurements of the jets, it was checked that events with low BDT values (low-\(E_T\) BDT \(< -0.2\) or high-\(E_T\) BDT \(< -0.35\)) have the typical characteristics of BIB, while events with intermediate values (\(-0.2 < \text{low-} E_T \text{ BDT} < 0.05\) or \(-0.35 < \text{high-} E_T \text{ BDT} < 0.0\)) are most like multijet events.

Taking this separation between BIB and multijets into account, the per-event BDT serves two purposes in the analysis: first, it is used as part of the event cleaning described below in this section to reject BIB; second, it is used in the data-driven background estimation in the search region (see section 5).

The simulated distributions of the per-jet NN and other inputs to the per-event BDTs show mostly good agreement with data. A dedicated procedure to evaluate the impact of residual mismodelling of NN and BDT input variables on the signal efficiency is described in section 6.

Two selections are defined, referred to as the high-\(E_T\) selection and the low-\(E_T\) selection, which are optimised to give maximum sensitivity for high-\(E_T\) models and low-\(E_T\) models, respectively.

Event-cleaning selections are applied to remove as much BIB and detector-related background as possible: the per-event BDT output must satisfy low-\(E_T\) BDT \(> 0.05\) and high-\(E_T\) BDT \(> 0.0\) in the low-\(E_T\) and high-\(E_T\) selections, respectively; trigger matching, where at least one of the CalRatio jet candidates must be matched to the jet that fired the
trigger and fulfil the requirements applied at trigger level; a timing window of $-3 < t < 15$ ns, where $t$ is the jet time for any of the CalRatio jet candidates or BIB jet candidates (this selection helps to remove remaining BIB jets and jet candidates produced by detector noise while retaining signal jets originating from relatively slowly moving LLPs with timing up to 15 ns); a veto on events where one of the CalRatio jet candidates falls in the transition region between the ECal barrel and endcaps ($1.45 < |\eta| < 1.55$), where poor coverage by the electromagnetic calorimeter produces jets with artificially low fractions of energy in the ECal; and a veto on events where one of the CalRatio jet candidates or BIB jet candidates has $\log_{10}(E_H/E_{EM}) < -1.5$. These requirements ensure that in the final selection, the only remaining source of background is multijet events, as described in section 5.

The final selections are optimised to maximise the signal-to-background ratio in each search region. Variables with good signal-to-background discrimination at event level are used, such as $H_{miss}^T/H_T$ and the product of the NN signal scores of the two relevant CalRatio jet candidates. The quantity $H_{miss}^T/H_T$ has high values for BIB events but it has a softer distribution for signal. Hence it is a good discriminator, especially in the high-$p_T$ regime. Making a selection based on the product of the NN signal scores of the two relevant CalRatio jet candidates helps to further reduce the background while keeping most signal events where one of the LLPs decays in the ATLAS calorimeters and the other one decays between the IP and the outer edge of the calorimeters. The low-$E_T$ NN product is defined as the product of the two highest low-$E_T$ NN signal-scores for clean jets in a given event. The high-$E_T$ NN product is defined analogously. The requirements shown in table 2 are applied for the low-$E_T$ and high-$E_T$ selections.

The signal efficiency for signal events to pass the low- and high-$E_T$ selections depends on the momenta, decay positions and decay products of the two LLPs in the event. For
the benchmark HS models which are analysed using the low-$E_T$ selection, the efficiencies range between 0.5% and 0.005%. Benchmark models analysed using the high-$E_T$ selection have signal efficiencies varying between nearly 9.3% and 1.3%. The efficiencies typically increase with the mean LLP $p_T$ in the sample. Further, the efficiencies are highest for samples where most decays take place in the calorimeter. Finally, the efficiency depends on what particles the LLPs decay to. LLP decays to bottom quarks are typically dominant in the samples considered by this analysis. Hence, the efficiency for events where the LLPs decay to $b$-quarks lead to similar efficiencies to the ranges mentioned above. Events where LLPs decay to $c$-quarks also have a similar efficiency. LLP decays to pairs of taus typically have an efficiency around half of the nominal efficiency. Finally samples where the LLPs decay to top quarks lead to a reduction in efficiency by about an order of magnitude. The signal efficiencies parameterized as a function of kinematic quantities and the decay mode are available in HEPData [66].

5 Background estimation

A data-driven ABCD method is used to estimate the contribution from the dominant background (SM multijet events) to the final selection. The ABCD method relies on the assumption that the distribution of background events can be factorised in the plane of two relatively statistically independent variables. In this plane, the method uses three control regions (B, C and D) to estimate the contribution of background events in the search region (A). In the case of no signal contamination in regions B, C and D, the number of background events in region A can be predicted using $N_A = (N_B \cdot N_C)/N_D$, where $N_X$ is the number of background events in region $X$. In reality, there is non-zero signal contamination in the control regions. This is accounted for by using a modified ABCD method, which involves fitting to background and signal models simultaneously. The background component of the yield in each of the regions A, B, C and D is constrained to obey the standard ABCD relation, within the bounds of the ABCD method uncertainty (described below), while a signal strength parameter uniformly scales the signal yield in each region.

Events passing the high-$E_T$ and low-$E_T$ selections defined in the previous section are divided into four subregions according to two variables: $\sum \Delta R_{\text{min}}(\text{jet, tracks})$ and high-$E_T$ BDT or low-$E_T$ BDT, depending on the selection. The variables are uncorrelated (Pearson

<table>
<thead>
<tr>
<th>Low-$E_T$ selection</th>
<th>High-$E_T$ selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_T^{\text{miss}}/H_T &lt; 0.6$</td>
<td>$H_T^{\text{miss}}/H_T &lt; 0.6$</td>
</tr>
<tr>
<td>$\left(\sum_{\text{jet}^{\text{sig}1,\text{jet}^{\text{sig}2}}} \log_{10}(E_T^{E_H}/E_T^{E_{EM}})\right) &gt; 2$</td>
<td>$\left(\sum_{\text{jet}^{\text{sig}1,\text{jet}^{\text{sig}2}}} \log_{10}(E_T^{E_H}/E_T^{E_{EM}})\right) &gt; 1$</td>
</tr>
<tr>
<td>$p_T^{(\text{jet}^{\text{sig}1})} &gt; 80 \text{ GeV}$</td>
<td>$p_T^{(\text{jet}^{\text{sig}1})} &gt; 70 \text{ GeV}$</td>
</tr>
<tr>
<td>$p_T^{(\text{jet}^{\text{sig}2})} &gt; 80 \text{ GeV}$</td>
<td>$p_T^{(\text{jet}^{\text{sig}2})} &gt; 80 \text{ GeV}$</td>
</tr>
<tr>
<td>low-$E_T$ NN product $&gt; 0.7$</td>
<td>high-$E_T$ NN product $&gt; 0.5$</td>
</tr>
</tbody>
</table>

Table 2. Final selection requirements in the low-$E_T$ and high-$E_T$ selections.
In the ABCD plane must be negligible after the object and event selection requirements, leaving only the contribution of SM multijet events to be estimated by the ABCD method. It is therefore necessary to make sure that the selection above has a high rejection power for two of the three sources. Specifically, the contribution from BIB and cosmics in the ABCD plane must be negligible after the object and event selection requirements, leaving only the contribution of SM multijet events to be estimated by the ABCD method.

Two checks were performed to confirm that the contribution of background events from non-collision background is negligible after the selection.

Figure 8. The distributions of $\sum \Delta R_{\text{min}}$ (jet, tracks) vs high-$E_T$ BDT for (a) BIB events, (b) main data and (c) a signal sample after event cleaning for the high-$E_T$ selection. Panels (d,e,f) show the equivalent distributions for the low-$E_T$ selection. The signal sample with $m_\Phi = 600$ GeV and $m_s = 150$ GeV is shown for the high-$E_T$ selection, while the sample with $m_\Phi = 200$ GeV and $m_s = 50$ GeV is shown for the low-$E_T$ selection.

correlation coefficient $|r| < 0.03$ in the main dataset after the event cleaning, with additional tests for correlations described below) and have good separation between signal and multijet background, as shown in figure 8.

Region A is defined by high-$E_T$ BDT $\geq 0.36$ and $\sum \Delta R_{\text{min}} \geq 1.5$ for the high-$E_T$ analysis and by low-$E_T$ BDT $\geq 0.27$ and $\sum \Delta R_{\text{min}} \geq 1.0$ for the low-$E_T$ analysis. Regions B, C, and D are obtained by reversing one or both of these selections.

In cases where more than one background population is included in the final selection, the condition that the two variables defining the plane are statistically independent is only guaranteed if their contributions have the same shape in the ABCD plane. In this search there are three major sources of background (BIB, cosmics and SM multijets) with different distributions in the plane, which results in the two variables defining it having some correlation. It is therefore necessary to make sure that the selection above has a high rejection power for two of the three sources. Specifically, the contribution from BIB and cosmics in the ABCD plane must be negligible after the object and event selection requirements, leaving only the contribution of SM multijet events to be estimated by the ABCD method.
First, the number of events satisfying each stage of the selection for the main dataset and the BIB dataset is shown in table 3 for the high-$E_T$ and low-$E_T$ selections, along with the fraction of signal events passing each cut for several benchmark samples. For both the high-$E_T$ and low-$E_T$ selections, the number of BIB events satisfying all selection criteria is well within the uncertainty in the number of events passing all selections in the main dataset. Considering that the efficiency of BIB identification in this dataset is approximately 70%, the BIB contamination in the main dataset can be considered negligible. Furthermore, the events from the BIB dataset that pass the selection were checked and found to display properties of multijet events. In particular, their $\phi$ and $z$ vs time distributions do not show the typical shape of BIB. The events from the main dataset that pass the event cleaning were also checked and were found not to display the properties of BIB.

The second check is to confirm that the contamination from cosmic rays in the ABCD plane is negligible. This is done using the cosmics dataset. The number of events in this dataset passing the full selections is checked after weighting by the following two factors. The first factor takes into account the difference between the number of bunch crossings where protons are present in both beams while the CalRatio triggers were enabled, and the number of empty bunch crossings, during which the cosmics dataset was collected. The protons-to-empty live-time ratio depends on the beam conditions, and its value lies in the range 2 to 3.5. The second factor takes into account the fact that events in the cosmics dataset will have no related collision activity, and therefore will be largely trackless: events in the cosmics dataset will be far more likely than events from the main dataset to pass the requirements on jet $\Delta R_{\text{min}}$ at the HLT and in the analysis preselection. The second factor is calculated as the ratio of the number of events entering the ABCD plane in the main dataset to the number passing the selection if all tracks in the event are ignored. This factor is $<0.1$ in all data-taking periods. The final estimated number of events in the ABCD plane is $2.5 \pm 0.8$ (none of which are in region A) in the high-$E_T$ selection and $1.8 \pm 0.5$ ($0.3 \pm 0.2$ in region A) in the low-$E_T$ selection. These yields are negligible in comparison with the expected number of data events in the ABCD plane in the main dataset, which is over 300 in both selections.

The validity of the ABCD method is tested by applying it to a number of validation regions (VRs), which are orthogonal to region A. A first set of validation regions is defined using the nominal event selections but looking only into part of regions B, C and D. Restricting the VR ABCD plane to intermediate values of the BDT output (using part of nominal regions C and D) allows a test of the background estimation method in the whole $\sum \Delta R_{\text{min}}$ range in the absence of signal contamination. Likewise, restricting the VR ABCD plane to low values of $\sum \Delta R_{\text{min}}$ and allowing any BDT output value (using part of nominal regions B and D) permits a test of the method at large values of the BDT output. As an example, the nominal high-$E_T$ event selection is validated in region VRCD$_{\text{high-ET}}$, defined using the nominal high-$E_T$ selection but restricted to the range $0.0 < \text{high-}E_T \text{BDT} < 0.3$ and $0.5 < \sum \Delta R_{\text{min}} < 5.0$. First, the ABCD method is tested by dividing this region into four subregions defined by the boundaries of high-$E_T$ BDT = 0.15 and $\sum \Delta R_{\text{min}} = x$, where the value of $x$ is varied from 1 to 4. Then the ABCD method is tested again in this same region by setting the $\sum \Delta R_{\text{min}}(\text{jet, tracks})$ boundary at 1.5 and allowing the high-$E_T$ BDT boundary to take values from 0.05 to 0.25.
<table>
<thead>
<tr>
<th>Region</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>22</td>
<td>7</td>
<td>233</td>
<td>131</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>0.63</td>
<td>1.1</td>
<td>0.10</td>
</tr>
<tr>
<td>G</td>
<td>0.93</td>
<td>11.6</td>
<td>1.4</td>
<td>0.14</td>
</tr>
<tr>
<td>H</td>
<td>0.09</td>
<td>0.54</td>
<td>1.2</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 3. Sequential impact of each requirement on the number of events passing the selection for the high-\(E_T\) (top table) and low-\(E_T\) (bottom table) selections. The signal columns represent the cumulative fraction of events passing the selection than the number of events.

The ABCD plane defined in this VR can be seen in figure 9 for VRCD_{high-\(E_T\)} and the low-\(E_T\) selection, VRCD_{low-\(E_T\)}, along with the level of agreement between the expected and observed numbers of events in region A in all the tested boundary selections. It should be noted that the tests in each VR are statistically correlated. Therefore, statistical fluctuations can affect several tests.

A second set of VRs is defined by inverting the selection on the product of the signal NN scores of the two CalRatio jets in the analysis selections.
Figure 9. (a) The per-event BDT vs $\sum \Delta R_{\text{min}}(\text{jet, tracks})$ distribution for the main dataset in the VRCD_{high-}E_T validation region. Region A in the nominal high-}E_T selection is shown for reference. Also shown is a comparison of the observed and expected events in region A of this VR. The x-axis shows the different (b) $\sum \Delta R_{\text{min}}(\text{jet, tracks})$ and (c) BDT boundaries used to test the ABCD method. Statistical uncertainties are shown in (b) and (c). Panels (d,e,f) show the equivalent distributions in the VRCD_{low-}E_T validation region.

Given that signal contamination is low in all the validation regions, the plain ABCD method assuming no signal is applied. The statistical precision of this closure test is better than the final statistical uncertainty of the number of events observed in region A. Good closure, within one standard deviation, is observed in all the validation regions defined in this section.

6 Systematic uncertainties

The uncertainty in the data-driven ABCD method for the background estimate is studied in the dijet control region. In this control region, an alternative ABCD plane is defined using the same variables as in the analysis, but adjusting the boundaries in regions A, B, C and D to reduce the effect of statistical fluctuations in the estimation of the number of dijet events in region A by this method. The observed number of events in region A is compared with the estimate given by the ABCD method for both the high-}E_T and the low-}E_T ABCD planes. In all the tests performed, the observed and expected numbers of events agree within the statistical uncertainties. This confirms that the two variables
forming the plane are statistically independent and therefore no systematic uncertainty is added for the background estimation method.

The uncertainty in the combined 2015–2018 integrated luminosity is 1.7% [67], obtained using the LUCID-2 detector [68] for the primary luminosity measurements.

Events in MC simulation are reweighted to obtain the correct pile-up distribution. The uncertainty in the pile-up reweighting of the reconstructed events in the MC simulation is estimated by comparing the distribution of the number of primary vertices in the MC simulation with the distribution in data as a function of the instantaneous luminosity. Differences between these distributions are reduced by scaling the mean number of \(pp\) interactions per bunch crossing in the MC simulation and the \(\pm 1\sigma\) uncertainties are assigned to these scaling factors [67, 68]. The effect on the signal event yields varies between 2% and 8% depending on the signal model.

The jet energy scale and jet energy resolution introduce uncertainties in the signal yield of <1% to 5% and 1% to 6%, respectively, depending on the signal model. These uncertainties are calculated using the procedure detailed in ref. [69]. Since the jets used in this analysis are required to have a low fraction of calorimeter energy in the ECal, the jet energy uncertainties are rederived as a function of ECal energy fraction as well as \(\eta\). The additional jet energy uncertainties are found to have an effect of 1% to 6% on the signal yield, and are conservatively taken in quadrature with the regular jet-energy uncertainties. The low-\(E_T\) models are more sensitive to all jet energy uncertainties than the high-\(E_T\) models.

The uncertainty in the signal trigger efficiency is estimated by studying how well modelled the three most important trigger variables (jet \(E_T\), \(\log_{10}(E_H/E_{EM})\), and \(p_T\) of tracks within the jet) are between HLT-reconstructed quantities and offline-reconstructed quantities in data and MC simulation. A tag-and-probe technique is applied to a pure sample of multijet events obtained using standard jet triggers in both data and MC simulation. Scale factors that represent the degree of mismodelling in each variable are derived and then applied in an emulation of the CalRatio triggers. The change in yield relative to the nominal (unscaled) trigger emulation after the full analysis selection is taken as the size of the systematic uncertainty, which is between 1% and 7% depending on the signal model.

A systematic uncertainty is included to account for potential mismodelling of input variables used in the machine-learning techniques applied in the analysis. Using the same control sample of dijet events defined for the evaluation of the systematic uncertainty in the data-driven background estimate, the distributions of the inputs and outputs of the per-jet NNs and the per-event BDTs are studied. They are found to agree fairly well between data and MC simulation. The residual differences are translated into a systematic uncertainty in the signal efficiency, using the following procedure. For each mis-modeled variable, the residual differences are quantified through a transfer factor between simulation and data. In an ensemble of pseudo-experiments, the NN and BDT input variables for each signal event are varied by this transfer factor; in each pseudo-experiment the transfer factor is modulated by a random Gaussian, with a mean of zero and a width determined by measurements in the control region. The final per-event BDTs are then re-evaluated, and the overall signal
efficiency of the sample is evaluated for each pseudo-experiment. This sequence of steps allows for the statistical determination of the systematic uncertainty in the final signal efficiency. The value of the resulting uncertainty can thus be obtained from the distributions of efficiencies for the ensemble, and can be as large as 6% depending on the signal model.

Finally, an uncertainty due to the NLO-reweighting of the signal samples is obtained by comparing the NLO MadGraph predictions for a 125 GeV Higgs boson mediator with predictions at next-to-next-to-leading-order accuracy in QCD from Powheg Box v2 [70–74]. This results in an additional signal efficiency uncertainty of 1%–7% for most samples.

7 Results and statistical interpretation

A data-driven background estimation and signal hypothesis test is performed simultaneously in all regions. The procedure for the simultaneous fit is explained in detail in the previous iteration of this analysis [37] and is summarised in the following. A profile likelihood function is constructed from the product of the probabilities of observing a given number of events in each region of the ABCD plane, given the expected number of events in that region. This expected number of events is given by the sum of the predicted signal yield in that region (scaled by a parameter of interest called the signal strength) and the expected number of background events. The background component of the expected yield in each region is constrained to satisfy the ABCD relation introduced in section 5. The introduction of a signal component will therefore dynamically modify the remaining allowed background prediction in this set-up. Additionally, a nuisance parameter which represents the total uncertainty in the signal efficiency, and consequently the signal yield, is introduced.

The background estimates before (a priori) and after (a posteriori) unblinding region A are shown in table 4. After unblinding, the modified ABCD method takes account of the observed number of events in region A when making the final fit. As a consequence, the background estimate changes relative to the a priori case, since the fit now uses all available information to determine the expected background in each region. As shown in the table, there is a slight excess in the observed number of events in region A over the a priori expected background in both the high- and low-\(E_T\) selections. In the a priori fit, these excesses correspond to background-only hypothesis p-values of 0.083 and 0.076 in the high-\(E_T\) and low-\(E_T\) selections, respectively. The significance is reduced to slightly under one standard deviation in the a posteriori (background-only fit) estimate.

The CLs method [75], using the “alternative test statistic” \(\tilde{q}\) [76] is used to set upper limits on the production cross-section times BR of the LLP signals considered. The pyhf [77, 78] framework is used to implement the likelihood function and extract the upper limits. An asymptotic approximation [76] is used for these results. This approximation was tested against the full frequentist pseudo-experiment-based method for a variety of signal samples and was found to give consistent limits. Since each signal sample was generated for a given lifetime assumption, it is necessary to extrapolate the limits across lifetimes. This extrapolation is performed using a reweighting method, which is described in detail in the previous iteration of this analysis [37] and summarised briefly here. LLPs follow an exponential decay distribution, where the decay constant is given by the mean lifetime \(\tau_{\text{gen}}\).
<table>
<thead>
<tr>
<th>High-$E_T$ selection</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed data</td>
<td>22</td>
<td>7</td>
<td>233</td>
<td>131</td>
</tr>
<tr>
<td>a priori</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated background</td>
<td>12.4 ± 4.7</td>
<td>7 ± 2.6</td>
<td>233 ± 15</td>
<td>131 ± 11</td>
</tr>
<tr>
<td>a posteriori (background-only fit)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fitted background</td>
<td>18.8 ± 3.5</td>
<td>10.2 ± 3.2</td>
<td>236 ± 15</td>
<td>128 ± 11</td>
</tr>
<tr>
<td>a posteriori (signal-plus-background fit)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fitted background</td>
<td>10.0 ± 6.0</td>
<td>5.7 ± 2.4</td>
<td>230 ± 15</td>
<td>131 ± 11</td>
</tr>
<tr>
<td>Fitted signal</td>
<td>12.2 ± 8.7</td>
<td>1.4 ± 1.0</td>
<td>3.4 ± 2.5</td>
<td>&lt; 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Low-$E_T$ selection</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed data</td>
<td>23</td>
<td>3</td>
<td>220</td>
<td>61</td>
</tr>
<tr>
<td>a priori</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimated background</td>
<td>10.8 ± 6.6</td>
<td>3 ± 1.7</td>
<td>220 ± 15</td>
<td>61 ± 7.8</td>
</tr>
<tr>
<td>a posteriori (background-only fit)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fitted background</td>
<td>20.6 ± 4.0</td>
<td>5.4 ± 2.3</td>
<td>222 ± 15</td>
<td>59 ± 7.7</td>
</tr>
<tr>
<td>a posteriori (signal-plus-background fit)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fitted background</td>
<td>8.4 ± 7.7</td>
<td>2.4 ± 1.5</td>
<td>217 ± 15</td>
<td>61 ± 7.8</td>
</tr>
<tr>
<td>Fitted signal</td>
<td>14.6 ± 9.9</td>
<td>&lt; 1</td>
<td>3.2 ± 2.2</td>
<td>&lt; 1</td>
</tr>
</tbody>
</table>

Table 4. Application of the modified ABCD method to the final high-$E_T$ and low-$E_T$ selections. The a priori estimate refers to the ‘pre-unblinding’ case, where the data in region A are ignored by removing the Poisson constraint in that region and the signal strength is fixed to zero. This matches the simple $N_{A}^{bkg} = (N_{B}^{bkg} \cdot N_{C}^{bkg})/N_{D}^{bkg}$ relation. The a posteriori estimate refers to the ‘post-unblinding’ case, including the observed data in region A in the background-only global fit, obtained by fixing the signal strength to 0 (background-only fit) or allowing it to float (signal-plus-background fit). The table also shows one set of representative signal yields in each selection for the signal-plus-background fit. Only statistical uncertainties are included in the quoted error of the background, while the uncertainties in the signal include those from both statistical and experimental sources.

in a given sample. To extrapolate to a different mean proper lifetime $\tau_{\text{new}}$, a weight $w$ is calculated as a function of the given particle’s proper decay time $t$:

$$w(t) = \frac{\tau_{\text{gen}}}{\exp(-t/\tau_{\text{gen}})} \cdot \frac{\exp(-t/\tau_{\text{new}})}{\tau_{\text{new}}}.$$  

The per-event weight is the product of the weights obtained for each LLP in a given event.

Figure 10 shows a summary of the observed limits for the samples where the mediator is assumed to be the SM Higgs boson, with mass 125 GeV. SM Higgs boson BRs to long-lived neutral scalars above 10% are excluded for $c\tau$ values between a few centimetres and about 20 metres, depending on the mass of the LLP. Figure 11 shows a summary of observed limits
on the cross-section times BR for other models, where the mediator is heavier or lighter than the SM Higgs boson. The excluded ranges for each model are summarised in table 5. For mediators of mass 60 GeV, cross-sections above 100 pb are excluded between about 30 mm and 5 m, depending on the LLP mass. For mediators heavier than the SM Higgs boson, cross-section times BR values above 0.1 pb are typically excluded between a few tens of millimetres and up to 20 metres. For high-$E_T$ models where the scalar mass is much smaller than the mediator mass, cross-section times BR values above 0.01 pb can be excluded for $c\tau$ values of a few hundreds of millimetres to a few metres, depending on the model.

The analysis improves on the limits from the previous iteration of the analysis for many models. The improvements are typically between a factor of two and a factor of five. The low-$E_T$ models are the ones for which the limits are most improved. The previous analysis used only a fraction of the 13 TeV dataset: 11 fb$^{-1}$ and 33 fb$^{-1}$ for the analysis of low-$E_T$ and high-$E_T$ models, respectively. Part of the gain is therefore explained by the increase in the integrated luminosity of the data sample. The rest of the gain, particularly for the low-$E_T$ selection, comes from improvements in the low-$E_T$ trigger and in the displaced-jet identification.

<table>
<thead>
<tr>
<th>Model</th>
<th>Excluded $c\tau$ range [m]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$B_{H\rightarrow ss} = 1$</td>
</tr>
<tr>
<td></td>
<td>$\sigma = 48.6 , \text{pb}$</td>
</tr>
<tr>
<td>$m_\Phi = 125 , \text{GeV}, m_s = 5 , \text{GeV}$</td>
<td>0.01–3.47</td>
</tr>
<tr>
<td>$m_\Phi = 125 , \text{GeV}, m_s = 16 , \text{GeV}$</td>
<td>0.06–11.4</td>
</tr>
<tr>
<td>$m_\Phi = 125 , \text{GeV}, m_s = 35 , \text{GeV}$</td>
<td>0.09–24.8</td>
</tr>
<tr>
<td>$m_\Phi = 125 , \text{GeV}, m_s = 55 , \text{GeV}$</td>
<td>0.26–40.5</td>
</tr>
<tr>
<td>$m_\Phi = 60 , \text{GeV}, m_s = 5 , \text{GeV}$</td>
<td>0.02–1.88</td>
</tr>
<tr>
<td>$m_\Phi = 60 , \text{GeV}, m_s = 16 , \text{GeV}$</td>
<td>0.07–12.1</td>
</tr>
<tr>
<td>$m_\Phi = 1000 , \text{GeV}, m_s = 50 , \text{GeV}$</td>
<td>0.04–4.54</td>
</tr>
<tr>
<td>$m_\Phi = 1000 , \text{GeV}, m_s = 275 , \text{GeV}$</td>
<td>0.24–29.8</td>
</tr>
<tr>
<td>$m_\Phi = 1000 , \text{GeV}, m_s = 475 , \text{GeV}$</td>
<td>0.99–18.8</td>
</tr>
<tr>
<td>$m_\Phi = 600 , \text{GeV}, m_s = 50 , \text{GeV}$</td>
<td>0.08–5.52</td>
</tr>
<tr>
<td>$m_\Phi = 600 , \text{GeV}, m_s = 150 , \text{GeV}$</td>
<td>0.23–19.4</td>
</tr>
<tr>
<td>$m_\Phi = 600 , \text{GeV}, m_s = 275 , \text{GeV}$</td>
<td>0.70–24.4</td>
</tr>
<tr>
<td>$m_\Phi = 400 , \text{GeV}, m_s = 100 , \text{GeV}$</td>
<td>0.31–13.7</td>
</tr>
<tr>
<td>$m_\Phi = 200 , \text{GeV}, m_s = 50 , \text{GeV}$</td>
<td>0.63–4.62</td>
</tr>
</tbody>
</table>

Table 5. Summary of the $c\tau$ values which are excluded at 95% CL for the models considered in this analysis.
Figure 10. (a) 95% CL expected and observed limits on the BR of SM Higgs bosons to pairs of neutral LLPs ($B_{H \rightarrow s\bar{s}}$), showing the ±1σ (green) and ±2σ (yellow) expected limit bands, as well as a comparison with the results from previous ATLAS searches [36, 37], and (b) summary of 95% CL expected and observed limits on the BR of a SM Higgs boson mediator to pairs of neutral LLPs considered in this analysis. The cross-section for SM Higgs boson gluon-gluon fusion production is assumed to be $48.6 \text{ pb}$.

Figure 11. Summary of 95% CL expected and observed limits for the hidden sector (HS) models considered in this analysis, where the mediator is not the SM Higgs boson: (a) shows the case where $m_\phi = 60 \text{ GeV}$; (b) shows intermediate masses of $m_\phi = 200 \text{ GeV}$ and $m_\phi = 400 \text{ GeV}$; (c) shows the case where $m_\phi = 600 \text{ GeV}$; (d) shows the highest $m_\phi$ hypothesis with $m_\phi = 1 \text{ TeV}$.
8 Conclusion

In this paper, an analysis targeting pair-produced neutral long-lived particles decaying in the ATLAS calorimeters is presented. The analysis used the full LHC Run 2 dataset of 13 TeV proton-proton collisions, the size of which is $139\, \text{fb}^{-1}$. The analysis is structured in a very similar way to the one described in a previous paper using only the 2016 dataset but has several improvements, including a new method to select the most signal-like displaced jets with the help of a deep neural network trained using an adversary network. No significant excess of events in the signal region is observed relative to the data-driven background prediction. The CL$_s$ method is used to set 95% CL limits on the cross-section times branching ratio as a function of $c$ times the long-lived particle mean proper lifetime. Limits are set on hidden-sector benchmark models with mediator masses ranging from 60 GeV to 1 TeV and long-lived scalar masses ranging from 5 GeV to 475 GeV. The improvements to the analysis and additional data lead to an improvement on the limits for mediator masses above and below 200 GeV by a factor of around 1.5–2 and 3–5, respectively, compared with the previous version of the analysis. For models with a SM Higgs boson mediator, branching ratios to neutral scalars above 10% are excluded for long-lived particle mean proper lifetimes times $c$ between approximately 20 mm and 10 m, depending on the model.

Acknowledgments

We thank CERN for the very successful operation of the LHC, as well as the support staff from our institutions without whom ATLAS could not be operated efficiently.

We acknowledge the support of ANPCyT, Argentina; YerPhI, Armenia; ARC, Australia; BMWFW and FWF, Austria; ANAS, Azerbaijan; SSTC, Belarus; CNPq and FAPESP, Brazil; NSERC, NRC and CFI, Canada; CERN; ANID, Chile; CAS, MOST and NSFC, China; Minciencias, Colombia; MEYS CR, Czech Republic; DNRF and DNSRC, Denmark; IN2P3-CNRS and CEA-DRF/IRFU, France; SRNSFG, Georgia; BMBF, HGF and MPG, Germany; GSRI, Greece; RGC and Hong Kong SAR, China; ISF and Benoziyo Center, Israel; INFN, Italy; MEXT and JSPS, Japan; CNRST, Morocco; NWO, Netherlands; RCN, Norway; MEiN, Poland; FCT, Portugal; MNE/IFA, Romania; JINR; MES of Russia and NRC KI, Russian Federation; MESTD, Serbia; MSSR, Slovakia; ARRS and MIZŠ, Slovenia; DSI/NRF, South Africa; MICINN, 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, CANARIE, Compute Canada and CRC, Canada; COST, ERC, ERDF, Horizon 2020 and Marie Skłodowska-Curie Actions, European Union; Investissements d’Avenir Labex, Investissements d’Avenir Idex and 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; Norwegian Financial Mechanism 2014–2021, Norway; NCN and NAWA, Poland; La Caixa Banking Foundation, CERCA Programme Generalitat de Catalunya and PROMETEO and GenT.
Programmes Generalitat Valenciana, Spain; Göran Gustafsson's Stiftelse, Sweden; 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 (U.K.) and BNL (U.S.A.), the Tier-2 facilities worldwide and large non-WLCG resource providers. Major contributors of computing resources are listed in ref. [79].

Open Access. This article is distributed under the terms of the Creative Commons Attribution License (CC-BY 4.0), which permits any use, distribution and reproduction in any medium, provided the original author(s) and source are credited.

References


[26] D0 collaboration, Search for Resonant Pair Production of long-lived particles decaying to $b\bar{b}$ in $p\bar{p}$ collisions at $\sqrt{s}=1.96$ TeV, Phys. Rev. Lett. 103 (2009) 071801 [arXiv:0906.1787] [nSPIRE].

[27] CDF collaboration, Search for heavy metastable particles decaying to jet pairs in $p\bar{p}$ collisions at $\sqrt{s}=1.96$ TeV, Phys. Rev. D 85 (2012) 012007 [arXiv:1109.3136] [nSPIRE].


[60] ATLAS collaboration, Characterisation and mitigation of beam-induced backgrounds observed in the ATLAS detector during the 2011 proton-proton run, 2013 *JINST* **8** P07004 [arXiv:1303.0223] [insPIRE].


-- 31 --
ATLAS collaboration, Jet energy scale measurements and their systematic uncertainties in
072002 [arXiv:1703.09665] [inSPIRE].

K. Hamilton, P. Nason, E. Re and G. Zanderighi, NNLOPS simulation of Higgs boson

K. Hamilton, P. Nason and G. Zanderighi, Finite quark-mass effects in the NNLOPS

S. Alioli, P. Nason, C. Oleari and E. Re, A general framework for implementing NLO
calculations in shower Monte Carlo programs: the POWHEG BOX, JHEP 06 (2010) 043
[arXiv:1002.2581] [inSPIRE].

P. Nason, A New method for combining NLO QCD with shower Monte Carlo algorithms,

S. Frixione, P. Nason and C. Oleari, Matching NLO QCD computations with Parton Shower

[inSPIRE].

G. Cowan, K. Cranmer, E. Gross and O. Vitells, Asymptotic formulae for likelihood-based tests
[arXiv:1007.1727] [inSPIRE].

[https://github.com/scikit-hep/pyhf/releases/tag/v0.6.3].

L. Heinrich, M. Feickert, G. Stark and K. Cranmer, pyhf: pure-python implementation of
histfactory statistical models, J. Open Source Softw. 6 (2021) 2823.

ATLAS collaboration, ATLAS Computing Acknowledgements, ATL-SOFT-PUB-2021-003
(2021).