Study of the rare decays of $B_s^0$ and $B^0$ into muon pairs from data collected during 2015 and 2016 with the ATLAS detector

The ATLAS Collaboration

A study of the decays $B_s^0 \rightarrow \mu^+\mu^-$ and $B^0 \rightarrow \mu^+\mu^-$ has been performed using 26.3 fb$^{-1}$ of 13 TeV LHC proton–proton collisions collected with the ATLAS detector in 2015 and 2016. For $B_s^0$, the branching fraction $\mathcal{B}(B_s^0 \rightarrow \mu^+\mu^-) = (3.2^{+1.1}_{-1.0}) \times 10^{-9}$ is measured. For the $B^0$, an upper limit on the branching fraction is set at $\mathcal{B}(B^0 \rightarrow \mu^+\mu^-) < 4.3 \times 10^{-10}$ at 95% confidence level. The result is combined with the full Run 1 ATLAS result, yielding $\mathcal{B}(B_s^0 \rightarrow \mu^+\mu^-) = (2.8^{+0.3}_{-0.2}) \times 10^{-9}$ and $\mathcal{B}(B^0 \rightarrow \mu^+\mu^-) < 2.1 \times 10^{-10}$. The combined result is consistent with the Standard Model within 2.4 standard deviations.

© 2018 CERN for the benefit of the ATLAS Collaboration.
Reproduction of this article or parts of it is allowed as specified in the CC-BY-4.0 license.
1 Introduction

Flavour-changing neutral-current (FCNC) processes are highly suppressed in the Standard Model (SM), and their study is relevant to indirect searches for physics beyond the SM. The branching fractions of the decays $B_{(s)}^0 \rightarrow \mu^+\mu^-$ are of particular interest because of the additional helicity suppression and since they are accurately predicted in the SM: $\mathcal{B}(B^0 \rightarrow \mu^+\mu^-) = (3.65 \pm 0.23) \times 10^{-9}$ and $\mathcal{B}(B^0 \rightarrow \mu^+\mu^-) = (1.06 \pm 0.09) \times 10^{-10}$ [1]. Deviations at the level of these values - or even larger - can arise in models involving non-SM heavy particles, such as those predicted in the Minimal Supersymmetric Standard Model [2–6], in extensions such as Minimal Flavour Violation [7, 8], Two-Higgs-Doublet Models [6], and others [9, 10]. The CMS and LHCb collaborations have reported the observation of $B^0_s \rightarrow \mu^+\mu^-$ [11, 12] and evidence for $B^0 \rightarrow \mu^+\mu^-$, with combined values: $\mathcal{B}(B^0_s \rightarrow \mu^+\mu^-) = (2.8^{+0.7}_{-0.6}) \times 10^{-9}$ and $\mathcal{B}(B^0 \rightarrow \mu^+\mu^-) = (3.9^{+1.6}_{-1.4}) \times 10^{-10}$ [13]. The LHCb collaboration updated its Run 1 result with part of the data collected in Run 2, measuring $\mathcal{B}(B^0_s \rightarrow \mu^+\mu^-) = (3.0 \pm 0.6^{+0.3}_{-0.2}) \times 10^{-9}$ (where the first uncertainty is statistical and the second systematic) and $\mathcal{B}(B^0 \rightarrow \mu^+\mu^-) < 3.4 \times 10^{-10}$ at 95% CL [14]. ATLAS has measured, with the Run 1 dataset [15], $\mathcal{B}(B^0_s \rightarrow \mu^+\mu^-) = (0.9^{+1.1}_{-0.8}) \times 10^{-9}$, while setting an upper bound on $B^0 \rightarrow \mu^+\mu^-$ of 4.2 $\times 10^{-10}$ at 95% CL.

This paper reports the result of a search for $B^0_s \rightarrow \mu^+\mu^-$ and $B^0 \rightarrow \mu^+\mu^-$ decays performed using $pp$ collision data corresponding to an integrated luminosity of 36.2 fb$^{-1}$, collected at 13 TeV during the first two years of the LHC Run 2 data-taking period using the ATLAS detector. The analysis strategy follows the approach employed in the previous ATLAS measurement [15], using data collected with one rather than three separate sets of trigger thresholds, standard ATLAS muon selection criteria (Section 4) and an improved statistical treatment of the result (Section 11).

2 Outline

The notation used throughout the paper refers to the combination of processes and their charge-conjugates, unless otherwise specified. The $B^0_s \rightarrow \mu^+\mu^-$ and $B^0 \rightarrow \mu^+\mu^-$ branching fractions are measured relative to the normalisation decay $B^+ \rightarrow J/\psi(\rightarrow \mu^+\mu^-)K^+$ which is abundant and has a well measured branching fraction $\mathcal{B}(B^+ \rightarrow J/\psi K^+) \times \mathcal{B}(J/\psi \rightarrow \mu^+\mu^-)$. The $B^0 \rightarrow \mu^+\mu^-$ ($B^0_s \rightarrow \mu^+\mu^-$) branching fraction can be extracted as:

$$
\mathcal{B}(B^0 \rightarrow \mu^+\mu^-) = \frac{N_{d(s)}}{\epsilon_{\mu^+\mu^-}} \times \left[ \mathcal{B}(B^+ \rightarrow J/\psi K^+) \times \mathcal{B}(J/\psi \rightarrow \mu^+\mu^-) \right] \frac{\epsilon_{J/\psi K^+}}{N_{J/\psi K^+}} \times \frac{f_u}{f_d}
$$

$$
= N_{d(s)} \mathcal{D}_{\text{norm}} \frac{\mathcal{B}(B^+ \rightarrow J/\psi K^+) \times \mathcal{B}(J/\psi \rightarrow \mu^+\mu^-)}{\epsilon_{\mu^+\mu^-}} \times \frac{f_u}{f_d},
$$

(1)

where $N_d$ ($N_s$) is the $B^0 \rightarrow \mu^+\mu^-$ ($B^0_s \rightarrow \mu^+\mu^-$) signal yield, $N_{J/\psi K^+}$ is the $B^+ \rightarrow J/\psi K^+$ normalisation yield, $\epsilon_{\mu^+\mu^-}$ and $\epsilon_{J/\psi K^+}$ are the corresponding values of acceptance times efficiency (measured in fiducial volumes defined in Section 10), and $f_u/f_d$ ($f_u/f_s$) is the ratio of the hadronisation probabilities of a $b$-quark into $B^+$ and $B^0$ ($B^0_s$). The quantity $\mathcal{D}_{\text{norm}} = N_{J/\psi K^+} \frac{\epsilon_{\mu^+\mu^-}}{\epsilon_{J/\psi K^+}}$, where the $\epsilon$ ratio takes into account relative differences in efficiencies, integrated luminosities and the trigger selections used for the signal and the normalisation modes. Signal and reference channel events are selected with similar dimuon triggers. When determining the control channel normalisation, the half of the statistics used to tune the kinematic distributions of simulated events is excluded.
The event selection uses variables related to the candidates decay time, thus introducing a dependence of the efficiency on the signal lifetime. The relation between the measured branching fraction and the corresponding value at production is established assuming the decay time distribution predicted in the SM, where the decay occurs predominantly through the heavy eigenstate $B_s/d_H$ of the $B^0(s)\rightarrow\bar{B}^0(s)$ system. Models for new physics [16, 17] can predict modification to the decay time distribution of $B^0(s)\rightarrow\mu^+\mu^-$ and a comparison with the experimental result may require a correction to the ratio of the time-integrated efficiencies.

The ATLAS inner tracking system, muon spectrometer and, for precise identification of muons, also calorimeters, are used to reconstruct and select the event candidates. Details of the detector, trigger, data sets, and preliminary selection criteria are discussed in Sections 3 and 4. A blind analysis was performed in which data in the dimuon invariant mass region from 5166 to 5526 MeV were removed until the procedures for event selection and the details of signal yield extraction were completely defined. Section 5 introduces the three main categories of background. Section 6 describes the strategy used to reduce the probability of hadron misidentification. The final sample of candidates is selected using a multivariate classifier, designed to enhance the signal relative to the dominant di-muon background component, as discussed in Section 7. Checks on the distributions of the variables used in the multivariate classifier are summarised in Section 8. They are based on the comparison of data and simulation for dimuon events, for $B^+\rightarrow J/\psi K^+$ candidates and for events selected as $B_s^0\rightarrow J/\psi \phi \rightarrow \mu^+\mu^-K^+K^-$, which provide an additional validation of the procedures used in the analysis. Section 9 details the fit procedure to extract the yield of $B^+\rightarrow J/\psi K^+$ events. The ratio of efficiencies in the signal and the normalisation channels is presented in Section 10. Section 11 describes the extraction of the signal yield, obtained with an unbinned maximum-likelihood fit performed on the dimuon invariant mass distribution. In this fit events are separated in classifier intervals to maximise the fit sensitivity. The results for the branching fractions $\mathcal{B}(B^0_s\rightarrow\mu^+\mu^-)$ and $\mathcal{B}(B^0\rightarrow\mu^+\mu^-)$ are reported in Section 12, with the likelihood combination of the present and full Run 1 results reported in Section 13.

3 ATLAS detector, data and simulation samples

The ATLAS detector\(^1\) consists of three main components: an inner detector (ID) tracking system immersed in a 2 T axial magnetic field, surrounded by electromagnetic and hadronic calorimeters and by the muon spectrometer (MS). A full description can be found in Ref. [18], complemented by Refs. [19, 20] for details on the Run 2 upgraded innermost silicon pixel layer.

This analysis is based on the Run 2 data recorded in 2015 and 2016 from $pp$ collisions at the LHC at $\sqrt{s} = 13$ TeV. Data used in the analysis were recorded during stable LHC beam periods. Data quality requirements were imposed, notably on the performance of the MS and ID systems. The total integrated luminosity used in this analysis is 36.2 fb\(^{-1}\). The uncertainty on this integrated luminosity is 2.1\%. It is derived, following a methodology similar to that detailed in Ref. [21], and using the LUCID-2 detector for the baseline luminosity measurements [22], from calibration of the luminosity scale using x-y beam-separation scans.

\(^1\) ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point. The z-axis is along the beam pipe, the x-axis points to the centre of the LHC ring and the y-axis points upward. Cylindrical coordinates $(r, \phi)$ are used in the transverse plane, $r$ being the distance from the origin and $\phi$ being the azimuthal angle around the beam pipe. The pseudorapidity $\eta$ is defined as $\eta = -\ln[\tan(\theta/2)]$ where $\theta$ is the polar angle.
Samples of simulated Monte-Carlo (MC) events are used for training and validation of the multivariate analyses, for the determination of the efficiency ratios, and for guiding the signal extraction fits. Exclusive MC samples were produced for the signal channels $B^0 \rightarrow \mu^+ \mu^-$ and $B^0 \rightarrow \mu^+ \mu^-$, the normalisation channel $B^+ \rightarrow J/\psi K^+$ ($J/\psi \rightarrow \mu^+ \mu^-$), and the control channel $B^0_s \rightarrow J/\psi \phi$ ($\phi \rightarrow K^+ K^-$). In addition, background studies employ MC samples of inclusive semileptonic decays $B \rightarrow \mu X$, samples of $B^0_s \rightarrow K^- \mu^+ \nu$, $B^0 \rightarrow \pi^- \mu^+ \nu$, $\Lambda_b \rightarrow p \mu^- \nu$, $B^{(s)}_c \rightarrow hh'$ decays with $h^{(')}$ being a charged pion or kaon, inclusive decays $B \rightarrow J/\psi X$ as well as the exclusive $B^+ \rightarrow J/\psi \pi^+$ decay.

Most of the dimuon candidates in the data sample originate from the decays of hadrons produced in the hadronisation of $b\bar{b}$ pairs. To describe this background, a $b\bar{b}$ MC sample equivalent to roughly two to three times the luminosity integrated in data was generated requiring the presence of two muons in the final state, originating with any topology from the $b\bar{b}$ pair.

The MC samples were generated with Pythia 8 [23]. The ATLAS detector and its response are simulated using Geant4 [24, 25]. Additional $pp$ interactions in the same and nearby bunch crossings (pile-up) are included in the simulation. Muon reconstruction and triggering efficiencies are corrected in the simulated samples using data-driven scale factors. The scale factors for the trigger efficiencies are obtained comparing the efficiencies between data and simulation, using a $J/\psi$ tag-and-probe method to determine them. This procedure yields the scale factors as a function of the muon transverse momentum and pseudorapidity which are applied throughout the analysis [26]. Reconstruction and selection efficiencies are obtained from simulation and similarly corrected according to data-driven comparisons. In addition to these efficiency corrections, simulated events are reweighted to reproduce the pile-up multiplicity observed in data, and according to the equivalent integrated luminosity associated with each trigger selection.

Using the iterative reweighting method described in Ref. [27], the simulated samples of the exclusive decays considered are adjusted with two-dimensional data-driven weights (DDW) to correct for the differences between simulation and data observed in the $B$ meson transverse momentum and pseudorapidity distributions. DDW obtained from $B^+ \rightarrow J/\psi K^+$ decays are used to correct the simulation samples in the signal and normalisation channels. DDW obtained from the $B^0 \rightarrow J/\psi \phi$ control channel are found to agree with those from $B^+ \rightarrow J/\psi K^+$, showing the consistency of the corrections.

Residual differences between data and simulation studied on the $B^+$ and $B^0_c$ reference channels are treated as sources of systematic uncertainties in the evaluation of the signal efficiency, as discussed in Section 10. The one exception to this treatment is the $B$ meson isolation ($I_{0.7}$ in Section 7 Table 1), where residual differences are used to reweight the signal MC and the corresponding uncertainties propagated to account for residual systematic effects.

Similarly to the exclusive decays, the large continuum background MC sample kinematic distributions are reweighted via corrections obtained from its comparison with the data in the sidebands of the signal region.

4 Data selection

For data collected during the LHC Run 2, the ATLAS detector uses a two-level trigger system, consisting of a hardware-based Level-1 trigger and a software-based High Level trigger. A dimuon trigger [28] selects events at Level-1 requiring that one muon has $p_T > 4$ GeV and the other has $p_T > 6$ GeV. A full track reconstruction of the muon candidates was performed at the software trigger level, where an additional loose selection was applied to the dimuon invariant mass $m_{\mu\mu}$, accepting candidates in the
range 4 GeV to 8.5 GeV. Due to the increased pile-up in 2016 data, an additional selection was added at this trigger stage, requiring a positive transverse displacement projection along the dimuon momentum of the dimuon vertex with respect to the primary vertex ($L_{xy}$). The effect of this selection is accounted for in the analysis but has no real consequences since stricter requirements are applied to the signal discriminant (see Section 7).

Both the signal and reference channels $B^+ \rightarrow J/\psi K^+$ and $B_s^0 \rightarrow J/\psi \phi$ events have been selected with prescales which vary during the data taking period. In the 36.2 fb$^{-1}$ of data analysed, the effective integrated luminosity for the signal sample is 26.3 fb$^{-1}$, corresponding to an effective prescale of 1.4, while for the reference channels 15.1 fb$^{-1}$ are collected due to an effective prescale of 2.4. These effects are taken into account in the extraction of the signal branching fraction, through the $\varepsilon$ parameters in Eq. (1).

After offline reconstruction, a preliminary selection is performed on candidates for $B^{0(s)}_s \rightarrow \mu^+\mu^-$. $B^+ \rightarrow J/\psi K^+ \rightarrow \mu^+\mu^- K^+$ and $B^0_s \rightarrow J/\psi \phi \rightarrow \mu^+\mu^- K^+K^-$. In the ID system, muons are required to have at least one hit in the pixel detector and two measured intercepts in the semiconductor tracker. They are also selected to be reconstructed in the MS, and to have $|\eta| < 2.5$. The offline muon pair must pass the $p_T > 4$ GeV and $p_T > 6$ GeV thresholds imposed also by the trigger. Furthermore the muons are required to fulfil tight muon quality criteria [29], relaxed to loose for the hadron misidentification studies in Section 6. Kaon candidates have to satisfy similar requirements in the ID, except for a looser requirement of $p_T > 1$ GeV.

$B$ meson properties are computed based on a decay vertex fitted to two, three or four tracks, depending on the decay process to be reconstructed. $B$ candidates are excluded when producing a $\chi^2$ per degree of freedom of the vertex fit above 6 for the $B$ vertex, and above 10 for the $J/\psi \rightarrow \mu\mu$ vertex. The selections $2915 < m(\mu\mu) < 3275$ MeV and $1005 < m(KK) < 1035$ MeV are applied to the $J/\psi \rightarrow \mu\mu$ and the $\phi \rightarrow KK$ vertices, respectively. In the $B^+ \rightarrow J/\psi K^+$ and $B^0_s \rightarrow J/\psi \phi$ fits the reconstructed dimuon mass is constrained to the $J/\psi$ world average value [30].

Reconstructed $B$ candidates are retained if they satisfy $p_T^B > 8.0$ GeV and $|\eta^B| < 2.5$. The invariant mass for all candidates is calculated using the muons combined ID and MS information, in order to improve the mass resolution with respect to using ID information only [31].

The invariant mass range considered for the $B^{0(s)}_s \rightarrow \mu^+\mu^-$ decay starts at 4766 MeV and is 1200 MeV wide. Within this range a 360 MeV wide signal region is defined, starting at 5166 MeV. The remainder of the range defines the upper and lower mass sidebands of the analysis.

For the reference channels, the mass range considered is 4930–5630 (5050–5650) MeV for $B^+ \rightarrow J/\psi K^+$ ($B^0_s \rightarrow J/\psi \phi$) in which the 5180–5380 (5297–5437) MeV range is the peak region and the two low and high mass ranges are the mass sidebands used for background subtraction.

The coordinates of primary vertices (PV) are obtained from charged-particle tracks not used in the decay vertices, and are transversely constrained to the luminous region of the colliding beams. The matching of a $B$ candidate to a PV is made by propagating the candidate trajectory to the point of closest approach to the beam axis, and choosing the PV with the smallest separation along $z$. Simulation shows that this method achieves a correct matching probability of better than 99%$, for all relevant pile-up conditions.

To reduce the large background in the $B^{0(s)}_s \rightarrow \mu^+\mu^-$ channel before the final selection based on multivariate classifiers, a loose collinearity requirement is applied between the momentum of the $B$ candidate ($\vec{p}^B$) and the spatial separation between the PV and the decay vertex ($\Delta x$). The absolute value of the difference in
azimuthal angle $\alpha_{2D}$ between these two vectors is required to be smaller than 1.0 rad. Using the difference in rapidity $\Delta \eta$ for the same vectors, an upper cut of 1.5 is applied to the combination $\Delta R = \sqrt{\alpha_{2D}^2 + \Delta \eta^2}$.

After the preliminary selection, approximately $3.5 \times 10^6$ candidates are found in the $B^0_{(s)} \rightarrow \mu^+\mu^-$ fit region, with about $1.0 \times 10^6$ falling in the blinded range [5166, 5526] MeV.

5 Background composition

The background to the $B^0_{(s)} \rightarrow \mu^+\mu^-$ signal originates from three main sources:

Continuum background, the dominant combinatorial component, consisting of muons originating from uncorrelated hadron decays and characterised by a small dependence on the dimuon invariant mass;

Partially reconstructed decays, characterised by non-reconstructed final-state particles ($X$) and thus accumulating in the low dimuon invariant mass sideband;

Peaking background, due to $B^0_{(s)} \rightarrow h h'$ decays, with both hadrons misidentified as muons.

The continuum background consists mainly of muons independently produced in the fragmentation and decay chains of a $b$ and a $\bar{b}$ quark (opposite-side muons). It is studied in the signal mass sidebands, and it is found to be well described by the inclusive MC sample of semileptonic decays of $b$ and $c$ hadrons.

The partially reconstructed decays consist of several topologies: (a) same-side combinatorial background from decay cascades ($b \rightarrow c\mu\nu \rightarrow s(d)\mu\mu\nu\nu$); (b) same-vertex background from $B$ decays containing a muon pair (e.g. $B^0 \rightarrow K^{*0}\mu\mu$ or $B \rightarrow J/\psi\mu X \rightarrow \mu\mu X$); (c) $B_c$ decays (e.g. $B_c \rightarrow J/\psi\mu\nu \rightarrow \mu\mu\nu$); (d) semileptonic $b$-hadron decays where a final-state hadron is misidentified as a muon. The remainder of this paper will implicitly exclude categories (c) and (d) when referring to partially reconstructed or $b \rightarrow \mu\mu X$ decays, since these categories are treated separately.

An inclusive MC sample of $b\bar{b} \rightarrow \mu\mu X$ decays was generated and the background composition after the analysis selection was investigated. All backgrounds in this sample have a dimuon invariant mass distribution accumulating below the mass range considered in this analysis, with a high-mass tail extending beyond the signal region. The simulation does not contemplate sources other than muons from $b\bar{b}$ decays: $c\bar{c}$ and prompt contributions are not included. All possible origins of two muons in the $b\bar{b}$ decay tree are however analysed, after classification in the mutually exclusive continuum and partially reconstructed categories described above. These samples are used exclusively to identify suitable functional models for the corresponding background components, and as benchmark for these models. No shape or normalization constraints are derived from this simulation. This makes the analysis largely insensitive to mismatches between background simulation and data.

The semileptonic decays with final-state hadrons misidentified as muons consist mainly of three-body charmless decays $B^0 \rightarrow \pi\mu\nu$, $B^0_s \rightarrow K\mu\nu$ and $\Lambda_b \rightarrow p\mu\nu$ in which the tail of the invariant mass distribution extends to the signal region. Due to branching fractions of the order of $10^{-6}$, this background is not large, and is further reduced by the muon identification requirements, discussed in Section 6. The MC invariant mass distributions of these partially reconstructed decay topologies are shown, together with the signals SM predictions, in Figure 1(a) after applying the preliminary selection criteria described in Section 4.

Finally, the peaking background is due to $B^0_{(s)}$ decays containing two hadrons misidentified as muons, which populate the signal region as shown in the distributions obtained from simulation of Figure 1(b). Section 6 will discuss this component.
Figure 1: (a) Dimuon invariant mass distribution for the partially reconstructed background, from simulation, before
the final selection against continuum is applied but after all other requirements. The different components are shown as
stacked histograms, normalised according to world-averaged measured branching fractions. The SM expectations for the $B^0(s) \rightarrow \mu^+ \mu^-$ signals are also shown for comparison. Continuum background is not included here. (b) Invariant mass distribution of the $B^0(s) \rightarrow hh'$ peaking background components, after the complete signal
selection is applied. The $B^0 \rightarrow \pi^\pm \pi^\mp$ and $B^0 \rightarrow K^\pm K^\mp$ contributions are negligible on this scale. In both plots the
vertical dashed lines indicate the analysis blinded region. Distributions are normalised to the expected yield for the
integrated luminosity of 26.3 fb$^{-1}$.

6 Hadron misidentification

In the preliminary selection, muon candidates are formed from the combination of tracks reconstructed
independently in the ID and MS. The performance of the muon reconstruction in ATLAS is presented in
Ref. [29]. Additional studies were performed to evaluate the amount of background related to hadrons
erroneously identified as muons.

Detailed simulation studies were performed for the $B^0(s) \rightarrow hh'$ channel with a full Geant4-based
simulation [24] in all systems of the ATLAS detector. The vast majority of background events from
particle misidentification are due to decays in flight of kaons and pions, in which the muon receives most
of the energy of the meson. Hence this background is generally related to true muons measured in the
MS, but not produced promptly in the decay of a $B$ meson.

The muon candidate is required to pass tight muon requirements in the preliminary selection, which are
based on the profile of energy deposits in the calorimeters as well as on tighter ID-MS matching criteria
than the loose requirements. For two-body $B$ decays the muon efficiency of this selection is 90%, with an
average reduction of the hadron misidentification by a factor 0.39, verified on control data samples. The
resulting final value of the misidentification probability is equal to 0.08% for kaons and 0.1% for pions.
Efficiencies and fake rates are relative to the analysis preselections, including tracking but excluding any
muon requirement.

The background due to $B^0(s) \rightarrow hh'$, with double misidentification $hh' \rightarrow \mu\mu$, has a distribution in the
reconstructed invariant mass peaking at $5240$ MeV, close to the $B^0$ mass and is effectively indistinguishable
from the $B^0$ signal (see Figure 1(b)). Beyond the muon selection, these events have the same acceptance and selection efficiency as the $B^0_{(s)} \rightarrow \mu^+\mu^-$ signal. Therefore, the expected number of peaking-background events can be estimated from the number of observed $B^+ \rightarrow J/\psi K^+$ events, in a way analogous to that for the signal, using Eq. (1). World average [30] values for the branching fractions of $B^0$ and $B^0_s$ into $K\pi$, $KK$ and $\pi\pi$ are used, together with the hadron misidentification probabilities obtained from simulation. This results in $2.7 \pm 1.3$ total expected peaking-background events, after the reference multivariate selection.

Inverting the tight muon selection, the number of events containing real muons is largely reduced, while the number of peaking-background events is approximately two times larger than in the sample obtained with the nominal selection. A fit to this background-enhanced sample returns $6.8 \pm 3.7$ events, which translates into a peaking background yield in the signal region of $2.9 \pm 2.0$ events, in good agreement with the simulation.

Besides the peaking background, the tight muon selection also reduces the semileptonic contributions with a single misidentified hadron. Simulation yields $30 \pm 3$ events expected from $B^0 \rightarrow \pi\mu\nu$ and $B^0_s \rightarrow K\mu\nu$ in the final sample, with a distribution kinematically constrained to be mostly below the signal region. The $\Lambda_b \rightarrow p\mu\nu$ contribution is negligible due to the smaller production cross section and the low rate at which protons fake muons.

7 Continuum background reduction

A multivariate analysis, implemented as a boosted decision tree, is employed to enhance the signal relative to the continuum background. This classifier, referred to as the BDT, is based on the 15 variables described in Table 1. The discriminating variables can be classified into three groups: (a) $B$ meson variables, related to the reconstruction of the decay vertex and to the collinearity between $\vec{p}^B$ and the separation between the production and decay vertices $\Delta \chi$; (b) variables describing the muons forming the $B$ meson candidate; and (c) variables related to the rest of the event. The selection of the variables aims to optimise the discrimination power of the classifier, while minimising the dependence on the invariant mass of the muon pair.

The same discriminating variables were used in the previous analysis based on the full Run 1 dataset [15]. A significant reduction of the BDT separation power has been observed when excluding discriminating variables. To minimise the dependence of the classifier on the effects of pile-up, requirements of compatibility with the same vertex matched to the dimuon candidate are placed on the additional tracks considered for the variables $I_{0.7}$, $DOCA_{\text{extrk}}$ and $N_{\text{close}}$.

The correlation between the discriminating variables was studied in the MC samples for signal and continuum background discussed in Section 3, and on data from the sidebands of the $\mu^+\mu^-$ invariant mass distribution. There are significant linear correlations among the variables $\chi^2_{\text{PV,DV,xy}}$, $L_{xy}$, $|d_0|_{\text{max-sig.}}$, $|d_0|_{\text{min-sig.}}$ and $\chi^2_{\mu,xy \nu}$. The variables $IP_{\mu}^{3D}$, $DOCA_{\mu\mu}$ and $I_{0.7}$ have negligible correlation with any of the others used in the classifier.

The simulated signal sample and the data from the dimuon invariant mass sideband regions are used for training and testing the classifier. As discussed in Section 3, simulated signal samples are corrected for muon reconstruction efficiency discrepancies between simulation and data, and reweighted according to

---

2 This selection, corresponding to 54% signal efficiency, has been also applied to derive all other quantities quoted in this section and includes a multivariate cut against $\mu^+\mu^-$ continuum background, the BDT discussed in Section 7.
Table 1: Description of the 15 variables used in the discrimination between signal and continuum background. When the BDT classifier is applied to $B^+ \to J/\psi K^+$ and $B_s^0 \to J/\psi \phi$ candidates, the variables related to the decay products of the $B$ mesons refer only to the muons from the decay of the $J/\psi$. Horizontal lines identify the classification in groups (a), (b) and (c) respectively, as described in the text. For category (c), additional tracks are required to have $p_T > 500$ MeV unless otherwise specified.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_T^B$</td>
<td>Magnitude of the $B$ candidate transverse momentum $\vec{p}_T^B$.</td>
</tr>
<tr>
<td>$\chi^2_{PV, DV_{xy}}$</td>
<td>Compatibility of the separation $\Delta x$ between production (i.e. associated PV) and decay (DV) vertices in the transverse projection: $\Delta x_T \cdot \Sigma^{-1} \Delta x_T$, where $\Sigma$ is the covariance matrix.</td>
</tr>
<tr>
<td>$\Delta R$</td>
<td>Three-dimensional opening between $\vec{p}_T^B$ and $\Delta x_T$: $\sqrt{\Delta \phi^2 + \Delta \eta^2}$</td>
</tr>
<tr>
<td>$</td>
<td>\alpha_{2D}</td>
</tr>
<tr>
<td>$L_{xy}$</td>
<td>Projection of $\Delta x_T$ along the direction of $\vec{p}_T^B$: $(\Delta x_T \cdot \vec{p}_T^B) /</td>
</tr>
<tr>
<td>IP$^{3D}_B$</td>
<td>Three-dimensional impact parameter of the $B$ candidate to the associated PV.</td>
</tr>
<tr>
<td>DOCA$_{\mu\mu}$</td>
<td>Distance of closest approach (DOCA) of the two tracks forming the $B$ candidate (three-dimensional).</td>
</tr>
<tr>
<td>$\Delta \phi_{\mu\mu}$</td>
<td>Difference in azimuthal angle between the momenta of the two tracks forming the $B$ candidate.</td>
</tr>
<tr>
<td>$</td>
<td>d_0</td>
</tr>
<tr>
<td>$</td>
<td>d_0</td>
</tr>
<tr>
<td>$p_{L\text{min}}^T$</td>
<td>Value of the smaller projection of the momenta of the muon candidates along $\vec{p}_T^B$.</td>
</tr>
<tr>
<td>$I_{0.7}$</td>
<td>Isolation variable defined as ratio of $</td>
</tr>
<tr>
<td>DOCA$_{\text{extrk}}$</td>
<td>DOCA of the closest additional track to the decay vertex of the $B$ candidate. Tracks matched to a PV different from the $B$ candidate are excluded.</td>
</tr>
<tr>
<td>$N_{\text{close extrk}}$</td>
<td>Number of additional tracks compatible with the decay vertex (DV) of the $B$ candidate with $\ln(\chi^2_{\text{extrk,DV}}) &lt; 1$. The tracks matched to a PV different from the $B$ candidate are excluded.</td>
</tr>
<tr>
<td>$\chi^2_{\mu,xPV}$</td>
<td>Minimum $\chi^2$ for the compatibility of a muon in the $B$ candidate with any PV reconstructed in the event.</td>
</tr>
</tbody>
</table>

the distributions of $p_T$ and $|\eta|$ of the dimuon and of the pile-up observed in data. The BDT training is done using TMVA [32].

Since sideband data are used for the BDT training and optimisation, the sample is subdivided into three randomly selected separate and equally populated sub-samples used in turn to train, verify and calibrate the selection efficiency of three independent BDTs. The resulting BDTs are verified to have performances that are statistically compatible, and are combined in one single classifier in such a way that each BDT is applied only to the part of the data sample not involved in the BDT training.

Figure 2 shows the distribution of the BDT output variable for simulated signal and backgrounds, separately
for continuum background and partially reconstructed events. Also shown is the BDT distribution for dimuon candidates from data, from the sidebands of the invariant mass distribution. The BDT has been verified not to show any significant correlation with the dimuon invariant mass. The final selection requires a BDT output value larger than 0.1439, corresponding to signal and continuum background efficiencies of 72% and 0.3% respectively. The analysis will use all candidates after this selection, however the $B_0(s) \rightarrow \mu^+\mu^-$ signal fit has been verified to be insensitive to the signal for candidates with BDT output value smaller than 0.2455 (corresponding to 54% and 0.03% efficiencies for signal and background respectively).

Signal and reference channel yields and efficiencies will be - for this reason - measured relative to the latter signal reference cut, while maintaining the looser selection in order to improve the background modelisation.

8 Data–simulation comparisons

Figure 3 compares the distributions of two discriminating variables in the continuum background MC sample with data in the dimuon sidebands. Agreement with the sideband data is fair and the discrepancies observed do not compromise the use of this MC background sample for the choice of the background models in the signal fit. The functional parameterization of the backgrounds as a function of the BDT selection has been investigated in data and shown to be consistent with simulation, even for lower BDT values, where prompt backgrounds contribute significantly. The normalization of the backgrounds is not tied to the simulation and is extracted purely from data. The continuum MC simulation is not used for BDT training, computation of efficiencies or normalisation purposes.

The distributions of the discriminating variables are also used for the comparison of $B^+ \rightarrow J/\psi K^+$ and
Figure 3: Data and continuum MC distributions of the (a) $|\alpha_{2D}|$ and (b) $\ln(\chi^2_{\mu,xPV})$ variables (defined in Table 1). The points correspond to the sideband data, while the continuous-line histogram corresponds to the continuum MC distribution, normalised to the number of data events. The filled-area histogram shows the signal MC distribution for comparison. The bottom insets report the data/MC ratio, zoomed in order to highlight discrepancies in the region that is most relevant for the analysis.

$B^0 \to J/\psi \phi$ events between simulation and data. To perform such comparison, for each variable the contribution of the background is subtracted from the signal. For this purpose, a maximum-likelihood fit is performed to the invariant mass distribution, separately in bins of rapidity and transverse momentum. The fit model used is simpler than the one employed for the extraction of the $B^+$ signal for normalisation as described in Section 9, but is sufficient for the purpose discussed here.

Figure 4 shows examples of the distributions of the discriminating variables obtained from data and simulation for the reference samples. Observed differences are accounted for as systematic effects with the procedure described in Section 10. The discrepancy shown for the isolation variable $I_{0.7}$ in the $B^+ \to J/\psi K^+$ channel is the most significant one among all variables and both reference channels.

9 Yield extraction for the normalisation channel $B^+ \to J/\psi K^+$

The $B^+$ yield for the normalisation channel is extracted with an unbinned extended maximum-likelihood fit to the $J/\psi K^+$ invariant mass distribution of half of the available events. The functional forms used to model both the signal and the backgrounds are obtained from studies of MC samples. All the yields are extracted from the fit to data, while the shape parameters are determined from a simultaneous fit to data and MC samples. Free parameters are introduced for the mass scale and mass resolution to accommodate data–MC differences. The fit yields a negligible increase in resolution and a mass shift at the level of 2 MeV.

The fit includes four components: $B^+ \to J/\psi K^+$ events, Cabibbo-suppressed $B^+ \to J/\psi \pi^+$ events on the right tail of the main peak, partially reconstructed $B$ decays (PRD) where one or more of the final-state particles are missing, and the continuum background composed mostly of $b\bar{b} \to J/\psi X$ events. All
components other than the last one have shapes constrained from MC simulation as described below, with
the data fit including an additional gaussian convolution to account for possible data-MC discrepancies in
mass scale and resolution. The shape of the $B^+ \rightarrow J/\psi K^+$ distribution is parameterised using a
Johnson $S_U$ function [33, 34]. The final $B^+ \rightarrow J/\psi K^+$ yield includes the contribution from radiative $B^\pm$
decays (i.e. where photons are emitted from the $B$ decay products). The $B^+ \rightarrow J/\psi \pi^+$ events are modelled
by the sum of a Johnson $S_U$ and a Gaussian function, where all parameters except the normalisation are
determined from the simulation. The decay modes contributing to the PRD are classified in simulation
on the basis of their mass dependence. Each of the three resulting categories is contributing to the

Figure 4: Data and MC distributions in $B^+ \rightarrow J/\psi K^+$ events for the discriminating variables: (a) $|\alpha_{2D}|$, (b)
$\ln(\chi^2_{PV,DV,xy})$ and (c) $I_{0.7}$. The variable $I_{0.7}$ is also shown in (d) for $B^0_s \rightarrow J/\psi \phi$ events. The black points
correspond to the sideband-subtracted data, while the red histogram corresponds to the MC distribution, normalised
to the number of data events. Differences in shape between MC events and data are accounted for as systematic
effects, as discussed in the text. The discrepancy shown for $I_{0.7}$ in the $B^+ \rightarrow J/\psi K^+$ channel is the most significant
among all variables and both reference channels. The bottom insets report the data/MC ratio, zoomed in order to
highlight discrepancies in the region that is most relevant for the analysis.
Figure 5: Result of the fit to the $J/\psi K^+$ invariant mass distribution for all $B^+$ candidates in half of the data events. The various components of the spectrum are described in the text. The inset at the bottom of the plot shows the bin-by-bin pulls for the fits, where the pull is defined as the difference between the data point and the value obtained from the fit function, divided by the error from the fit.

The overall PRD shape with combinations of Fermi–Dirac and exponential functions, contributing differently in the low-mass region. Their shape parameters are determined from simulation. Finally, the continuum background is modelled with an exponential function with the shape parameter extracted from the fit. The normalisation of each component is unconstrained in the fit, which is therefore mostly independent of external inputs for the branching fractions. The residual dependence of the PRD model shapes on the relative branching fractions of the contributing decays will be considered as a source of systematic uncertainty. The resulting fit, shown in Figure 5, yields $334351$ $B^+ \rightarrow J/\psi K^+$ events with a statistical uncertainty of 0.3%. The ratio of yields of $B^+ \rightarrow J/\psi \pi^+$ and $B^+ \rightarrow J/\psi K^+$ is $(3.71 \pm 0.09)$% (where the uncertainty reported is statistical only), in agreement with the expectation from the PDG branching fractions [30] of $(3.84 \pm 0.16)$%.

Some of the systematic effects are included automatically in the fit: the effect of the limited MC sample size, for example, is included in the uncertainties through a simultaneous fit to data and MC samples. Scaling factors determined in the fit to data account for the differences in mass scale and resolution between data and simulation. Additional systematic uncertainties are evaluated by varying the default fit model described above: they take into account the kinematic differences between data and the MC samples used in the fit, differences in efficiency between $B^+$ and $B^-$ decays, uncertainties in the relative fractions and shapes of PRD, and in the shape of the various fit components, as well as the high-statistics fit stability as a function of the fit parameters starting values. In each case, the difference with respect to the default fit is recorded, symmetrised and used as an estimate of the systematic uncertainty. The main contributions to the systematic uncertainty come from the functional models of the background components, the relative fractions of PRD and the signal charge asymmetry. The total systematic uncertainty in the $B^+$ normalisation yield amounts to 4.8%.
10 Evaluation of the $B^+ \to J/\psi K^+ \to B^0_{(s)} \to \mu^+\mu^-$ efficiency ratio

The ratio of efficiencies $R_e = \frac{\varepsilon(B^+\to J/\psi K^+)}{\varepsilon(B^+\to J/\psi K^+)}$ enters the $D_{\text{norm}}$ term defined in Section 2: $D_{\text{norm}} = \frac{N_{J/\psi K^+}}{R_e}$. Both channels are measured in the fiducial volume of the $B$ meson defined as $p_T^B > 8.0$ GeV and $|\eta_B| < 2.5$. The total efficiencies within the fiducial volume include acceptance and trigger, reconstruction and selection efficiencies. The acceptance is defined by the selection placed on the particles in the final state: $|\eta_\mu| < 2.5$ and $p_T^\mu > 6.0$ (4.0) GeV for the leading (trailing) $p_T$ muons, $p_T^K > 1.0$ GeV and $|\eta_K| < 2.5$ for kaons. The signal reference BDT selection applied to the signal and reference channel corresponds to an efficiency of about 54% and 51% respectively. The resulting efficiency ratio $R_e$ is $0.1176\pm0.0009$ (stat.)$\pm0.0047$ (syst.), with uncertainties determined as described below.

$R_e$ is computed using the mean lifetime of $B_s^0$ [30, 35] in the MC generator. The same efficiency ratios apply to the $B_s^0 \to \mu^+\mu^-$ and $B^0 \to \mu^+\mu^-$ decays, within the MC statistical uncertainty of $\pm0.8\%$. The statistical uncertainties in the efficiency ratios come from the finite number of events available for the simulated samples. The systematic uncertainty affecting $R_e$ comes from five sources.

A first contribution is due to the uncertainties in the DDW, and amounts to 0.8%. This term is assessed from pseudo-MC studies, performed by varying the corrections within their statistical uncertainties. The RMS value of the distribution of $R_e$ obtained from pseudo-MC samples is taken as the systematic uncertainty.

A second contribution of 1% is related to the muon trigger and reconstruction efficiencies. The effect of the uncertainties in the data-driven efficiencies is evaluated with pseudo-MC studies.

A 3.2% systematic uncertainty contribution arises from the differences between data and simulation observed in the modelling of the discriminating variables used in the BDT classifier (Table 1). For each of the 15 variables, the MC samples for $B_0^{(s)} \to \mu^+\mu^-$ and $B^+ \to J/\psi K^+$ are reweighted according to the distribution of the variable observed in $B^+ \to J/\psi K^+$ events from the data sample, after background subtraction. The isolation variable $I_{0.7}$ is computed using charged-particle tracks only, and differences between $B^+$ and $B_0^{(s)}$ are expected and were observed in previous studies [27]. Hence for this variable the reweighting procedure for the $B_0^{(s)} \to \mu^+\mu^-$ MC sample is based on $B_0^{(s)} \to J/\psi \phi$ data. For all discriminating variables except $I_{0.7}$, the value of the efficiency ratio is modified by less than 2% by the reweighting procedure. For these variables, each variation is taken as an independent contribution to the systematic uncertainty in the efficiency ratio. For $I_{0.7}$ the reweighting procedure changes the efficiency ratio by about 6%. Because of the significant mis-modelling, the MC samples obtained after reweighting on the distribution of $I_{0.7}$ are taken as a reference, thus correcting the central value of the efficiency ratio. The 1% uncertainty in the correction is added to the sum in quadrature of the uncertainties assigned to the other discriminating variables. The total uncertainty in the modelling of the discriminating variables is the dominant contribution to the systematic uncertainties on $R_e$.

A fourth source of systematic uncertainty arises from differences between the $B^0 \to \mu^+\mu^-$ and the $B^+ \to J/\psi K^+$ channels related to the reconstruction efficiency of the kaon track and of the $B^+$ decay vertex. These uncertainties are mainly related to inaccuracy in the modelling of passive material in the ID system. The corresponding systematic uncertainty is estimated varying the detector model in simulations, observing effects that vary between 0.4% and 1.5% depending on the $\eta$ range considered. A conservative systematic effect of $\pm1.5\%$ is assumed.

$^3$ The requirement is for the output of the BDT to be larger than 0.2455.
Finally, the propagation of the uncertainty on the reweighting of the simulated events as a function of the pile-up multiplicity distribution contributes an additional 0.6% to the total.

Table 2 summarises the contribution of these systematic uncertainties to the total.

<table>
<thead>
<tr>
<th>Source</th>
<th>Contribution (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical</td>
<td>0.8</td>
</tr>
<tr>
<td>BDT Input Variables</td>
<td>3.2</td>
</tr>
<tr>
<td>Kaon Tracking Efficiency</td>
<td>1.5</td>
</tr>
<tr>
<td>Muon trigger and reconstruction</td>
<td>1.0</td>
</tr>
<tr>
<td>Kinematic Reweighting (DDW)</td>
<td>0.8</td>
</tr>
<tr>
<td>Pile-up Reweighting</td>
<td>0.6</td>
</tr>
</tbody>
</table>

The efficiency ratio enters in Eq. (1) with the $D_{\text{norm}}$ term defined in Section 2, multiplied by the number of observed $B^\pm$ candidates. The total uncertainty on $D_{\text{norm}}$ is ±6.3%.

A correction to the efficiency ratio for $B_{s}^{0} \rightarrow \mu^{+}\mu^{-}$ is needed because of the width difference $\Delta \Gamma_s$ between the $B_{s}^{0}$ eigenstates. According to the SM, the decay $B_{s}^{0} \rightarrow \mu^{+}\mu^{-}$ proceeds predominantly through the heavy state $B_{s, H}$ [1, 16], which has width $\Gamma_{s, H} = \Gamma_{s} - \Delta \Gamma_{s}/2$, which is 6.6% smaller than the average $\Gamma_{s}$ [30]. The variation in the value of the $B_{s}^{0} \rightarrow \mu^{+}\mu^{-}$ mean lifetime was tested with simulation, and found to change the $B_{s}^{0}$ efficiency by +2.7%, and consequently the $B_{s}^{0}$ to $B^{+}$ efficiency ratio. This correction is applied to the central value of $D_{\text{norm}}$ used in Section 12 for the determination of $B(\Bosm \rightarrow \mu^{+}\mu^{-})$. Due to the small value of $\Delta \Gamma_{s}$, no correction needs to be applied to the $B_{s}^{0} \rightarrow \mu^{+}\mu^{-}$ decay.

11 Extraction of the signal yield

Dimuon candidates passing the preliminary selection and the selections against hadron misidentification and continuum background are classified according to four intervals (with boundaries at 0.1439, 0.2455, 0.3312, 0.4163 and 1) in the BDT output. Repeating the Run 1 analysis approach, each interval is chosen to insure an equal efficiency of 18% for signal events, and they are ordered according to increasing signal-to-background ratio.

An unbinned extended maximum-likelihood fit is performed on the dimuon invariant mass distribution simultaneously across the four BDT intervals. The first two bins contribute much less to the signal determination and are included for background modelling. They have been verified to have negligible relevance for signal extraction. The result of the fit is the total yield of $B_{s}^{0} \rightarrow \mu^{+}\mu^{-}$ and $B_{s}^{0} \rightarrow \mu^{+}\mu^{-}$ events in the three most sensitive BDT intervals. The parameters describing the background are allowed to vary freely and are determined by the fit. The normalisations of the individual fit components, including the signals, are completely unconstrained and allowed to take negative values. Signal yields in the individual BDT bins are constrained to be the same within the relative signal efficiency in each of the BDT bins.

### Footnote

4 The decay time distribution of $B_{s}^{0} \rightarrow \mu^{+}\mu^{-}$ is predicted to be different from the one of $B_{s, H}$ in scenarios of new physics, with the effect related to the observable $\mathcal{A}_{\Delta t}^{\mu\mu}$. The maximum possible deviation from the SM prediction of $\mathcal{A}_{\Delta t}^{\mu\mu} = +1$, is for $\mathcal{A}_{\Delta t}^{\mu\mu} = -1$, for which the decay time distribution of $B_{s}^{0} \rightarrow \mu^{+}\mu^{-}$ corresponds to the distribution of the $B_{s,1}$ eigenstate. In the comparison with new physics predictions, the value of $\mathcal{B}(B_{s}^{0} \rightarrow \mu^{+}\mu^{-})$ obtained from this analysis should be corrected by $+2.7\% \times (1 - \mathcal{A}_{\Delta t}^{\mu\mu})$. 

15
discussed in Section 11.1, where the signal and background fit models are described. The systematic uncertainties related to the BDT intervals, to the signal and to the background model are discussed in Sections 11.1 and 11.2, and are included in the likelihood with Gaussian multiplicative factors with width equal to the systematic uncertainty.

11.1 Signal and background model

The signal and background models are derived from simulations and data collected in the mass sidebands of the search region.

The invariant mass distribution of the $B_0^{(s)} \rightarrow \mu^+ \mu^-$ signal is described by two double-Gaussian distributions, centred respectively at the $B_0^0$ or $B_0^0$ mass. The shape parameters are extracted from simulation, where they are verified to be uncorrelated with the BDT output. Systematic uncertainties in the mass scale and resolutions are considered separately. Figure 6 shows the invariant mass distributions for $B_0^0$ and $B_0^0$, obtained from MC events and normalised to the SM expectations. Section 10 shows how systematic uncertainties affect the selection of all signal candidates considered in the extraction of the signal yield. Their separation in BDT bins however requires also the knowledge of the relative signal efficiency in each of these bins, which must be accompanied by the uncertainty in its determination.

Two different procedures are explored. First, the distribution of the BDT output is compared between MC and background-subtracted data. The differences observed in the ratio of data over simulation are described with a linear dependence on the BDT output. The linear dependencies observed for $B^+ \rightarrow J/\psi K^+$ and $B_s^0 \rightarrow J/\psi \phi$ are in turn used to reweight the BDT-output distribution in the $B_0^{(s)} \rightarrow \mu^+ \mu^-$ MC sample. The maximum corresponding absolute variations in the efficiencies are equal to $+1.7\%$ and $-2.3\%$ respectively in the second and fourth BDT intervals, with the third interval basically unaffected. A second assessment of the systematic uncertainties in the relative efficiency of the BDT intervals is obtained with a procedure similar to the one used for the event selection (Section 10). For each discriminating variable, the MC sample is reweighted according to the difference between simulation and data observed in the

Figure 6: Dimuon invariant mass distribution for the $B_0^0$ and $B_0^0$ signals from simulation. The double Gaussian fits are overlaid. The two distributions are normalised to the SM prediction for the expected yield with an integrated luminosity of 26.3 fb$^{-1}$. 

uncertainties affect the selection of all signal candidates considered in the extraction of the signal yield. Their separation in BDT bins however requires also the knowledge of the relative signal efficiency in each of these bins, which must be accompanied by the uncertainty in its determination.

Two different procedures are explored. First, the distribution of the BDT output is compared between MC and background-subtracted data. The differences observed in the ratio of data over simulation are described with a linear dependence on the BDT output. The linear dependencies observed for $B^+ \rightarrow J/\psi K^+$ and $B_s^0 \rightarrow J/\psi \phi$ are in turn used to reweight the BDT-output distribution in the $B_0^{(s)} \rightarrow \mu^+ \mu^-$ MC sample. The maximum corresponding absolute variations in the efficiencies are equal to $+1.7\%$ and $-2.3\%$ respectively in the second and fourth BDT intervals, with the third interval basically unaffected. A second assessment of the systematic uncertainties in the relative efficiency of the BDT intervals is obtained with a procedure similar to the one used for the event selection (Section 10). For each discriminating variable, the MC sample is reweighted according to the difference between simulation and data observed in the
reference channels. The variation in the efficiency of each BDT interval is taken as the contribution to the systematic uncertainty due to mis-modelling of that variable. The sum in quadrature of the variations due to all discriminating variables is found to be similar in the $B^+ \to J/\psi K^+$ and $B_0^s \to J/\psi \phi$ channels. Absolute variations of $\pm 1\%$, $\pm 2.4\%$ and $\pm 4.4\%$ are found respectively in the second, third and fourth BDT intervals.

Both these assessments are quite conservative, since they account for effects which are already included in the absolute factor $R_\varepsilon$. The first of these procedures is used as a baseline for inclusion of Gaussian terms in the signal extraction likelihood to account for the uncertainty on the relative signal efficiency in the three most sensitive BDT bins. Care is taken in constraining the sum of the efficiencies of the three intervals sensitive to the signal, since that absolute efficiency and the corresponding uncertainty is parameterised with the $R_\varepsilon$ term.

Figure 7 shows the distribution of the BDT output from data and simulation for the reference channels, after reweighting the MC sample. The MC distribution for $B_0^s \to \mu^+ \mu^-$ events is also shown, illustrating how the linear deviation obtained from the reference channels affects the simulated signal BDT output. When studying these effects, the linear fit to the ratios in Figures 7(a) and 7(b) are performed in the range corresponding to the three BDT bins with the highest S/B, since the remaining bin is insensitive to the signal contribution.

The background in the signal fit is composed of the types of events described in Section 5: (a) the continuum background; (b) the background from partially reconstructed $b \to \mu \mu X$ events, which is present mainly in the low-mass sideband; (c) the peaking background.

The non-peaking contributions have a common mass shape model, with parameters constrained across the fit BDT bins as described below, and independent yields across BDT bins and components.

Both in simulation and sideband data, the continuum background has a small linear dependence on the dimuon invariant mass. In the simulation, the slope of the distribution shows a roughly linear trend, of small amplitude, as a function of the BDT intervals, which is confirmed in sideband data within large statistical errors. This dependence is included in the fit model. The small systematic uncertainties due to deviations from this assumption are discussed below in Section 11.2.

The $b \to \mu \mu X$ background has a dimuon invariant mass distribution peaking below the low-mass sideband region. The mass dependence is derived from data in the low-mass sideband region, and described with an exponential function with equal shape in the BDT intervals. The value of the shape parameter is extracted from the fit to data.

The invariant mass distribution of the peaking background is very similar to the $B_0^0$ signal, as shown in Figure 1(b). The shape for this component is obtained from MC, which indicates that shape and normalisation are independent from the BDT bin considered. In the fit, this contribution is included with fixed mass shape and with a normalisation of $2.9 \pm 2.0$ events, as discussed in Section 6. This contribution is equally distributed among the three top intervals of the BDT output.

The fitting procedure is tested with pseudo-MC experiments, as discussed below.

11.2 Systematic uncertainties in the fit

Studies based on pseudo-MC experiments are used to assess the sensitivity of the fit to the input assumptions. Variations in the description of signal and background components are used in the generation of the
Figure 7: BDT value distributions in data and MC for (a) $B^+ \rightarrow J/\psi K^+$, (b) $B^0_s \rightarrow J/\psi \phi$. The MC samples are normalised to the number of data events passing the signal reference BDT selection (Section 7). Figure (c) illustrates the BDT output for the $B^0_s \rightarrow \mu^+ \mu^-$ signal. The dashed histogram in (c) illustrates the effect of the linear reweighting on the BDT output discussed in the text. The vertical dashed lines correspond to the boundaries of the BDT intervals used in the $B^0(s) \rightarrow \mu^+ \mu^-$ signal fit.

pseudo-MC samples. The corresponding deviations in the average numbers $N_s, N_d$ of $B^0$ and $B^0_s$ events returned by the fit, run in the nominal configuration, are taken as systematic uncertainties. The amplitude of the variations in the generation of the pseudo-MC samples is determined in some cases by known characteristics of the ATLAS detector (reconstructed momentum scale and momentum resolution), in others using MC evaluation (background due to semileptonic three-body $B^0_s$ decays and to $B_c \rightarrow J/\psi \mu$), and in others from uncertainties determined from data in the sidebands and from simulation (shapes of the background components and their variation across the BDT intervals).

The pseudo-MC experiments were generated with the normalisation of the continuum and $b \rightarrow \mu \mu X$ components obtained from the fit to the data in the sideband of the invariant mass distribution, and the
peaking background from the expectation discussed in Section 6. The signal was generated with different configurations, roughly covering the range between zero and twice the expected yield according to the SM prediction.

For all variations in the assumptions and all configurations of the signal amplitudes the distributions of the differences between results and generated values are used to evaluate systematic effects. In addition, distributions obtained from pseudo-MC samples generated and fitted according to the nominal fit model are used to study systematic biases deriving from the fit procedure. For both signals, the bias on the yield is smaller than 15% of the fit error, for true values of the $B^0_s \rightarrow \mu^+\mu^-$ BR above $5 \times 10^{-10}$.

The shifts in $N_s$ or $N_d$ are combined by considering separately the sums in quadrature of the positive and negative shifts and taking the larger as the symmetric systematic uncertainty. The total systematic uncertainty is found to increase with the assumed size of the signal, with a dependence $\sigma_{\text{syst}}^{N_s} = 3 + 0.05 N_s$ and $\sigma_{\text{syst}}^{N_d} = 2.9 + 0.05 N_s + 0.05 N_d$. Most of the shifts observed have opposite sign for $N_s$ and $N_d$, resulting in a combined correlation coefficient in the systematic uncertainties of $\rho_{\text{syst}} = -0.83$.

The fit to the yield of $B^0_s$ and $B^0$ events is modified by including in the likelihood two smearing parameters for $N_s$ and $N_d$ that are constrained by a combined Gaussian distribution parameterised by the values of $\sigma_{\text{syst}}^{N_s}$, $\sigma_{\text{syst}}^{N_d}$ and $\rho_{\text{syst}}$.

11.3 Results of the signal yield extraction

The numbers of background events contained in the signal region (5166–5526 MeV) are computed from the interpolation of the data observed in the sidebands. This procedure yields $2685 \pm 37, 330 \pm 14, 51 \pm 6$ and $7.9 \pm 2.6$ events respectively in the four intervals of BDT. For comparison, the total expected number of signal events according to the SM prediction is 91 and 10 respectively for $N_s$ and $N_d$, equally distributed among the three intervals with the highest signal to background ratio.

In those three BDT intervals, in the unblinded signal region, a total of 1951 events in the full mass range of 4766–5966 MeV are used for the likelihood fit to signal and background. Without applying any boundary on the values of the fitted parameters, the values determined by the fit are $N_s = 80 \pm 22$ and $N_d = -12 \pm 20$, where the uncertainties correspond to likelihood variations of $-2 \Delta \ln(L) = 1$. The likelihood includes the systematic uncertainties discussed above, but statistical uncertainties largely dominate. The result is consistent with the expectation from simulation. The uncertainties in the result of the fit are discussed in Section 12, where the measured values of the branching fractions are presented.

Figure 8 shows the dimuon invariant mass distributions in the four intervals of BDT, together with the projections of the likelihood fit.

12 Branching fraction extraction

The branching fractions for the decays $B^0_s \rightarrow \mu^+\mu^-$ and $B^0 \rightarrow \mu^+\mu^-$ are extracted from data using a maximum-likelihood fit. The likelihood is obtained from the one used for $N_s$ and $N_d$ replacing the fit parameters with the corresponding branching fractions divided by normalisation terms in Eq. (1), and including Gaussian multiplicative factors for the normalisation uncertainties.
Figure 8: Dimuon invariant mass distributions in the unblinded data, in the four intervals of BDT output. Superimposed is the result of the maximum-likelihood fit. The total fit is shown as a black continuous line, with the dashed red line corresponding to the observed signal component, dashed blue to the \( b \to \mu\mu X \) background, and the dashed green to the continuum background. The signal components are grouped in one single curve, including both the \( B_s^0 \to \mu^+\mu^- \) and the negative \( B^0 \to \mu^+\mu^- \) component. The peaking \( B_s^0(\psi) \to hh' \) background is illustrated by the brown dashed line which is, for all BDT bins, very close to the x axis.

The normalisation terms include external inputs for the \( B^+ \) branching fraction and the relative hadronisation probability. The first is obtained from world averages [30] as the product of \( \mathcal{B}(B^+ \to J/\psi K^+) = (1.010 \pm 0.029) \times 10^{-3} \) and \( \mathcal{B}(J/\psi \to \mu^+\mu^-) = (5.961 \pm 0.033)\% \). The second is equal to one for \( B^0_s \), while for \( B^0 \) it is taken from the latest available HFLAV average [35] \( f_s/f_d = 0.256 \pm 0.013 \), assuming \( f_u/f_d = 1 \).

The efficiency- and luminosity-weighted number of events for the normalisation channel enters in Eq. (1) with the denominator \( \mathcal{D}_{\text{norm}} \). The values \( \mathcal{D}_{\text{norm}} = (5.69 \pm 0.36) \times 10^6 \) for \( B^0_s \) and \( (5.84 \pm 0.37) \times 10^6 \) for \( B^0 \) are obtained from the results of Sections 10 and 9, and include the +2.7% correction to the \( B_s^0 \to \mu^+\mu^- \) efficiency due to the lifetime difference between \( B_{s,H} \) and \( B^0 \).

The combination of \( B^+ \) branching fraction, hadronisation probabilities and \( \mathcal{D}_{\text{norm}} \), i.e. the single-event
sensitivity, is equal to \((4.02 \pm 0.35) \times 10^{-11}\) for \(B^0_s \rightarrow \mu^+\mu^-\) and \((1.059 \pm 0.074) \times 10^{-11}\) for \(B^0 \rightarrow \mu^+\mu^-\).

The values of the branching fractions that maximise the profile-likelihood are \(B(B^0_s \rightarrow \mu^+\mu^-) = (3.2 \pm 0.9) \times 10^{-9}\) and \(B(B^0 \rightarrow \mu^+\mu^-) = (-1.3 \pm 2.1) \times 10^{-10}\), where the uncertainties, both statistical and systematic combined, are estimated with a Gaussian approximation of the logarithm of the likelihood function maximum. Figure 9(a) shows likelihood contours for the simultaneous fit to \(B(B^0_s \rightarrow \mu^+\mu^-)\) and \(B(B^0 \rightarrow \mu^+\mu^-)\), for values of \(-2\Delta \ln (L)\) equal to 2.3, 6.2 and 11.8, relative to the maximum of the likelihood.

Table 3 gives a breakdown of the estimated contributions of systematic and statistical uncertainties: the results are still dominated by statistical uncertainties, with the most prominent source of systematic uncertainty coming from fit uncertainties (where the largest contributors are the mass scale and \(b \rightarrow \mu\mu X\) background parameterization).

Given the statistical regime of the analysis, the likelihood contours of Figure 9(a) cannot be immediately translated into contours with the conventional coverage of 1, 2 and 3 Gaussian standard deviations.

Moreover, the contours extend to regions of negative branching fractions, which are of less straightforward physical interpretation. In order to address these points, a Neyman construction [36] is employed to obtain the 68.3%, 95.5% and 99.7% confidence intervals in the 2D plane \(B(B^0_s \rightarrow \mu^+\mu^-)\) and \(B(B^0 \rightarrow \mu^+\mu^-)\). This construction yields the contours shown in Figure 9(b). The same construction is used to determine the 68.3% confidence interval for \(B(B^0_s \rightarrow \mu^+\mu^-)\) with pseudo-MC experiments, obtaining:

\[
B(B^0_s \rightarrow \mu^+\mu^-) = \left(3.21^{+0.96+0.49}_{-0.91-0.30}\right) \times 10^{-9} = \left(3.2^{+1.1}_{-1.0}\right) \times 10^{-9}.
\]

Statistical and systematic uncertainties (shown separately in the first expression, and combined in the last) are separated by repeating the likelihood fit after setting all systematic uncertainties to zero. The fitted branching fractions (obtained from the unconstrained likelihood maximum) are in all cases inputs to the Neyman constructions, yielding by design a physical range for the branching fractions extracted.

The 95% upper limit on \(B(B^0 \rightarrow \mu^+\mu^-)\) is determined with the same Neyman procedure, yielding:

\[
B(B^0 \rightarrow \mu^+\mu^-) < 4.3 \times 10^{-10}.
\]

For comparison, the predicted SM values are \(B(B^0_s \rightarrow \mu^+\mu^-) = (3.65 \pm 0.23) \times 10^{-9}\) and \(B(B^0 \rightarrow \mu^+\mu^-) = (1.06 \pm 0.09) \times 10^{-10}\) [1]. Assuming these predictions, the analysis expected results are \(\left(3.6^{+1.1}_{-1.0}\right) \times 10^{-9}\) for \(B(B^0_s \rightarrow \mu^+\mu^-)\) and an upper bound of \(7.1 \times 10^{-10}\) for \(B(B^0 \rightarrow \mu^+\mu^-)\).
Figure 9: (a) Likelihood contours for the simultaneous fit to $\mathcal{B}(B^0_s \to \mu^+ \mu^-)$ and $\mathcal{B}(B^0 \to \mu^+ \mu^-)$, for values of $-2\Delta \ln (L)$ equal to 2.3, 6.2 and 11.8, with the SM projection and its uncertainties are included. (b) Neyman contours in the plane $\mathcal{B}(B^0_s \to \mu^+ \mu^-)$, $\mathcal{B}(B^0 \to \mu^+ \mu^-)$ for 68.3%, 95.5% and 99.7% coverage. The red (inner side) contours are statistical uncertainty only, while the blue (outer side) set is including statistical and systematic uncertainties.

13 Combination with the Run 1 result

The likelihood of the current result is combined with the likelihood from the Run 1 result [15]. The only common parameters in the combination are the fitted $\mathcal{B}(B^0_s \to \mu^+ \mu^-)$ and the combination of external inputs $F_{\text{ext}} = \mathcal{B}(B^+ \to J/\psi K^+) \times \mathcal{B}(J/\psi \to \mu^+ \mu^-) \times \mu_{\psi^{''}} / \mu_{\psi} = (2.35 \pm 0.14) \times 10^{-4}$. Except for $F_{\text{ext}}$, all nuisance parameters are treated as uncorrelated between the two likelihoods, with both likelihoods including their individual parameterizations of systematic effects. A negligible change in the results is found when all sources of systematic uncertainty are assumed to be fully correlated.

The maximum of the combined likelihood is found for:

\[
\mathcal{B}(B^0_s \to \mu^+ \mu^-) = (2.8 \pm 0.7) \times 10^{-9}, \\
\mathcal{B}(B^0 \to \mu^+ \mu^-) = (-1.9 \pm 1.6) \times 10^{-10}.
\]

Figure 10 shows the likelihood contours for the combination with the Run 1 result for the simultaneous fit to $\mathcal{B}(B^0_s \to \mu^+ \mu^-)$ and $\mathcal{B}(B^0 \to \mu^+ \mu^-)$, for values of $-2\Delta \ln (L)$ equal to 2.3, 6.2 and 11.8, relative to the maximum of the likelihood. The contours for the Run 2 2015-2016 result are overlaid for comparison.

When applying the same 1D Neyman construction of section 12 to this combined likelihood, the 68.3% confidence interval for $\mathcal{B}(B^0_s \to \mu^+ \mu^-)$ obtained is:

\[
\mathcal{B}(B^0_s \to \mu^+ \mu^-) = (2.8^{+0.8}_{-0.7}) \times 10^{-9}.
\]

The maximum of the likelihood is unconstrained and allowed to access the unphysical (negative) region. The upper limit on $\mathcal{B}(B^0 \to \mu^+ \mu^-)$ is determined with the same Neyman procedure, yielding:

\[
\mathcal{B}(B^0 \to \mu^+ \mu^-) < 2.1 \times 10^{-10}.
\]
For comparison, the predicted SM values are $\mathcal{B}(B_s^0 \to \mu^+\mu^-) = (3.65 \pm 0.23) \times 10^{-9}$ and $\mathcal{B}(B^0 \to \mu^+\mu^-) = (1.06 \pm 0.09) \times 10^{-10}$ [1]. Assuming these predictions, the analysis expected results are $\left(3.6^{+0.9}_{-0.8}\right) \times 10^{-9}$ for $\mathcal{B}(B_s^0 \to \mu^+\mu^-)$ and an upper bound of $5.6 \times 10^{-10}$ for $\mathcal{B}(B^0 \to \mu^+\mu^-)$.

The variations of the combined and separate log-likelihoods are considered for the current and the full Run 1 result. The Run 1 and Run 2 results are 1.2 standard deviations apart, while the combined significance of the $B_s^0 \to \mu^+\mu^-$ signal is estimated to be 4.6 sigma, with the combined result 2.4 standard deviations apart from the SM value. These evaluations are obtained purely from the evaluation of likelihood ratios.

14 Conclusions

A study of the rare decays of $B_s^0$ and $B^0$ mesons into oppositely charged muon pairs is presented, based on 36.2 fb$^{-1}$ of 13 TeV LHC proton–proton collision data collected by the ATLAS experiment in 2015 and 2016.
For the $B_0^s$ the result is $\mathcal{B}(B_0^s \rightarrow \mu^+\mu^-) = (3.2^{+1.1}_{-1.0}) \times 10^{-9}$, where the uncertainty includes both the statistical and systematic components. The result is consistent with the analysis expectation in the SM hypothesis of $(3.6^{+1.1}_{-1.0}) \times 10^{-9}$.

For the $B^0$ an upper limit $\mathcal{B}(B^0 \rightarrow \mu^+\mu^-) < 4.3 \times 10^{-10}$ is placed at the 95% confidence level, with an expected upper bound of $7.1 \times 10^{-10}$ in the SM hypothesis. The limit is compatible with the SM prediction.

The present result is combined with the ATLAS Run 1 full dataset result, obtaining: $\mathcal{B}(B_s^0 \rightarrow \mu^+\mu^-) = (2.8^{+0.8}_{-0.7}) \times 10^{-9}$ and $\mathcal{B}(B^0 \rightarrow \mu^+\mu^-) < 2.1 \times 10^{-10}$.

All the results presented are compatible with the branching fractions predicted by the SM as well as the available experimental results.

References


LHCb Collaboration, Measurement of the $B^0_s \rightarrow \mu^+\mu^-$ Branching Fraction and Search for $B^0 \rightarrow \mu^+\mu^-$ Decays at the LHCb Experiment, Phys. Rev. Lett. 111 (2013) 101805, arXiv: 1307.5024 [hep-ex].

CMS and LHCb Collaborations, Observation of the rare $B^0_s \rightarrow \mu^+\mu^-$ decay from the combined analysis of CMS and LHCb data, Nature 522 (2015) 68, arXiv: 1411.4413 [hep-ex].

LHCb Collaboration, Measurement of the $B^0_s \rightarrow \mu^+\mu^-$ Branching Fraction and Effective Lifetime and Search for $B^0 \rightarrow \mu^+\mu^-$ Decays, Phys. Rev. Lett. 118 (2017) 191801, arXiv: 1703.05747 [hep-ex].


