Recording and reconstructing 10 billion unbiased b hadron decays in CMS

Robert Bainbridge[1], on behalf of the CMS Collaboration*

[1]Imperial College London, SW7 2AZ, UK

Abstract. The CMS experiment has recorded a high-purity sample of 10 billion unbiased b hadron decays. The CMS trigger and data acquisition systems were configured to deliver a custom data stream at an average throughput of 2 GB s^{-1}, which was "parked" prior to reconstruction. The data stream was defined by Level-1 and high level trigger algorithms that operated at peak trigger rates in excess of 50 and 5 kHz, respectively. New algorithms have been developed to reconstruct and identify electrons with high efficiency at transverse momenta as low as 0.5 GeV. The trigger strategy and electron reconstruction performance were validated with pilot processing campaigns. The accumulation and reconstruction of this data set, now complete, were delivered without significant impact on the core physics programme of CMS. This unprecedented sample provides a unique opportunity for physics analyses in the flavour sector and beyond.

1 Introduction

In recent years, a number of experimental results related to lepton universality tests in b hadron decays have yielded measurements [] that are in tension with expected values from the standard model. The cited measurements, performed by the BaBar [1], Belle [2], and LHCb [3] Collaborations, are for both $b\to s\ell\ell$ and $b\to c\ell\nu$ transitions and the individual measurements exhibit deviations in the range 2–4σ. Collectively, they may be the first indications of the violation of lepton flavour universality (LFU) [4, 5]. The confirmation of LFU violation would be a striking proof of the existence of physics beyond the SM. A key experimental observable R_{K} is defined by the double ratio

$$R_{\mathrm{K}} = \frac{\mathcal{B}[\mathrm{B}^+\to\mathrm{K}^+\mu^+\mu^-] \,/\, \mathcal{B}[\mathrm{B}^+\to\mathrm{K}^+(\mathrm{J}/\Psi\to)\mu^+\mu^-]}{\mathcal{B}[\mathrm{B}^+\to\mathrm{K}^+\mathrm{e}^+\mathrm{e}^-] \,/\, \mathcal{B}[\mathrm{B}^+\to\mathrm{K}^+(\mathrm{J}/\Psi\to)\mathrm{e}^+\mathrm{e}^-]}, \tag{1}$$

where the numerator and denominator are ratios of the branching fractions for the resonant $\mathrm{B}^+\to\mathrm{K}^+(\mathrm{J}/\Psi\to)\ell^+\ell^-$ and nonresonant $\mathrm{B}^+\to\mathrm{K}^+\ell^+\ell^-$ decays in the muonic (electronic) channel, respectively.[1] The R_{K^*} observable is similarly defined using the branching fractions for the resonant $\mathrm{B}^0\to\mathrm{K}^*\ell^+\ell^-$ and nonresonant $\mathrm{B}^0\to\mathrm{K}^*(\mathrm{J}/\Psi\to)\ell^+\ell^-$ decays. The R_{K} and R_{K^*} observables are known with high theoretical precision [] as a function of the 4-momentum transfer of the dilepton system, $q^2(\ell\ell)$, and thus are ideal probes for the presence of new-physics processes in rare decays due to $b\to s\ell\ell$ transitions.

*e-mail: cms-phys-conveners-BPH@cern.ch

[1]The $\mathrm{B}^{+(0)}\to\mathrm{K}^{(*)}\ell^+\ell^-$ and $\mathrm{B}^{+(0)}\to\mathrm{K}^{(*)}(\mathrm{J}/\Psi\to)\ell^+\ell^-$ notation implies charge conjugation.

Results related to lepton universality from the CMS experiment [6] are thus far limited: examples include the measurements of branching fractions for $B^0_{(s)} \to \mu^+\mu^-$ [7] and angular analyses of $B^0 \to K^*\mu^+\mu^-$ decays [8]. The CMS trigger system [9] comprises a two-tier system that enables algorithms on a level-1 (L1) subsystem of custom hardware processors and a software-based high-level trigger (HLT) subsystem that runs on a farm of processors. The system already implements the required algorithms to efficiently record samples of b hadron decays in muonic final states with high purity. However, there is no corresponding trigger logic that can be used to collect an adequate sample of $B^{+(0)} \to K^{(*)}e^+e^-$ decays. This limitation has thus far prevented the measurements of R_K and R_{K^*} by the CMS Collaboration.

A novel trigger and "B parking" strategy was deployed during the data taking period in 2018, which has enabled the accumulation and reconstruction of 10 B unbiased b hadron decays from which the measurements of R_K and R_{K^*} may be derived. The data streams that serve the core physics programme of CMS are promptly reconstructed at the CERN Tier0 data centre, and are generally available within 48 hours for physics analysis. The new data stream has a trigger rate of several kHz, which is beyond the standard processing capabilities of the Tier0 centre. However, the trigger and data acquisition (DAQ) systems have the ability to record nonstandard parked data streams to extend the CMS physics programme [10]. These data streams, typically defined by relaxed inclusive trigger requirements, are not processed immediately by the CMS reconstruction software. Instead, the data are temporarily stored in local buffers at Point 5 before being transferred—unprocessed—to permanent tape storage. These data streams are processed at a later point in time, e.g. during an end-of-year or long shutdown of the LHC. The parked data streams serve analyses with complementary or extended coverage (e.g. Ref. [11]) with respect to the core CMS physics programme.

This sample of unbiased b hadron decays, unprecedented in its size, provides a unique opportunity for the discovery of new-physics processes, in the flavour sector and beyond, and it is complementary to the high-p_T new-physics search programme of CMS. The trigger and B parking strategy, a new electron reconstruction algorithm, and some preliminary validation studies are described in the following sections.

2 Trigger strategy

The selection of $b\bar{b}$ events using a "tag-side" trigger logic in order to accumulate a sample of unbiased "signal-side" b hadron decays has been an important technique for analyses at B factories, LEP, and hadron colliders. The natural decay channels for the signal-side b hadron are unbiased by the trigger logic requirements imposed on the tag-side decay. The logic is based on the presence of a single muon, as semileptonic decays to muonic final states, $b \to (c \to) \mu X$, account for ≈20% of all b hadron decays.

In CMS, the same tag-side technique, coupled with existing trigger logic for muons, is used to record both the (signal-side) muonic and electronic final states required by the R_K and R_{K^*} measurements. The CMS trigger logic has been tuned to record $b \to (c \to) \mu X$ events with a purity of ≈80%, as described below. The $B^{+(0)} \to K^{(*)}\ell^+\ell^-$ decays have branching fractions of $\mathcal{O}(10^{-7})$. Assuming an acceptance times efficiency ($\mathcal{A}\epsilon$) of ≈10%, a large sample of $\mathcal{O}(10^{10})$ $b\bar{b}$ events is therefore required to obtain $\mathcal{O}(100)$ events containing $B^{+(0)} \to K^{(*)}\mu^+\mu^-$ or $B^{+(0)} \to K^{(*)}e^+e^-$ decays. The expected yield, $N(B^{+(0)} \to K^{(*)}\ell^+\ell^-)$, after the application of a muon-based L1 trigger algorithm during data taking in 2018 can be estimated by

$$N(B^{+(0)} \to K^{(*)}\ell^+\ell^-) = f_B \times \mathcal{B}[B^{+(0)} \to K^{(*)}\ell^+\ell^-] \times R_{L1} \times P_{L1} \times t_{LHC}, \quad (2)$$

where f_B is the fractional production rate of a particular type of b hadron relative to all b hadrons (e.g. 0.4 for B^0 and $B^\pm$); R_{L1} is the rate [kHz] of positive decisions by the L1 trigger

logic; P_{L1} is the purity of the event sample recorded by the L1 trigger logic, assumed here to be 0.3; and t_{LHC} is the duration of the data taking period in 2018, assumed to be 7.8×10^6 s (i.e. six months of LHC operation with a duty cycle of 50%). The branching fraction for $\mathrm{B}^+ \to \mathrm{K}^+\ell^+\ell^-$ ($\mathrm{B}^0 \to \mathrm{K}^*\ell^+\ell^-$) is 4.5 (6.7) $\times 10^{-7}$ []. Hence, assuming a L1 trigger rate of 10 kHz, the total number of events with a positive L1 decision that contain a signal-side $\mathrm{B}^+ \to \mathrm{K}^+\ell^+\ell^-$ ($\mathrm{B}^0 \to \mathrm{K}^*\ell^+\ell^-$) decay is estimated to be ≈4000 (≈6000).

The purity of the data stream is substantially improved through the use of tailored muon algorithms in the HLT. Studies have identified the two variables with the highest discriminating power to improve purity while maintaining acceptance to the signal processes: the muon p_{T} and the muon impact parameter (defined as the spatial distance between the primary pp collision and the muon at its point of closest approach), expressed in terms of its measurement significance, $\mathrm{IP}_{\mathrm{sig}}$. The latter variable leverages the lifetime of the $\mathrm{B}^{\pm(0)}$ meson and the characteristic displacement of the muon. The improved purity provided by the HLT algorithm is an important factor in controlling the total rate at which events are recorded by the CMS trigger system and written to tape.

The trigger strategy aims to maximise the number of $\mathrm{B}^{+(0)} \to \mathrm{K}^{(*)}\ell^+\ell^-$ events recorded during 2018 while ensuring that the ability of the CMS trigger and DAQ systems to deliver the core physics programme is unaffected. This is achieved by taking advantage of an increase in idle online computing resources as the instantaneous luminosity $\mathcal{L}_{\mathrm{inst}}$ decreases during each LHC fill. Specifically, as $\mathcal{L}_{\mathrm{inst}}$ decreases, the L1 and HLT trigger rates decrease, and the per-event processing load also decreases as a consequence of a reduced number of additional pp interactions within the same bunch crossing as the primary interaction (pileup).

Table 1 summarises the tag-side muon trigger requirements imposed by the L1 and HLT algorithms. The L1 logic requires the presence of a muon that satifies $|\eta| < 1.5$, which helps to control rate and also improves the acceptance for the signal-side $\mathrm{B}^{+(0)} \to \mathrm{K}^{(*)}\ell^+\ell^-$ decays. Both the L1 and HLT requirements are relaxed through a series of settings that progressively increase the rate at which the CMS trigger system results in a positive decision with only a moderate reduction in purity. The purity, estimated from simulation, is found to be in the range 0.59–0.92, with an average value of ≈0.75 that has been validated against data by reconstructing D^{*+} candidates from the decay $\mathrm{B}^0 \to \mathrm{D}^{*+}\mu\nu \to (\mathrm{D}^0\pi_{\mathrm{soft}})\mu\nu \to (\mathrm{K}^+\pi\pi_{\mathrm{soft}})\mu\nu$. The trigger rates of the L1 and HLT system peak at values of ≈50 kHz and 5.4 kHz, respectively. The highest rates are observed late in an LHC fill, which results in a pileup value of ≈20 when averaged over an entire LHC fill. This value is a factor ≈2 lower than that typically observed for the standard physics data streams of CMS.

Figure 1 shows the trigger rate of the CMS L1 system as a function of time during an LHC fill in 2017 (left) and 2018 (right). The left panel illustrates how the total rate decreases with time, as a consequence of the decreasing $\mathcal{L}_{\mathrm{inst}}$ during the LHC fill. The right panel illustrates

Table 1. Summary of the tag-side muon trigger requirements imposed by the L1 and HLT algorithms: the L1 and HLT muon p_{T} thresholds, and the HLT muon impact parameter significance $\mathrm{IP}_{\mathrm{sig}}$. Also shown are the trigger purity and peak trigger rate. All values are shown as a function of the peak $\mathcal{L}_{\mathrm{inst}}$. Table contents taken from Ref. [12].

Settings	Peak $\mathcal{L}_{\mathrm{inst}}$ [10^{34} cm^{-2}s^{-1}]	L1 μ p_{T} thresh. [GeV]	HLT μ p_{T} thresh. [GeV]	HLT μ $\mathrm{IP}_{\mathrm{sig}}$ threshold	Purity	Peak Rate [kHz]
1	1.7	12	12	6	0.92	1.5
2	1.5	10	9	6	0.87	2.8
3	1.3	9	9	5	0.86	3.0
4	1.1	8	8	5	0.83	3.7
5	0.9	7	7	4	0.59	5.4

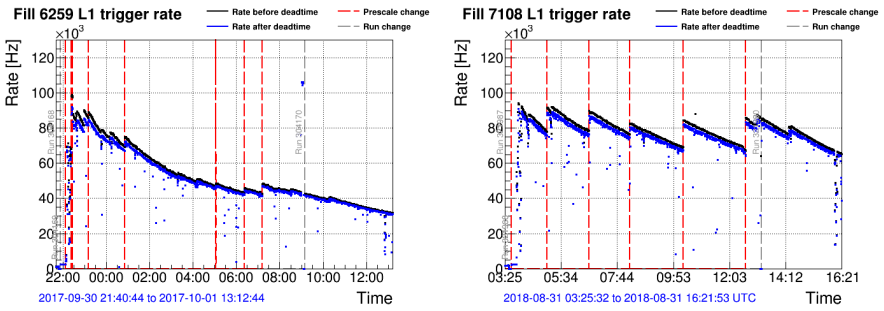

Figure 1. Rate of the CMS L1 trigger (blue data points), as a function of time, during an LHC fill in (**left panel**) 2017 and (**right panel**) 2018. The time intervals cover 13–15 hours. Changes in the run number and settings (prescale column) are indicated by vertical grey and red dashed lines, respectively. Taken from Ref. [12].

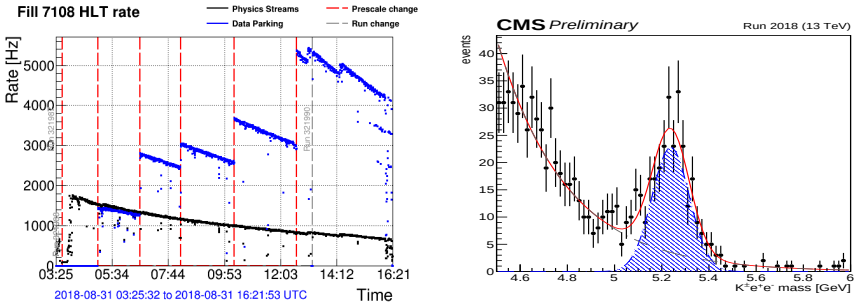

Figure 2. Left panel: the trigger rates of the CMS HLT system as a function of time during an LHC fill in 2018, for the physics (black data points) and B parking (blue data points) streams. **Right panel:** Invariant mass distribution for candidate $B^+ \to K^+(J/\Psi \to)e^+e^-$ decays, obtained from a small fraction of the B parking data sample. Taken from Ref. [12].

how the total rate is maintained close to the optimum value of ≈90 kHz by evolving the settings, as defined in Table 1. The left panel of Fig. 2 shows the trigger rates of the CMS HLT system as a function of time during an LHC fill in 2018, for both the standard physics and B parking streams. Sharp increases in the rate for the B parking stream occur throughout the LHC fill, as the settings are evolved, while the rate decreases monotonically for the standard physics data streams.

3 Data parking

The DAQ system is able to handle the additional load from the B parking stream up to a limitation determined primarily by the transfer of the from local storage buffers at Point 5 to tape resources available via the Tier0 centre. The trigger strategy outlined in Sec. 2 delivers a rate of ≈2 kHz when averaged over an LHC fill, which corresponds to a throughput of ≈2 GB s^{-1}. This throughput, when averaged over a timescale of several days, can be sustained without compromising the performance of the CMS DAQ system. The allocation of higher rates later in the LHC fills helps to load-balance the DAQ system.

Table 2. Summary of the expected yields for important signal-side b hadron production (and decay) modes in the B parking data sample. The quoted yields are based on the same assumptions used to evaluate Equation 2 in Sec. 2, and do not account for $\mathcal{A}\epsilon$. The branching fraction $\mathcal{B}$ is stated where appropriate, and f_B is defined in the text (Sec. 2). Table contents taken from Ref. [12].

Mode	Expected yield	f_B	$\mathcal{B}$
Generic b hadrons			
B^0_d	4.0×10^9	0.4	–
$B^\pm$	4.0×10^9	0.4	–
B_s	1.2×10^9	0.1	–
b baryons	1.2×10^9	0.1	–
B_c	1.0×10^7	0.001	–
Total	1.0×10^{10}	1.0	–
Events for R_K and R_{K^*} analyses			
$B^0 \to K^* \ell^+ \ell^-$	2600	0.4	6.6×10^{-7}
$B^+ \to K^+ \ell^+ \ell^-$	1800	0.4	4.5×10^{-7}

At the beginning of the LHC Run 2, CMS allocated tape resources to accommodate the parking of data (and a copy) at an average rate of ≈500 Hz during 2016, 2017, and 2018 to support the analysis of the scouting data stream [10]. The resources for 2017 and 2018 were reallocated to accommodate the new B parking proposal. Assuming a single copy, these resources are sufficient to permanently store the B parking data stream.

4 Event reconstruction and validation

The B parking sample was accumulated during the period June–November 2018. The sample comprises 12 B events, recorded with high purity triggers, and contains ≈10 B unbiased b hadron decays. The size of the single-copy unprocessed data sample is 7.6 PB. The reconstruction of the B parking sample occurred during the LHC long shutdown, in the period May–December 2019. The sample is permanently available as an analysis-level data format (MINIAOD) with a reduced footprint. Table 2 summarises the composition of the sample.

Approximately 7% of the data sample, enriched in dielectron final states from $b \to s\ell\ell$ transitions, is also temporarily available in the raw and AOD data formats, which permits further developments of algorithms and validation studies. A "pilot" reconstruction campaign, comprising a small fraction of the full data set, $O(1\%)$, was undertaken early in the data taking period to allow the validation of the trigger and parking strategies. The right panel of Fig. 3 shows the invariant mass distribution obtained from candidate $B^+ \to K^+(J/\Psi \to)e^+e^-$ decays using the standard CMS reconstruction software. This is the first observation from CMS of $b \to s\ell\ell$ transitions in the dielectron final state, obtained from the pilot campaign, which demonstrates the rich physics potential of the B parking sample. The trigger purity studies, based on the reconstructed D^{*+} candidates, were also based on the pilot campaign.

5 Electron reconstruction

A crucial component of the R_K and R_{K^*} measurements is the ability to efficiently identify electrons down to very low transverse momenta. The left panel of Fig. 3 shows the generator-level p_T distributions for the daughter particles from $B^+ \to K^+ \ell^+ \ell^-$ decays. The p_T distributions are very soft, with those for the kaon and subleading lepton peaking at ≈1 GeV. The right panel of Fig. 3 shows the efficiency to reconstruct electrons as a function of the generator-level p_T, as obtained with the CMS default electron reconstruction algorithm (blue square markers).

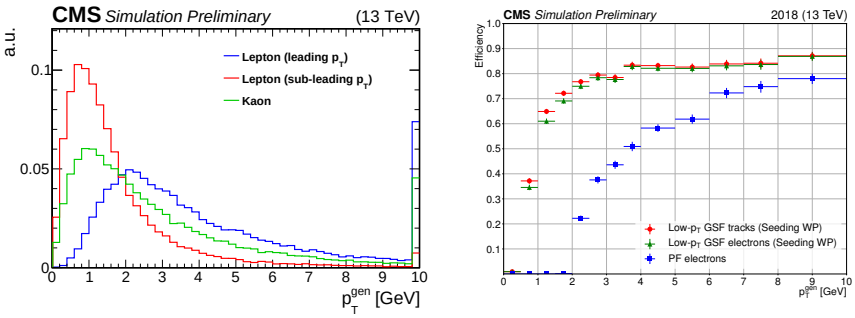

Figure 3. Left panel: Generator-level p_{T} distributions of the daughter particles from signal-side $\mathrm{B}^{+}\to\mathrm{K}^{+}\ell^{+}\ell^{-}$ decays. **Right panel:** Reconstruction efficiency curves for standard CMS electron candidates (blue squares) and for GSF tracks (red circles) and electron candidates (green triangles) from the new algorithm as a function of the generator-level electron p_{T}. Taken from Ref. [12].

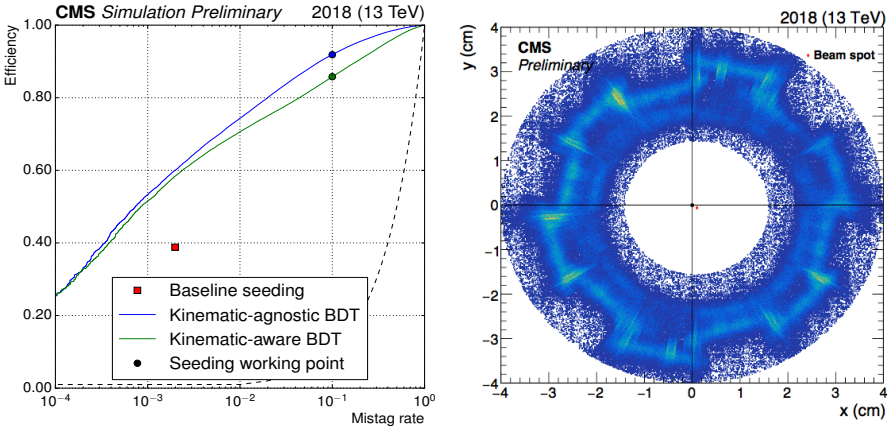

Figure 4. Left panel: ROC curves for a "kinematically agnostic" BDT (green curve) and a model-dependent "kinematically aware" BDT (blue curve), obtained from simulated $\mathrm{B}^{+}\to\mathrm{K}^{+}\mathrm{e}^{+}\mathrm{e}^{-}$ events. Various working points, described in the text, are also indicated. **Right panel:** Vertex positions of photon conversion candidates in the transverse plane for the region $|\eta| < 1$. Taken from Ref. [12].

The efficiency is essentially zero for the region $0 < p_{\mathrm{T}} < 2$ GeV and in the range 0.2–0.8 for the region $2 < p_{\mathrm{T}} < 10$ GeV.

A custom electron reconstruction algorithm, optimised for the low-p_{T} regime, has been developed for the B parking data set. As for the standard CMS electron algorithm, the determination of the charged-particle track parameters for electron candidates, in the presence of bremsstrahlung energy loss, relies on the use of a Gaussian sum filter (GSF) approach []. The "GSF tracking" stage is computationally expensive, and therefore it is seeded by a more computationally efficient logic that identifies potential electron candidates. The trajectory of each GSF track is used to identify a compatible "seed" cluster of energy in the CMS electromagnetic calorimeter. Additional clusters of energy, consistent with the bremsstrahlung energy loss pattern of the electron candidate, are associated with the seed cluster as part of a "super cluster", which can be used with the tracking information to identify genuine electron

candidates with high efficiency and purity. The right panel of Fig. 3 illustrates the increase in efficiency obtained with the new electron reconstruction algorithm with respect to the standard algorithm with only minimal identification quality criteria applied.

The seeding logic implements two independent models based on boosted decision trees (BDT). The first BDT provides signal-to-background discrimination based on a "kinematically agnostic" approach that exploits only tracking and calorimeter information. The second BDT provides a (model-dependent) "kinematically aware" model that also uses the p_{T}, η, and track impact parameter of an electron candidate to discriminate signal from background. The left panel of Fig. 4 shows the ROC curves obtained for the two BDTs based on simulated $\mathrm{B}^{+}\to\mathrm{K}^{+}\mathrm{e}^{+}\mathrm{e}^{-}$ events. A loose working point is defined for each BDT that yields a 10% mistag rate while providing a factor $\approx$2 gain in efficiency over that obtained from the baseline seeding logic of the standard CMS algorithm. These working points were used to seed the new electron reconstruction sequence as part of the reconstruction campaign described in Sec. 4 and the electrons are available for analysis in the MINIAOD data format.

A large, high purity sample of electrons with $0.5 < p_{\mathrm{T}} < 10$ GeV can be obtained from converted photons resulting from interactions with the beam pipe and inner tracking structures. This sample is being used to study and tune the identification algorithm for low-p_{T} electrons. The right panel of Fig. 4 shows the vertex positions of photon conversion candidates in the transverse plane for the region $|\eta| < 1$. The structures of the beam pipe and inner layer of the CMS pixel barrel subdetector are clearly visible.

6 Summary

The CMS experiment has recorded and reconstructed a high-purity sample of 10 billion unbiased b hadron decays. This sample was recorded with minimal impact on the core CMS physics programme, as the strategy exploited the use of existing infrastructure, trigger algorithms, and idle resources available during the latter part of LHC fills. The data stream was parked during 2018 and processed during 2019. A new electron reconstruction algorithm was deployed as part of the processing campaign, which provides the potential for highly efficient electron identification at transverse momenta as low as 0.5 GeV. This unprecedented sample provides a unique opportunity for physics analyses in the flavour sector and beyond.

References

[1] BaBar Collaboration, Nucl. Instrum. Meth. A **479** 1 (2002)

[2] Belle Collaboration, Nucl. Instrum. Meth. A **479** 117 (2002)

[3] LHCb Collaboration, JINST **3** S08005 (2008)

[4] HFLAV Collaboration, https://hflav-eos.web.cern.ch/hflav-eos/semi/spring19/r_dtaunu/rdrds_spring2019.pdf (2019)

[5] W. Altmannshofer *et al*, Phys. Rev. D **96** 055008 (2017)

[6] CMS Collaboration, JINST **3** S08004 (2008)

[7] CMS and LHCb Collaborations, Nature **522** 68 (2015)

[8] CMS Collaboration, Phys. Lett. B **781** 517 (2018)

[9] CMS Collaboration, JINST **12** P01020 (2017)

[10] CMS Collaboration, CMS-DP-2012-022, https://cds.cern.ch/record/1480607 (2012)

[11] CMS Collaboration, Phys. Lett. B **767** 403 (2017)

[12] CMS Collaboration, CMS-DP-2019-043, https://cds.cern.ch/record/2704495 (2019)