EPFL LPHE

Testing lepton flavour universality in $B \to K\pi\pi\ell\ell$ decays using 2017 LHCb data

A master's thesis written by

Géraldine Räuber

under the supervision of

Prof. Lesya Shchutska

and

Dr. Elena Graverini

submitted on

January 14, 2021

Contents

Abstract **1**

Introduction **1**

1 Probing the Standard Model's limits **3**
1.1 Lepton flavour universality tests . . . 3
1.2 $B \to K\pi\pi\ell\ell$ decays and $R_{K\pi\pi}$ measurement . . . 5

2 The LHCb experiment **8**
2.1 Detector and trigger system . . . 8
2.2 Data and Monte Carlo simulations . . . 12

3 Reweighting of Monte Carlo simulation **16**
3.1 ${}_s\mathscr{P}lot$ technique and background-subtracted data . . . 16
3.2 Resampling of particle identification variables . . . 19
3.3 Reweighting of the pp event and B^+ kinematics . . . 20

4 Event classification with multivariate analysis **21**
4.1 MVA classification methods . . . 21
4.1.1 Algorithms . . . 22
4.1.2 Performance assessment . . . 24
4.2 Test and training sets of MVA method . . . 25
4.3 Discriminating features of MVA method . . . 27
4.3.1 Selection of relevant features . . . 27
4.3.2 Features score . . . 30
4.4 MVA responses . . . 35
4.5 Comparison with previous MVA responses . . . 37
4.6 Optimisation of classifier output . . . 40

Conclusions and outlook **45**

References **48**

Appendix **52**

A Background and signal PDFs used in the ${}_s\mathscr{P}lot$ technique **52**

B *s*Fit intermediate plots **54**

C PID variables resampling **57**

D GBR hyperparameters **60**

E Kinematics and multiplicity variables plots **61**

F General plots of features score **63**

G Background and signal distributions **69**

H Cross-checks **74**

I Comparison of MVA performances **77**

Abstract

This present work proposes a study on $B \to K\pi\pi\ell\ell$ decays using 2017 LHCb data in the context of Lepton Flavour Universality tests. In view of the determination of the ratio of branching fractions $R_{K\pi\pi}$, that complements the actual investigations around the Standard Model shortcomings, several intermediate computations have been carried out. A previous work [1] provided two corrections, concerning the particle identification, and the kinematics and multiplicity variables, associated to the Monte Carlo simulation. These are useful in the accurate separation of events from a background of incorrectly reconstructed candidates; a task accomplished by using the boosted decision tree (BDT), a multivariate analysis (MVA) method, trained and tested on specific data sets defined for a selected set of discriminating features. The resulting classifier output, providing the classification, has an optimised threshold determined by a figure of merit that yields a signal efficiency of up to 82.1% while removing more than 92.5% of background.

Introduction

Developed in stages as a monumental collaborative effort of generations of scientists, and with the current formulation being finalised in the latter half of the 20th century, the Standard Model (SM) still stands out as the theory describing matter and forces at the most fundamental level, backed with large-scale experiments. The SM contains twelve spin-$\frac{1}{2}$ particles, as well as their corresponding antiparticles, know as fermions, four spin-1 particles, known as gauge bosons, and one spin-0 particle known as the famous Higgs boson. Fermions, which are matter particles, are divided into two categories: quarks and leptons. There are three generations of quarks (u, d, c, s, t, b) and three flavours of leptons (e^-, ν_e, μ^-, ν_μ, τ^-, ν_τ). Bosons, on the other hand, are forces mediator. The photon (γ) mediates the electromagnetic interaction, the weak bosons $W^\pm$ and Z the weak interaction and the gluon g the strong interaction. The quarks interact under the electromagnetic, weak and strong forces, whereas the charged leptons interact only with the first two and the neutrinos only interact weakly.

Yet, as particle accelerators worldwide perform more and more ambitious experiments, the SM's limits slowly but surely reveal themselves, allowing for the uncharted territory of New Physics (NP) to be modestly scouted, part by part, experiment by experiment. One of these concerns a pillar of the Standard Model, taken for granted until then and therefore left aside for years: the lepton flavour universality (LFU). This prediction, which stands out from the others by its simplicity, contains a key assumption; It assumes that the three lepton flavours are identical with the exception of phase-space effects related to their different masses. This means that the weak bosons are expected to couple to them with the same strength. One way to test this prediction is to use rare decays where effects of NP could more easily emerge.

As $b \to s\ell^+\ell^-$ transitions are suppressed in the SM, because they require a flavour changing neutral current (FCNC), they represent an appropriate choice of decay channel. This implies that NP effects can be sizeable compared to its SM amplitudes, which would allow the discovery of new virtual mediators. Moreover, because they would appear virtually, it would be possible to probe heavier scales than those accessible via direct searches.

However, the use of rare decays also brings its share of difficulties, especially in terms of uncertainties related to the quantity we want to compute. This is the main reason why it is essential to focus on LFU-related computations where the uncertainties are

reduced to some extent. In particular, one computation demonstrated that this issue may be overcome: the ratio of branching fraction. It effectively establishes a relationship between two branching fractions that describe each the probability that a particle chooses a specific decay mode among all the others.

Recent measurements of ratio of branching fractions demonstrated a consistent pattern of anomalies. In fact, the ratios R_K and $R_{K^{*0}}$ determined by the LHCb experiment [2] [3], a forward spectrometer studying particles containing heavy flavours at the Large Hadron Collider (LHC) at CERN, revealed 2.1–2.3 and 2.4–2.5 σ deviations from the values predicted by the SM. Other measurements lead to a similar conclusion, e.g. branching fractions with muons measured to be consistently below the SM computation. Deviations from SM predictions are also observed in both angular observables and charged current $b \to c\ell\nu$. But none of these observations alone possesses the significance to claim a discovery, which prompts us to look further in this direction using other particle systems to eventually achieve a first evidence, and, possibly, an observation of NP phenomena.

In any case, the computation of the ratio of branching fractions is not a trivial task. One of the difficulties arises from the Monte Carlo simulations. These simulations, relying on statistical methods to obtain numerical results, are commonly used in particle physics experiments. They are particularly useful when no other approaches can solve problems that are usually deterministic, or when systems with many coupled degrees of freedom should be simulated. But even with the most advanced technologies, their computation inevitably presents imperfections. A combination of several considerations allows us to make continuous corrections as to rectify their inaccuracies, and strive for perfection, in order to perform high precision measurements.

In this analysis, $B \to K\pi\pi\ell\ell$ decays are considered in the context of computation of ratios of branching fractions. Section 1 introduces the main concept surrounding this computation, and describes the strategy put in place for the reweighting of the Monte Carlo simulation and the identification of the combinatorial background in data. The LHCb experiment, where the data were collected in 2017, is meticulously detailed in Section 2, from the detectors constituting the detection system to the data sets production. Then, a brief overview of part of the corrections applied on Monte Carlo simulations done in a previous work [1] is given in Section 3. These results are useful for the classification described in Section 4. The latter is performed using machine learning techniques to predict the signal- or background-likeliness of an event. A comparison with previous results [4] is discussed and an optimisation of the resulting classifier ouput is carried out to select events classified as signal.

1 Probing the Standard Model's limits

1.1 Lepton flavour universality tests

The Standard Model (SM) represents a synthesis of our understanding of the properties and interactions of elementary particles; this theory describes three of the four known fundamental forces in the universe, and it classifies all known elementary particles (as depicted in Figure 1), including all flavours of leptons and quarks, and the gauge and scalar bosons.

Although the SM has demonstrated great successes in providing experimental predictions, it leaves some phenomena unexplained, such as neutrino oscillations and their non-zero masses or the existence of dark matter. These few shortcomings motivate an intensive examination of the self-consistency of the SM in other circumstances in order to decide if the SM needs to be complemented with New Physics (NP).

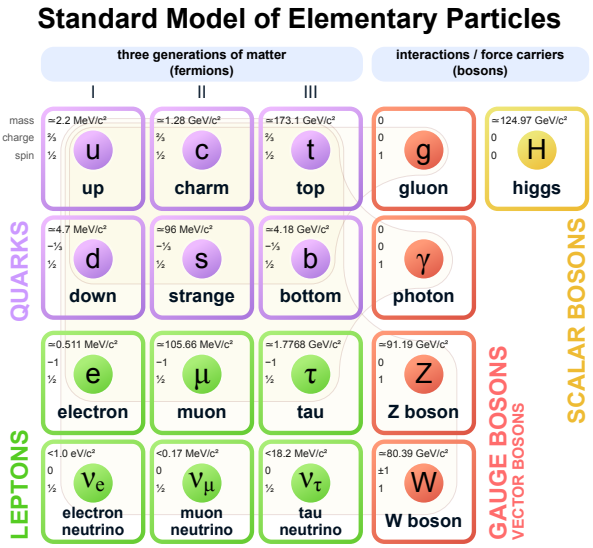
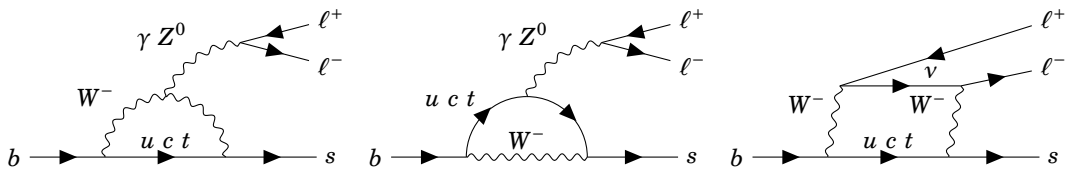


Figure 1: Elementary particles of the Standard Model [5].

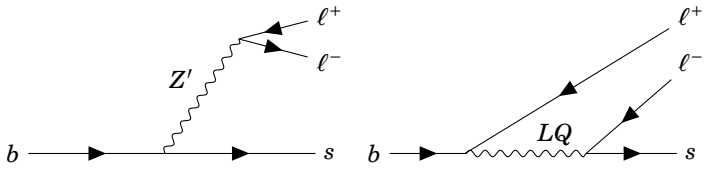
Such a thorough investigation forces us to question the fundamental elements of the SM which have, for the most part, been taken for granted until recently. There is one particular feature of the SM that stands out for its simplicity: the lepton flavour universality (LFU). This assumption states that the three lepton flavours are identical, except for their mass, which implies that the weak bosons are expected to couple to them with the same strength.

There are several possible approaches that are used to increase the probability of observing NP phenomena, such as high-intensity beam dump or direct searches in high-energy collisions. Another possibility is to use flavour-changing neutral currents (FCNC) as they are suppressed in the SM, revealing the NP effects. Moreover, as NP particles would appear virtually, it is possible to probe heavier NP scales than those accessible via direct searches.

These currents proceed only though loop-level in the SM, in decays including a flavour transition. In this thesis, $b \to s\ell^+\ell^-$ transitions are examined. Figure 2 displays Feynman diagrams of these transitions according to both the SM and NP predictions.



(a) Standard Model



(b) New physics

Figure 2: Lowest order Feynman diagrams of $b \to s\ell^+\ell^-$ transition according to (a) the Standard Model and (b) New Physics. The Z' boson arises from extensions of the electroweak symmetry of the SM, and the leptoquark LQ may be encountered in various extensions of the SM, such as technicolor theories or GUTs.

The branching fraction $\mathscr{B}$ of such transitions, all known with precision of the order of the percent, are in fact in the order of 10^{-7} to 10^{-6}. Given these small values, it is therefore essential, when performing LFU tests, to focus on quantities where the uncertainties are reduced to some extent, such as angular observables or ratios of branching fractions. Assuming X_s being a generic particle system with a strange quark, e.g. K, K^{*0}, or $K\pi\pi$, the latter, denoted as R_{X_s}, is computed as:

$$R_{X_s} = \frac{\mathscr{B}(B \to X_s \mu^+ \mu^-)}{\mathscr{B}(B \to X_s e^+ e^-)} \overset{\text{SM}}{=} 1 \tag{1}$$

This assumption was tested in direct and indirect measurements so far for two kinds of decays containing a $b \to s\ell^+\ell^-$ transition: $B^+ \to K^+\ell^+\ell^-$ and $B^0 \to K^{*0}\ell^+\ell^-$. Their respective branching fractions R_K and $R_{K^{*0}}$ were measured by different high energy experiments. These are presented in Figure 3, where a good compatibility between the values reported by the LHCb [2] [3], the Belle [6] [7] and the BaBar [8] collaborations can be seen.

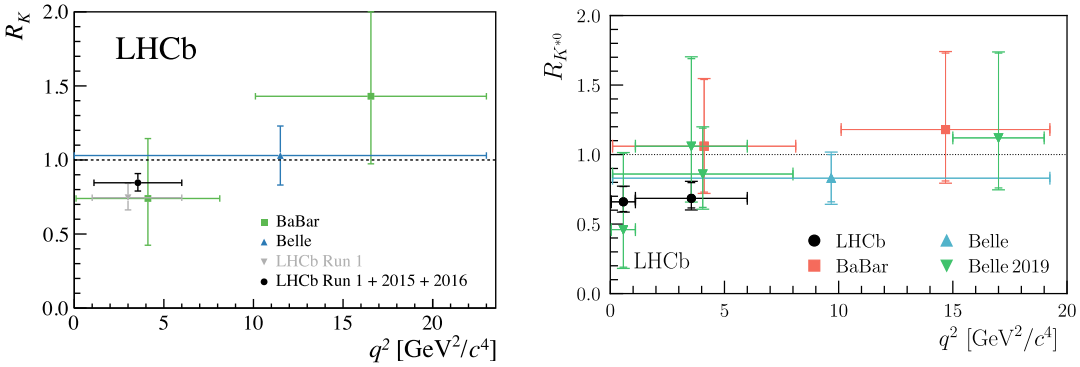


Figure 3: Results reported by different high energy experiments (LHCb [2], [3]) (Belle [6], [7]) (BaBar [8]). Evidence of a 2.1–2.3 and 2.4–2.5 σ deviations from the values predicted by the SM observed for $X_s = K$ and K^{*0} by the LHCb Collaboration.

The LHCb measurement of R_K is of major interest. Its value, which corresponds to

$$R_K = 0.846^{+0.060}_{-0.054}\,(\text{stat})\,^{+0.016}_{-0.014}\,(\text{syst}) \quad 1.1 < q^2 < 6\ \text{GeV}^2/c^4 \tag{2}$$

shows small uncertainties that raise a tension within the SM given its central value. This deviation, quantified as 2.5σ deviation, is below the threshold for a first evidence but still participate in the search of anomalies. A similar conclusion can be drawn for the LHCb measurement of $R_{K^{*0}}$ in two q^2 regions, given by the following:

$$R_{K^{*0}} = \begin{cases} 0.66^{+0.11}_{-0.07}\,(\text{stat}) \pm 0.03\,(\text{syst}) & 0.045 < q^2 < 1.1\ \text{GeV}^2/c^4 \\ 0.69^{+0.11}_{-0.07}\,(\text{stat}) \pm 0.05\,(\text{syst}) & 1.1 < q^2 < 6\ \text{GeV}^2/c^4 \end{cases} \tag{3}$$

These measurements, which nevertheless do not have the significance to claim a discovery, point to a coherent pattern of anomalies that could soon turn into the first observation of physics beyond the SM. For this purpose, other particle systems have to be explored within the scope of branching fractions measurements. And for that, $B \to K\pi\pi\ell\ell$ decays seem to be a good step in this direction.

1.2 $B \to K\pi\pi\ell\ell$ decays and $R_{K\pi\pi}$ measurement

In this analysis, we decided to study the $B \to K\pi\pi\ell\ell$ rare decay, illustrated in Figure 4, with the aim of computing the ratio of branching fractions $R_{K\pi\pi}$.

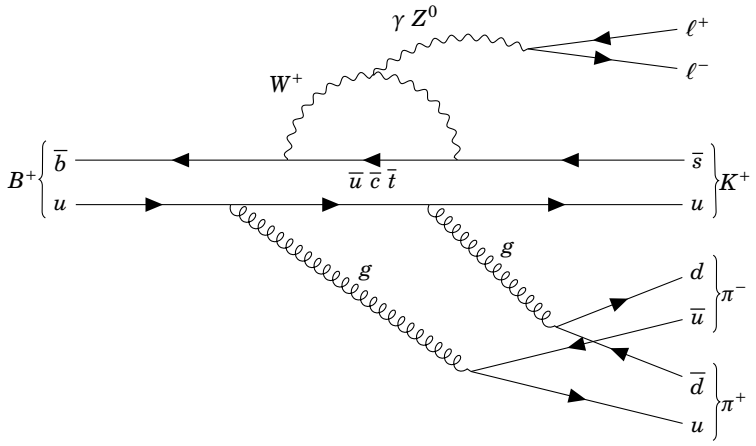


Figure 4: Example of lowest order Feynman diagram of the rare decay $B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$ according to the SM.

This decay includes three mesons in its final state. This specificity allows the production of complex spin structures that can be further studied. Indeed, if lepton flavour non-universality is observed, it will give further insight on how the new particles responsible for breaking this universality couple with SM particles of different spins. It is also very rich in resonances, such as K_1, K_2, $K^*\pi$, $K^+\rho$, ω, η [9], that may end up in the decay final state because the measurement is inclusive.

$B^+ \to K^+\pi^+\pi^-\mu^+\mu^-$ has already been observed by the LHCb collaboration [10] as demonstrated in Figure 5.

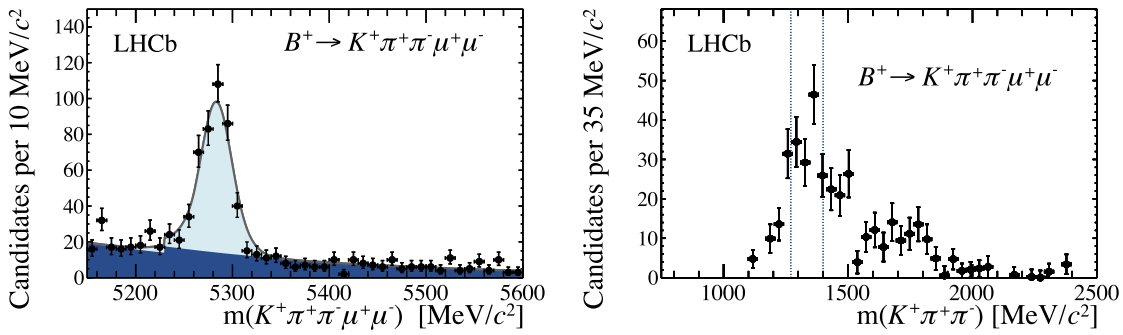


Figure 5: Observation of the $B^+ \to K^+\pi^+\pi^-\mu^+\mu^-$ decay at LHCb [10].

The branching fraction computed by the LHCb collaboration [10] for this decay using data taken during the LHC Run 1 corresponding to an integrated luminosity of 3.0 fb^{-1} and centre-of-mass energies of 7 and 8 TeV is given by

$$\mathcal{B}(B^+ \to K^+\pi^+\pi^-\mu^+\mu^-) = \left(4.36^{+0.29}_{-0.27}\ (\text{stat}) \pm 0.21\ (\text{syst}) \pm 0.18\ (\text{norm})\right) \cdot 10^{-7} \tag{4}$$

It is worth emphasising that the computation of $R_{K\pi\pi}$ would also provide the first observation of $B^+ \to K^+\pi^+\pi^-e^+e^-$. By using Equation (5), $R_{K\pi\pi}$ is defined as:

$$\begin{aligned} R_{K\pi\pi} &= \frac{\mathcal{B}(B^+ \to K^+\pi^+\pi^-\mu^+\mu^-)}{\mathcal{B}(B^+ \to K^+\pi^+\pi^-e^+e^-)} \\ &= \underbrace{\frac{N_{(B^+\to K^+\pi^+\pi^-\mu^+\mu^-)}}{N_{(B^+\to K^+\pi^+\pi^-e^+e^-)}}}_{\text{Yields}} \cdot \underbrace{\frac{\varepsilon_{(B^+\to K^+\pi^+\pi^-e^+e^-)}}{\varepsilon_{(B^+\to K^+\pi^+\pi^-\mu^+\mu^-)}}}_{\text{Efficiencies}} \end{aligned} \tag{5}$$

where N represents the yields, and ε the efficiencies. To suppress a majority of systematic uncertainties due to the event reconstruction [2], a double ratio using a decay with the exactly same topology as for the signal channel is preferred. In this analysis, we use the resonant decay $B \to K\pi\pi(J/\psi \to \ell\ell)$, illustrated in Figure 6, and the computation of $R_{K\pi\pi}$ becomes:

$$\begin{aligned} R_{K\pi\pi} &= \frac{\mathcal{B}(B \to K\pi\pi\mu\mu)}{\mathcal{B}(B \to K\pi\pi(J/\psi \to \mu\mu))} \cdot \frac{\mathcal{B}(B \to K\pi\pi(J/\psi \to ee))}{\mathcal{B}(B \to K\pi\pi ee)} \\ &= \underbrace{\frac{N_{(B\to K\pi\pi\mu\mu)}}{N_{(B\to K\pi\pi(J/\psi\to\mu\mu))}} \cdot \frac{N_{(B\to K\pi\pi(J/\psi\to ee))}}{N_{(B\to K\pi\pi ee)}}}_{\text{Yields from simultaneous fit}} \cdot \underbrace{\frac{\varepsilon_{(B\to K\pi\pi(J/\psi\to\mu\mu))}}{\varepsilon_{(B\to K\pi\pi\mu\mu)}} \cdot \frac{\varepsilon_{(B\to K\pi\pi ee)}}{\varepsilon_{(B\to K\pi\pi(J/\psi\to ee))}}}_{\text{Efficiencies from corrected simulation}} \end{aligned} \tag{6}$$

Several advantages have led to the use of this decay in particular. The reconstruction of events is possible because the final state is the same in both decays; The only way to distinguish the resonant from the rare decay is by calculating the invariant dilepton mass squared $q^2 = m^2(\ell\ell)$. Unlike the rare decay, it does not involve FCNC, resulting in a much more abundant decay. Its use for the validation and the correction of the Monte Carlo simulations, thus allowing a proper computation of the efficiencies, is therefore evident.

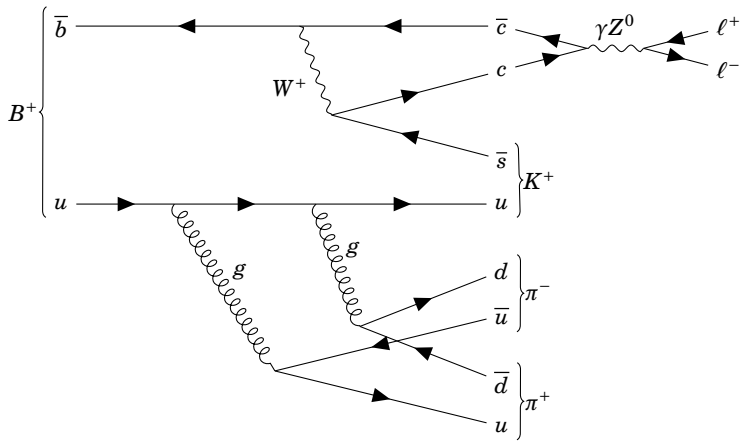


Figure 6: Example of lowest order Feynman diagram of the resonant decay $B^+ \to K^+\pi^+\pi^-(J/\psi \to \ell^+\ell^-)$ according to the SM. The J/ψ meson is represented by the charm quark pair $c\bar{c}$.

The analysis procedure applied to obtain $R_{K\pi\pi}$ is not trivial and involves several successive measurements. We present in this analysis the results obtained for the reweighting of the Monte Carlo, covered in Section 3, and the identification of the combinatorial background in data, described in Section 4. The reweighting part mainly refers to a previous work [1] based on a correction chain used for the analysis of 2011-2016 LHCb data [11], which is similar to the strategy opted for the R_K and $R_{K^{*0}}$ measurement. The second part focuses on the separation of signal and background in real data. It uses a machine learning method to establish an event classification, and cross-checks the performance of a previous work [4] that used a method developed when only Run 1 data was available. It also employs, among other inputs, a set of discriminating features determined using an algorithm specifically designed for this analysis.

In order to complete the analysis, the remaining steps of the correction chain and several cross-checks, which will be addressed in subsequent work, need to be carried out. One of the latter correspond to the verification that the value of the single-ratio measurement, known to be lepton universal with a precision of 5 ‰, denoted as $r_{(J/\psi)}$ and computed as:

$$r_{(J/\psi)} = \frac{\mathscr{B}(B \to K\pi\pi(J/\psi \to \mu\mu))}{\mathscr{B}(B \to K\pi\pi(J/\psi \to ee))} = 1 \tag{7}$$

is indeed equal to 1. And only then, the ensuing systematic uncertainties will be estimated, and the measurement of $R_{K\pi\pi}$ performed.

2 The LHCb experiment

The LHCb experiment is a specialised b-physics experiment collecting data generated by pp collisions produced in the Large Hadron Collider (LHC) at CERN. This facility has already operated several runs of data collections, denoted as Run 1 (2009-2013) and Run 2 (2015-2018). It is currently upgraded for its next run, known as Run 3, that will span from May 2021, to the end of 2024. The LHCb detection system is designed primarily to measure the parameters of CP violation in the interactions of b-hadrons. The different detectors constituting it as well as the trigger system are detailed in Subsection 2.1. Then, the data and the Monte Carlo simulations of the particular decay $B \to K\pi\pi\ell\ell$ are described in Subsection 2.2.

2.1 Detector and trigger system

Unlike the other experiments at CERN, the LHCb experiment uses a series of detectors that are placed along the beam pipe and optimised to detect forward particles. This particular design is greatly motivated by the predicted angular distribution of b quark production at the LHC that fall within the LHCb acceptance of $2 < \eta < 5$ [12]. The first detector is mounted a few millimetres from the pp collision point, with the others following one behind the other over a length of 19 meters [12], as illustrated in Figure 7.

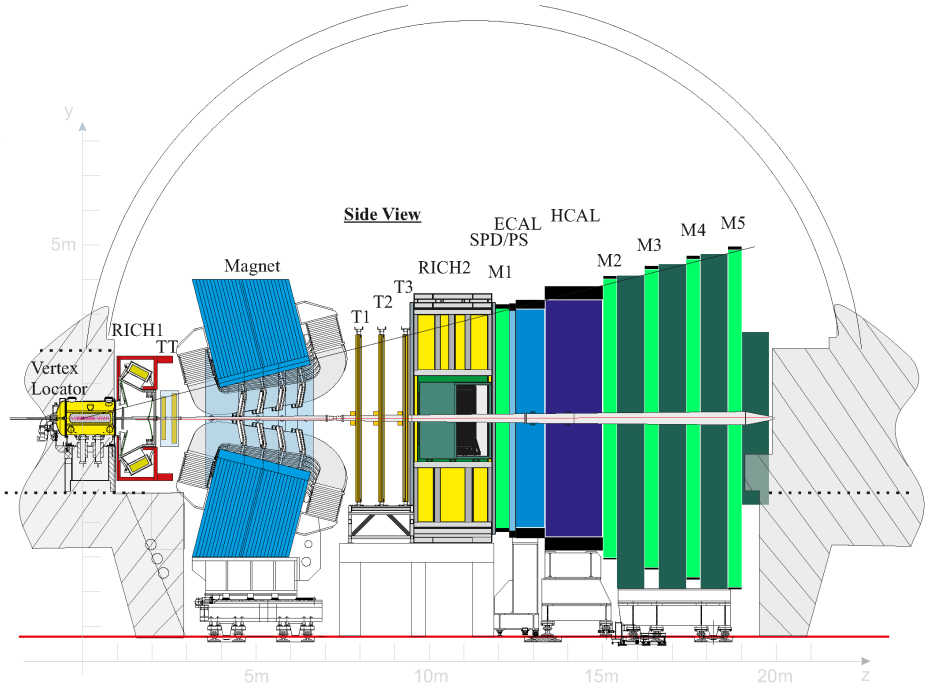


Figure 7: Illustration of the LHCb detection system shown side on, with the different detectors highlighted along the beam pipe [12].

The detection system comprises the following detectors, described from the nearest to the farthest from the pp interaction point.

Vertex Locator (VELO)

The Vertex Locator (VELO) is a silicon microstrip detector that surrounds the pp interaction point. By measuring the track coordinates with a resolution of a few tens of μm [13], this detector aims at separating the primary vertex (PV) defined as the pp interaction point from the secondary vertex (SV) defined as the B displaced decay vertex. Because decays of B mesons are typically characterised by their relatively long lifetimes, this detector provides a key parameter for the identification of B decays.

Ring imaging Cherenkov detectors (RICH1 and RICH2)

The Ring imaging Cherenkov (RICH) detectors are divided into two detection points: RICH-1 and RICH-2. The first detector is located right after the vertex locator and is used for particle identification of low-momentum tracks, whereas the latter is located after the tracking system and allows the identification of high-momentum tracks [12]. The physics principle underlying the RICH detectors is known as the Cherenkov effect: when a charged particle with a velocity v travels faster than the speed of light c in a medium with a reflective index n called radiator, it emits light. The angle θ_c at which the light is emitted is given by

$$\cos\theta_c = \frac{1}{n\frac{v}{c}} \equiv \frac{1}{n\beta}$$

In order to match as much as possible the momentum spectrum of particles from B decays at LHCb, fluorocarbon gases C_4F_{10} and CF_4 are used as radiators for RICH-1, respectively RICH-2 [12].

These detectors play a significant role in particle identification (PID) in general, because combining the Cherenkov angle θ_c, and therefore the particle speed v, with the momentum p allows the determination of the rest mass m of the particle thus establishing its identity.

Magnet

Even though the magnet is not strictly considered as a detector, it also plays a role in PID. It consists of two coils that bend the charged particles trajectory with a power of 4 Tm [12] in the up or down direction, and allows the tracking system to measure their momentum p, which is an essential parameter. It is placed after the RICH1 detector, in between the tracking system.

Tracking system (TT, IT and OT)

The tracking system comprises two detectors: the Tracker Turicensis (TT) located in between RICH1 and the LHCb magnet, which is a large-surface silicon microstrip detectors, and the Inner Tracker (IT), which is the inner part of the Outer Tracker (OT), that plays a major role in the downstream tracking. Its layers are made of straw tubes whose inner parts are composed of silicon strips. The IT and OT are combined together in all T1, T2, T3 detectors, that are shown in Figure 7. The whole system is used to reconstruct the trajectories of charged particles and to measure their momenta p. TT, IT and OT are being replaced for the LHC Run 3 with another silicon tracker called UT and a SciFi Tracker, respectively for TT and IT.

Preshower (PS) and Scintillating Pad Detector (SPD)

Before both calorimeters there is a double detector made of three layers: the Scintillator Pad Detector (SPD), a 2.5 radiation lengths lead wall, and the Preshower (PS) [14]. The PS provides geometrical information to the calorimeter, required to match the tracks and the clusters in the calorimeter. The SPD helps the calorimeter on the requirement of good background rejection and reasonable efficiency on the detection of photons with enough precision to enable the reconstruction of B-decay channels containing a prompt photon or π^0 and on the electron identification. It identifies charged particles, which allows electrons to be separated from photons.

Electromagnetic (ECAL) and Hadronic (HCAL) Calorimeters

Calorimeters are usually placed at the end of the detection system. Their alternating structure made of metal and plastic plates is designed to stop particles as they pass through the detector, measuring the amount of UV light produced during their shower process to get their energy.

There are two types of calorimeters. The electromagnetic calorimeter is responsible solely for electrons and photons, whereas the hadronic calorimeter takes care of hadrons, such as protons or neutrons. Their difference lies in the fact that the stopping power needed for light and heavy particles is not the same, leading to different detector conceptions. Another thing that can be added is that these detectors are useful for identifying neutral particles, such as photons and neutrons.

Muon system (M1, M2, M3, M4 and M5)

Because muons are minimum interacting particles at GeV energies, they interact very little with matter, which makes them penetrate far deeper into the detector. Therefore, muons detectors are placed at the end of the detection system, and comprise around a thousand multi-wire proportional chambers. Muon triggering and offline muon identification are fundamental requirements of the LHCb experiment because muons are present in the final states of many CP-sensitive B decays. They play a major role in CP asymmetry and oscillation measurements, since muons from semi-leptonic b decays provide a tag of the initial state flavor of the accompanying neutral B mesons [15].

This meticulous detection system allows the measurement of sensible features related to particles produced by a phenomenal number of collisions. But not only this number is unreachable in terms of data processing, there is only a small percentage of all decays measured by the LHCb detector that are expected to produce the decays of interest. It is therefore essential to set up a trigger system capable of identifying potentially interesting events and setting them aside for offline analysis. Figure 8 illustrates the LHCb trigger strategy put in place for Run 1 and Run 2 data, and the one foreseen for Run 3 data.

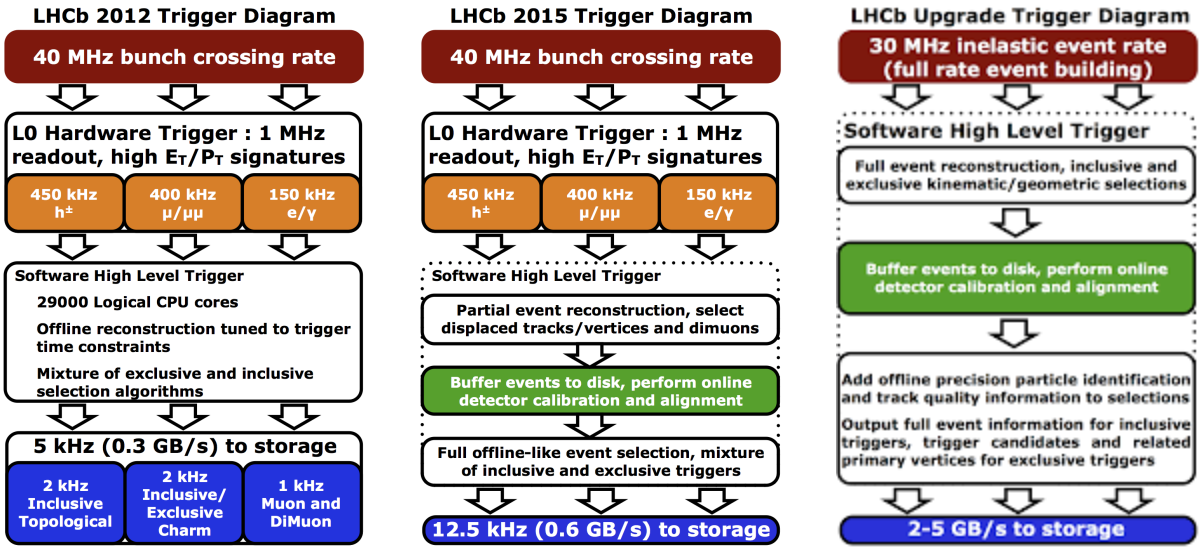


Figure 8: Diagram of the LHCb trigger structure used for Run 1 (left) and Run 2 (center), and the one foreseen for Run 3 (right). [16].

To do this crucial task, three main triggers acting on two different stages are used: the level 0 trigger (L0) implemented in hardware and the high level triggers (HLT1 and HLT2) implemented in software. We give here a brief description of these triggers.

Level 0 trigger (L0)

The level 0 trigger reduces the rate at which the LHCb detector is read out. To achieve such a goal, a decision has to be done in under 4μs by the hardware implementation [15]. The only detectors in the LHCb detection system capable of supplying information fast enough are the muon systems and the calorimeters. Events with either high transverse momentum muons or large transverse energy deposits in the calorimeters are selected.

First stage of high level trigger (HLT1)

The first stage of the high level trigger reduces the rate of events by performing a partial event reconstruction [15]. Only a partial event reconstruction is possible, and consists of a reconstruction of track segments in the vertex locator, an extrapolation of the track segments that either have a large impact parameter or can be matched to hits in the muon chambers, and finally an extension of the selected track segments in the tracking stations by searching for hits consistent with a high transverse momentum track.

Second stage of high level trigger (HLT2)

The second stage of high level trigger reconstructs all tracks with a transverse momentum $p_{\mathrm{T}} = 300$ MeV/c [15] from the events selected by HLT1, independently of their impact parameter or matching hits in the muon chambers.

The trigger system represents only one step of the LHCb data flow. Moreover, data are generally present in analysis along complex Monte Carlo simulations that reproduce almost perfectly the processes happening in the LHC and measured by the LHCb detec-

tor. The next subsection discusses these two computational types of events throughout the whole production of analysable data, especially in the context of the specific decay $B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$.

2.2 Data and Monte Carlo simulations

Events exist in two different forms: real data that are collected at the LHCb detector, and Monte Carlo simulations that are generated with statistical methods.

Data are acquired by the LHCb detector and are filtered through the three triggers detailed in the previous subsection. The resulting events that correspond physically to detector hits are reconstructed into tracks or clusters by a specific application called Brunel. These objects are stored in an output file in a DST format to get filtered further through a set of selections based on fully reconstructed events. This process provides separate data sets called stripping lines with special characteristics, that are used by different analysis working groups. Unlike the trigger, the stripping does not delete the deselection; data can always be stripped again if needed. After this, LHCb software tools, such as DaVinci, allow physicists to take data off these stripping lines, only selecting the decays of interest, and putting them in ROOT ntuple format suitable for the analysis.

On the other hand, Monte Carlo simulations are, by definition, generated computationally. Using a computer simulation program called Pythia8, the pp collision that happens in the LHC at CERN are simulated. Then, a toolkit named EVTGEN adds the simulation of the event of interest. It is configured to match as much as possible the kinematics of the latter. The particles involved in the events are propagated through a simulated copy of the LHCb detector implemented in Geant4. The detector response is simulated and signals are digitalised in the same way as real data. Once this step is achieved, the Monte Carlo simulations follow the same data flow as for the collected data, and end up in another file stored in ROOT ntuple format.

The whole procedure is summarized in Figure 9. In what follows, these two multidimensional data sets are referred to as data sample, and Monte Carlo sample respectively. Because the LHCb magnet has two polarities, being up and down, these two samples are generally separated into two ROOT ntuple. In this analysis, we merged the up and down magnet polarities ROOT ntuple, and for what follows, only the merged files are considered, unless otherwise specified.

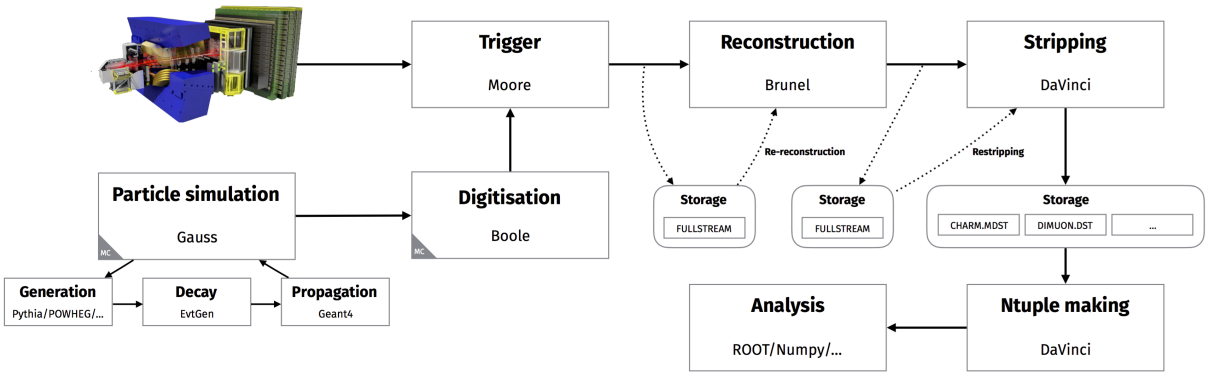


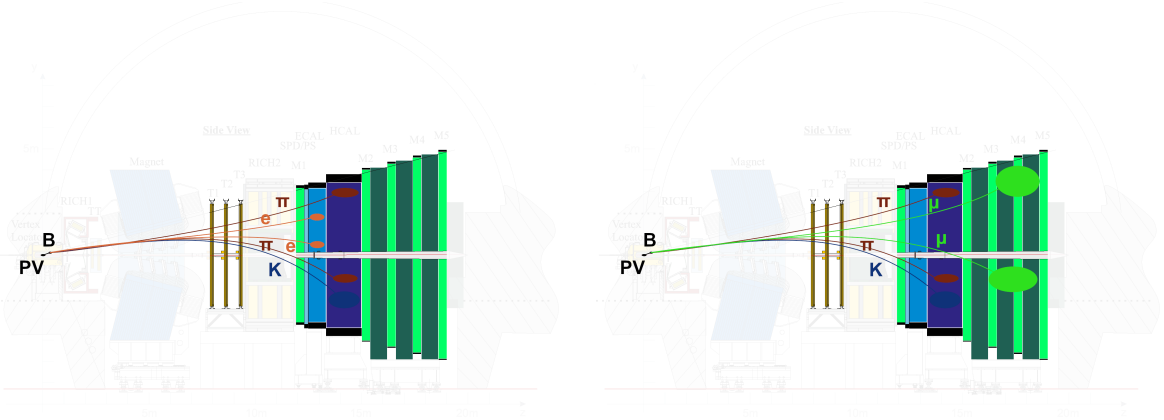
Figure 9: LHCb data flow [17].

The data and Monte Carlo samples focus in this analysis on the decays $B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$, whose specific mesons and leptons are detailed in Tables 1. In what follows, charge conjugates of each decay mode are implied, except where explicitly stated, e.g. $B^+ \to K^+\pi^+\pi^-\mu^+\mu^-$ refers to both $B^+ \to K^+\pi^+\pi^-\mu^+\mu^-$ and $B^- \to K^-\pi^+\pi^-\mu^+\mu^-$.

Particle	Quark content	Mass [MeV/c^2]	Mean life [s]
$B^\pm$	$u\bar{b}$, $\bar{u}b$	5279.3 ± 0.1	$(1638 \pm 4) \times 10^{-15}$
$K^\pm$	$u\bar{s}$, $\bar{u}s$	493.68 ± 0.02	$(1.238 \pm 0.002) \times 10^{-8}$
$\pi^\pm$	$u\bar{d}$, $d\bar{u}$	139.5704 ± 0.0002	$(2.6033 \pm 0.0005) \times 10^{-8}$
$\mu^\pm$	-	105.658375 ± 0.000002	$(2.196981 \pm 0.000002) \times 10^{-6}$
$e^\pm$	-	$0.510998946 \pm 0.000000003$	$> 6.6 \times 10^{28}$

Table 1: Details of mesons and leptons involved in $B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$ decays [18].

Because the final states of the rare decays differ in the leptons flavour, their trigger signature in the LHCb detector is different. Indeed, not all particles interact with matter in the same way. The dominant process for energy loss of charged particles, such as K^+, $\pi^\pm$ or even μ^+, in the LHC range of energy is the electromagnetic interaction with electrons of atoms. They loose energy through ionisation processes happening in matter according to the Bethe-Bloch formula [18]. With the exception of muons, which are very penetrating, and of electrons, that are completely stopped by the ECAL, all charged particles present in the $B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$ decays will therefore deposit their energies in the hadronic calorimeter. Muons interact very little with matter at these energies, which makes them penetrate far deeper into the detector. They will follow a rather continuous path towards the muon station, where they will deposit their total energy. Because of their high momenta, lighter particles, such as $e^\pm$, release their energy in matter by emitting a large amount of Bremsstrahlung radiation, which alters their path in the detector. The stopping power of the electromagnetic calorimeter is sufficient to stop them and collect their deposited energy. Figure 10 illustrates the trigger signatures of $B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$ decays in the LHCb detector.



(a) $B^+ \to K^+\pi^+\pi^-e^+e^-$ (b) $B^+ \to K^+\pi^+\pi^-\mu^+\mu^-$

Figure 10: $B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$ decays schematised in the LHCb detector.

Data were collected by the LHCb detector in 2017 with an integrated luminosity of 1.7 fb^{-1} [19] and a centre-of-mass energy of 13 TeV. The analysis uses full LHCb dataset and a region of $1.1 < q^2 < 7\text{GeV}^2$, with a veto on the J/ψ constrained B mass.

A preselection comprising different steps summarised in Table 2 is applied on both data and Monte Carlo samples. This preselection is essential in ensuring that only the relevant events are retained, while the insignificant ones are discarded. There is e.g. a preselection of actual electrons, or at least to misidentify as few particles as possible, which is ensured by the consideration of events with a neural network probability ProbNN$e(e^+)$ greater than 0.2.

Variables	Requirements
Fiducial	
Detector	InAccMuon ($\pi^\pm$, K), hasRich ($e^\pm$, $\pi^\pm$, K), hasCalo($e^\pm$)
Transverse momentum	$p_{\mathrm{T}}(\pi^\pm,K^+)$ >250MeV/c, $p_{\mathrm{T}}(e^\pm,J/\psi)$ >500MeV/c
	$p_{\mathrm{T,\,true}}(e^\pm)$ >200MeV/c
Momentum	$p(\pi^\pm,K^+)$ > 2000 MeV/c, $p(e^\pm)$ > 3000 MeV/c
Ghost probability	GhostProb($e^\pm$, $\pi^\pm$, K)<0.3
Projection	$\lvert x_{\mathrm{proj}}(e^\pm_{\mathrm{L0Calo\ ECAL}})\rvert > 363.6$ or $\lvert y_{\mathrm{proj}}(e^\pm_{\mathrm{L0Calo\ ECAL}})\rvert > 282.6$
Angle between particles	$\alpha > 0.0005$
PID	
NN probability	ProbNN$\alpha(\alpha) > 0.2$ with $\alpha = e^\pm,\pi^\pm,K$
Particle ID	$e^\pm$PID$e > 0$, KPID$K > -5$
J/ψ veto	
Mass veto	$5139 < m_{\mathrm{constr}} < 5476$ MeV/c^2
MC truth-matching	
True ID	$\lvert$TRUEID(α)$\rvert$ = ID(α) or $\lvert$TRUEID(α)$\rvert$ = 0 with $\alpha = e^\pm,\pi^\pm,K$
	(TRUEID($e^\pm$) = 0) + (TRUEID($\pi^\pm$) = 0) + (TRUEID(K) = 0) < 1.5
Background categories	(BKGCAT(B) < 11 or BKGCAT(B) = 50 or BKGCAT(B) = 60) and (hasLowest BKGCAT(B))

Table 2: Preselection applied. The Fiducial and PID preselection apply on both data and Monte Carlo samples, whereas the J/ψ veto concerns only the data and the MC truth-matching affects the Monte Carlo samples.

Another selection is further applied on the data sample to suppress peaking backgrounds. These are due to several facts, such as resonances or misidentification, and therefore need different vetoes, described in Table 3, to be removed.

Veto name	Responsible decays	Variables and requirements
$\psi(2S)K$	$B \to (\psi(2S) \to (J/\psi \to \ell\ell)\pi\pi)K$	$m_{J/\psi\ \mathrm{constr.}}(\pi\pi\ell\ell) \notin [3663,3709]$ MeV/c^2
$\chi_{c1}(3872)$	$B \to K\pi\pi(\chi_{c1}(3872) \to \ell\ell)$	$m_{J/\psi\ \mathrm{constr.}}(\pi\pi\ell\ell) \notin [3848,3894]$ MeV/c^2
$J/\psi K^*$	$B \to (K^* \to K\pi)\pi(J/\psi \to \ell\ell)$	$m_{J/\psi\ \mathrm{constr.}}(K\pi\ell\ell) < 5175$ MeV/c^2
D^0	$B \to (D^0 \to K\pi)\pi\pi\pi$	$m(K\pi) \notin [1838,1898]$ MeV/c^2,
	$B \to (D^0 \to K\pi\pi)\pi\pi$	$m(K\pi\pi) \notin [1838,1898]$ MeV/c^2,
	$B \to (D^0 \to K\pi\pi\pi)\pi$	$m(K\pi\pi\pi) \notin [1838,1898]$ MeV/c^2
Double MisID	$B \to K_{\to\ell}\pi\pi\ell\ell_{\to\pi}$	$m(\pi\ell) \notin [3047,3147]$ MeV/c^2,
		$m(K\ell) \notin [3047,3147]$ MeV/c^2

Table 3: Vetoes applied on the data sample to suppress peaking backgrounds. The responsible decays are indicated along the vetoes values.

The first three vetoes are used to suppress resonances produced by $\psi(2S)$, $\chi_{c1}(3872)$ or even K^* that result in the same final state as $B \to K\pi\pi\ell\ell$ decays. Then, the fourth veto concerns fully hadronic decay with $\pi\pi$ reconstructed as $\ell\ell$. And finally, the last veto takes into account double misidentification that cannot be neglected considering rare decays.

Owing to the similarities between rare and resonant decays, the same preselection strategy is applied to the latter. The resulting data and Monte Carlo samples contain initially a certain number of events $n_{\text{ev., initial}}$ and features $n_{\text{feat., initial}}$. These numbers are reported in Table 4 for up, down and merged magnet polarities.

	$B^+ \to K^+\pi^+\pi^-(J/\psi \to e^+e^-)$		$B^+ \to K^+\pi^+\pi^-(J/\psi \to \mu^+\mu^-)$	
	MC	Data	MC	Data
$n_{\text{ev., initial}}$ (up)	64962	103080	99939	452812
$n_{\text{ev., initial}}$ (down)	65414	108317	99540	442932
$n_{\text{ev., initial}}$ (merged)	130376	211397	199479	895744
$n_{\text{feat., initial}}$ (up, down, merged)	1736	1558	1702	1526

Table 4: Monte Carlo (MC) and data samples initial size. The number of events $n_{\text{ev., initial}}$ and features $n_{\text{feat., initial}}$ are indicated. The magnet polarity is indicated in parenthesis.

These data and Monte Carlo samples are used for the rest of the analysis, that is, for the Monte Carlo reweighting presented in Section 3, as well as for the identification of combinatorial background in data discussed in Section 4.

3 Reweighting of Monte Carlo simulation

The multiple corrections applied on the Monte Carlo simulations follow a correction chain that was put in place for the analysis of 2011-2016 LHCb data [11]. It consists of several steps that are used to erase the slightest imperfections that remain from the original computation. Summarised in Table 5, they each correct a specific default that occurs during the simulation production. First, the PID variables must be corrected because their response in the Monte Carlo simulation differs from what we have in real data, and the data selection relies heavily on them. Because the electron long-track reconstruction efficiency is affected by material simulation, a tracking correction is also added. Then, a phase space correction is performed to reweight the Monte Carlo simulation with flat $K\pi\pi$ mass distribution, known as phase space. It is needed because the $K\pi\pi$ system contains many resonances (K_1, K_2, $K^*\pi$, $K^+\rho$, ω, η ...) in a proportion that is unknown with enough accuracy. The pp underlying event is not well simulated either, which implies another correction. Then, due to slight detector and DAQ mismodelling, some events that pass the trigger selections in collision data may not pass them in simulation, or vice versa, so a trigger correction is applied. And finally, to account for reconstruction differences, another reweighting is added.

Step	Correction	Technique	Variables
1	Particle identification	Resampling	PID variables
2	Tracking	Tag&Probe	$p_{\mathrm{T}}(e)$, $\phi(e)$, $\eta(e)$
3	Phase space	BDT reweighter	$m(K\pi\pi)$, $m(K\pi)$, $m(\pi\pi)$
4	Kin. and mult.	GB reweighter	$p_{\mathrm{T}}(B)$, $\eta(B)$, nTracks, nHits
5	Trigger	Tag&Probe	L0 correction, HLT alignment and correction
6	Reconstruction	BDT reweighter	$p_{\mathrm{T}}(B)$, $\eta(B)$, χ^2_{VTX}/NDOF, $\chi^2_{\mathrm{IP}}(B)$

Table 5: Correction chain put in place for the analysis of 2011-2016 LHCb data. In Section 3, only the 1st and 4th steps are performed on the 2017 LHCb data.

In this section, an overview of the main results obtained in Reference [1] is given. More precisely, this concerns two of the six steps of the correction chain: the validation of the particle identification variables correction, as described in Subsection 3.2, and the reweighting of the pp event and the B^+ kinematics, discussed in Subsection 3.3. Background-subtracted data is needed in both steps, and is obtained using the ${}_s\mathcal{P}lot$ technique, as explained in Subsection 3.1.

3.1 ${}_s\mathcal{P}lot$ technique and background-subtracted data

The ${}_s\mathcal{P}lot$ statistical technique is a powerful tool to unfold the contribution of different components to the data distributions with a result similar to a likelihood projection plot, but without cuts [20]. Indeed, instead of assigning a category to each event in the data set, all events are weighted and contributes to the distribution. Its principle is the following: By knowing the distribution of discriminating variables associated to the various sources of events that compose the sample, but not necessarily all, the ${}_s\mathcal{P}lot$ technique

aims at reconstructing the unknown true distribution of a control variable [21]. A crucial assumption is that the control variable is uncorrelated with the discriminating variables.

The first step involved in the application of the ${}_s\mathscr{P}lot$ technique is the computation of an extended maximum likelihood fit of the probability distributions for the discriminating variables, e.g. mass distribution with a peak, to the various sources of events in the data. This fit, called also *s*Fit, returns the yields of the sources and determines the parameters of the distributions. The log-Likelihood is expressed as [22]:

$$\mathscr{L} = \sum_{e=1}^{N} \ln \left[\sum_{i=1}^{N_s} N_i f_i(y_e) \right] - \sum_{i=1}^{N_s} N_i \tag{8}$$

where N is the total number of events in the data sample, N_s the number of events species populating the data sample, N_i the number of events expected on average for the i^{th} species, y the set of discriminating variables, f_i the PDF of the discriminating variables for the i^{th} species, $f_i(y_e)$ the value taken by the PDFs f_i for event e, the later being associated with a set of values y_e for the set of discriminating variables, and x the set of control variables which, by definition, do not appear in the above expression of $\mathscr{L}$.

Then, all events in a data subset are weighted with a function of the distribution of the discriminating variables and their estimated covariance matrix V_{nj}. These weights are called *s*Weight, and are defined computationally as [23]:

$$W_n(y_e) = \sum_{j=1}^{N_s} V_{nj} f_j(y_e) \left[\sum_{k=1}^{N_s} N_k f_k(y_e) \right]^{-1} \tag{9}$$

where the inverse of the covariance matrix V_{nj}^{-1} is defined as [22]:

$$V_{nj}^{-1} = \frac{\partial^2(-\mathscr{L})}{\partial N_n \partial N_j} = \sum_{e=1}^{N} f_n(y_e) f_j(y_e) \left[\sum_{i=1}^{N_s} N_i f_i(y_e) \right]^{-2} \tag{10}$$

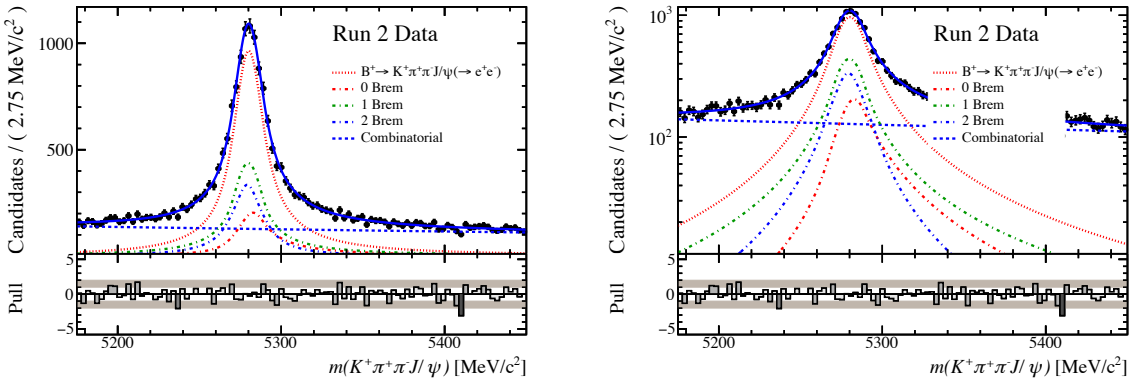
In this analysis, the ${}_s\mathscr{P}lot$ technique is used to computed background-subtracted data distributions of any given feature in our data samples. This implies that two contributions are needed for the ${}_s\mathscr{P}lot$ technique, one for the background, indicated by the letter b, and one for the signal, denoted by the letter s. By substituting these in Equations (8)-(10), we get the following expressions:

$$\begin{cases} \mathscr{L} &= \sum_{e=1}^{N} \ln \left[N_s f_s(y_e) + N_b f_b(y_e) \right] - (N_s + N_b) \\ W_s(y_e) &= \dfrac{V_{ss} f_s(y) + V_{sb} f_b(y)}{N_s f_s(y) + N_b f_b(y)} \\ V_{nj}^{-1} &= \sum_{e=1}^{N} f_n(y_e) f_j(y_e) \left[N_s f_s(y_e) + N_b f_b(y_e) \right]^{-2}, \text{ with } n, j = s, b \end{cases} \tag{11}$$

The number of events expected on average $N_{s,b}$ are given with the computation of the PDFs $f_{s,b}$ of a discriminating variable. The latter is defined as the reconstructed B^+ meson mass given by $m_{\text{constr}} = m(K^+\pi^+\pi^- J/\psi(\rightarrow \ell^+\ell^-))$ where the two leptons are constrained to come from a J/ψ decay and the B^+ to point back to the primary pp collision vertex. To accurately describe this discriminating variable, an Exponential PDF is assumed for the background PDF f_b, and a Hypatia PDF is used for the signal PDF f_s. More detail on these two PDFs can be found in Appendix A.

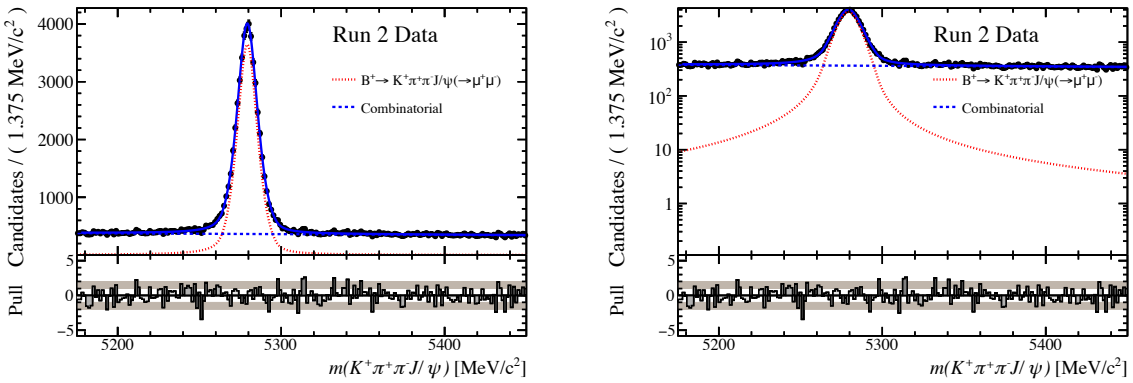
To ease the fit process of the data, the signal PDF f_s is first computed using the Monte Carlo simulation to extract relevant parameters that can be fixed later for the data fit. These parameters are in this case the ones that describe the right and left tails of the signal distribution.

A preselection is applied on the events to select only those that fall within the interval [5175, 5450] MeV/c^2 to surround the B^+ mass peak that occurs at approximately 5280 MeV/c^2. One PDF is assumed for $\ell = \mu$, and separate PDFs are considered for $\ell = e$ modelling events with zero (Brem0), one (Brem1) or two (Brem2) added Bremsstrahlung photons. The results of these *s*Fits are shown in Figures 12 and 11. All intermediate plots can be found in Appendix B.



(a) linear scale (b) log scale

Figure 11: *s*Fit (plain blue) of the reconstructed B^+ meson mass using $\ell = e$, with the signal component (dashed red) and the background component (dashed blue) as well as all Brem categories merged (Brem0 in dashdotted red, Brem1 in dashdotted green and Brem2 in dashdotted blue) in (a) linear and in (b) logarithmic scale.



(a) linear scale (b) log scale

Figure 12: *s*Fit (plain blue) of the reconstructed B^+ meson mass using $\ell = \mu$, with the signal component (dashed red) and the background component (dashed blue) in (a) linear and in (b) logarithmic scale.

The *s*Weights are extracted from these *s*Fits and are applied to particle identification variables, as explained in Subsection 3.2, and both kinematics and multiplicity variables, as demonstrated in Subsection 3.3, to evaluate the two different Monte Carlo reweighting steps used in Section 4.

3.2 Resampling of particle identification variables

As briefly mentioned in the description of the different parts of the LHCb detector, the particle identification (PID) is performed combining information from RICH-1, RICH-2, ECAL, HCAL and the muon system. In each event, mass hypotheses are assigned to every particle, and the relative likelihoods are calculated via multivariate likelihood classifiers combining information from the PID detectors. The classifiers are trained to compare the various particle hypotheses against the pion hypothesis, e.g. the estimation of the belonging of a reconstructed track to a kaon or pion based on the difference between the log-likelihoods of the kaon and pion hypotheses, respectively.

PID information is further refined by the use of a neural network combining the output of the PID classifiers and information from the other LHCb detectors, e.g. the tracking stations. The output of the neural network is a variable, called ProbNNx, taking values between 0 and 1. It can be interpreted as the absolute probability of a particle to be of a certain species when its suffix x matches the particle, e.g. ProbNNμ of muons.

Due to slight imperfections in the detector description and in the simulation of the detector responses, PID-related quantities are not perfectly well simulated in the Monte Carlo samples. This means that potential mistakes must be corrected if the Monte Carlo sample are used to calculate the efficiency of a certain selection.

A specific package called Meerkat [24], which stands for multidimensional efficiency estimation using the relative kernel approximation technique, is used to resample PID variables. Such resampling method is based on kernel density estimation with corrections to account for boundary effects [25]. The PID variables are assumed to depend only on the track kinematics and the detector occupancy, which correspond to the transverse momentum p_T, the pseudorapidity η, and the total number n_{tracks} of reconstructed tracks in the whole event. Thus, the PID response can be expressed as a function of those variables, i.e. $f(p_T, \eta, n_{\text{tracks}})$.

Calibration data samples, that is, well-known and well reconstructed decays, such as $D^* \to D^0\pi$ or $B \to K(J/\psi \to \ell\ell)$, are used to evaluate the probability distribution function of the PID variable as a function of the kinematics and detector occupancy. Concretely, once the PDF $g(\text{PID}) = f(p_T, \eta, n_{\text{tracks}})$ is known, it is possible to sample from these PDFs the PID response for simulated tracks given the values of p_T, η, and n_{tracks}. This approach allows to have correct PID responses in Monte Carlo on a statistical basis.

In this analysis, the PID variables listed in Table 6 were already resampled and evaluated. The evaluation consist in a comparison of the resampled PID variables distributions with the PID variables distributions from Monte Carlo simulation, and background-subtracted data, and demonstrated that the resampled PID variables were more accurate. These comparisons can be found in Appendix C.

Particle	PID variables
K^+	ProbNNk, ProbNNp, PIDk, PIDp
$\pi^\pm$	ProbNNπ, ProbNNp, PIDk, PIDp
$e^\pm$	ProbNNe, PIDe
$\mu^\pm$	ProbNNμ

Table 6: PID resampled variables.

3.3 Reweighting of the pp event and B^+ kinematics

Imperfections are also present in the pp event modelling, and the reproduction of the B^+ kinematics. These key steps in the production of Monte Carlo simulation are sufficiently well described by four variables, listed in Tables 7, that includes the numbers of reconstructed tracks recorded in the vertex locator and hits detected in the scintillating pad detector, and the B^+ transverse momentum and pseudorapidity, respectively.

Variables	Description
Kinematics	
$p_{\mathrm{T}}(B^+)$	B^+ transverse momentum. It is described by the contribution of the momentum vector $p(B^+)$ in the x-y plane perpendicular to the beam pipe (z-axis): $p_{\mathrm{T}}(B^+) = \sqrt{p_x^2(B^+) + p_y^2(B^+)}$
$\eta(B^+)$	B^+ pseudorapidity. It is a measure of how far the track direction lies with respect to the beam axis using $\theta(p(B^+), z)$ as the angle between the B^+ momentum and the positive direction of the beam axis z: $\eta(B^+) = -\ln\left(\tan\frac{\theta(p(B^+),z)}{2}\right)$
Multiplicity	
$n_{\mathrm{VELO,\ Tracks}}$	Number of reconstructed tracks recorded in the vertex locator surrounding the pp interaction point of the LHCb detector.
$n_{\mathrm{SPD,\ Hits}}$	Number of hits detected in the scintillating pad detector situated at the entrance of the calorimeter system.

Table 7: Kinematics and multiplicity variables.

Hence these variables are used as input for the multivariate analysis (MVA) reweighter that corrects the Monte Carlo simulation. For this particular task, a gradient boosted reweighter (GBR) is opted. As Reference [26] thoroughly details this technique, only a short description is given here. GBR is a machine learning reweighting algorithm based on ensemble of regression trees, where parameters have the same role as in gradient boosting. One of its interesting feature is that it is applicable to data of high dimensionality. Main parameters of GBR are indicated in Appendix D.

This reweighter takes as input Monte Carlo simulations and background-subtracted data as data sets, a set of features and hyperparameters to compute weights. The GBR is trained on four different combinations of inputs, as the electrons and muons data samples and the kinematics and multiplicity variables are separated, using a 10-fold cross-validation. Once these weights are computed, they are used to reweight variables distributions. This reweighted Monte Carlo distribution is validated by comparing it with the background-subtracted data and the initial Monte Carlo simulation distributions of the kinematics and multiplicity variables. The results obtained for this part of the reweighting can be found in Appendix E.

This step in the Monte Carlo reweighting process allows a first computation of an effective correction. And this correction is represented by weights associated to every event contained in the data sample. Even if these will be further refined by the other steps of the correction chain, they can still be used as a first estimate. In the next section in particular, these weights are used to reweight the Monte Carlo simulations.

4 Event classification with multivariate analysis

As any other data collected from high-energy experiments, they consist of a mixture of two categories of events, that are qualified as background and signal. In this analysis, signal defines the decay $B \to K^+\pi^+\pi^-\ell^+\ell^-$ while different categories of events fall under the background category, such as B decays with misidentified particles, partially reconstructed decays, and events reconstructed from random combinations of final state tracks. This section addresses the removal of the latter events. When cut-based selection is not sufficient anymore, powerful computing techniques, such as multivariate analysis (MVA) methods, are needed to do so. These machine learning algorithms are able to assign a category to each event. More precisely, they compute weights to characterise signal-likeliness of every events present in a data sample. Applying a selection based on the value of the event's weight is then equivalent to identifying events considered as background, making it easier to remove them from data. But before achieving this result, different parameters need to be introduced.

Subsection 4.1 presents the multivariate analysis methods used in this analysis, as well as the different ways of assessing the methods performance. Apart from hyperparameters, these methods require test and training sets of all classes, that is, background and signal, as detailed in 4.2, and discriminating features, whose selection, detailed in Subsection 4.3, requires a particular attention. The results obtained using these considerations are displayed in Subsection 4.4, and a comparison is done in Subsection 4.5 with a previous work performed when only Run 1 data was available. It is worth mentioning that a cross check of its performance or even a new method using Run 2 data was long overdue. Finally, the optimisation of the classifier output is explained in Subsection 4.6.

4.1 MVA classification methods

Often compared to a black box, multivariate analysis methods employing machine learning algorithms have a fairly clear objective, which is to recognise patterns in data humanly unfeasible and to use this information as a prediction of unknown similar data sets.

In order to achieve such a great task, machine learning algorithms and mutivariate analysis (MVA) methods, use three main types of input: sets of events on which the MVA method is trained and test to recognise the pattern found, named test and training sets, decisive variables defining the previous events, called discriminating features, and a set of computational settings referred as hyperparameters.

Several machine learning methods can be considered for so-called supervised learning classification, among which, Boosted Decision Trees (BDT), Multilayer Perceptrons (MLP), k-Nearest Neighbours (kNN) or even Support Vector Machine (SVM). Considering the nature of these MVA methods, which is to classify events, their output, corresponding to a prediction, is often called a classifier.

In this analysis, we decided to focus on BDT, and we explored two different kind of boosting methods, detailed in Subsection 4.1.1. To assess the performance of classification methods, ROC curves and overtraining checks are required. These particular computations are detailed in Subsection 4.1.2. They will particularly be useful to select the best combination of MVA method and parameters. The results obtained are shown in Subsection 4.4, whereas some of its variants are reported in Appendix I.

4.1.1 Algorithms

Decision trees are powerful classification methods. They take as main inputs a set of discriminating features and their corresponding data sets. Its working principle is the following: The data is split recursively according to a criteria imposed on one input feature at a time until a stop condition is met. More precisely, this procedure generates a decision at each stage, denoted as nodes, and when these lead to an end point, the result, called a leaf, represents a class label or a probability. Each split at a node is chosen to maximise information gain or minimise entropy [27]. The whole procedure, illustrated in Figure 13, draws a tree, hence the name of this method.

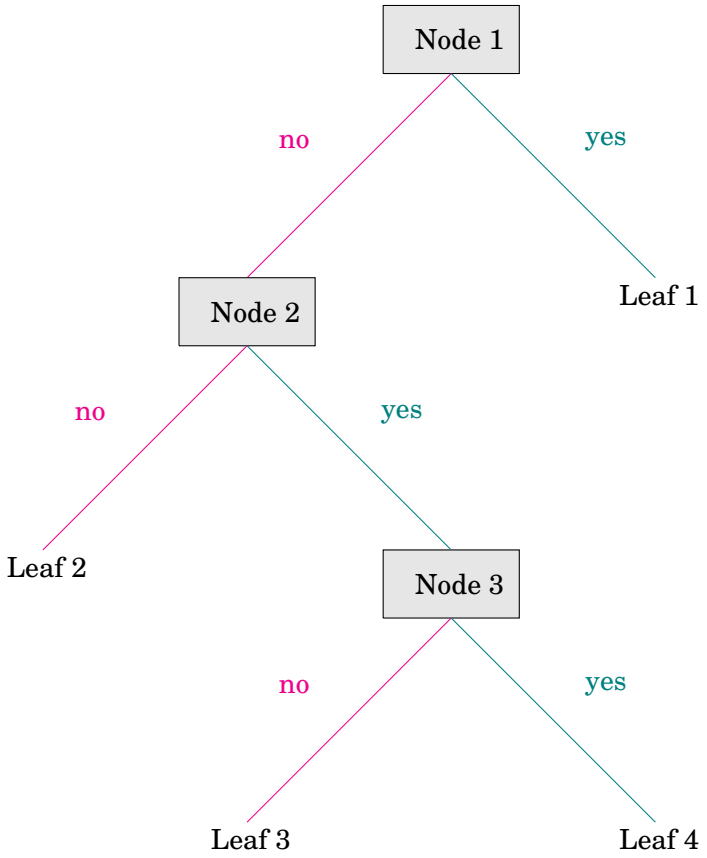


Figure 13: Illustration of an example of decision tree. Two branches arise from each decision, denoted as node, corresponding to whether or not the criterion is met. This process progressively draws branches constituting a so-called decision tree. The branches that result in a decision is called a leaf.

The ensemble of trees that leads to a classification is interpreted as boosting [28]. In supervised learning, the prediction $\hat{y}_i$ is computed using the input x_i. Generally, predictions are assumed to be proportional to inputs, which gives:

$$\hat{y}_i = \sum_j \theta_j x_{ij} \tag{12}$$

The θ_j coefficients are determined by data. More precisely, the task of training the model allows to find the coefficients that best fit inputs and predictions. To do so, the objective function $O(\theta)$ is defined to measure the accuracy with which the model fits the training

data. This function is the sum of two terms: training loss $L(\theta)$ and regularisation $\Omega(\theta)$:

$$O(\theta) = L(\theta) + \Omega(\theta) \tag{13}$$

The regularisation term $\Omega(\theta)$ handles overtraining issues that can arise when a model is too complex. The training loss measures the predictability of a model, which is represented mathematically by the distance between y_i and $\hat{y}_i$. The mean squared error is a common mathematical expression for the latter, which is given by:

$$L(\theta) \equiv \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{14}$$

Within the scope of decision tree ensembles, the model used to determine the predictions $\hat{y}_i$ is computed as the following:

$$\hat{y}_i = \sum_{k=1}^{n_\mathrm{T}} f_k(x_i) \tag{15}$$

where n_T is the number of trees, and f_k is a function. The objective function to be optimised becomes:

$$O(\theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{n_\mathrm{T}} \Omega(f_k) \tag{16}$$

The objective function is further refined by the type of boosting chosen for the classification; it optimises the whole process of decision trees creation. In this analysis, two types of boosting inspired by the well-known AdaBoost [29] and Gradient Boosting [30] techniques are mainly employed: Stagewise Additive Modeling using a Multi-class Exponential loss function (SAMME) [31] and eXtreme Gradient (XG) [32] boosting algorithms. The former was proposed as a first attempt, and its main results are summarised in Appendix I, whereas the latter showed better performance and was therefore kept for the rest of the analysis. The XG-BDT method was trained separately on electrons and muons, using the hyperparameters listed in Table 8. Monte Carlo weights computed in Section 3 were used for the test and training samples of the signal. The performance is evaluated through statistical criteria explained in the next subsection.

Hyperparameters	$B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$	Default	Description
learning_rate	0.1	0.3	step size shrinkage used in update to prevent overfitting
max_depth	3	6	maximum depth of a tree
nthread	None	Max	number of parallel threads used to run XGBoost

Table 8: Hyperparameters used for the XG-BDT [33].

4.1.2 Performance assessment

Since there are several classification algorithms, it is necessary to have a baseline to efficiently compare them in terms of performance. For that, test and training sets are essential to verify the predictions of MVA methods and therefore measure their separability power. In what follows, two main assessments are presented.

Receiver Operating Characteristics (ROC) curve

The receiver operating characteristics (ROC) curve illustrates graphically the classification ability of a MVA binary classification method. More precisely, it is the area under the curve (AUC), normalised to unity, that numerically represents this measurement. An excellent classification has an AUC close to 1, while a poor one has an AUC close to 0. The AUC value of 0.5 represents a classification made randomly.

In the case of binary classification, the computation of the ROC curve is solely based on the classification interpretation [34]. Indeed, considering that the classes are labelled either as positive (P) or negative (N), there are four possible outcomes: A true positive (TP) occurs when both the prediction outcome and the actual value are P; however if the actual value is N, then it is a false positive (FP). Similarly, a true negative (TN) and a false negative (FN) are defined. Using these definitions, a true positive rate (TPR) and a false positive rate (FPR) are computed as the following:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \qquad \text{and} \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{17}$$

The ROC curve is thus obtained by plotting the latter as a function of the former. Figure 14 illustrates an example of a ROC curve.

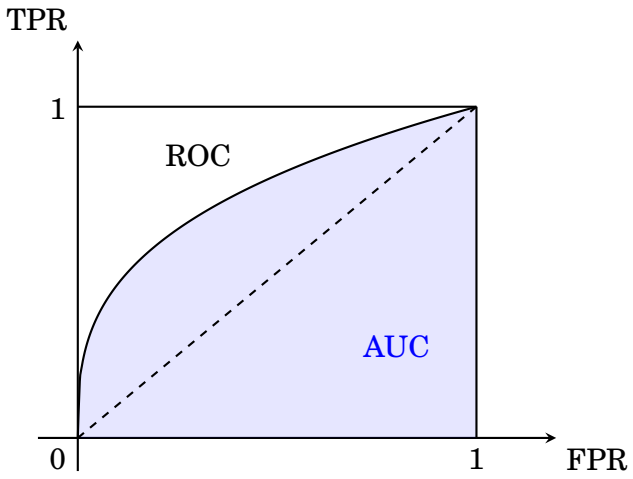


Figure 14: Example of a ROC curve indicating a result needing improvements.

Overtraining Check and Kolmogorov–Smirnov test

Overtraining occurs when the complexity of a model is so high that it reproduces every single feature of the specific data set used for training, therefore computing a not necessarily reproducible fit. An example of overtraining is illustrated in Figure 15, where two

classes of a data set are represented in blue and red dots. A fit (Fit 1), qualified as suitable, separates globally the measurements, leaving some of them misclassified, but being easily reproducible on another data set. Another fit (Fit 2), considered as overtrained, classifies exactly the measurements, but cannot be applied on another data set due to its highly specific behaviour.

Some MVA methods are more likely to experience overtraining, but the general consensus is that simplified models can prevent this issue during the training phase. In the case of BDT, they usually suffer from at least partial overtraining, owing to their large number of nodes [35]. Moreover, the application of the fit on a test set allows a verification of the fit reproducibility. Indeed, if a notable difference arises in the comparison of the MVA method performance on the test and training sets, it indicates that overtraining is occuring.

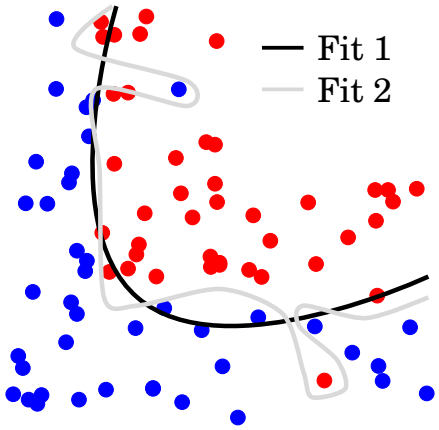


Figure 15: Illustration of overtraining. Fit 2 represents an overtrained model and Fit 1 represents a regularized model.

A common measurement of the overtraining is the Kolmogorov-Smirnov test [36]. This test gives an uniformly distributed value between 0 and 1, and verifies if the outcomes of both test and training sets are equivalent. A value of 0 is associated to MVA methods that show no overtraining.

4.2 Test and training sets of MVA method

The MVA method allows an event classification based on a set of events defining the categories to be assigned, i.e. background and signal. For the latter, the Monte Carlo samples are used with a specific cut on the background category to remove ghost events: B_BKGCAT < 59. As for the background, the data with a cut lying above the reconstructed B^+ meson mass m_{constr} allows the selection of potential combinatorial background. This cut, denoted as $m_{\ell,\text{limit}}$ differs for the two leptons data set: for $\ell = e$, the selection is such that $m_{\text{constr}} > m_{e,\text{limit}} = 5420$ MeV/c^2 and for $\ell = \mu$, it is such that $m_{\text{constr}} > m_{\mu,\text{limit}} = 5340$ MeV/c^2. The difference between these two values comes from the fact that the peak of the distribution using the electrons data set is wider than the one using the muons data set, which makes the signal spread over a wider range, as shown in Figures 11 and 12 from Subsection 3.1. The cuts applied on the up and down magnet polarities of both distributions is displayed in Figure 16. Approximately 18% and 26% of the initial number of events remains for $\ell = e, \mu$, respectively.

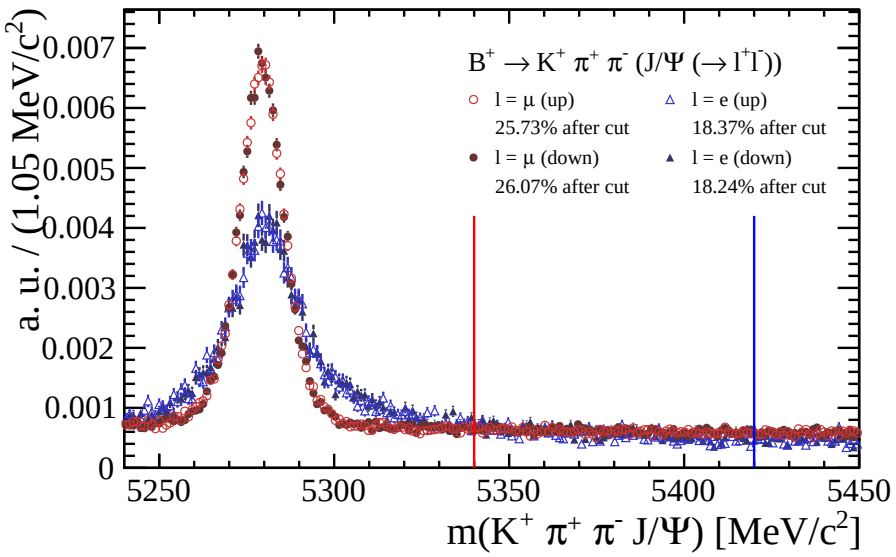


Figure 16: Distribution of the reconstructed B^+ meson mass using $\ell = e, \mu$ data sets and differentiating the up and down magnet polarities. A sub-interval [5240, 5450] MeV/c^2 was preferred to highlight the mass peak, even though the events cover an wider range going from 4196 to 7552 MeV/c^2. Both cuts happening at 5420 and 5340 MeV/c^2 for $\ell = e, \mu$ respectively are represented.

The size of the samples of the two categories influences the predictions of the MVA method, e.g. if one sample is much larger than the other, the MVA method is more likely to predict events in that category than the other. The size not only depends on the number of events $n_{\text{ev.}}$, but also on the sum of the weights $w_{\text{ev.}}$ assigned to the event, if any. Therefore, by using weighted events for the Monte Carlo sample, coming from Subsection 3.3, the balance between the two samples is achieved when the sum of the selected background events equals the sum of the weights of the selected signal events. The number of events of the different mentioned stages is shown in Table 9. Thereafter, we keep the samples which have a balanced number of events for the signal and a selected number of events for the background, called signal and background data set, respectively.

	$B^+ \to K^+\pi^+\pi^-(J/\psi \to e^+e^-)$			$B^+ \to K^+\pi^+\pi^-(J/\psi \to \mu^+\mu^-)$		
	up	down	merged	up	down	merged
Monte Carlo (sig.)						
$n_{\text{ev., initial}}$	64962	65414	130376	99939	99540	199479
$w_{\text{ev., initial}}$	114210	123205	237415	172013	170334	342347
$n_{\text{ev., selected}}$	60112	60542	120654	95273	94821	190094
$w_{\text{ev., selected}}$	105375	113902	219277	163511	161821	325333
$n_{\text{ev., balanced}}$	10884	10552	21436	67787	67620	135407
$w_{\text{ev., balanced}}$	19113	19814	38927	116243	115328	231570
Data (bkg.)						
$n_{\text{ev., initial}}$	103080	108317	211397	452812	442932	895744
$n_{\text{ev., selected}}$	18936	19755	38691	116517	115465	231982

Table 9: Number of events and weights of data and Monte Carlo samples at different stages of selection: initial samples, samples with a selection operating on the reconstructed B^+ meson and the background category respectively, and the Monte Carlo samples balanced with the consideration of its weights.

Given these two data sets, it is necessary to define on which events the MVA method will be trained. Selecting all events is not the optimal solution, as it becomes impossible to evaluate the performance of the MVA method. For this reason, it is not only necessary to define a set of events for training, but also a complementary set, which will be useful for testing the MVA method. There are several possible splits between training and testing, e.g. the k-fold technique detailed in Reference [37], but we decided to use 70% of the events for training and the remaining 30% for test.

4.3 Discriminating features of MVA method

The previously selected events are described by a variety of features, e.g. the J/ψ transverse momentum $p_{\mathrm{T}}(J/\psi)$, the vertex reconstruction quality of the B^+, or even its decay time $\tau(B^+)$. These features are measured for each event by the detectors that compose the LHCb, as detailed in Section 2. But some of these features are more useful in helping the MVA method to distinguish the signal from the background; they are not only meaningful from a physics point of view, but they also have a strong discriminating power. Therefore, of all possible features contained in the signal and background data sets, a selection is made according to these two criteria to select the most relevant ones.

A supplementary constraint is added to the selection of discriminating features: the MVA method only takes a limited number of features. Since the first selection of relevant features gives a number exceeding this limit by far, a further selection is operated. To reduce the number of features, their correlations are taken into account. Moreover, by sampling the features, in such a way that they appear in several samples, a rank, determined by a recursive feature elimination (RFE) algorithm, can be assigned to each feature in each sample. This rank is then transformed into a score normalised to unity. Based on these quantities, two other scores are established for each feature. The first one is computed by taking the mean of their score computed in the samples where it appears, and is referred later as mean score. The second one takes into account a weight, calculated according to the value of the correlations between this feature and all those present in each sample, which is multiplied to the mean score. It is referred later as weighted score. Having two different scores makes a comparison of the top features performance possible, and only the one of the two that shows the best results is kept for the analysis.

4.3.1 Selection of relevant features

The selection starts by considering the features that are present in both background and signal data sets. Indeed, the MVA method can only provide a classification using a feature present in both data sets. Because muons and electrons data sets are trained separately, similar features but applied to different leptons, e.g. ProbNN$e(e^+)$ and ProbNN$\mu(\mu^+)$, are taken into account. From these possible features, only those of interest from a physics point of view are selected; A first approach is to consider that combinatorial background is related to the reconstruction of the B meson decay. This means that B variables related to the kinematics, distance of closest approach (DOCA), impact parameter (IP), and the goodness of reconstruction of its vertex should be tested. Because the rare decay can be split into two resonances, i.e. $B \to (K_1 \to K\pi\pi)(J/\psi \to \ell\ell)$, the same considerations as for the B meson stand for both K_1 and J/ψ mesons. Some features related to both primary (PV) and secondary (SV) vertices are especially illustrated in Figure 17.

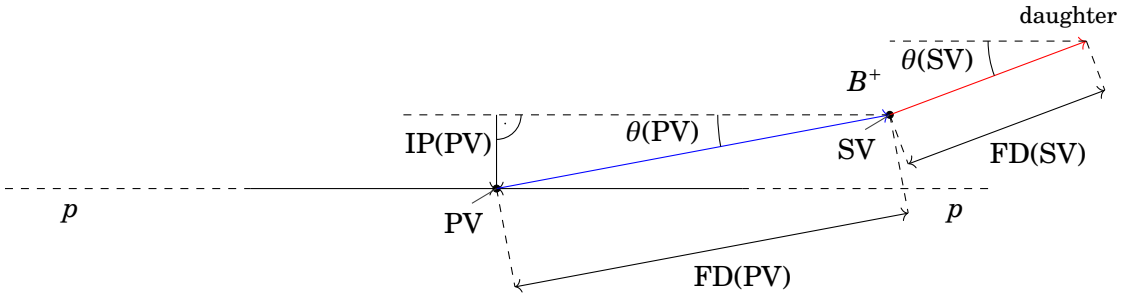


Figure 17: Vertex reconstruction variables. The cosine of the direction angle θ(PV) (θ(SV)) gives the feature denoted as DIRA(x,PV) (DIRA(x,SV)).

The PID features of the decay daughters complete the list. In particular, the resampled PID features from Subsection 3.2 are used for the signal data set, with the exception of PID$\mu^{\pm}$, which is used without correction. The complete list of relevant features, derived from the ones present simultaneously in both background and signal data sets, is reported in Table 11. For each feature, a description as well as some specifications are provided.

Following what was done in Reference [4], other features were subsequently added. They mainly derive from operations involving already existing features, e.g. the sum of all daughters transverse momentum, which is defined as

$$p_{\mathrm{T}}(K^+\pi^+\pi^-\ell^+\ell^-) = p_{\mathrm{T}}(K^+) + p_{\mathrm{T}}(\pi^+) + p_{\mathrm{T}}(\pi^-) + p_{\mathrm{T}}(\ell^+) + p_{\mathrm{T}}(\ell^-) \tag{18}$$

or the ratio of K_1 vertex location χ^2 and its number of degrees of freedom, which is computed as

$$\chi^2_{\mathrm{VTX}}/\mathrm{ndf}(K_1,PV) = \frac{\chi^2_{\mathrm{VTX}}(K_1,PV)}{\mathrm{ndf}(K_1,PV)} \tag{19}$$

by following the notation suggested in Table 11. By removing features with low discriminating power, i.e. too similar background and signal distributions, the final number of features, which represent good candidates for the MVA method, is 92 and 94 for the electrons and muons data sets respectively. The main numbers involved in the selection are summarised in Table 10.

	$B^+ \to K^+\pi^+\pi^-(J/\psi \to e^+e^-)$		$B^+ \to K^+\pi^+\pi^-(J/\psi \to \mu^+\mu^-)$	
	MC	Data	MC	Data
Up, down, merged				
$n_{\text{feat., initial}}$	1748	1584	1714	1550
$n_{\text{feat., shared}}$		1578		1544
$n_{\text{feat., relevant}}$		122		122
$n_{\text{feat., selected}}$		92		94

Table 10: Number of features through the selection, comprising the initial number of features present in signal and background data sets, the number of features shared by both sets respectively for $\ell = e,\mu$ data sets, the number of relevant features extracted, and the number of selected features for the MVA method. The number of features is the same for all up, down and merged polarities.

Variable description	Units	Symbol	LOKI Variable Name
Decay time (B^+, K_1, J/ψ)			
Particle decay time	[ns]	τ	TAU
Particle decay time χ^2	[ns]	χ^2_τ	TAUCHI2
Kinematics (B^+, K_1, J/ψ)			
Particle momentum vector norm	[MeV/c]	p	P
Particle transverse momentum vector norm	[MeV/c]	p_T	PT
Particle momentum vector α (= X, Y, Z) component	[MeV/c]	p_α	Pα
Particle energy	[MeV]	E	PE
Primary vertex (PV) quality (B^+, K_1, J/ψ)			
Particle decay vertex loc. α (= X, Y, Z) component	[mm]	VTX(x,PV)	ENDVERTEX_α
Particle decay vertex loc. χ^2	[mm]	$\chi^2_\mathrm{VTX}(x,PV)$	ENDVERTEX_CHI2
Particle decay vertex loc. degrees of freedom	[-]	ndf(x,PV)	ENDVERTEX_NDOF
Particle PV location α (= X, Y, Z) component	[mm]	PV$_\alpha(x)$	OWNPV_α
Particle PV location χ^2	[mm]	$\chi^2_\mathrm{PV}(x)$	OWNPV_CHI2
Particle flight distance w.r.t. PV	[mm]	FD(x,PV)	FD_OWNPV
Particle flight distance χ^2 w.r.t. PV	[mm]	$\chi^2_\mathrm{FD}(x,PV)$	FDCHI2_OWNPV
Particle direction angle w.r.t. PV	[-]	DIRA(x,PV)	DIRA_OWNPV
Particle impact parameter w.r.t. PV	[mm]	IP(x,PV)	IP_OWNPV
Particle impact parameter χ^2 w.r.t. PV	[mm]	$\chi^2_\mathrm{IP}(x,PV)$	IPCHI2_OWNPV
Secondary vertex (SV) quality (K_1, J/ψ)			
Particle origin SV α (= X, Y, Z) component	[mm]	SV$_\alpha(x)$	ORIVX_α
Particle origin SV χ^2	[mm]	$\chi^2_\mathrm{SV}(x)$	ORIVX_CHI2
Particle flight distance w.r.t. SV	[mm]	FD(x,SV)	FD_ORIVX
Particle flight distance χ^2 w.r.t. SV	[mm]	$\chi^2_\mathrm{FD}(x,SV)$	FDCHI2_ORIVX
Particle direction angle w.r.t. SV	[-]	DIRA(x,SV)	DIRA_ORIVX
Vertex reconstruction quality (B^+)			
Invariant mass χ^2 when the vertex is reconstructed with one missing track	[MeV/c^2]	$\mathrm{ISO}^{vtx}_{m,1}$	VTXISODCHI2MASSONETRACK
Invariant mass χ^2 when the vertex is reconstructed with two missing tracks	[MeV/c^2]	$\mathrm{ISO}^{vtx}_{m,2}$	VTXISODCHI2MASSTWOTRACK
Vertex reconstruction χ^2 with one missing track	[-]	ISO^{vtx}_{1}	VTXISODCHI2ONETRACK
Vertex reconstruction χ^2 with two missing tracks	[-]	ISO^{vtx}_{2}	VTXISODCHI2TWOTRACK
Number of vertices	[-]	ISO^{vtx}_{num}	VTXISONUMVTX
Particle identification (K^+, $\pi^\pm$, $\ell^\pm$)			
Particle NN probability to be an x particle	[-]	ProbNNx	ProbNNx
Particle identification of an x particle	[-]	PIDx	PIDx
Other features (B^+)			
Particle pseudorapidity	[-]	η	eta
Particle azimuthal angle	[rad]	ϕ	phi
Particle maximum distance of closest approach	[mm]	$\mathrm{max}_\mathrm{DOCA}$	AMAXDOCA
Particle minimum distance of closest approach	[mm]	$\mathrm{min}_\mathrm{DOCA}$	AMINDOCA

Table 11: Summary of the features relevant from a physics point of view. The particles concerned by a particular set of features are indicated in parenthesis.

Variable description	Units	Symbol	LOKI Variable Name
Sum of transverse momentum			
Sum of all daughters transverse momentum vector norm	[MeV/c]	$p_{\mathrm{T}}(K^+\pi^+\pi^-\ell^+\ell^-)$	`Daughters_PT`
Sum of electrons transverse momentum vector norm	[MeV/c]	$p_{\mathrm{T}}(e^+e^-)$	`electrons_PT`
Sum of muons transverse momentum vector norm	[MeV/c]	$p_{\mathrm{T}}(\mu^+\mu^-)$	`muons_PT`
Sum of all hadrons transverse momentum vector norm	[MeV/c]	$p_{\mathrm{T}}(K^+\pi^+\pi^-)$	`hadrons_PT`
Sum of impact parameter χ^2			
Sum of all daughters impact parameter χ^2 w.r.t. PV	[mm]	$\chi^2_{\mathrm{IP}}(K^+\pi^+\pi^-\ell^+\ell^-,PV)$	`Daughters_IPCHI2_OWNPV`
Sum of electrons impact parameter χ^2 w.r.t. PV	[mm]	$\chi^2_{\mathrm{IP}}(e^+e^-,PV)$	`electrons_IPCHI2_OWNPV`
Sum of muons impact parameter χ^2 w.r.t. PV	[mm]	$\chi^2_{\mathrm{IP}}(\mu^+\mu^-,PV)$	`muons_IPCHI2_OWNPV`
Sum of all hadrons impact parameter χ^2 w.r.t. PV	[mm]	$\chi^2_{\mathrm{IP}}(K^+\pi^+\pi^-,PV)$	`hadrons_IPCHI2_OWNPV`
Ratio ($\mathbf{B^+}$, $\mathbf{K_1}$, $\mathbf{J/\psi}$)			
Ratio of particle vertex location χ^2 and number of degrees of freedom	[mm]	$\chi^2_{\mathrm{VTX}}/\mathrm{ndf}(x,PV)$	`ENDVERTEX_CHI2DOF`

Table 12: Summary of the features added. The particles concerned by a particular set of features are indicated in parenthesis.

4.3.2 Features score

A number of about 100 features is inconceivable as input of a MVA method. Other conditions must be specified in order to select a more reasonable number situated in the $\mathcal{O}(10)$ order of magnitude. A score computation has been developed specifically to select features according to how strongly they depend on the others and how much more significant they are compared to the others. The suggested comparison is made on samples of 20 features created from the selected features. The dependency of a feature to another is computed by the Pearson's correlation, and the significance of a feature is estimated using a recursive feature elimination (RFE) algorithm on each samples. In what follows, we give more detail on the different components involved in the score computation process.

Features sampling

For the computation of the elements participating in the reduction of features, a limited number of features has to be considered. The idea is to create samples out of the whole set of selected features, and perform the computations on each of them. There are several ways to sample the features. One way to achieve a homogeneous result, especially for the computation of the significance of each feature, is to create enough samples so that each feature is evaluated at least once with all the others. We created an algorithm that randomly selects features that have not yet been evaluated together, and forms samples until all feature combinations have been tried. In our case, it produced approximately 130 samples of 20 features each for both electrons and muons data sets. A matrix representing the number of time each feature is evaluated with any other one is displayed in

Although this approach represents an improvement over the previous ones, it has two major disadvantages: some features may be used more than others and the number of samples is not negligible. To overcome these issues, some restrictions on the number of use of all features may be imposed, and a recursive approach consisting of several steps focusing on the best scores could decrease the number of samples.

Correlation computation

Correlation is a coefficient $C \in [-1,1]$ attributed to two features to describe how strongly one depends on the other. It can be positive or negative, depending on the relation of dependency, e.g. if one feature decreases while the other increases, then the correlation is negative. If two features happen to be highly correlated, only one of them should be used in the MVA method. Mathematically, the Pearson's correlation [38] between two features X and Y is computed using the signal data set reduced to these features, being A and B respectively, as the following:

$$C_{X,Y} = C_{Y,X} = \frac{\text{cov}(A,B)}{\sigma_A \sigma_B} \tag{20}$$

where $\text{cov}(\alpha,\beta)$ is the covariance and σ_α the standard deviation, for $\alpha,\beta = A,B$. Considering the events a_i, b_i contained in A,B respectively, Equation (20) becomes:

$$C_{X,Y} = \frac{\sum_{i=1}^{N}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{N}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{N}(b_i - \bar{b})^2}} \tag{21}$$

with

$$\bar{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i \tag{22}$$

where $\alpha = a,b$. Another definition of the correlation may be given in regards to the fact that the signal data set is weighted. Assuming a weight w_i attributed to each event a_i and b_i in Equation (21), the weighted correlation is defined as:

$$C_{X,Y}^{\text{weighted}} = \frac{\sum_{i=1}^{N}(a_i w_i - \bar{a})(b_i w_i - \bar{b})}{\sqrt{\sum_{i=1}^{N}(a_i w_i - \bar{a})^2}\sqrt{\sum_{i=1}^{N}(b_i w_i - \bar{b})^2}} \tag{23}$$

with

$$\bar{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i w_i \tag{24}$$

An example of the difference between the two computations is included in Appendix H. In this analysis, the computation corresponding to Equation (21) was used. Equation (23) constitutes an improvement that may be addressed in subsequent work.

Recursive feature elimination (RFE) technique

Given an initial set of features, this computing technique allows both the selection of the most relevant features and the computation of a rank R assigned to each of them. The

selection is operated by recursively considering sets of features where the least important features are pruned. To determine the latter, an estimator is trained each time of the actual set, and the importance of each feature is obtained. More detail on this technique may be found in Reference [39].

In this analysis, the estimator corresponds to the SAMME algorithm introduced in Subsection 4.1.1 using default parameters. In order to avoid statistical bias that can arise during the feature training phase, a system of repeated evaluation is set up. This cross-validation uses a technique called k-fold [37]. It essentially cuts each sample into k parts and keeps $k-1$ parts for training and only one for testing, and changes the test sample by withholding a different part of the sample. In our case, $k = 10$ was chosen. Moreover, each split is repeated 5 times with events selected each time randomly. The score of the cross-validation is based on the AUC, explained in Subsection 4.1.2.

Score computation

Two different scores, using the rank R_F attributed to each feature F and the correlation $C_{F,G}$ associated to two features F and G within each sample, are computed in this analysis. The algorithm that we propose consists of four steps:

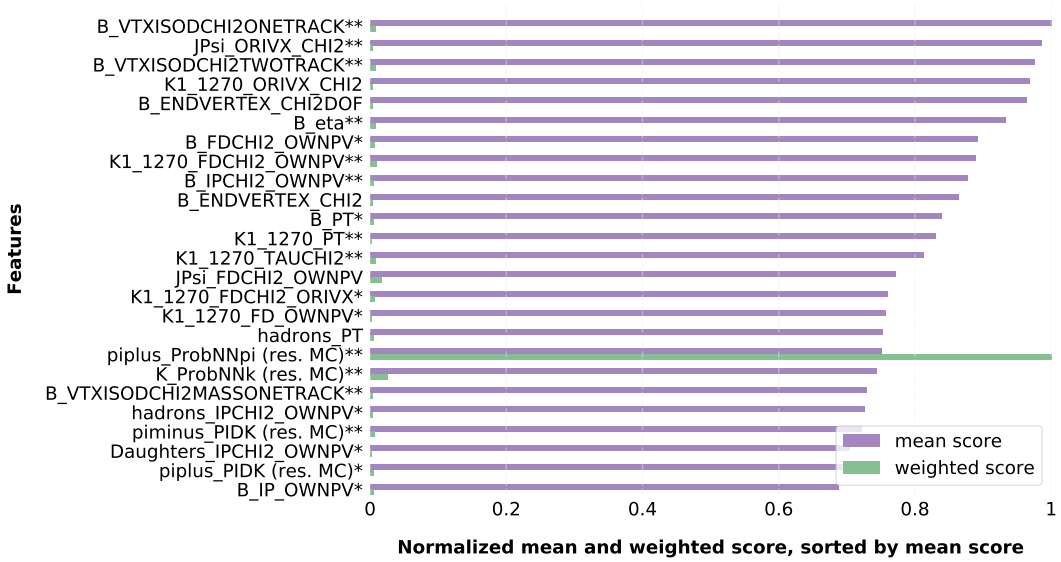
1. The rank R_F of each feature is an absolute number ranging from 1 to n_S, with 1 being the best. We decided first to normalise this rank to unity, which is referred later as score:
$$\rho_F = \frac{n_S - R_F}{n_S - 1} \tag{25}$$
It should be noted that $n_S = 20$ in our case, and corresponds to the number of features within each sample.

2. The introduction of the multiple use of features calls for a mean computation of the score, which is done for each feature F as:
$$\rho_F^{\mathrm{M}} = \frac{1}{n_F} \sum_{S_F} \rho_F \tag{26}$$
where n_F is the number of samples using the feature F, and S_F are the samples containing the feature F.

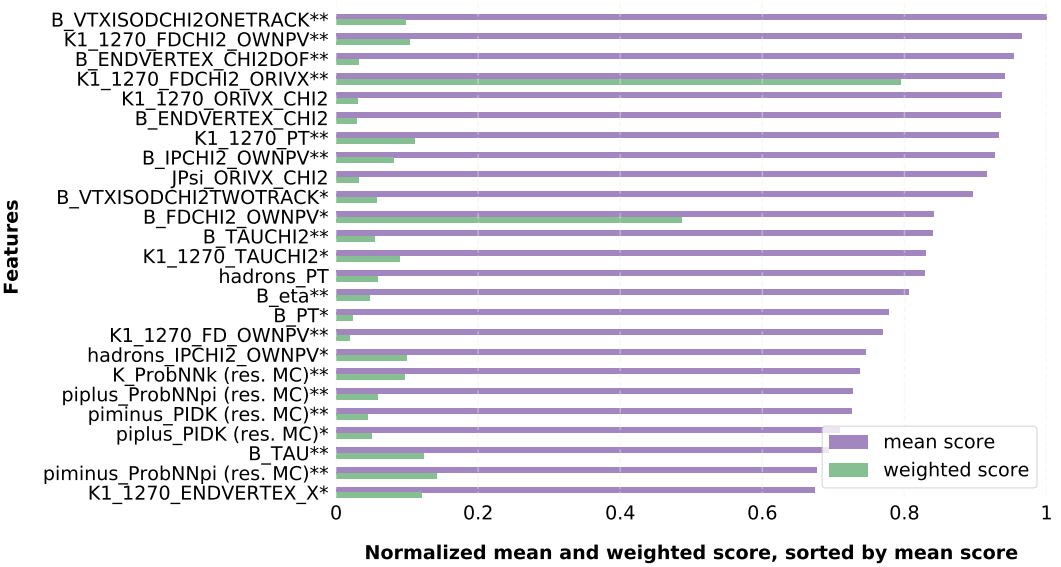
3. The mean score ρ_F^{M} does not take into account the correlation between the different features present within each sample. This motivates the use of weights that will tune the score depending on the correlations. These are computed by taking into account all other features G present in all the different samples a given feature F appears in, as:
$$W_F = \frac{1}{\sum_{S_F}(n_S - 1)} \sum_{S_F} \sum_{G \in S_F} \frac{\rho_G^{\mathrm{M}}}{|C_{F,G}|} \tag{27}$$
An absolute value on the correlation $C_{F,G}$ makes it possible to take into account only the effect of the latter, whether negative or positive.

4. These weights W_F are used to get a weighted score ρ_F^{W} of each feature F:
$$\rho_F^{\mathrm{W}} = W_F \cdot \rho_F^{\mathrm{M}} \tag{28}$$

A mean score and a weighted score are attributed to each selected feature present in both $\ell = e, \mu$ data sets provided by Subsection 4.3.1. Figure 18 shows a comparison between the best mean and weighted score, as well as the associated features. A general figure showing the scores of all features may be found in Appendix F. To select the final features to be used as input by the MVA method, the RFE technique is again employed. Several sets of features on which this technique has been applied were defined: two sets took the features associated to the 20 best mean and weighted score respectively, and two others used the features associated to the 20 best mean and weighted score respectively, which showed no correlation value of 1. The MVA method is then trained on all of them, and only the best performances have been retained.



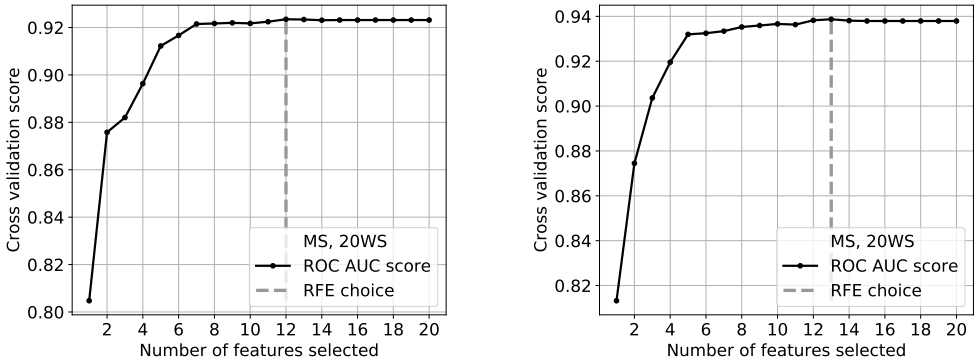
(a) $\ell = e$



(b) $\ell = \mu$

Figure 18: Comparison of features associated to the best mean and weighted score for both $\ell = e, \mu$ data sets. The features are sorted according to the mean score: the best feature is at the top of the graph with the highest value, and the worst is at the bottom. The features showing no correlations of 1 are indicated with an asterisk. The features selected by the RFE method are indicated with a double asterisk.

The choice of feature that stood out from the others are the one selecting the features associated to the 20 best mean score which showed no correlation value of 1. The RFE method evaluated the AUC score for these features, and selected 12 and 13 out of the 20, for electrons and muons data sets respectively, as demonstrated in Figure 19.



(a) $\ell = e$ (b) $\ell = \mu$

Figure 19: Evaluation of the AUC values performed by the RFE technique. The optimal number of features is 12 and 13 for $\ell = e, \mu$ data sets repectively.

The final discriminating features correspond to the ones indicated in Table 13 All the associated results are presented in Subsection 4.4. They are often referred as "MS 20WS", which is an abbreviation for the mean score and a selection of the 20 best score which showed no correlation value of 1.

$B^+ \to K^+\pi^+\pi^-(J/\psi \to e^+e^-)$	$B^+ \to K^+\pi^+\pi^-(J/\psi \to \mu^+\mu^-)$
B_VTXISODCHI2ONETRACK	B_VTXISODCHI2ONETRACK
JPsi_ORIVX_CHI2	K1_1270_FDCHI2_OWNPV
B _VTXISODCHI2TWOTRACK	B_ENDVERTEX_CHI2DOF
B_eta	K1_1270_FDCHI2_ORIVX
K1_1270_FDCHI2_OWNPV	K1_1270_PT
B_IPCHI2_OWNPV	B_IPCHI2_OWNPV
K1_1270_PT	B_eta
K1_1270_TAUCHI2	K1_1270_FD_OWNPV
piplus_ProbNNpi (res. MC)	K_ProbNNk (res. MC)
K_ProbNNk (res. MC)	piplus_ProbNNpi (res. MC)
B_VTXISODCHI2MASSONETRACK	piminus_PIDK (res. MC)
piminus_PIDK (res. MC)	B_TAU
	piminus_ProbNNpi (res. MC)

Table 13: Discriminating features selected by the RFE technique applied on the features associated to the 20 best mean score which showed no correlation value of 1.

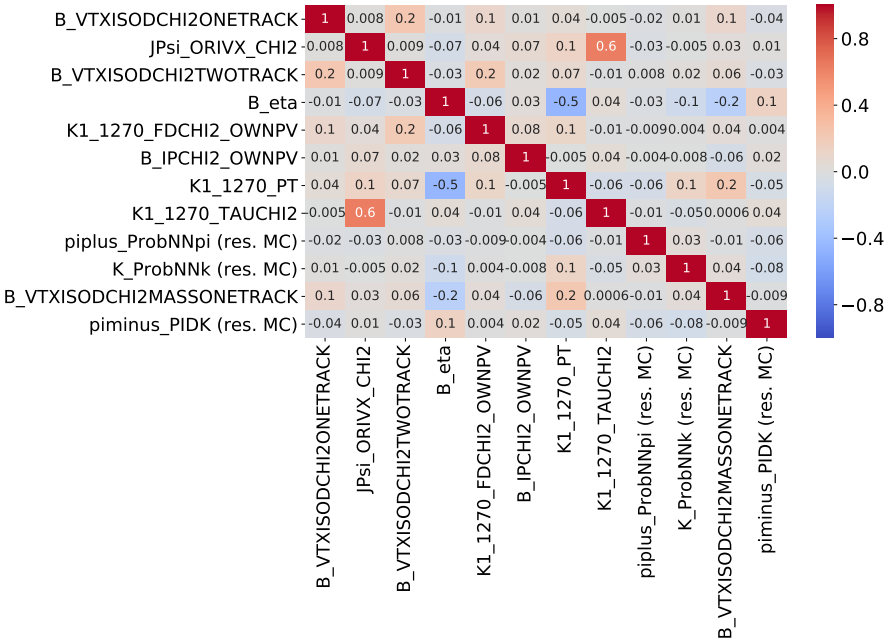
Several improvements that could be made have already been identified. The weighted score, for example, did not surpass the mean score results, although this was expected. This is certainly due to the fact that the weighted score does not take enough account of one of its two components, which could be done by incorporating tuning parameters a and b into Equation (28), as

$$\rho_F^{\mathrm{W}} = a \cdot W_F + b \cdot \rho_F^{\mathrm{M}} \tag{29}$$

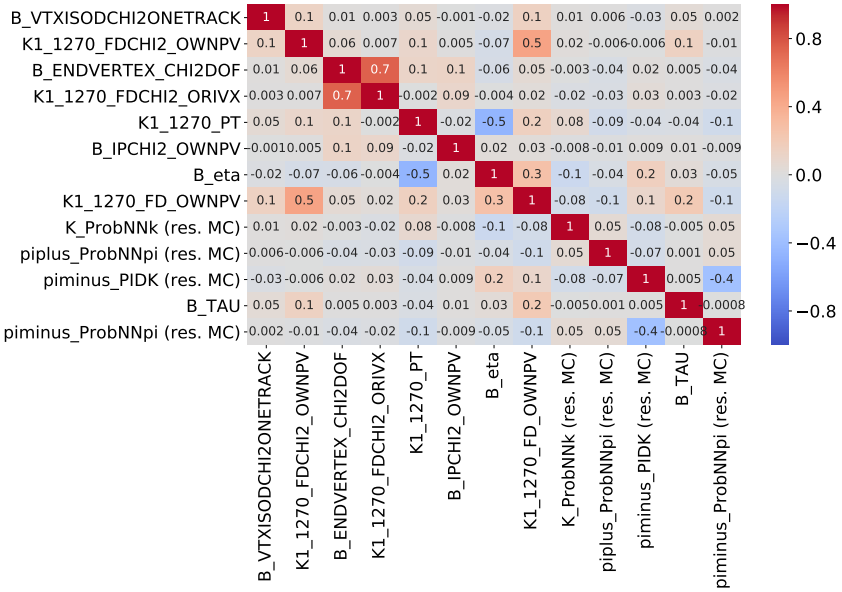
and optimising them. Another improvement concerns the RFE technique. It was found that not only a rank R, but a score ρ^{RFE} is computed by this technique, and is in fact automatically assigned to each feature. This score, which is not necessarily linear, could be used instead of the one computed in Equation (26), such as $\rho_F = \rho^{\text{RFE}}$.

4.4 MVA responses

The correlation matrices, displayed in Figure 20, demonstrate reasonable correlations, with an upper value of 0.6 between $\chi^2_\tau(K_1)$ and $\chi^2_{\text{SV}}(J/\psi)$ using electrons signal data set, and 0.7 between $\chi^2_{\text{FD}}(K_1, SV)$ and $\chi^2_{\text{VTX}}/\text{ndf}(B^+, PV)$ using muons ones.



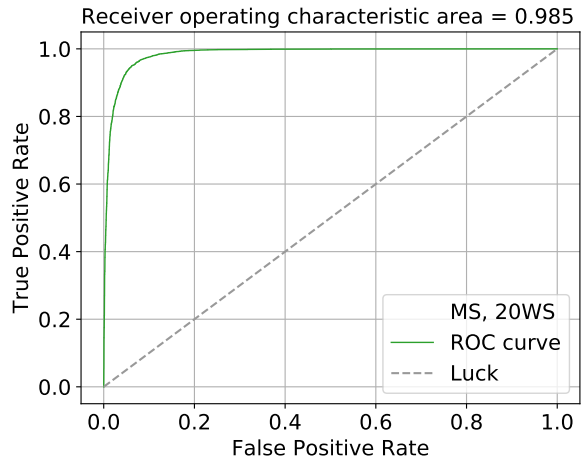
(a) $\ell = e$



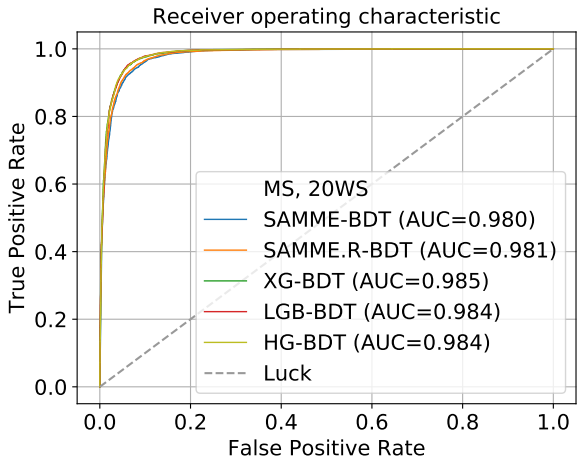
(b) $\ell = \mu$

Figure 20: Correlation matrices of the features associated to the 20 best mean score which showed no correlation value of 1, using electrons and muons signal data sets.

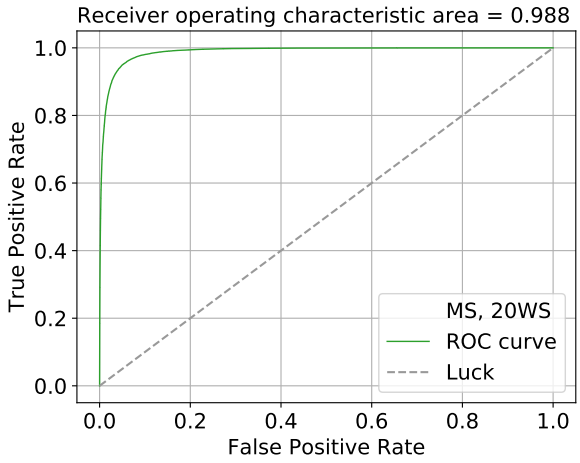
An area under curve (AUC) of 0.985 and 0.988 has been achieved for the electrons and muons data sets respectively, as reported in Figures 21. Several other boosting methods, such as SAMME and SAMME.R [31], Light Gradient Boosting Machine [40] (LGB) or Histogram-Based Gradient Boosting [41] (HG) were tested alongside the boosting opted in this analysis, which is the extreme gradient boosting (XG).



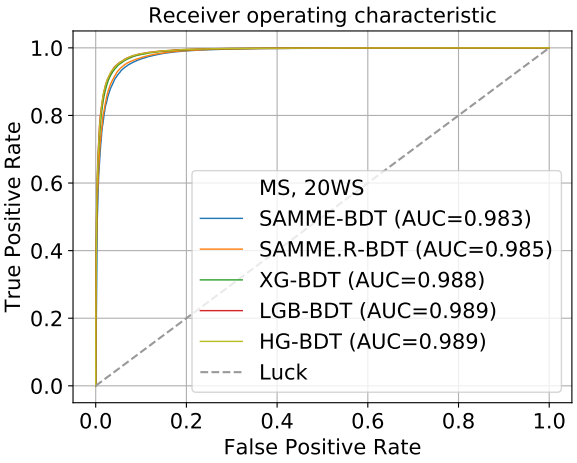
(a) $\ell = e$, XG-BDT



(b) $\ell = e$, Comparison



(c) $\ell = \mu$, XG-BDT

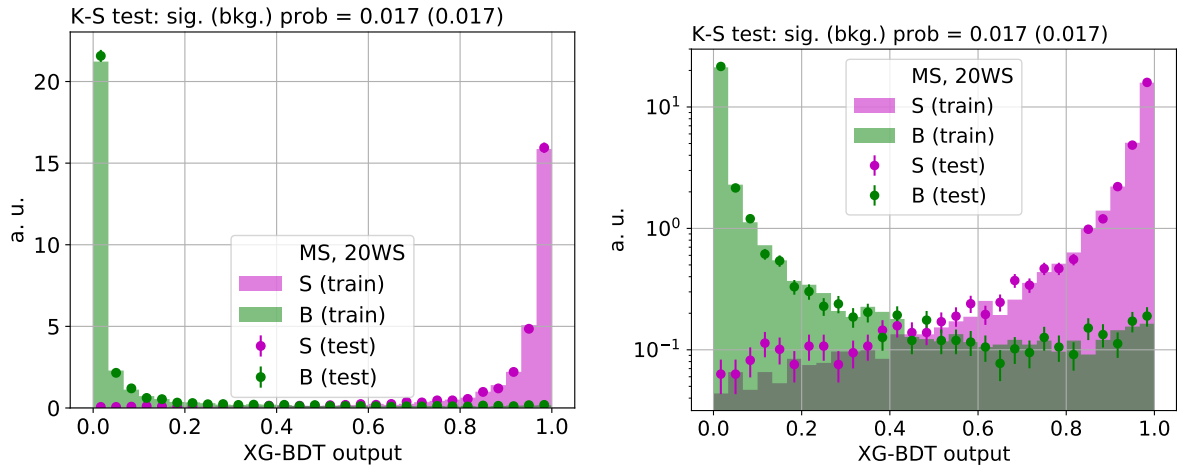


(d) $\ell = \mu$, Comparison

Figure 21: ROC curves obtained for the electrons and muons data sets using the XG-BDT method (left) and other methods (right).

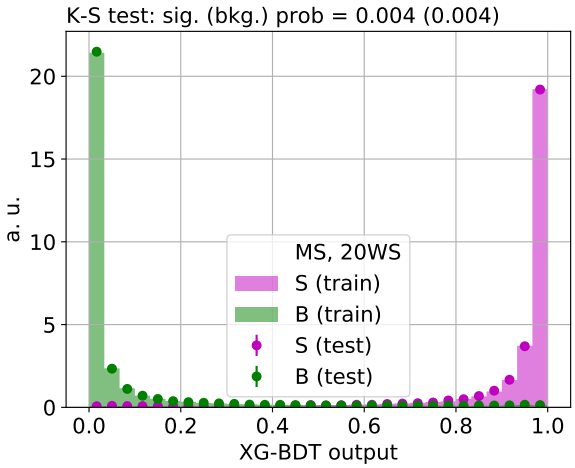
No excess of overtraining was detected neither for the electrons, nor for the muons data set, as displayed in Figures 22. The Kolmogorov–Smirnov test indicated a value of 0.017 and 0.004 for $\ell = e, \mu$ data sets respectively, which confirms the first assumption.

By combining all these results, we claim to use a very efficient MVA classification method. The small correlation values imply that the selected features are all useful, the ROC curve demonstrates a very high efficiency and the very low or non-existent overtraining shows that the fit is reproducible.

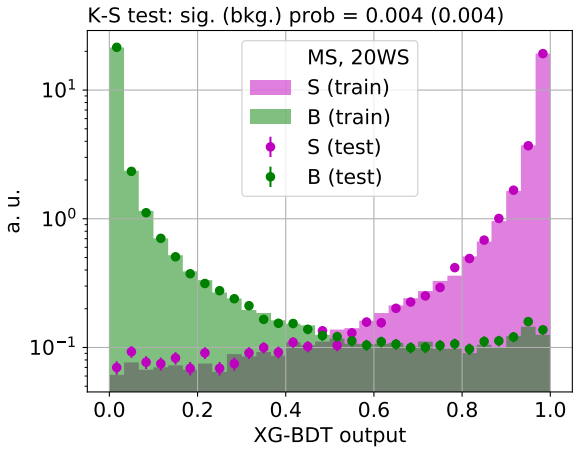


(a) $\ell = e$, linear

(b) $\ell = e$, logarithmic



(c) $\ell = \mu$, linear



(d) $\ell = \mu$, logarithmic

Figure 22: Overtraining check of the electrons and muons data sets. The classifier outputs are plotted in linear (left) and logarithmic (right) scales.

4.5 Comparison with previous MVA responses

By way of comparison, we repeat the same procedure with the features used in a previous study [4] with the identical purpose but for Run 1 LHCb data. The complete list of features is reported in Table 14.

$B^+ \to K^+\pi^+\pi^-(J/\psi \to \ell^+\ell^-)$	
B_PT	B_AMAXDOCA
JPsi_FDCHI2_OWNPV	B_P
B_ENDVERTEX_CHI2DOF	Daughters_PT
B_IPCHI2_OWNPV	min_daughter_IP
K1_1270_FDCHI2_OWNPV	abs_K1_1270_DIRA_OWNPV
B_VTXISODCHI2TWOTRACK	

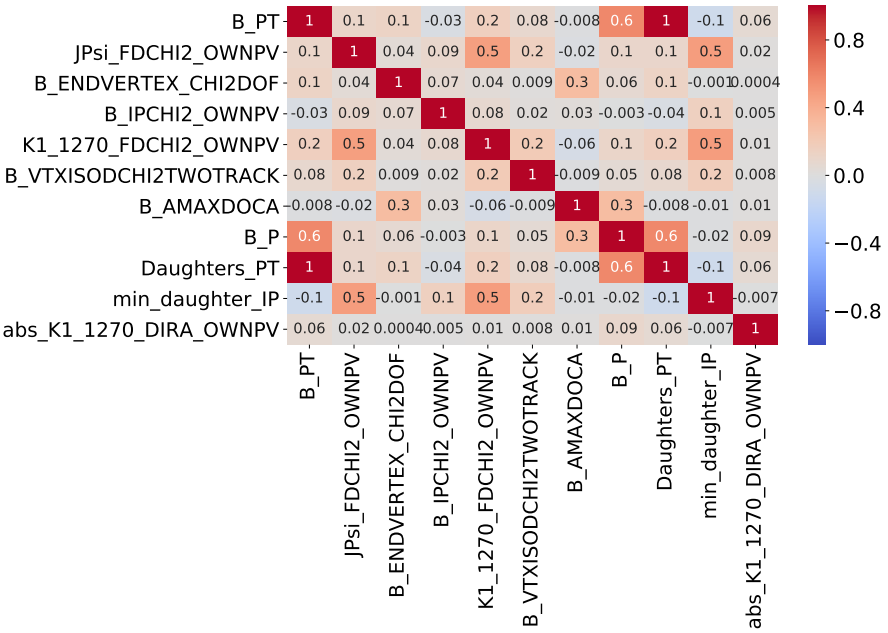
Table 14: Features used in the previous analysis, used for both electrons and muons data sets.

Most of the features used have already been introduced in a previous section, with the exception of two of them. The minimum of the daughter impact parameter $\min_{\text{IP}}$, which is defined as

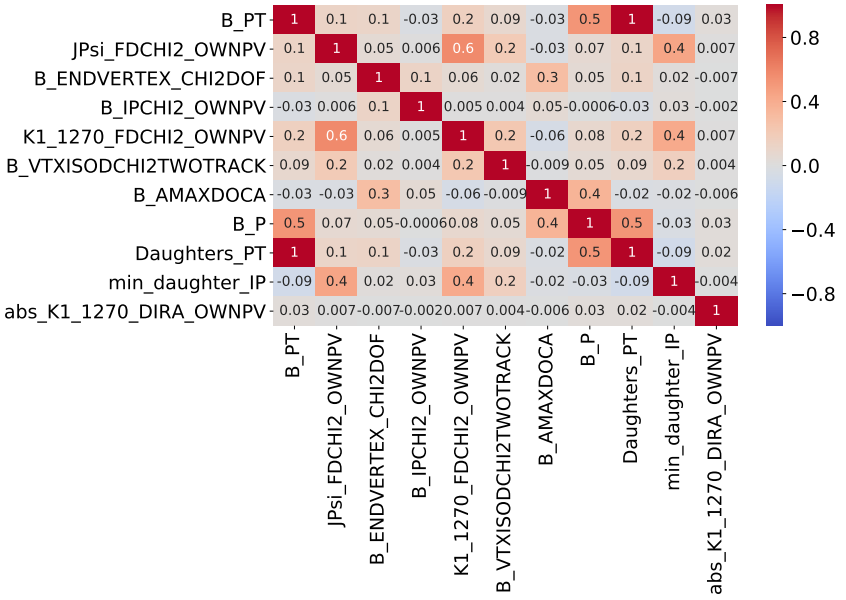
$$\min_{\text{IP}} = \min(\text{IP}(K^+), \text{IP}(\pi^+), \text{IP}(\pi^-), \text{IP}(\ell^+), \text{IP}(\ell^-)) \tag{30}$$

and the absolute value of the direction angle of the K_1 meson, denoted as $|\text{DIRA}(K_1, PV)|$.

The correlation matrices reveal in both cases that two features are highly correlated: $p_{\text{T}}(B^+)$ and $p_{\text{T}}(K^+\pi^+\pi^-\ell^+\ell^-)$. Several other features have a correlation value greater than or equal to 0.5, e.g. $\chi^2_{\text{FD}}(K_1, PV)$ and $\chi^2_{\text{FD}}(J/\psi, PV)$.

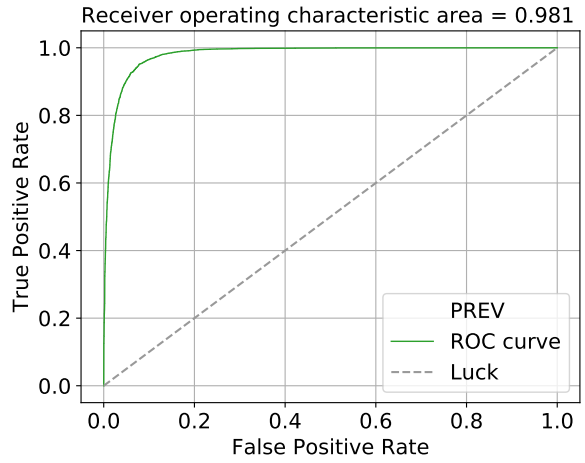


(a) $\ell = e$

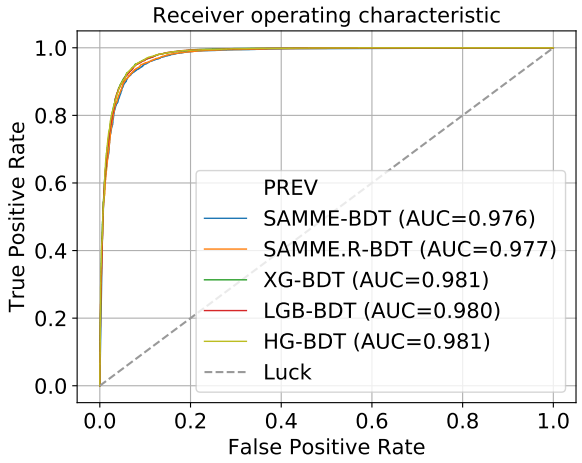


(b) $\ell = \mu$

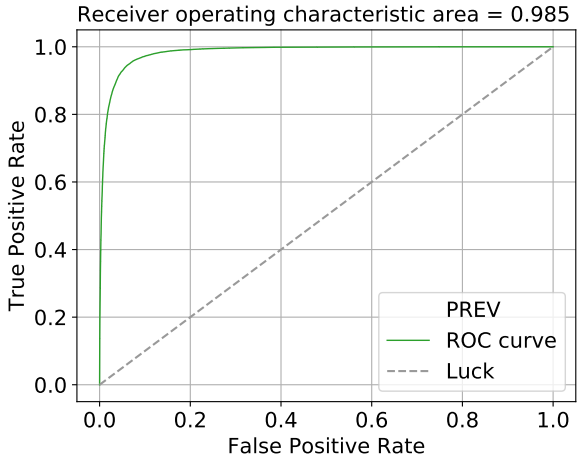
Figure 23: Correlation matrices of the features used in the previous analysis, using electrons and muons signal data sets.



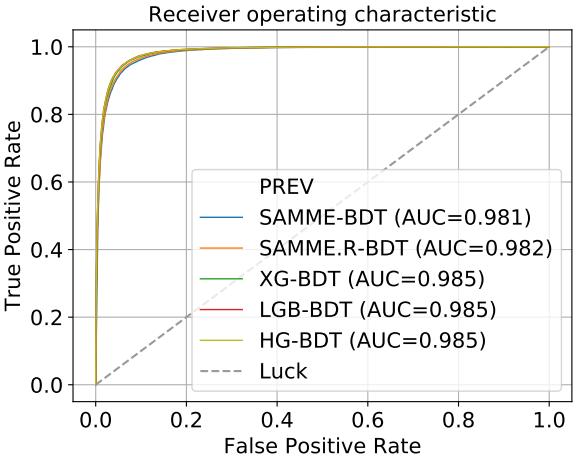
(a) $\ell = e$, XG-BDT



(b) $\ell = e$, Comparison



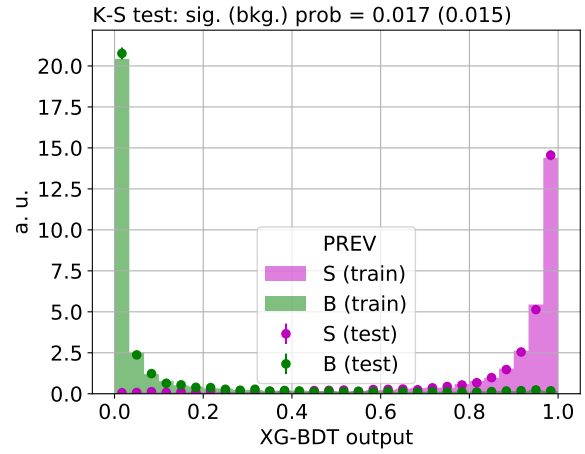
(c) $\ell = \mu$, XG-BDT



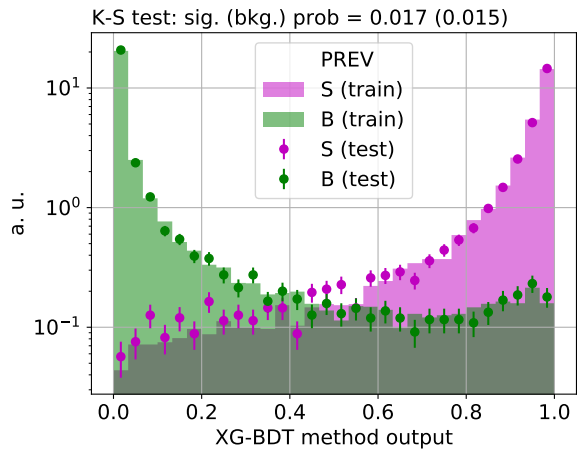
(d) $\ell = \mu$, Comparison

Figure 24: ROC curves obtained for the electrons and muons data sets using the XG-BDT method (left) and other methods (right).

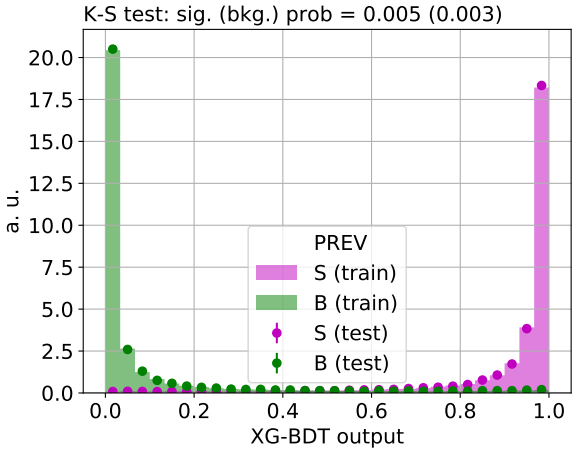
The AUC value of 0.981 and 0.985 for $\ell = e, \mu$ data sets respectively, shown in Figures 24, are in the same range as the values obtained previously. The same observation can be made for overtraining results. As Figures 25 proves, no excessive amount of overtraining is detected.



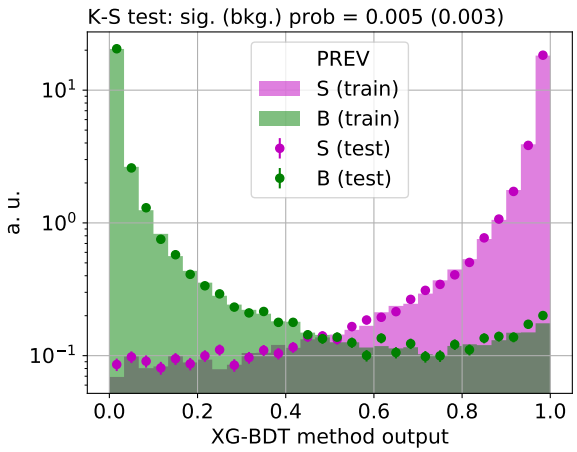
(a) $\ell = e$, linear



(b) $\ell = e$, logarithmic



(c) $\ell = \mu$, linear



(d) $\ell = \mu$, logarithmic

Figure 25: Overtraining check of the electrons and muons data sets. The classifier outputs are plotted in linear (left) and logarithmic (right) scales.

4.6 Optimisation of classifier output

The output of the MVA method is a single classifier on which a requirement can be placed to reject background events. This requirement corresponds to the maximum of the figure of merit (FoM):

$$\text{FoM} = \frac{S}{\sqrt{S+B}} \tag{31}$$

and can be found by the combination of an application of a scan on the XG-BDT output and an evaluation using a mass variable, i.e. the reconstructed B^+ meson mass. The variables defining the figure of merit are the expected signal (S) and background (B) yields. The computation of the former is based on the product of several components:

$$S = 2 \times \sigma_{b\bar{b}}(13\text{TeV}) \times \mathscr{L}_{\text{int}}(13\text{TeV}) \times f_u \times \mathscr{B}(B^+ \to K^+\pi^+\pi^-\ell^+\ell^-) \times \varepsilon_{\text{tot},\ell} \tag{32}$$

where $\sigma_{b\bar{b}}(13\text{TeV})$ and $\mathscr{L}_{\text{int}}(13\text{TeV})$ are the $b\bar{b}$ quark pair production cross section and the integrated luminosity respectively, for a centre-of-mass energy of the collisions of 13 TeV, f_u the hadronisation probability, $\mathscr{B}(B^+ \to K^+\pi^+\pi^-\mu^+\mu^-)$ the branching fraction introduced in Subsection 1.2, and used for both $B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$, $\varepsilon_{\text{tot},\ell}$ the total efficiency to detect, reconstruct and select the regarding decay. The numerical values of the inputs are summarised in Table 15.

Quantity	$B^+ \to K^+\pi^+\pi^-(J/\psi \to \ell^+\ell^-)$
$\sigma_{b\bar{b}}(13\text{TeV})$	560μb [42]
$\mathscr{L}_{\text{int}}(13\text{TeV})$	1.7fb^{-1} [19]
f_u	$(40.1 \pm 0.8)\%$ [18]
$\mathscr{B}(B^+ \to K^+\pi^+\pi^-\mu^+\mu^-)$	$(4.36 \pm 0.40) \times 10^{-7}$ [10]
ε_e	$(0.6985 \pm 0.0025)\%$
ε_μ	$(1.0277 \pm 0.0032)\%$

Table 15: Numerical values of the inputs used for the optimisation of the FoM.

The factor 2 is a direct consequence of the implication of charge conjugation. It comes from the fact that both b and $\bar{b}$ quarks can hadronise into a $B^{\pm}$ meson. It should be also noted that the optimisation should be performed for the rare decay signal i.e. $B^+ \to K\pi^+\pi^-\ell^+\ell^-$. While in this study we used simulated and collision data for $B^+ \to K\pi^+\pi^-(J/\psi \to \ell^+\ell^-)$, the use of $\mathcal{B}(B^+ \to K^+\pi^+\pi^-\ell^+\ell^-)$ in Equation (32) allows to approximate the final threshold values that will be obtained once rare decay simulation is fully reweighted. As Equation (33) shows, $\varepsilon_{\text{tot},\ell}$ is computed as the product of several efficiencies related to the geometrical acceptance of the LHCb detector defined as its solid angle coverage ($\varepsilon_{\text{geom}}$), the stripping and reconstruction ($\varepsilon_{\text{strip, reco}}$), the selection applied on the Monte Carlo sample (ε_{sel}), and the one applied on the XG-BDT classifier (ε_{BDT}).

$$\varepsilon_{\text{tot},\ell} = \varepsilon_{\text{geom}} \times \varepsilon_{\text{strip, reco}} \times \varepsilon_{\text{sel}} \times \varepsilon_{\text{BDT}} \equiv \varepsilon_{\ell} \times \varepsilon_{\text{BDT}} \tag{33}$$

All these efficiencies are computed as the ratio of number of events after the constraint over the number of events before, e.g. ε_{BDT} is defined as:

$$\varepsilon_{\text{BDT}} = \frac{n_{\text{ev., XG-BDT}}}{n_{\text{ev., tot}}} \frac{n_{\text{sig}}}{n_{\text{ev., XG-BDT}}} = \frac{n_{\text{sig}}}{n_{\text{ev., tot}}} \tag{34}$$

where $n_{\text{ev., tot}}$ is the total number of events, $n_{\text{ev., XG-BDT}}$ the number of events remaining after the lower cut on the XG-BDT during the scan process, and n_{sig} the number of signal events selected by the cut that are also included in the range signal of the mass variable. This range corresponds to the smallest interval containing 95% of the signal simulation of the electrons and the muons respectively, as shown in Figure 26.

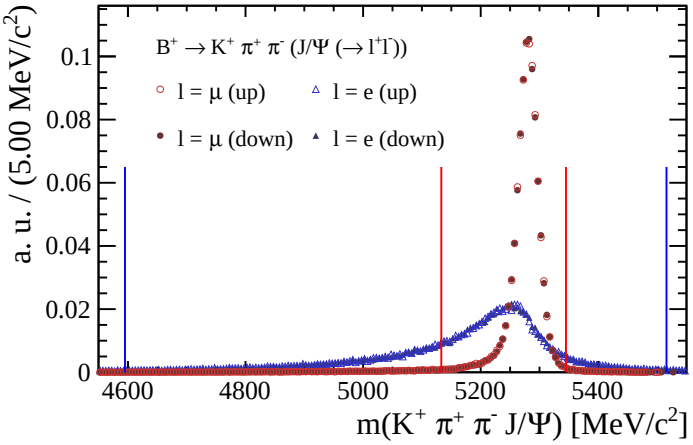
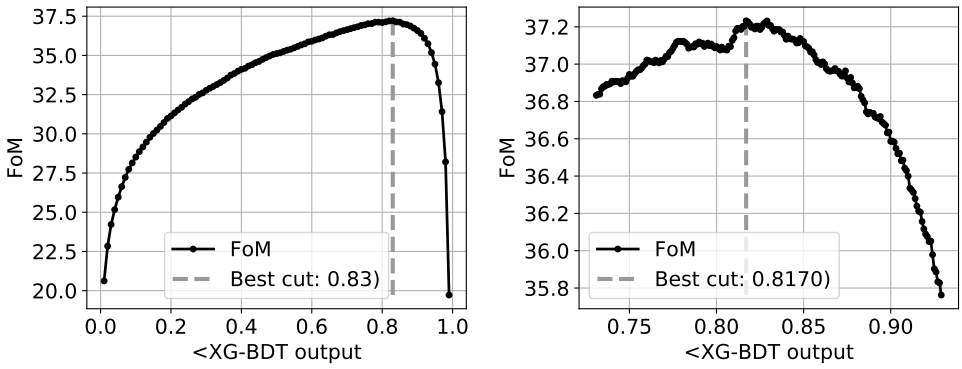


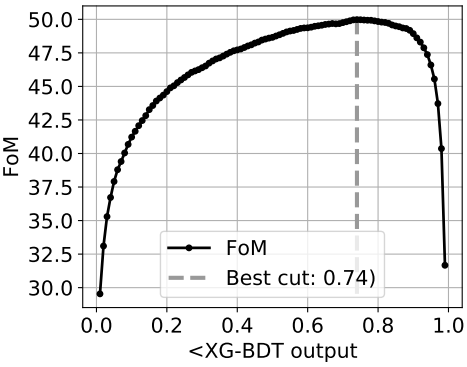
Figure 26: Distribution of the reconstructed B^+ meson mass using $\ell = e, \mu$ data sets and differentiating the up and down magnet polarities. A sub-interval [4550, 5550] MeV/c^2 was preferred to highlight the mass peak, even though the events cover an wider range. Both ranges [4595, 5516] MeV/c^2 and [5133, 5345] MeV/c^2 for $\ell = e, \mu$ data sets respectively are represented.

The expected background yields B is computed by taking the number of background events selected by the cut that is also included in the range signal of the mass variable. For both the efficiency computation and figure of merit optimisation, no weight was taken into account. A scan of 100 XG-BDT cuts on the range [0,1] allowed to find a first approximation of the maximum value of Equation (31) for both electrons and muons data sets, as displayed in Figure 27. Two more detailed scans of 200 XG-BDT cuts on [0.73, 0.93] and [0.64, 0.84] ranges for the electrons and muons data sets respectively allowed to find a maximum value reported in Table 16.

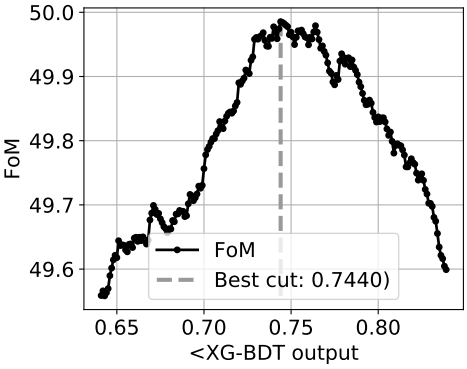


(a) $\ell = e$, 100 XG-BDT cuts

(b) $\ell = e$, 200 XG-BDT cuts



(c) $\ell = \mu$, 100 XG-BDT cuts



(d) $\ell = \mu$, 200 XG-BDT cuts

Figure 27: Optimisation of the FoM in 1D, using the classifier output. A scan over the whole interval [0, 1] (left) and a smaller region, surrounding the maximum value, [0.73, 0.93] and [0.64, 0.84] for $\ell = e, \mu$ data sets respectively (right) are considered.

	$B^+ \to K^+\pi^+\pi^-(J/\psi \to e^+e^-)$	$B^+ \to K^+\pi^+\pi^-(J/\psi \to \mu^+\mu^-)$
XG-BDT	0.817	0.744
FoM	37.23	49.99
ε_{sig}	0.826	0.886
ε_{bkg}	0.019	0.003

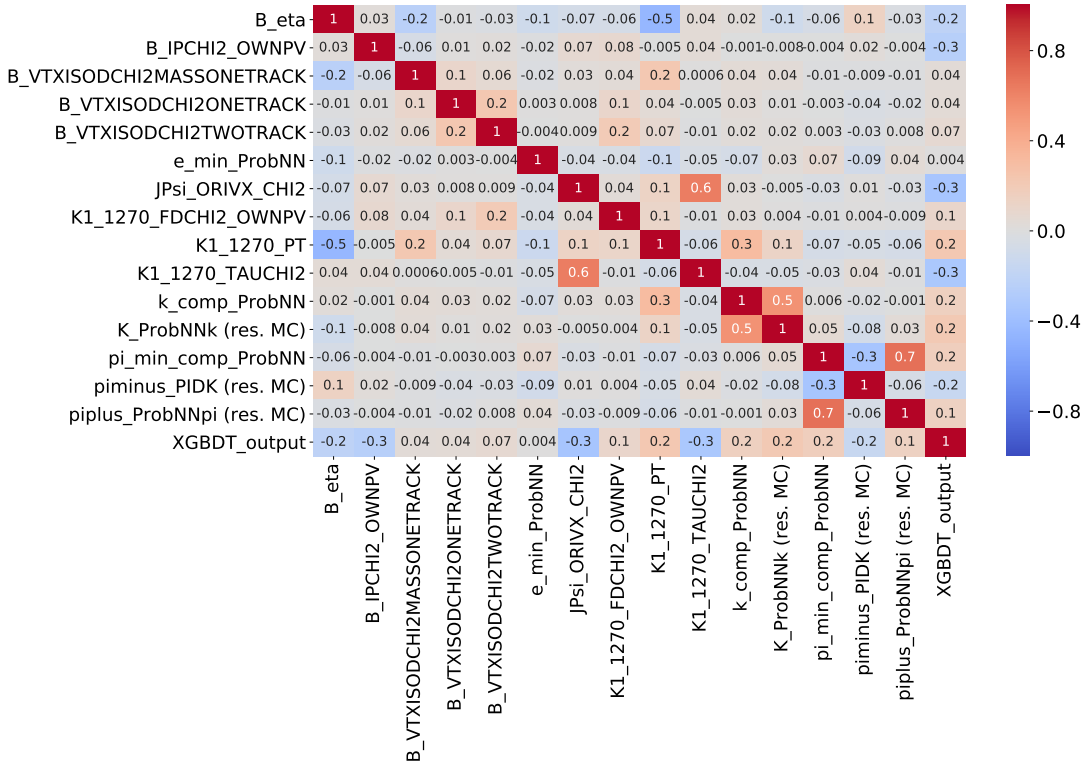
Table 16: Optimal values found in the smaller region scan of the classifier output.

The preselection applied on Monte Carlo and data samples only contains very loose PID cuts, i.e. a limit of 0.2 is imposed on tracks. In order to further discriminate between signal and background, and removing decays with misidentified particles, an optimisation of the figure of merit with requirements on the XG-BDT response and composed PID variables of the decay daughters is performed. The definition of the three composed PID variables is reported in Table 17, using non resampled PID variables.

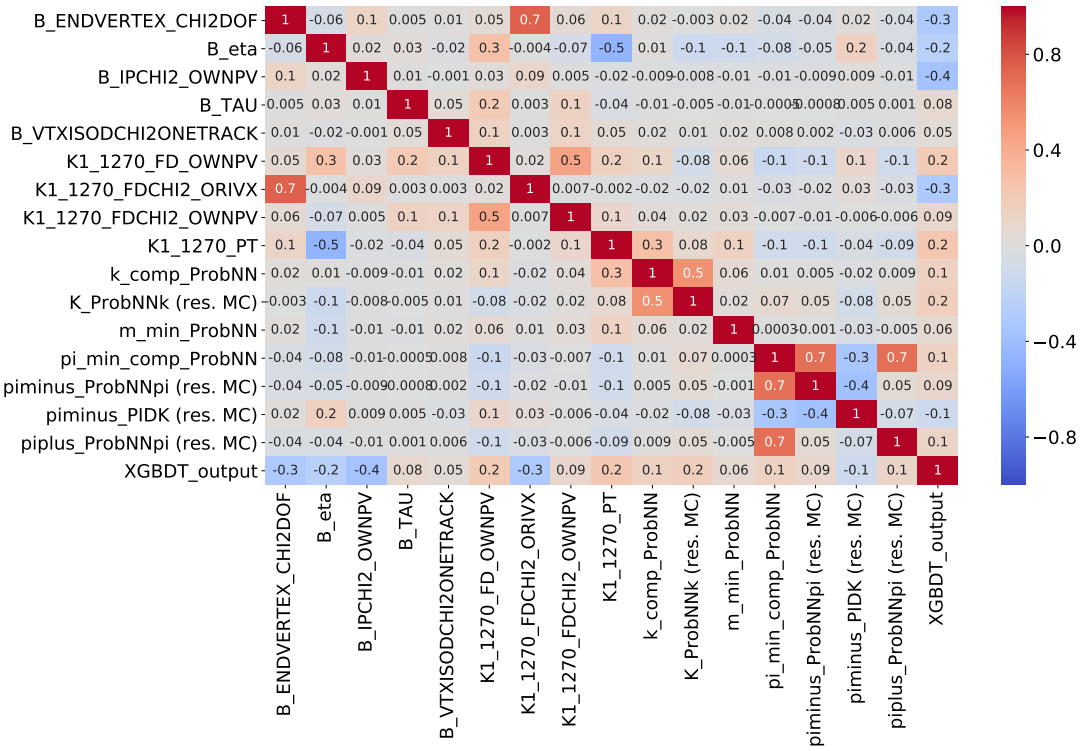
Name	Definition
PIDℓ	min(ProbNNℓ ($\ell^\pm$))
PIDK	ProbNNk(K^+) $\times$ (1$-$ProbNNp(K^+))
PIDπ	min(ProbNNπ($\pi^\pm$) $\times$ (1$-$ProbNNk($\pi^\pm$)) $\times$ (1$-$ProbNNp($\pi^\pm$))

Table 17: Definition of the PID variables used for the optimisation of the FoM in 4D.

The correlation matrix of the discriminating features and the four variables on which the figure of merit is optimised is displayed in Figure 28.



(a) $\ell = e$



(b) $\ell = \mu$

Figure 28: Correlation matrices of the discriminating, BDT and PID features, using electrons and muons signal data sets. The PID variables are referred as e_min_ProbNN, m_min_ProbNN, k_comp_ProbNN and pi_min_comp_ProbNN.

The use of these composed PID variables reduced to a minimum value allows the reduction of the dimensionality of the optimisation problem while using two oppositely charged variables. This 4D optimisation led to the values displayed in Table 18. It should be noted that the figure of merit value is negative. It is due to the fact that the minimum of the negative FoM was employed instead of the maximum to be able to use existing numerical methods. Moreover, its computation is constrained on a smaller range of the XG-BDT, given by [0.2,1], which makes the search for the optimum much more efficient.

	$B^+ \to K^+\pi^+\pi^-(J/\psi \to e^+e^-)$	$B^+ \to K^+\pi^+\pi^-(J/\psi \to \mu^+\mu^-)$
XG-BDT	0.8283	0.7446
PIDℓ	0.1152	0.1570
PIDK	0.0055	0.0020
PIDπ	0.0206	0.0200
FoM	-37.26	-50.02
ε_{sig}	0.820	0.889
ε_{bkg}	0.075	0.012

Table 18: Optimal values found for the 4D optimisation.

The optimised selections for both final states yield a signal efficiency of 82.1% (88.9%) while removing 92.5% (98.8%) of the combinatorial background for the electrons (muons) data sets.

Conclusions and outlook

In this thesis, we presented various computational procedures representing a subset of the required steps that need to be performed in the measurement of $R_{K\pi\pi}$ using LHCb 2017 data. More precisely, we aimed at improving the simulation and the data of these decays so that it is possible to calculate LHCb's detection efficiency for the rare $B \to K\pi\pi\ell\ell$ decays, a crucial ingredient to turn measured yields into a ratio of branching fractions. To achieve this, we used resonant $B \to K\pi\pi(J/\Psi \to \ell\ell)$ decays as a proxy. This approach is justified by the fact that the final state is identical to the signal we look for but it offers the advantage of being much more abundant.

It should be noted that the results obtained in previous work [1] already represented a step forward that should not be overlooked. Moreover, it provided resampled PID variables and Monte Carlo weights related to kinematics and multiplicity variables correction, notably thanks to the application of the ${}_s\mathscr{P}lot$ technique on the reconstructed B^+ meson mass.

These two corrections were very useful for the classification of events into background and signal categories. This crucial task, that represents the core of this analysis, was performed through several stages which enabled the process to be optimised. But all trials have one thing in common: they use a MVA method that takes as input, apart from hyperparameters, training and test sets, as well as a set of discriminating features.

The SAMME-BDT method was used mainly to test the machine learning procedure. Then, it was quickly detected that the XG-BDT method was outperforming its results, and the method was therefore changed. It was trained on 70% of both background and signal data sets and then tested on the remaining parts. These data sets were defined using data and Monte Carlo samples respectively; a background category cut for the signal and a reconstructed B^+ meson mass one for the background, complemented by a balance of number of events due to the use of Monte Carlo weights, were required for this task. The selection of discriminating features was made from candidates demonstrating a significance from a physics point of view and sufficient discriminating power. Moreover, a previous analysis [4] provided the computation of additional useful features. Scores calculation assigned to each feature has been set up to highlight and select the most promising ones. We more precisely defined two types of score that allowed to take into account the results of the RFE technique and the correlations calculated within samples gathering about twenty features. The features presenting the 20 best mean scores and that showed no correlation value of 1 were selected, and the application of the RFE technique once again allowed a further selection that defined the final set of discriminating features.

The results following these considerations revealed interesting performances; The correlations between features did not exceed the value of 0.7, there was no apparent overtraining and the AUC reached values of 0.985 and 0.988 for electrons and muons data sets respectively. These results, in light of those obtained by the features from the previous analysis, show an excellent separation of events, and proposes an alternative to the latter, which was long overdue.

The optimisation of the figure of merit completes the analysis. It was performed in 1 and 4 dimensions, considering the classifier output and three compositions of PID variables. This yielded a signal efficiency of 82.1% and 88.9% while removing 92.5% and 98.8% of the combinatorial background for the electrons and muons data sets respectively.

Even if the results achieved in this work are satisfactory, they also revealed some im-

provements that may be addressed in subsequent work. As already mentioned, the computation of the correlation could have taken into account the signal set weights. The feature sampling could contain additional constrains as to avoid using a feature more than the others, and reducing the number of samples by considering a recursive method. A major point would be to adapt the computation of the weighted score so that it manages to find a balance between the weights, resulting from the correlations, and the mean score. Some further optimisations of the MVA method may be conducted, such as the set up of a cross validation process that could avoid bias or the optimisation of the hyperparameters. Finally, concerning the optimisation of the figure of merit, it should and will use the rare decay $B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$ once the corresponding Monte Carlo has gone through the whole reweighting chain and becomes available, although a significant change is not expected because the various selection efficiencies are very similar for the resonant and rare decays.

The study of 2011-2016 LHCb data has been validated via the $r_{J/\psi}$ test, which comprises the efficiencies computation and the Monte Carlo reweighting procedure. This work, presenting a MVA-based signal selection that has been obtained for the last data set, will complement the ongoing analysis, and provides a path forward for the use of the full LHCb dataset (2011-2018). Once the efficiencies of 2017 and 2018 will be validated, and the corresponding remaining Monte Carlo reweightings performed, the ensuing systematic uncertainties will be estimated, and the measurement of $R_{K\pi\pi}$ performed.

Acknowledgements

Although this work is my own, it would likely not happen without the help of many people whom I would like to thank.

This flavoured adventure began in April 2020 during my specialisation semester, while the world was dealing with a virus that won't be named here. I was lucky enough to be able to work remotely under the remarkable supervision of Elena. Already at that moment she knew how to guide me through the numerous challenges that this great project concerning LFU tests represented.

My Master's thesis allowed me to meet other great minds. I had the occasion to work with Sara, who gave me very precious advice, particularly for the end of my project. And I had the pleasure to follow and listen on zoom every Tuesday morning the $R_{K\pi\pi}$ research group work.

Because I always liked my projects during my Master studies at the LPHE, I would like to thank Prof. Schneider and Prof. Shchutska as well. And even if I had to continue my work at home rather than at EPFL in the middle of the semester, I had the opportunity to be surrounded by other students, Jennifer and Fabrice, who knew how to make the days a little less gloomy on an empty campus.

Finally, I conclude this part by thinking of my family, my close friends, and Lucas. I could say a lot more, but I would just add that without you and your endless support, this would not have been possible.

References

[1] Géraldine Räuber. *LFU test with 2017 LHCb data: Agreement between simulation and data for the study of $B \to K\pi\pi\ell\ell$ decays*. 2020.

[2] R. Aaij et al. "Search for Lepton-Universality Violation in $B^+ \to K^+\ell^+\ell^-$ Decays". In: *Phys. Rev. Lett.* 122 (19 May 2019), p. 191801. DOI: 10.1103/PhysRevLett.122.191801. URL: https://link.aps.org/doi/10.1103/PhysRevLett.122.191801.

[3] R. Aaij et al. "Test of lepton universality with $B^0 \to K^{*0}\ell^+\ell^-$ decays". In: *Journal of High Energy Physics* 2017.8 (Aug. 2017), p. 55. ISSN: 1029-8479. DOI: 10.1007/JHEP08(2017)055.

[4] Tobias Tekampe. "Probing the standard model of particle physics with rare B decays". PhD thesis. Tech. U., Dortmund (main), 2019. DOI: 10.17877/DE290R-20178.

[5] Wikimedia Commons. *Standard Model of Elementary Particles*. 2019. URL: https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg.

[6] Belle Collaboration et al. *Test of lepton flavor universality in $B \to K^*\ell^+\ell^-$ decays at Belle*. 2019. arXiv: 1904.02440 [hep-ex].

[7] J.-T. Wei et al. "Measurement of the Differential Branching Fraction and Forward-Backword Asymmetry for $B \to K^{(*)}\ell^+\ell^-$". In: *Phys. Rev. Lett.* 103 (2009), p. 171801. DOI: 10.1103/PhysRevLett.103.171801. arXiv: 0904.0770 [hep-ex].

[8] J. P. Lees et al. "Measurement of branching fractions and rate asymmetries in the rare decays $B \to K^{(*)}\ell^+\ell^-$". In: *Phys. Rev. D* 86 (3 Aug. 2012), p. 032012. DOI: 10.1103/PhysRevD.86.032012. URL: https://link.aps.org/doi/10.1103/PhysRevD.86.032012.

[9] H. Guler et al. "Study of the $K^+\pi^+\pi^-$ final state in $B^+ \to J/\psi K^+\pi^+\pi^-$ and $B^+ \to \psi' K^+\pi^+\pi^-$". In: *Phys. Rev. D* 83 (3 Feb. 2011), p. 032005. DOI: 10.1103/PhysRevD.83.032005. URL: https://link.aps.org/doi/10.1103/PhysRevD.83.032005.

[10] Roel Aaij et al. "First observations of the rare decays $B^+ \to K^+\pi^+\pi^-\mu^+\mu^-$ and $B^+ \to \phi K^+\mu^+\mu^-$". In: *JHEP* 10 (2014), p. 064. DOI: 10.1007/JHEP10(2014)064. arXiv: 1408.1137 [hep-ex].

[11] Sara Celani. "Testing Lepton Flavour Universality with $B \to K\pi\pi\ell\ell$ decays". Dec. 2020. URL: https://indico.cern.ch/event/973067/contributions/4128919/attachments/2158520/3641774/LHCb_week.pdf.

[12] The LHCb Collaboration et al. "The LHCb Detector at the LHC". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08005–S08005. DOI: 10.1088/1748-0221/3/08/s08005. URL: https://doi.org/10.1088/1748-0221/3/08/S08005.

[13] R Aaij et al. "Performance of the LHCb Vertex Locator". In: *Journal of Instrumentation* 9.09 (Sept. 2014), P09007–P09007. ISSN: 1748-0221. DOI: 10.1088/1748-0221/9/09/p09007. URL: http://dx.doi.org/10.1088/1748-0221/9/09/P09007.

[14] Eduardo Picatoste Olloqui and the LHCb Collaboration. "LHCb Preshower(PS) and Scintillating Pad Detector (SPD): Commissioning, calibration, and monitoring". In: *Journal of Physics: Conference Series* 160 (Apr. 2009), p. 012046. DOI: 10.1088/1742-6596/160/1/012046. URL: https://doi.org/10.1088%5C%2F1742-6596%5C%2F160%5C%2F1%5C%2F012046.

[15] T Head. "The LHCb trigger system". In: *Journal of Instrumentation* 9.09 (Oct. 2014), pp. C09015–C09015. DOI: 10.1088/1748-0221/9/09/c09015. URL: https://doi.org/10.1088%5C%2F1748-0221%5C%2F9%5C%2F09%5C%2Fc09015.

[16] Renato Quagliani. "Novel real-time alignment and calibration of LHCb detector for Run II and tracking for the upgrade." In: *Journal of Physics: Conference Series* 762 (Oct. 2016), p. 012046. DOI: 10.1088/1742-6596/762/1/012046.

[17] LHCb Starterkit. *The LHCb data flow*. URL: https://lhcb.github.io/starterkit-lessons/first-analysis-steps/dataflow.html.

[18] P.A. Zyla et al. "Review of Particle Physics". In: *PTEP* 2020.8 (2020), p. 083C01. DOI: 10.1093/ptep/ptaa104.

[19] CERN. *Standard set of performance numbers*. URL: https://lhcb.web.cern.ch/lhcb/speakersbureau/html/PerformanceNumbers.html.

[20] *RooStats::SPlot Class Reference*. URL: https://root.cern.ch/doc/master/classRooStats_1_1SPlot.html.

[21] Francesco Spanò. "Unfolding in particle physics: A window on solving inverse problems". In: *EPJ Web of Conferences* 55 (July 2013), pp. 03002–. DOI: 10.1051/epjconf/20135503002.

[22] Muriel Pivk and Francois R. Le Diberder. "SPlot: A Statistical tool to unfold data distributions". In: *Nucl. Instrum. Meth. A* 555 (2005), pp. 356–369. DOI: 10.1016/j.nima.2005.08.106. arXiv: physics/0402083.

[23] Yuehong Xie. *sFit: a method for background subtraction in maximum likelihood fit*. 2009. arXiv: 0905.0724 [physics.data-an].

[24] A. Poluektov. "Kernel density estimation of a multidimensional efficiency profile". In: *Journal of Instrumentation* 10.02 (Feb. 2015), P02011–P02011. DOI: 10.1088/1748-0221/10/02/p02011. URL: https://doi.org/10.1088%5C%2F1748-0221%5C%2F10%5C%2F02%5C%2Fp02011.

[25] Quagliani Renato. "Study of double charm B decays with the LHCb experiment at CERN and track reconstruction for the LHCb upgrade". PhD thesis. Orsay, LAL, Sept. 2017.

[26] Alex Rogozhnikov. *Reweighting algorithms*. 2017. URL: https://arogozhnikov.github.io/hep_ml/reweight.html.

[27] Katherine Woodruff. "Introduction to boosted decision trees". Sept. 2017. URL: https://indico.fnal.gov/event/15356/contributions/31377/attachments/19671/24560/DecisionTrees.pdf.

[28] xgboost developers. *Introduction to Boosted Trees*. URL: https://xgboost.readthedocs.io/en/latest/tutorials/model.html#:~:text=XGBoost%5C%20stands%5C%20for%5C%20%5C%E2%5C%80%5C%9CExtreme%5C%20Gradient,Gradient%5C%20Boosting%5C%20Machine%5C%2C%5C%20by%5C%20Friedman..

[29] Yoav Freund and Robert E Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1504. URL: http://www.sciencedirect.com/science/article/pii/S002200009791504X.

[30] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." In: *Ann. Statist.* 29.5 (Oct. 2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.

[31] Trevor Hastie et al. “Multi-class AdaBoost”. In: *Statistics and Its Interface* 2.3 (2009), pp. 349–360. ISSN: 19387989, 19387997. DOI: 10.4310/SII.2009.v2.n3.a8.

[32] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Aug. 2016). DOI: 10.1145/2939672.2939785. URL: http://dx.doi.org/10.1145/2939672.2939785.

[33] xgboost developers. *XGBoost Parameters*. URL: https://xgboost.readthedocs.io/en/latest/parameter.html.

[34] Sarang Narkhede. *Understanding AUC - ROC Curve*. URL: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[35] A. Hoecker et al. “TMVA - Toolkit for Multivariate Data Analysis”. In: *arXiv:physics/0703039* (July 2009). arXiv: physics/0703039. URL: http://arxiv.org/abs/physics/0703039.

[36] In: *The Concise Encyclopedia of Statistics*. Springer New York, 2008, pp. 283–287. ISBN: 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1_214. URL: https://doi.org/10.1007/978-0-387-32833-1_214.

[37] Avrim Blum, Adam Kalai, and John Langford. “Beating the Hold-out: Bounds for K-fold and Progressive Cross-validation”. In: *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*. COLT ’99. Santa Cruz, California, USA: ACM, 1999, pp. 203–208. ISBN: 1-58113-167-4. DOI: 10.1145/307400.307439. URL: http://doi.acm.org/10.1145/307400.307439.

[38] Wikipedia. *Correlation and dependence*. URL: https://en.wikipedia.org/wiki/Correlation_and_dependence.

[39] Isabelle Guyon et al. “Gene Selection for Cancer Classification using Support Vector Machines”. In: *Machine Learning* 46.1 (Jan. 2002), pp. 389–422. ISSN: 1573-0565. DOI: 10.1023/A:1012487302797.

[40] Guolin Ke et al. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: *Advances in Neural Information Processing Systems* 30 (2017), pp. 3146–3154.

[41] Aleksei Guryanov. “Histogram-Based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees”. In: Dec. 2019, pp. 39–50. DOI: 10.1007/978-3-030-37334-4_4.

[42] R. Aaij et al. “Measurement of the b-Quark Production Cross Section in 7 and 13 TeV pp Collisions”. In: *Physical Review Letters* 118.5 (Feb. 2017). ISSN: 1079-7114. DOI: 10.1103/physrevlett.118.052002. URL: http://dx.doi.org/10.1103/PhysRevLett.118.052002.

[43] Michael Schmelling. *The sPlot method to separate signal and background*. Tech. rep. Heidelberg, May 2020.

[44] D. Martínez Santos and F. Dupertuis. “Mass distributions marginalized over per-event errors”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 764 (Nov. 2014), pp. 150–155. ISSN: 0168-9002. DOI: 10.1016/j.nima.2014.06.081. URL: http://dx.doi.org/10.1016/j.nima.2014.06.081.

[45] *RooExponential Class Reference*. URL: https://root.cern.ch/doc/master/classRooExponential.html.

[46] *RooHypatia2 Class Reference*. URL: https://root.cern.ch/doc/master/classRooHypatia2.html.

[47] scikit-learn developers. *Decision Tree Classifier*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.

[48] scikit-learn developers. *AdaBoost Classifier*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html.

[49] scikit-learn developers. *RFE*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html.

Appendix

A Background and signal PDFs used in the ${}_s\mathscr{P}lot$ technique

The background PDF f_b is computed by the method `RooExponential` provided by the library `Roofit` in the `ROOT` data framework as [45]:

$$f_b = f_{\exp}(m, e_0) = \mathscr{N} \exp(e_0 \cdot m) \tag{35}$$

where e_0 is the optimal exponent calculated by the method `RooExponential` provided by `Roofit`, m is the mass variable and $\mathscr{N}$ a normalisation constant that depends on the range and values of the arguments.

The signal PDF f_s follows a more sophisticated distribution called `Ipatia2` and implemented in `Roofit` as `RooIpatia2`. This PDF [46], corresponding to the Hypatia distribution described in [44] is defined as the following:

$$f_s = f_{\text{Ip2}} = \begin{cases} \dfrac{f_{\text{dcbl (l)}}}{\left(1 - \dfrac{m}{n_l g(f_{\text{dcbl (l)}}) - a_l \sigma}\right)^{n_l}} & \text{if } \frac{m-\mu}{\sigma} < -a_l \\ f_{\text{dcbl}} & \text{if } -a_l < \frac{m-\mu}{\sigma} < a_r \\ \dfrac{f_{\text{dcbl (r)}}}{\left(1 - \dfrac{m}{-n_r g(f_{\text{dcbl (r)}}) - a_r \sigma}\right)^{n_r}} & \text{if } \frac{m-\mu}{\sigma} > a_r \end{cases} \tag{36}$$

where $f_{\text{Ip2}} \equiv f_{\text{Ip2}}(m, \mu, \sigma, \lambda, \zeta, \beta, a_l, n_l, a_r, n_r)$, $g(f) \equiv f \cdot (f')^{-1}$, and $f_{\text{dcbl}} \equiv f_{\text{dcbl}}(m, \mu, \sigma, \lambda, \zeta, \beta)$, $f_{\text{dcbl (l)}} \equiv f_{\text{dcbl}}(\mu - a_l\sigma, \mu, \sigma, \lambda, \zeta, \beta)$, and $f_{\text{dcbl (r)}} \equiv f_{\text{dcbl}}(\mu + a_r\sigma, \mu, \sigma, \lambda, \zeta, \beta)$ are double crystal ball like distributions [46] defined as:

$$f_{\text{dcbl}} \equiv f_{\text{dcbl}}(m, \mu, \sigma, \lambda, \zeta, \beta) = \left((m-\mu)^2 A_\lambda^2(\zeta)\sigma^2\right)^{\frac{1}{2}\lambda - \frac{1}{4}} e^{\beta(m-\mu)} K_{\lambda - \frac{1}{2}}\left(\zeta \sqrt{1 + \left(\frac{m-\mu}{A_\lambda(\zeta)\sigma}\right)^2}\right) \tag{37}$$

where K_λ are the modified Bessel functions of the second kind and A_λ is the ratio of these, which are defined as:

$$K_\lambda(\zeta) = \int_0^\infty \cosh(\lambda t) e^{-\zeta \cosh t} \mathrm{d}t \quad \text{and} \quad A_\lambda^2(\zeta) = \frac{\zeta K_\lambda(\zeta)}{K_{\lambda+1}(\zeta)}$$

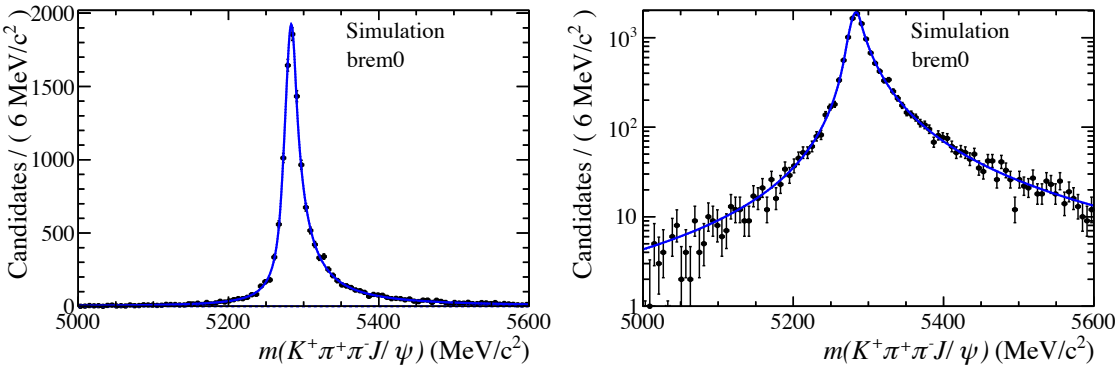
Because f_s contains many parameters, their description is detailed in Table 19.

Variable	Description	Remarks
m	Mass variable	
λ	Shape parameter	$\lambda \ll 0$ is required if $\zeta = 0$
β	Asymmetry parameter	symmetric case is $\beta = 0$
a_l	Start of the left tail	$a_l \geq 0$
a_r	Start of right tail	$a_r \geq 0$
n_l	Shape parameter of left tail	$n_l \geq 0$, with $n_l = 0$, the function is constant
n_r	Shape parameter of right tail	$n_r \geq 0$, with $n_r = 0$, the function is constant
μ	Location parameter	Shifts the distribution left/right
σ	Width parameter	if $\beta = 0$, σ is the RMS
ζ	Shape parameter	$\zeta \geq 0$

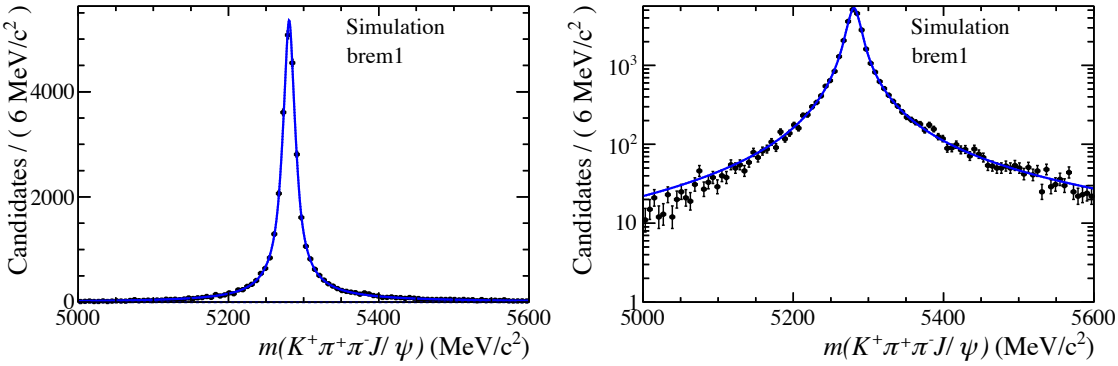
Table 19: Parameters defining the f_{Ip2} in Equation (36) according to [46].

B *s*Fit intermediate plots

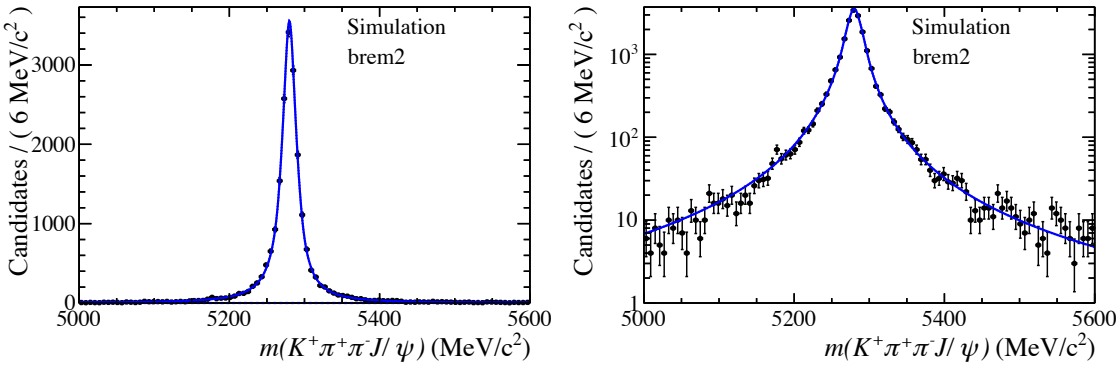
Figures 29 display the Monte Carlo sample *s*Fits that were used for the extraction of the tail parameters of $\ell = e$. These are fixed in Figures 30 for the data sample, that are later combined.



(a) 0Brem, linear (b) 0Brem, logarithmic

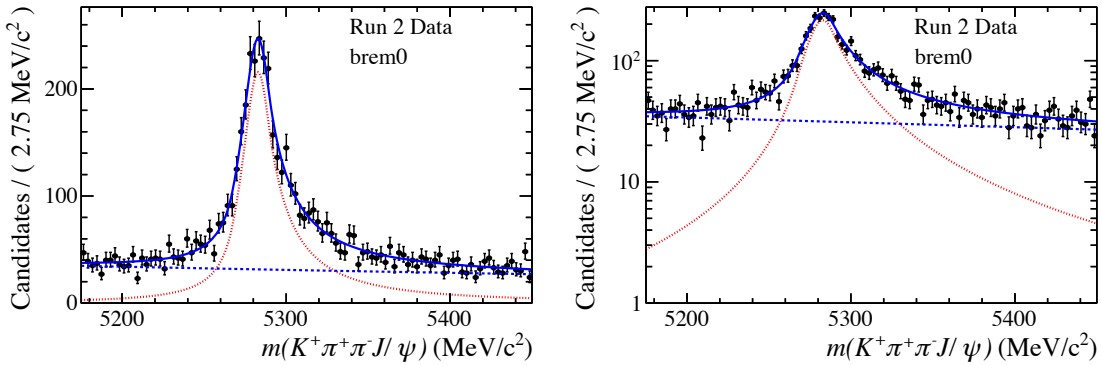


(c) 1Brem, linear (d) 1Brem, logarithmic

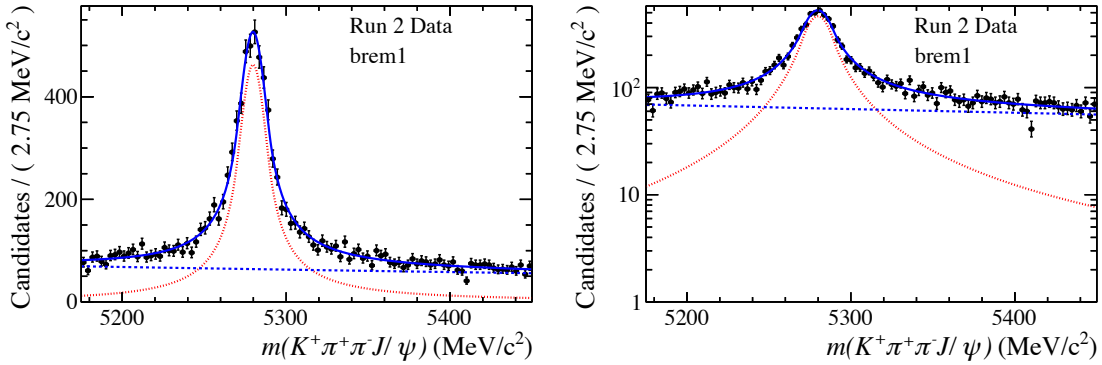


(e) 2Brem, linear (f) 2Brem, logarithmic

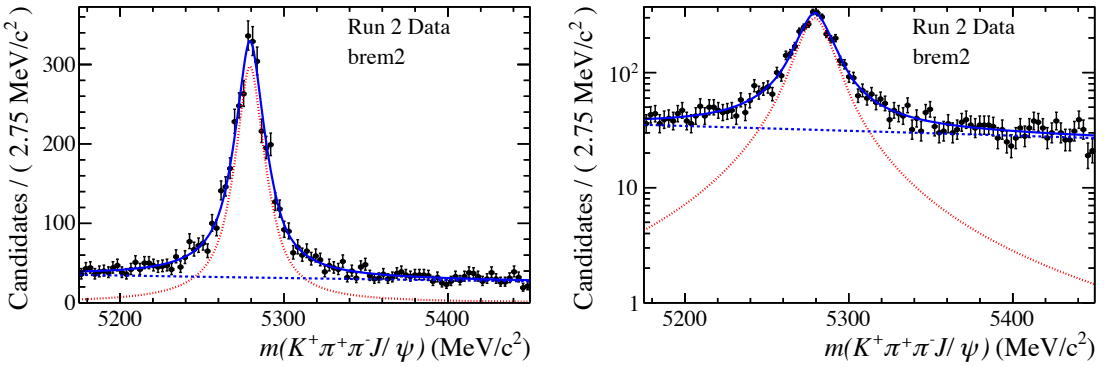
Figure 29: *s*Fit (plain blue) of the reconstructed B^+ meson mass using $\ell = e$ Monte Carlo sample for all Brem categories (Brem0 up, Brem1 middle and Brem2 down) in linear (left) and logarithmic (right) scale.



(a) 0Brem, linear (b) 0Brem, logarithmic



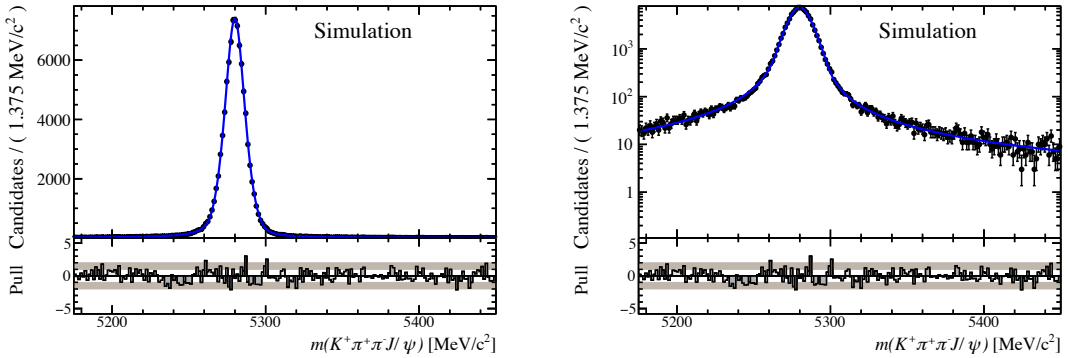
(c) 1Brem, linear (d) 1Brem, logarithmic



(e) 2Brem, linear (f) 2Brem, logarithmic

Figure 30: *s*Fit (plain blue) of the reconstructed B^+ meson mass using $\ell = e$ data sample with the signal component (dashed red) and the background component (dashed blue) for all Brem categories (Brem0 up, Brem1 middle and Brem2 down) in linear (left) and logarithmic (right) scale.

Figures 31 show the Monte Carlo sample *s*Fits that were used for the extraction of the tail parameters of $\ell = \mu$.

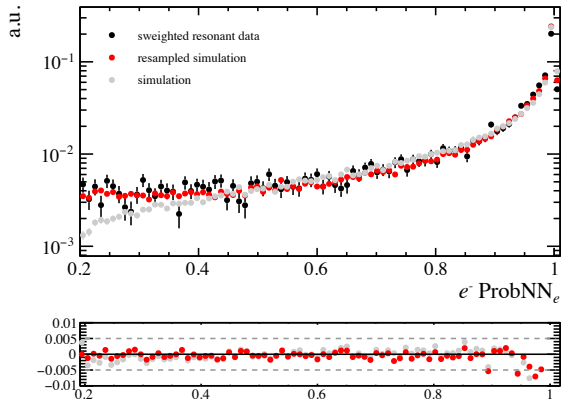


(a) linear (b) logarithmic

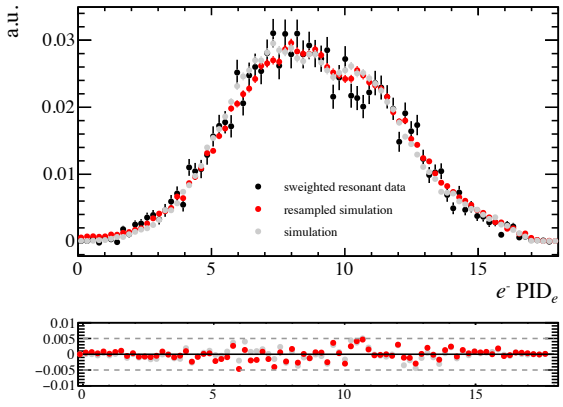
Figure 31: *s*Fit (plain blue) of the reconstructed B^+ meson mass using $\ell = e$ Monte Carlo sample in linear (left) and logarithmic (right) scale.

C PID variables resampling

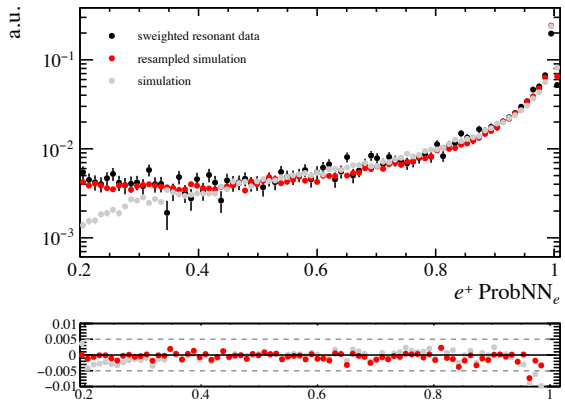
Figure 32 demonstrates the validation of the PID variables for $\ell = e$, and Figure 33 does the same, but for $\ell = \mu$.



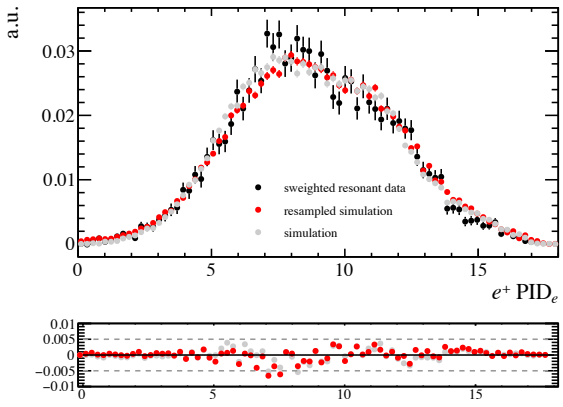
(a) ProbNNe (e^-)



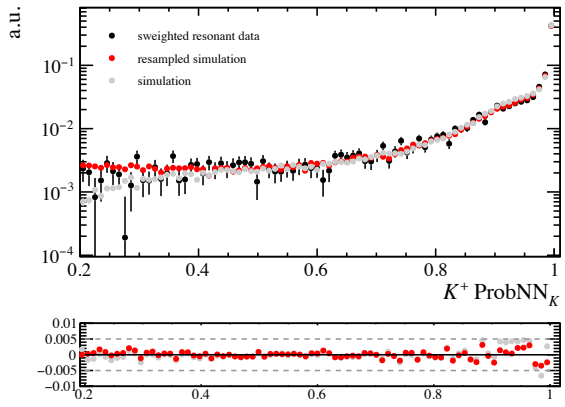
(b) PIDe (e^-)



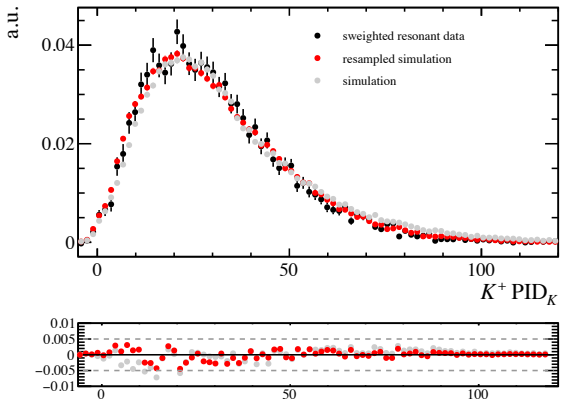
(c) ProbNNe (e^+)



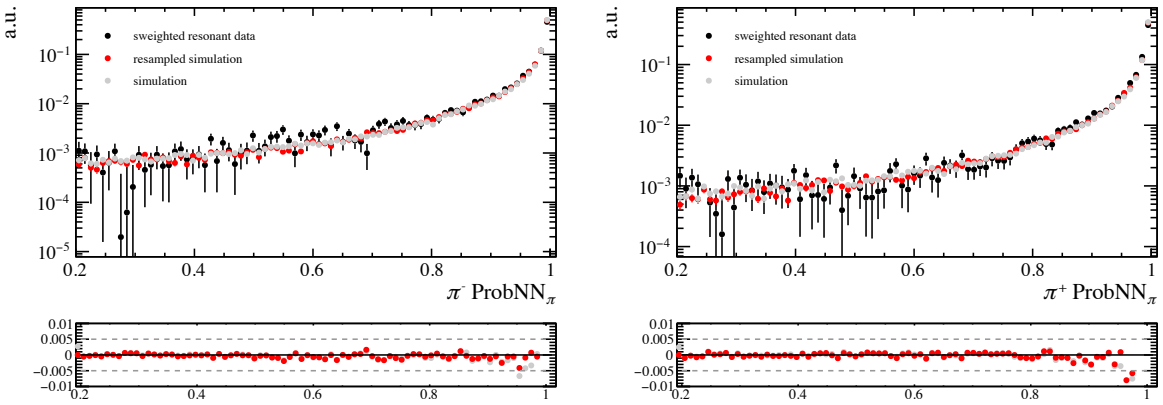
(d) PIDe (e^+)



(e) ProbNNk (K^+)



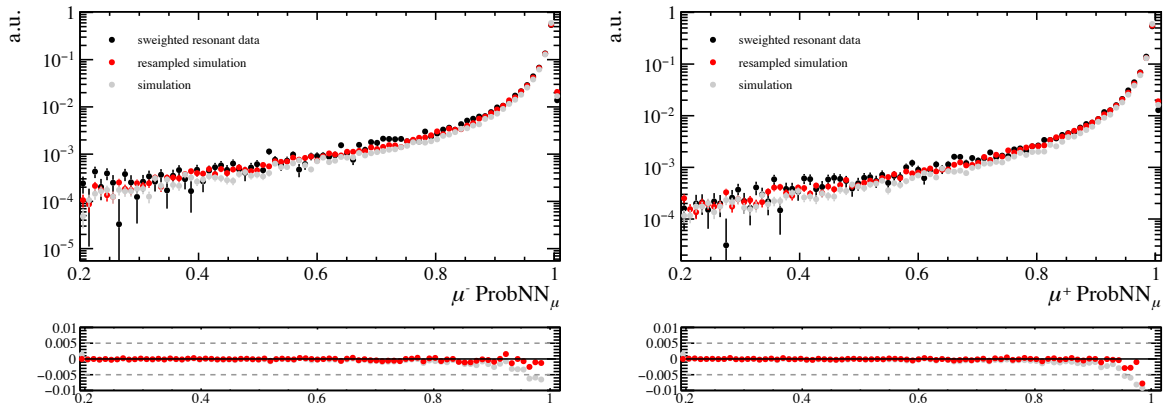
(f) PIDk (K^+)



(g) ProbNNπ (π^-)

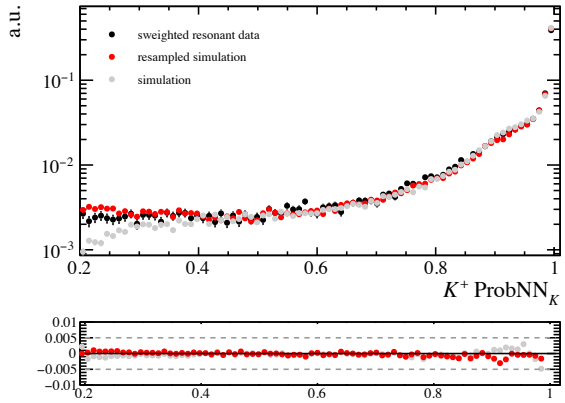
(h) ProbNNπ (π^+)

Figure 32: Validation of the PID resampling made by a comparison of the background-subtracted data (black) with the original (grey) and resampled Monte Carlo (red) distributions of several PID variables for $\ell = e$.

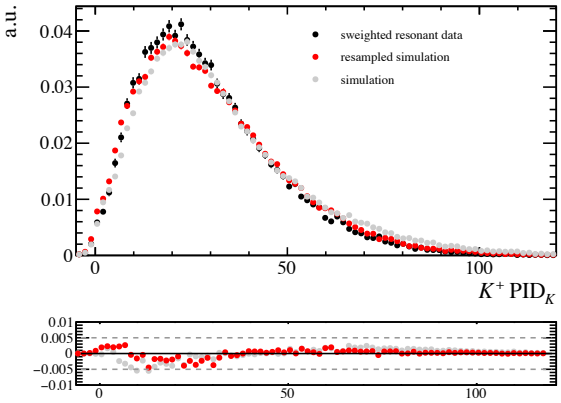


(a) ProbNNμ (μ^-)

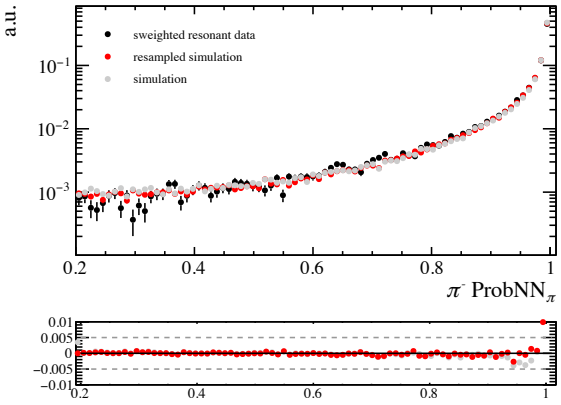
(b) ProbNNμ (μ^+)



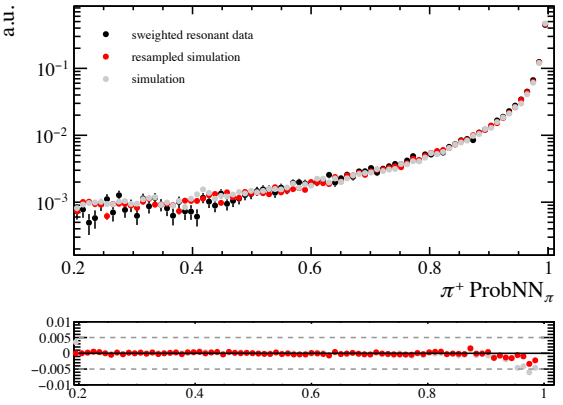
(c) ProbNNk (K^+)



(d) PIDk (K^+)



(e) ProbNNπ (π^-)



(f) ProbNNπ (π^+)

Figure 33: Validation of the PID resampling made by a comparison of the background-subtracted data (black) with the original (grey) and resampled Monte Carlo (red) distributions of several PID variables for $\ell = \mu$.

D GBR hyperparameters

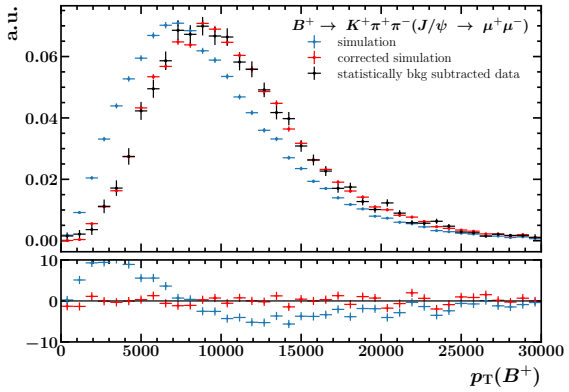
Table 20 summarises the value of the parameters involved in the gradient boosting reweighter used for the computation of kinematics and multiplicity weights.

Parameters	Kin.	Mult.	Default	Description
n_estimators	60	90	40	number of trees
learning_rate	0.5	0.05	0.2	float from [0, 1]
max_depth	2	3	3	maximal depth of trees
min_samples_leaf	200	200	200	minimal number of events in the leaf
loss_regularization	500	5	5	number of events that algorithm puts in each leaf to prevent exploding

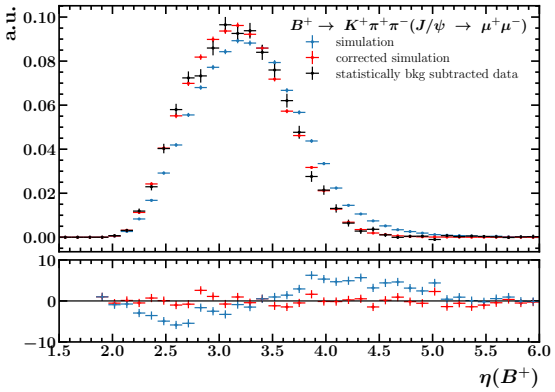
Table 20: Description of the parameters used in the implemented GBR [26] with their default value.

E Kinematics and multiplicity variables plots

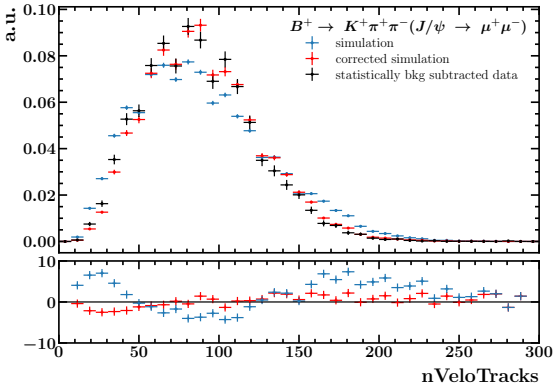
Figures 34 and 35 reveal that the corrected Monte Carlo sample follows reliably the background-subtracted data.



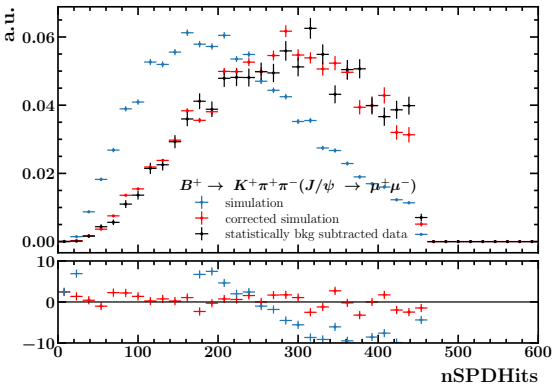
(a) $p_T(B^+)$ kinematics variable



(b) $\eta(B^+)$ kinematics variable

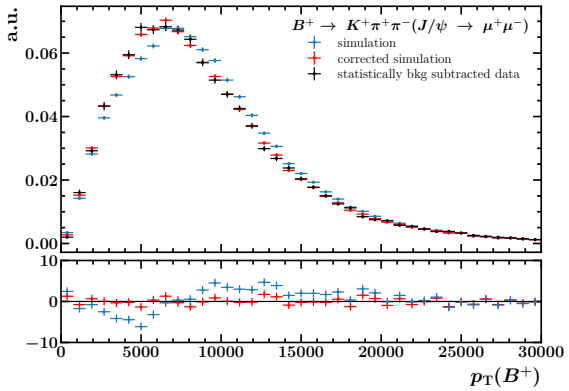


(c) $n_{\text{VELO, Tracks}}$ multiplicity variable

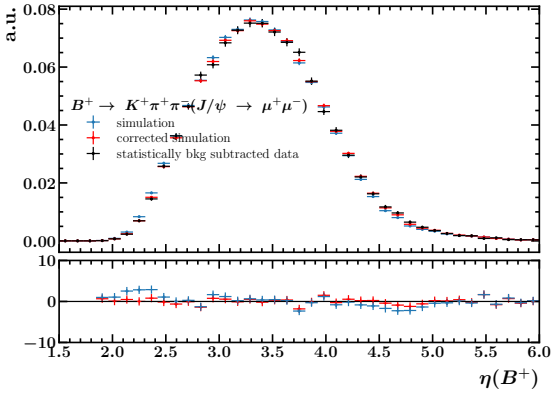


(d) $n_{\text{SPD, Hits}}$ multiplicity variable

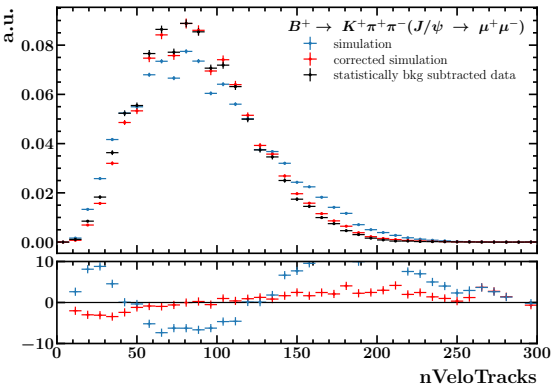
Figure 34: Comparison of weighted and non-weighted Monte Carlo simulation with background-subtracted data distributions of the kinematics (a) and (b) and multiplicity (c) and (d) variables using electrons data sample.



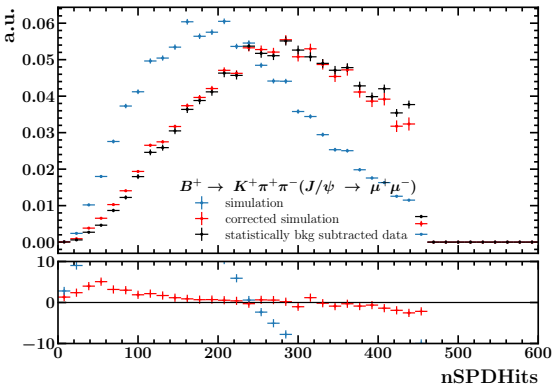
(a) $p_T(B^+)$ kinematics variable



(b) $\eta(B^+)$ kinematics variable



(c) $n_{\text{VELO, Tracks}}$ multiplicity variable



(d) $n_{\text{SPD, Hits}}$ multiplicity variable

Figure 35: Comparison of weighted and non-weighted Monte Carlo simulation with background-subtracted data distributions of the kinematics (a) and (b) and multiplicity (c) and (d) variables using muons data sample.

F General plots of features score

Figures 36 and 37 show a matrix containing the number of times two features are evaluated together in a sample.

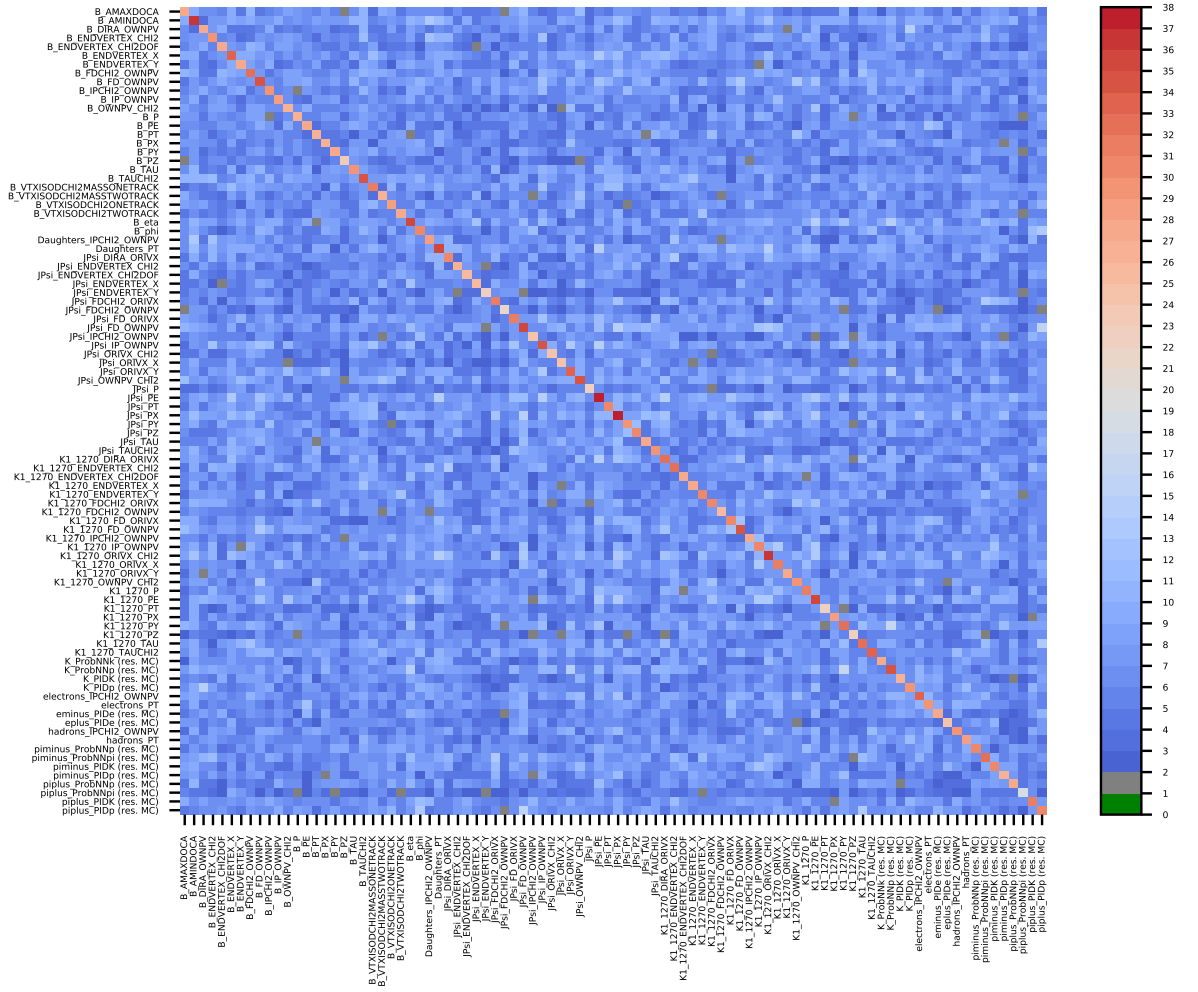


Figure 36: Representation of the number of times two features are evaluated together considering all given samples, for $\ell = e$ data set.

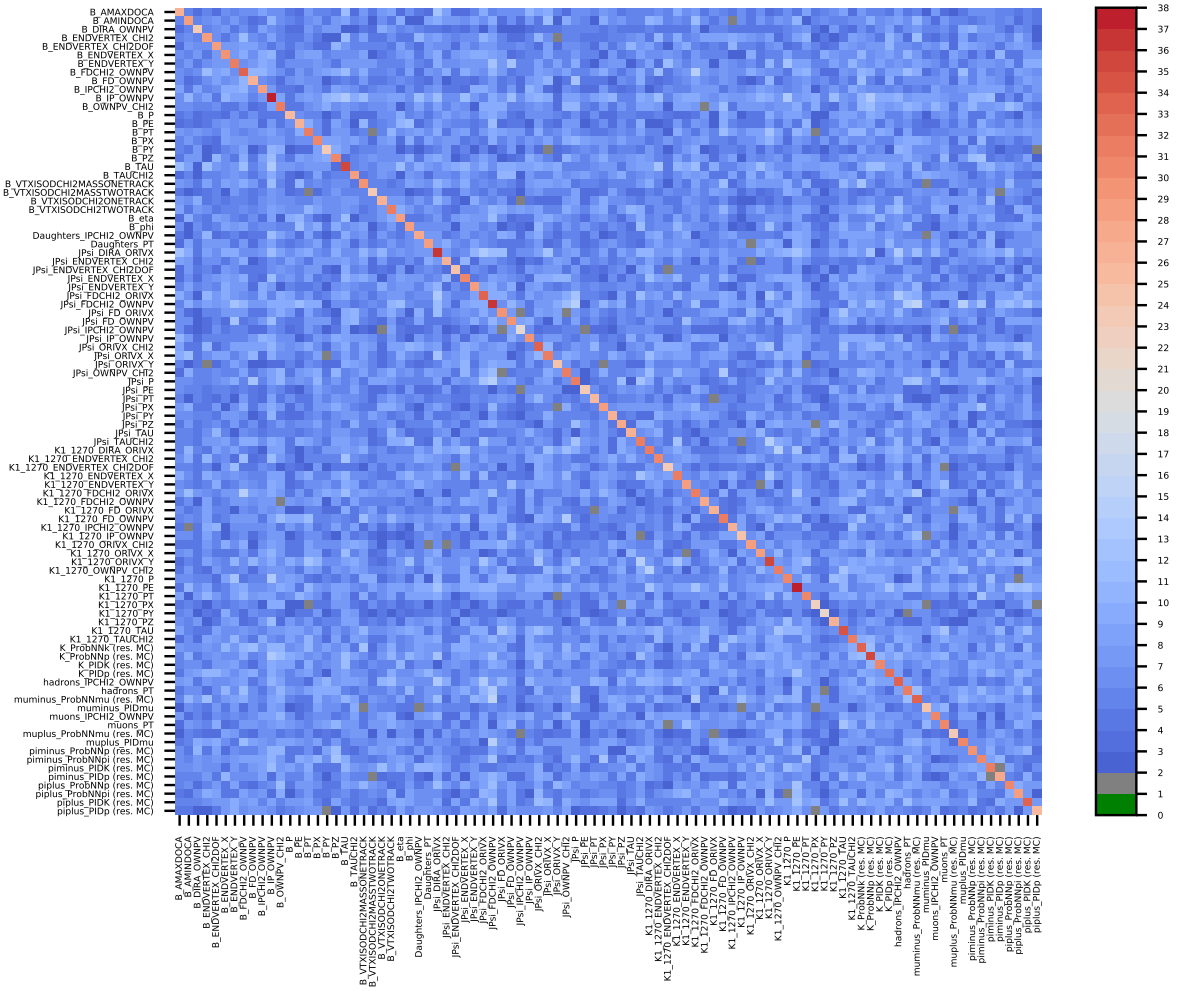


Figure 37: Representation of the number of times two features are evaluated together considering all given samples, for $\ell = \mu$ data set.

Figures 38 and 39 display the comparison of features associated to all mean and weighted scores, for both electrons and muons data sets.

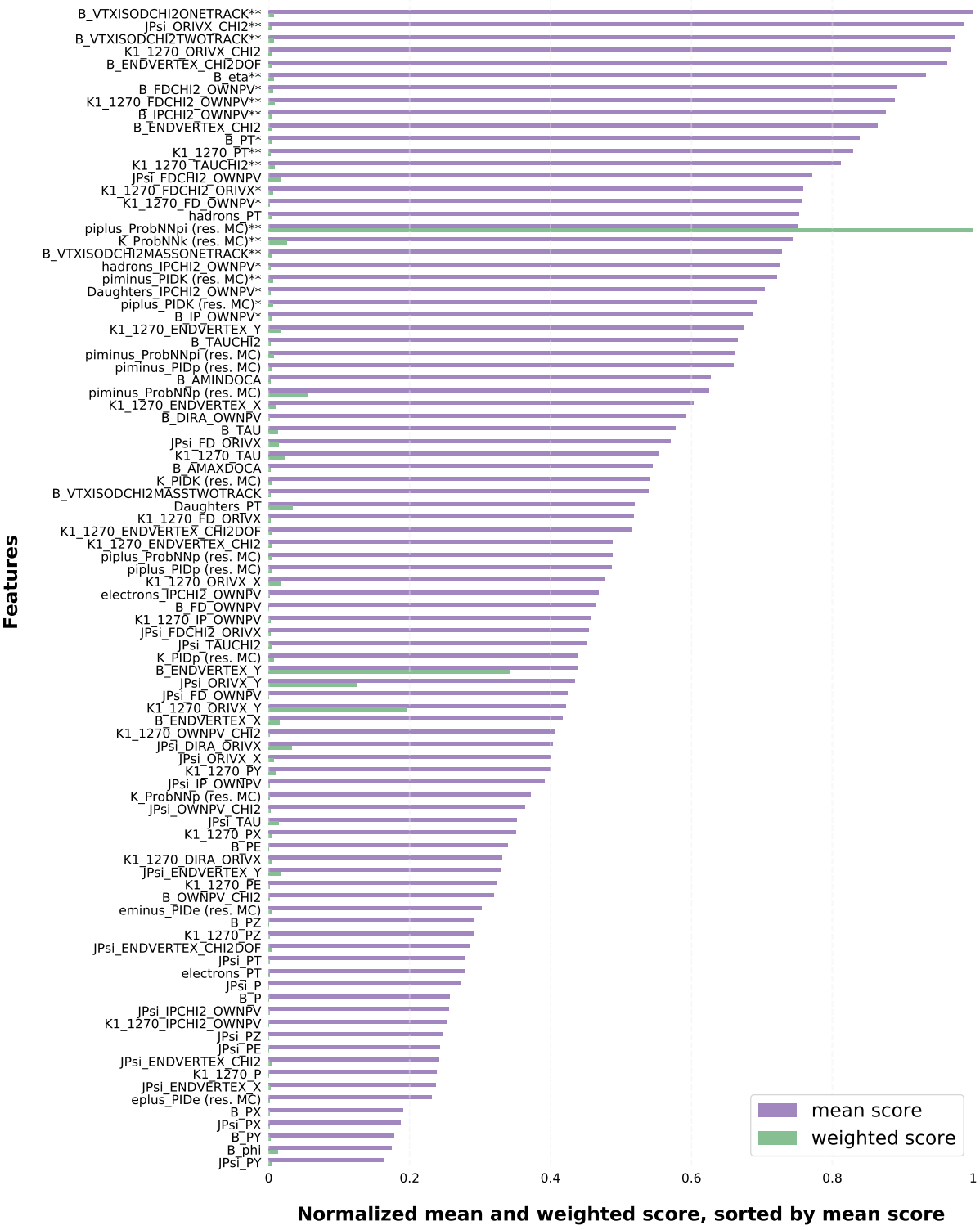


Figure 38: Comparison of features associated to all mean and weighted score for $\ell = e$ data set. The features are sorted according to the mean score: the best feature is at the top of the graph with the highest value, and the worst is at the bottom. The notation is the same used in Figure 18.

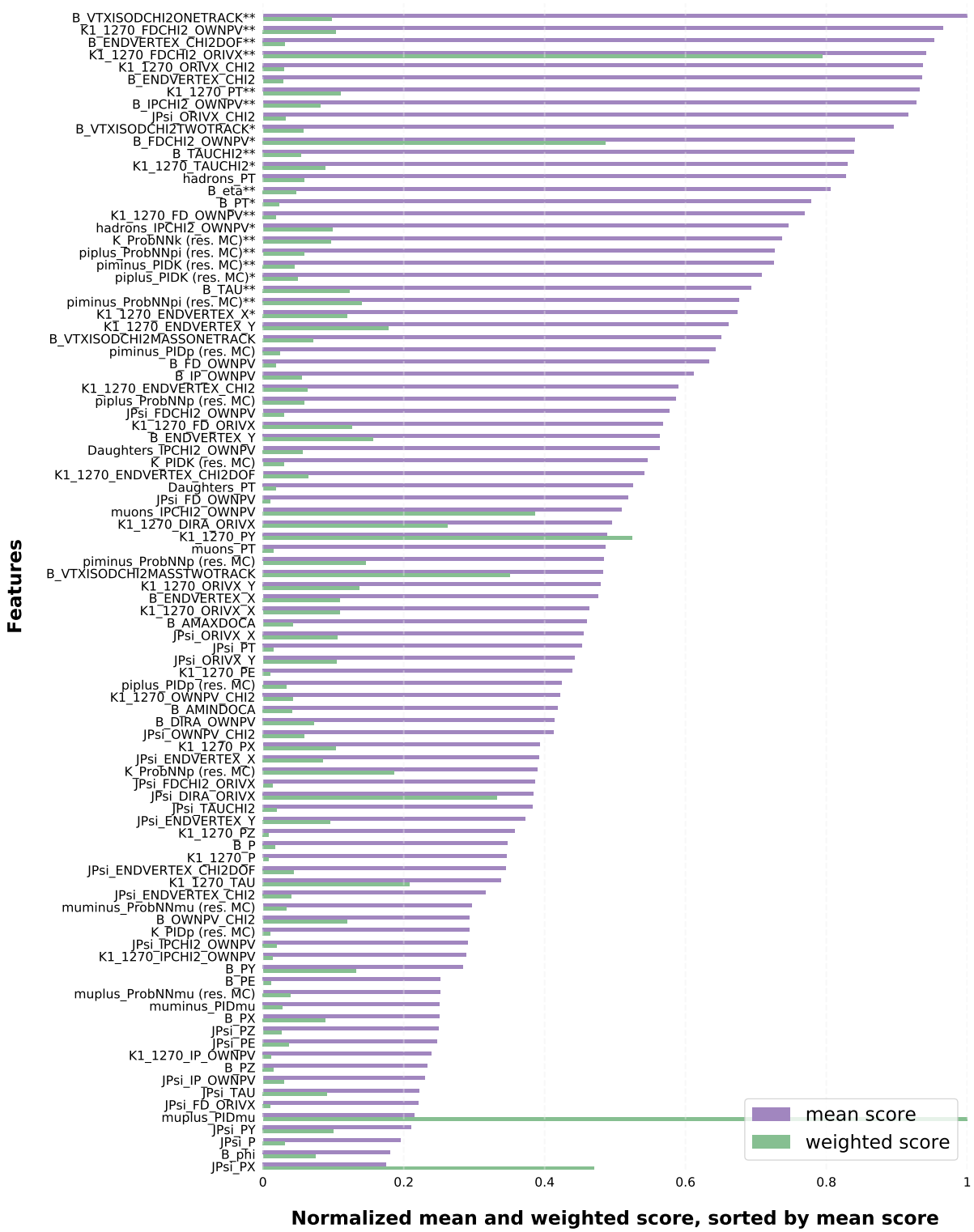
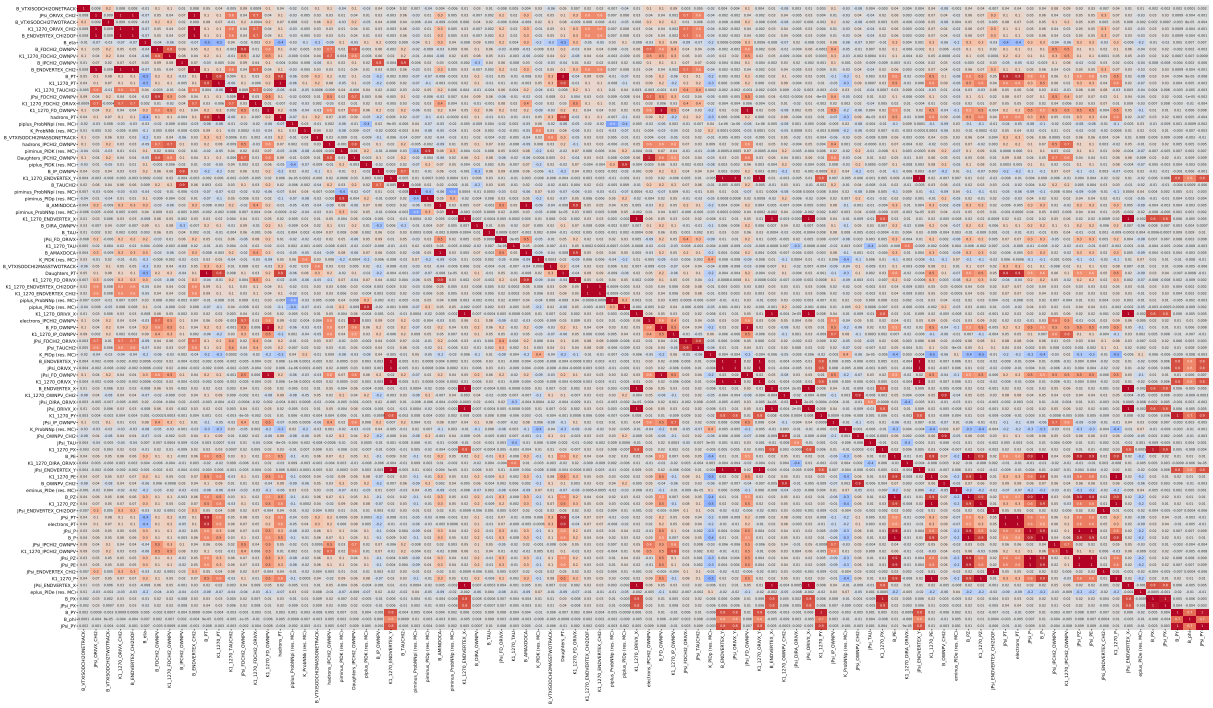
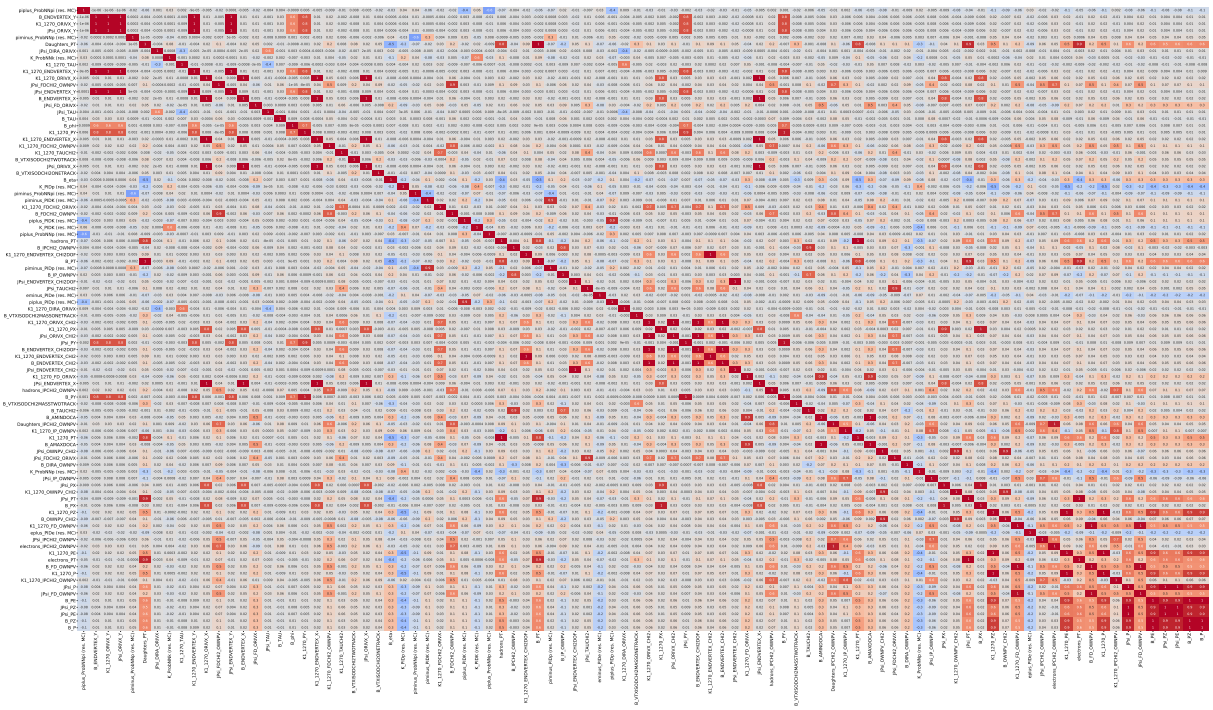


Figure 39: Comparison of features associated to all mean and weighted score for $\ell = \mu$ data set. The features are sorted according to the mean score: the best feature is at the top of the graph with the highest value, and the worst is at the bottom. The notation is the same used in Figure 18.

Figures 40 and 41 show the correlation matrices of all features sorted according to the mean and weighted scores, for both electrons and muons data set.

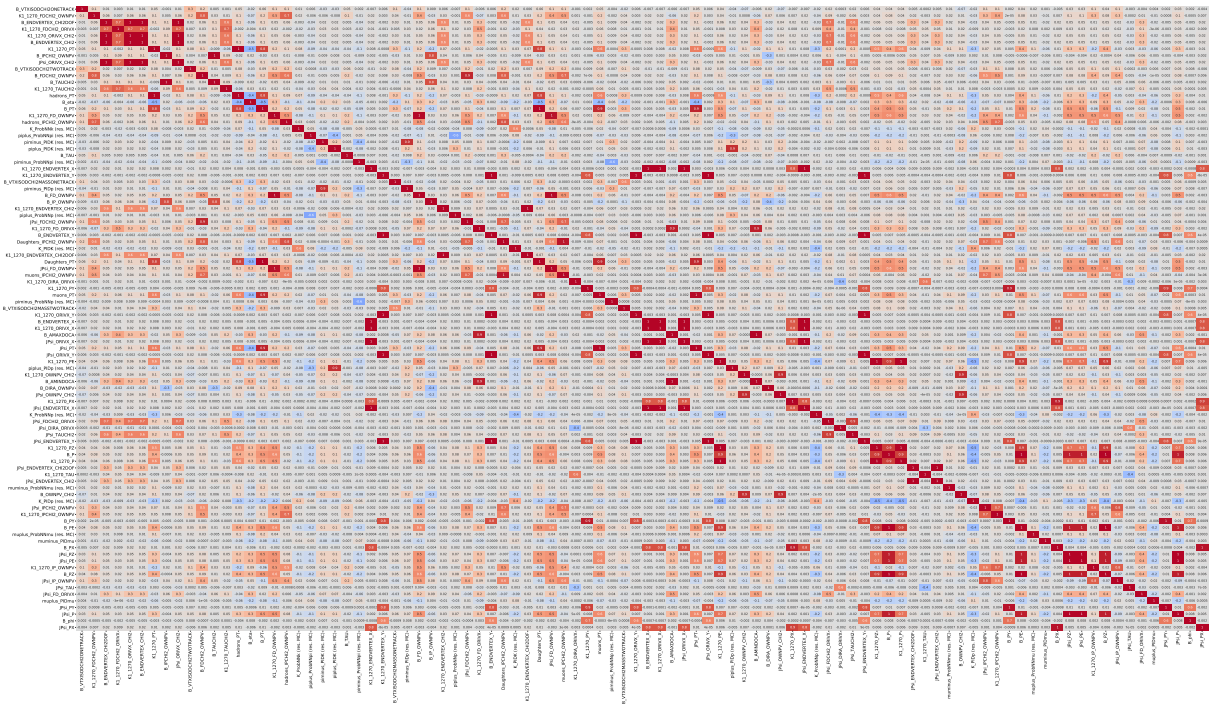


(a) $\ell = e$, mean score

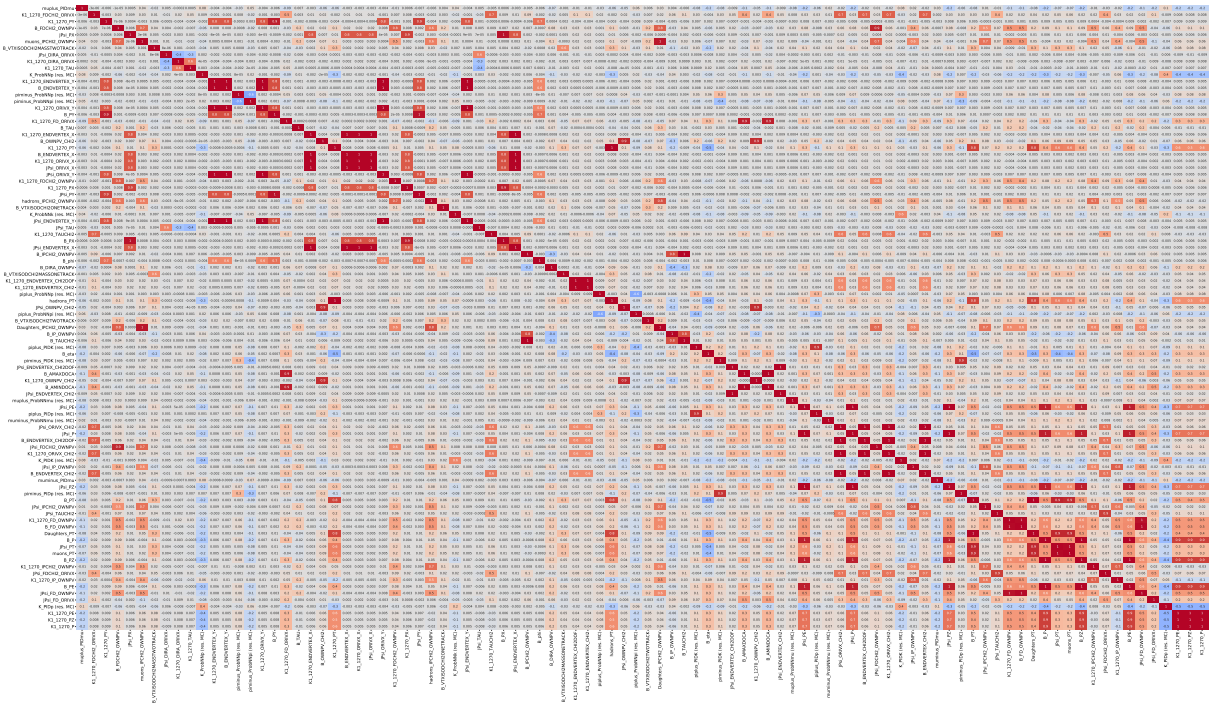


(b) $\ell = e$, weighted score

Figure 40: Correlation matrices of all features sorted according to the mean (up) and weighted (down) scores, with the best features on the top and the worst on the bottom, using electrons signal data set.



(a) $\ell = \mu$, mean score



(b) $\ell = \mu$, weighted score

Figure 41: Correlation matrices of all features sorted according to the mean (up) and weighted (down) scores, with the best features on the top and the worst on the bottom, using muons signal data set.

G Background and signal distributions

Figures 42 and 43 show the background and signal distributions of the discriminating features of both electrons and muons data sets.

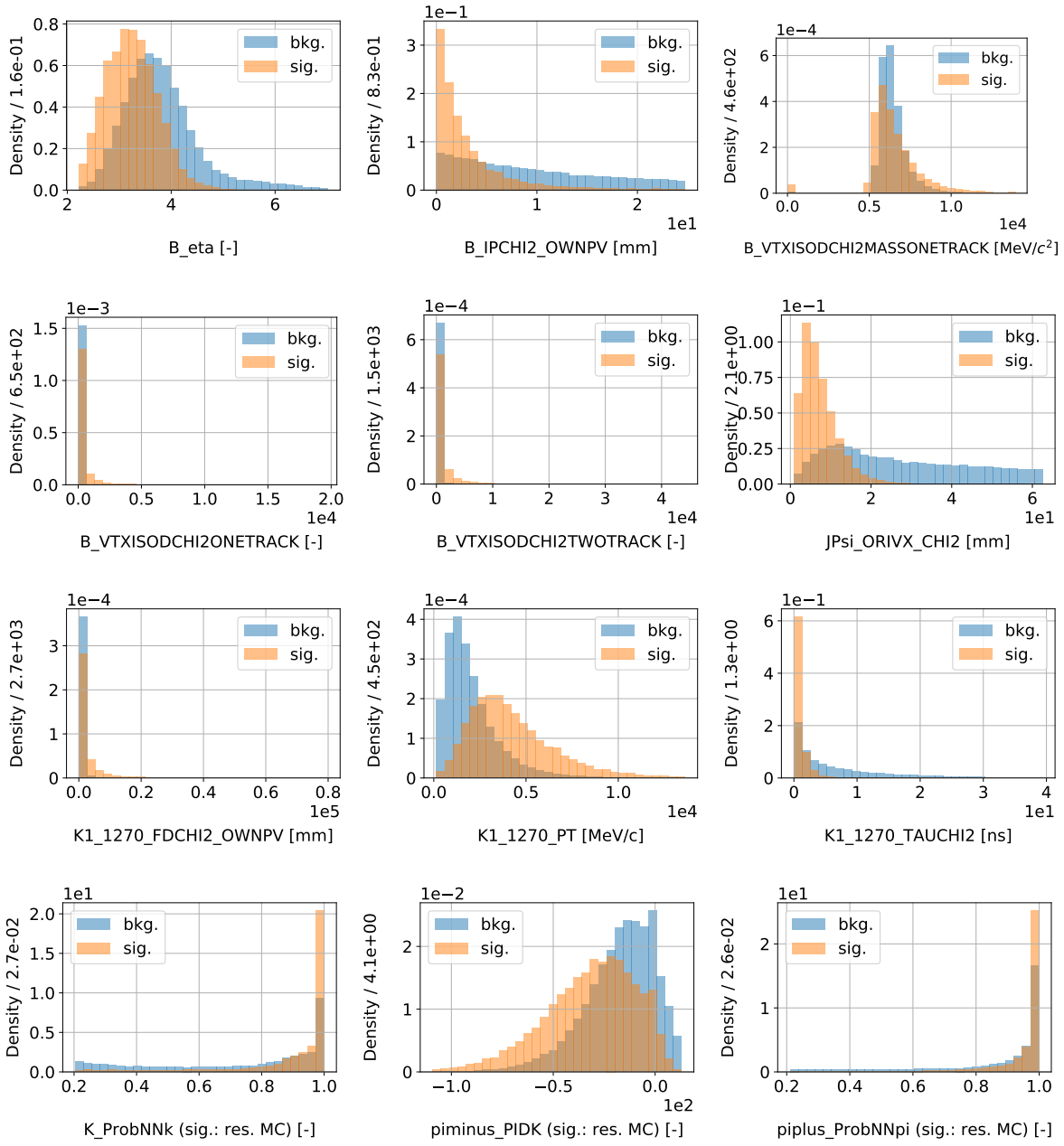


Figure 42: Background and signal distributions of the discriminating features used for the electrons data set.

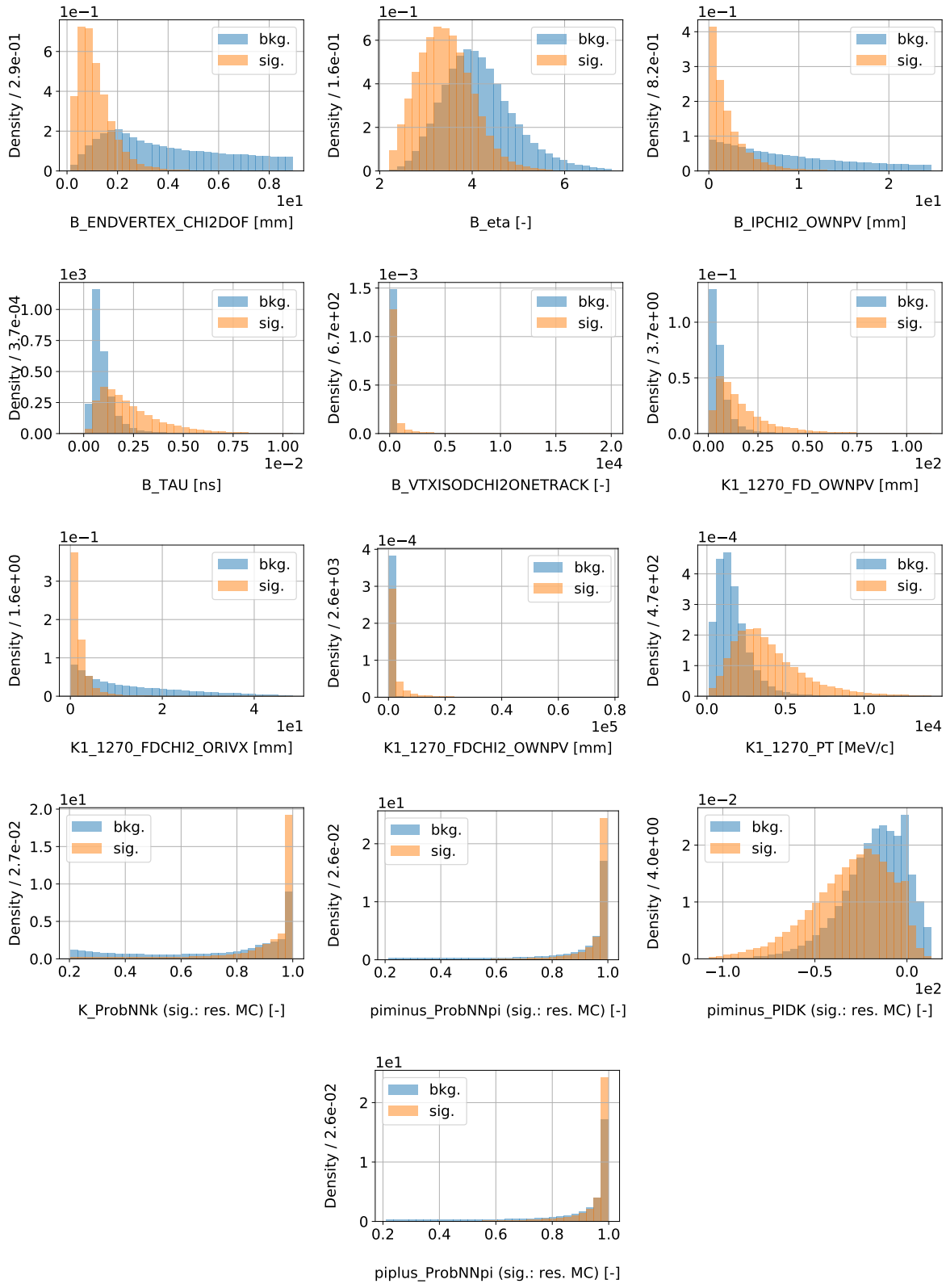


Figure 43: Background and signal distributions of the discriminating features used for the muons data set.

Figures 44 and 45 show the background and signal distributions of the features from the previous analysis [4] of both electrons and muons data sets.

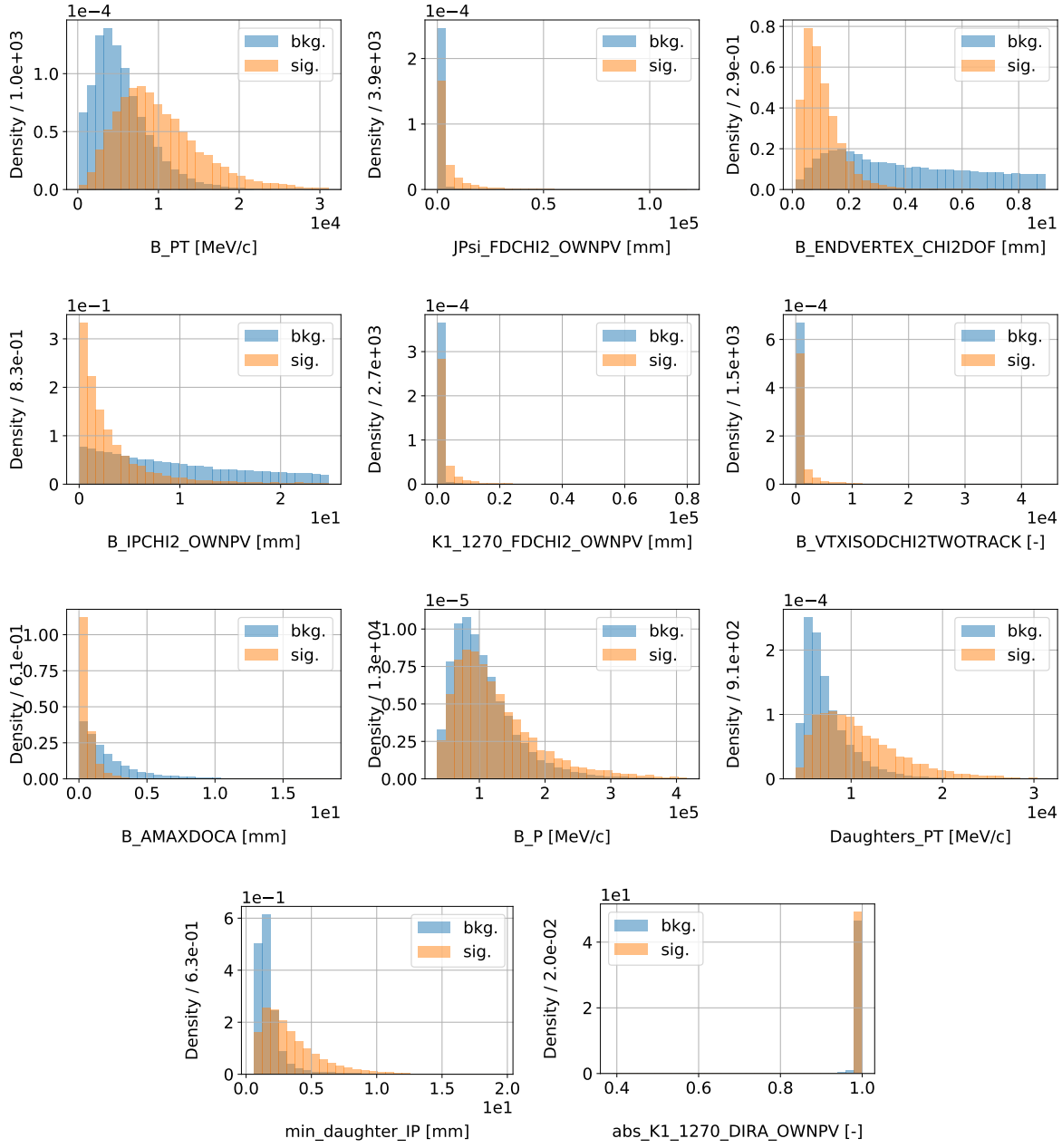


Figure 44: Background and signal distributions of the features provided by the previous analysis [4], using electrons data set.

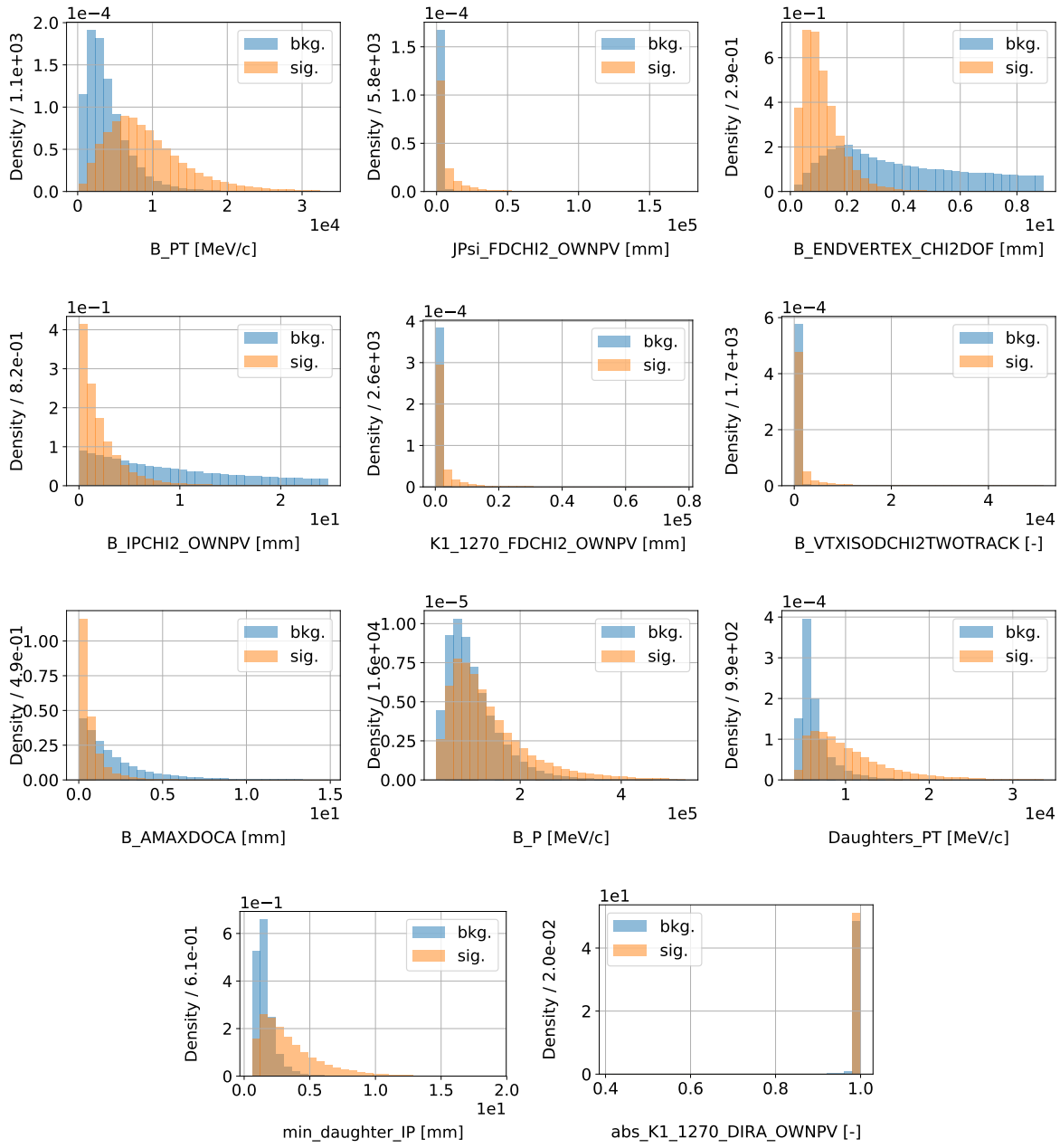
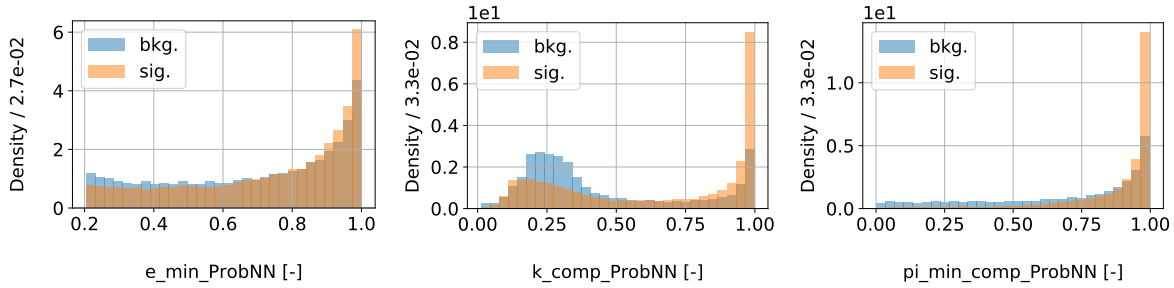
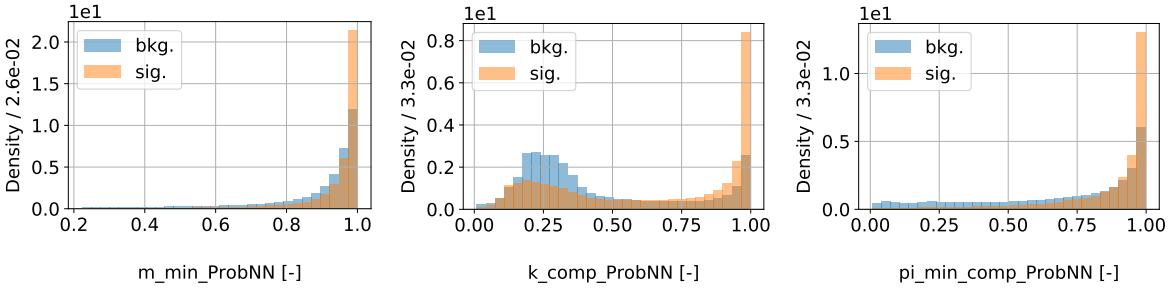


Figure 45: Background and signal distributions of the features provided by the previous analysis [4], using muons data set.

Figure 46 shows the background and signal distributions of the composed PID variables of both electrons and muons data sets.



(a) $\ell = e$



(b) $\ell = \mu$

Figure 46: Background and signal distributions of the PID features used for the 4D optimisation of the FoM, for the electrons (up) and muons (down) data sets.

H Cross-checks

Background and signal distributions

Figures 47 and 48 show the background and signal distributions for the discriminating features, but without taking into account the MC weights for the signal. It is meant as a comparison with Figures 42 and 43.

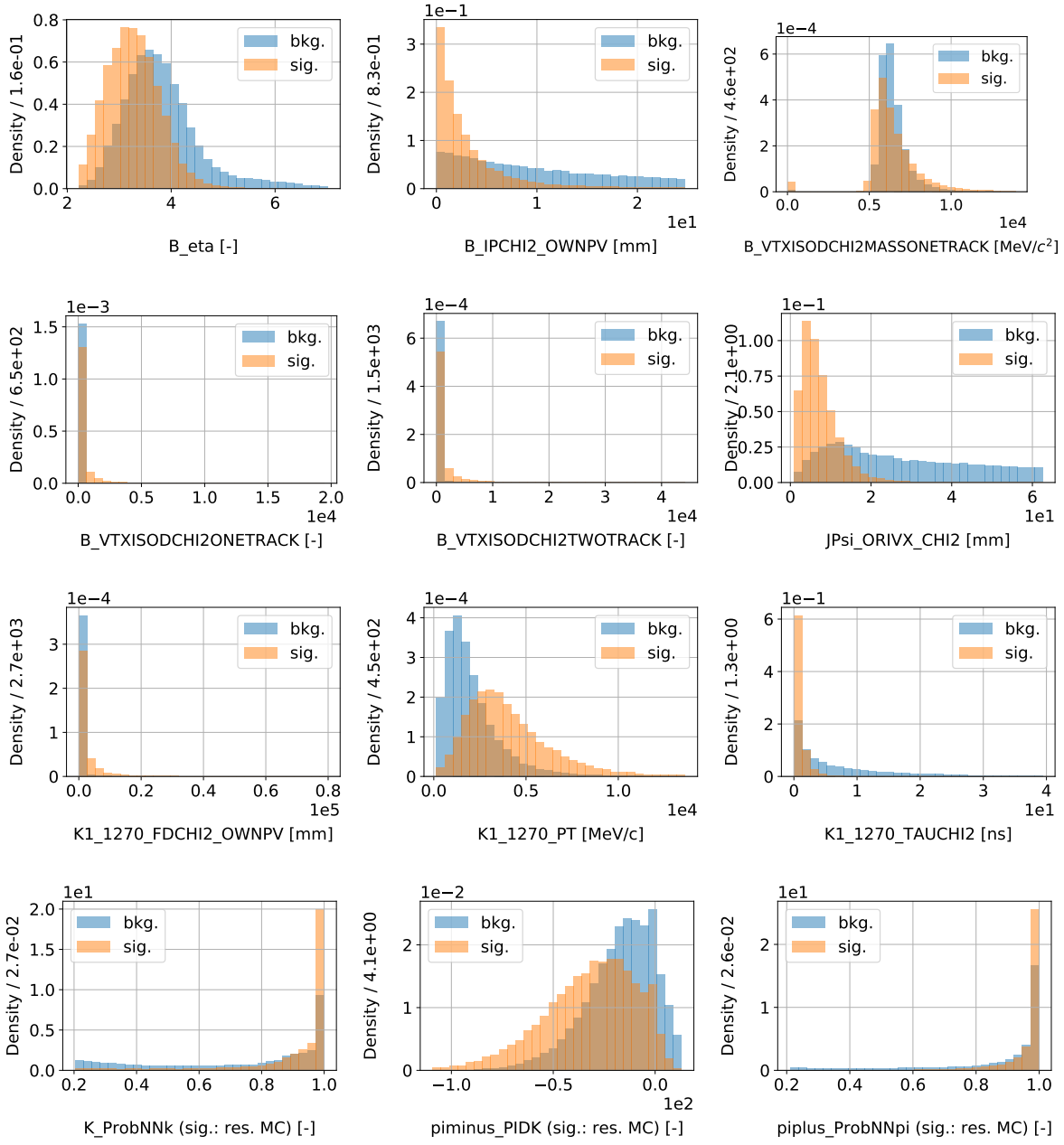


Figure 47: Background and signal distributions of the discriminating features used for the electrons data set, without the use of the signal weights.

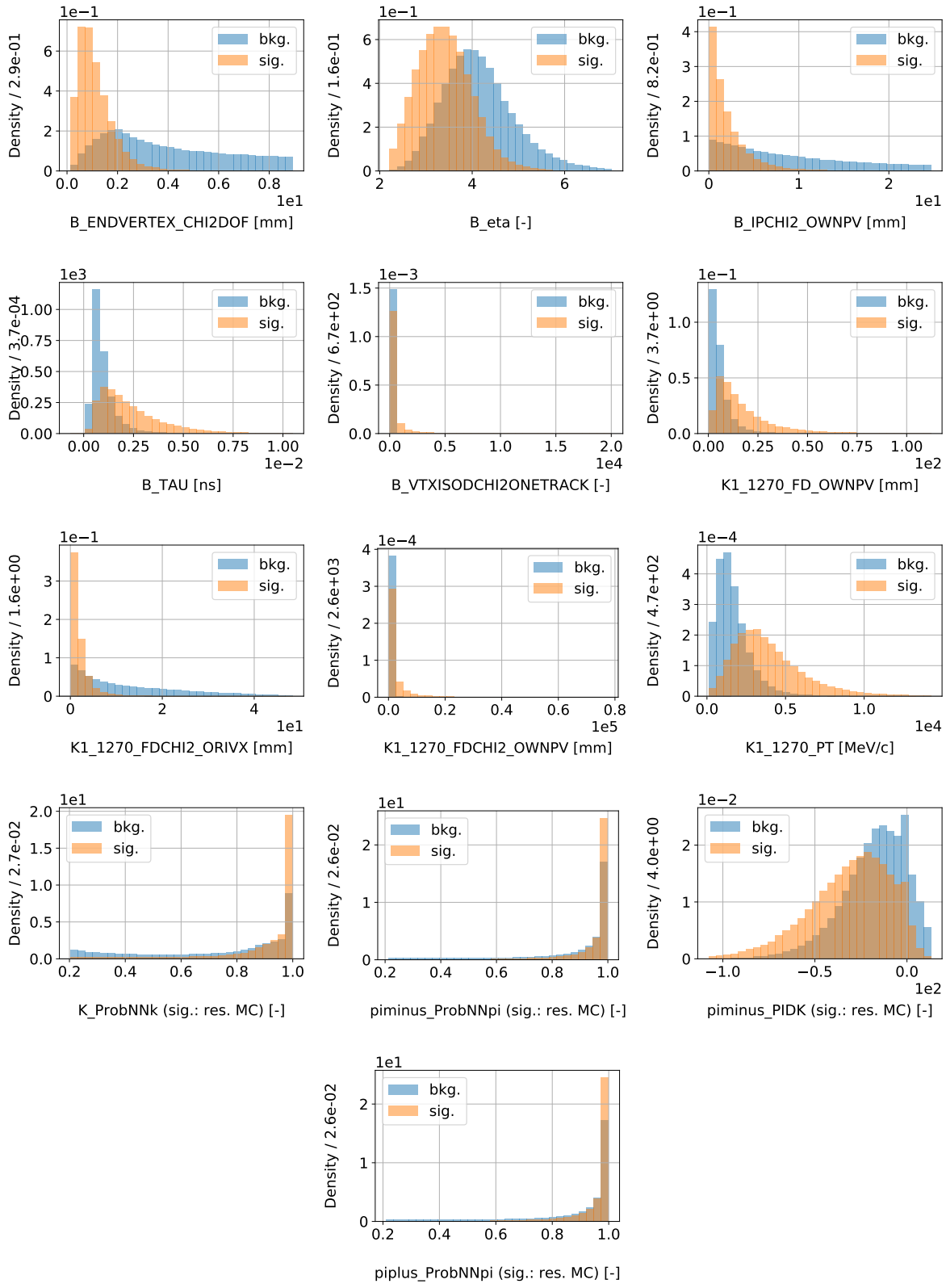
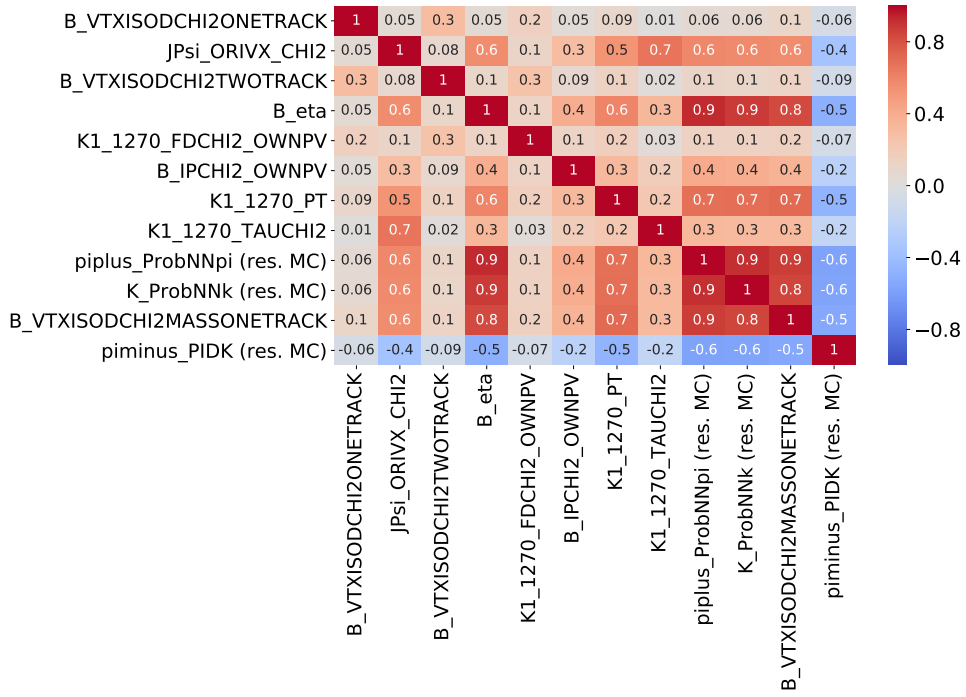


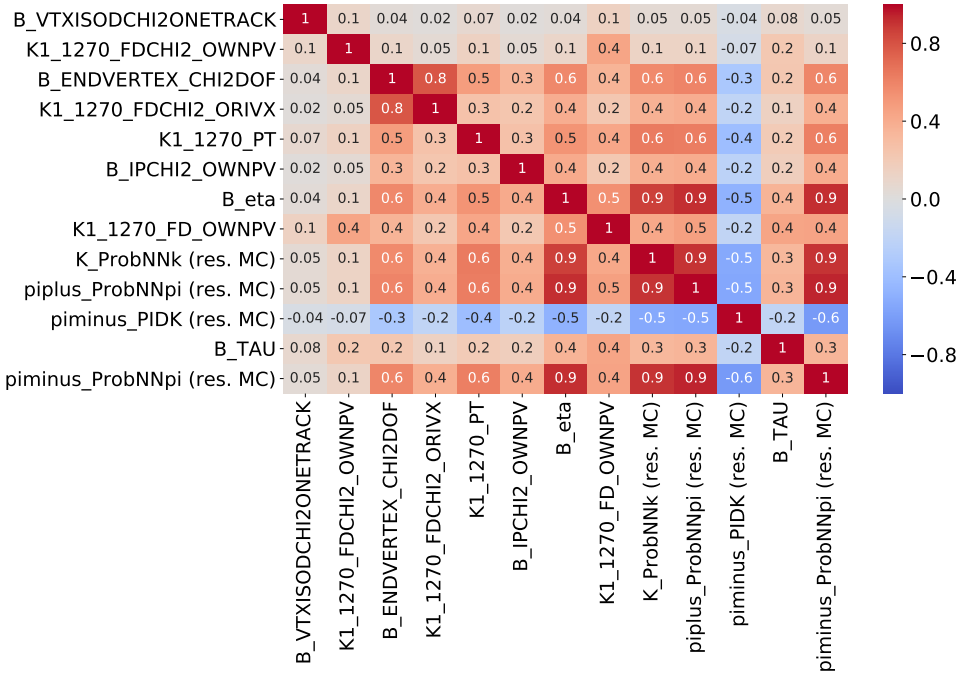
Figure 48: Background and signal distributions of the discriminating features used for the muons data set, without the use of the signal weights.

Correlation matrices

Figures 49 show the correlation matrices obtained for the discriminating features for both electrons and muons signal samples, but computed using Equation (23). It is meant as a comparison with Figures 20.



(a) $\ell = e$



(b) $\ell = \mu$

Figure 49: Correlation matrices of the features associated to the 20 best mean score which showed no correlation value of 1, using electrons (up) and muons (down) signal data sets, computed with Equation (23).

I Comparison of MVA performances

As a way of finding the best combination of MVA method and hyperparameters, several attempt were made. We illustrate here a reduced portion of them, which are detailed in Table 21.

Abbreviation	Description
SAMME-BDT, MS, 20WS, WMCW	SAMME boost (SAMME-BDT) using the features associated to the 20 best mean score (MS) which showed no correlation value of 1 (20WS), and taking into account the MC weights during the training (WMCW)
XG-BDT, MS, 20NS, WMCW	XG boost (XG-BDT) using the features associated to the 20 best mean score (MS) without a selection on them regarding their correlation (20NS), and taking into account the MC weights during the training (WMCW)
XG-BDT, MS, 20WS, NMCW	XG boost (XG-BDT) using the features associated to the 20 best mean score (MS) which showed no correlation value of 1 (20WS), without taking into account the MC weights during the training (NMCW)
XG-BDT, WS, 20WS, WMCW	XG boost (XG-BDT) using the features associated to the 20 best weighted score (WS) which showed no correlation value of 1 (20WS), and taking into account the MC weights during the training (WMCW)

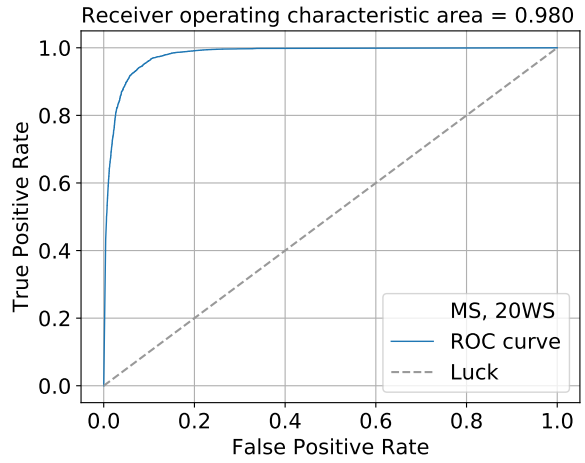
Table 21: Description of the different combinations tested.

The SAMME boost used for the comparison uses the hyperparameters listed in Table 22.

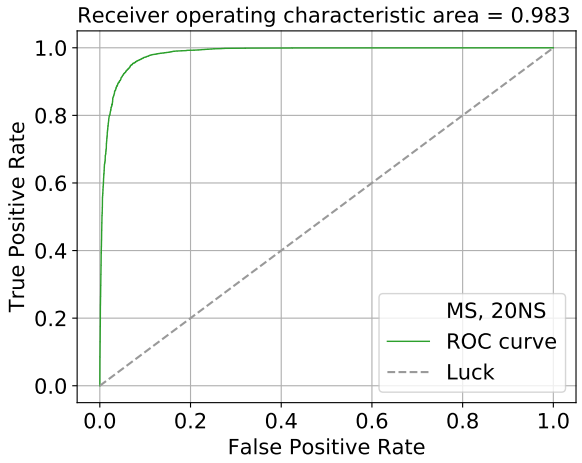
Hyperparameters	$B^+ \to K^+\pi^+\pi^-\ell^+\ell^-$	default	Description
algorithm	SAMME	SAMME.R	boosting algorithm
learning_rate	0.01	1	shrinks the contribution of each classifier by learning_rate
max_depth	5	None	maximum depth of the tree
max_leaf_nodes	300	None	grow a tree with it in best-first fashion
n_estimators	150	50	maximum number of estimators at which boosting is terminated

Table 22: Hyperparameters used for the SAMME-BDT method [47] [48].

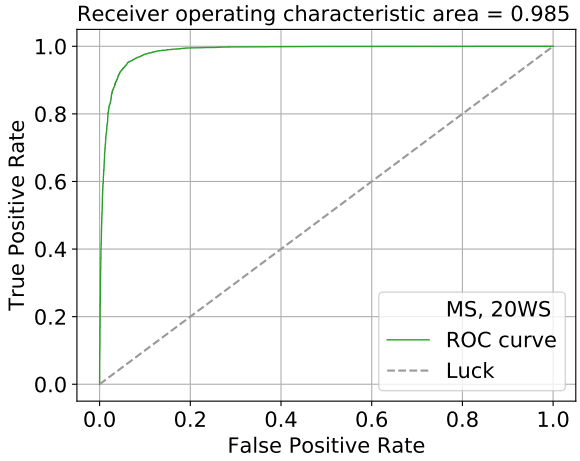
Figures 50 and 51 are meant to be a comparison of 21, whereas Figures 52 and 53 are a comparison of 22.



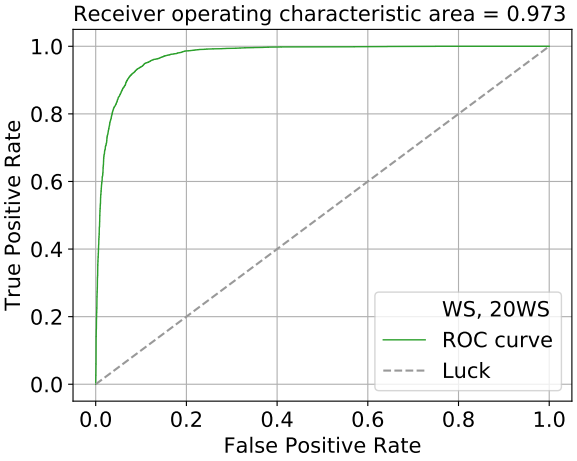
(a) SAMME-BDT, MS, 20SF, WMCW



(b) XG-BDT, MS, N20SF, WMCW

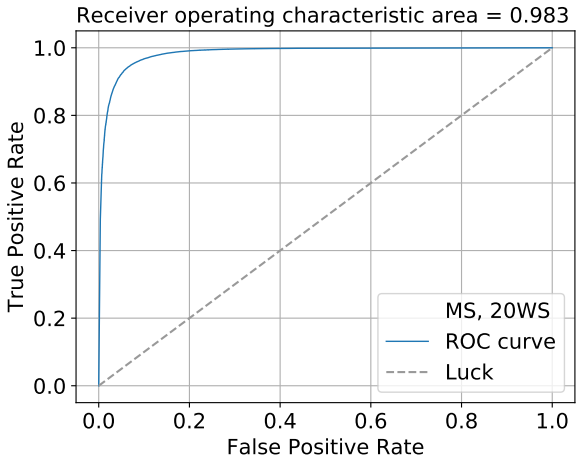


(c) XG-BDT, MS, 20SF, NMCW

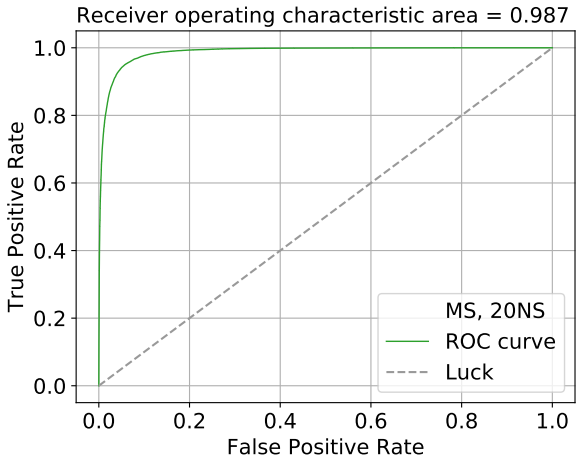


(d) XG-BDT, WS, 20SF, WMCW

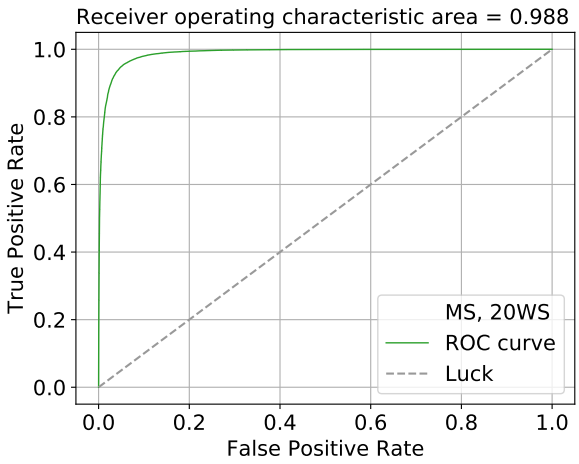
Figure 50: Comparison of ROC curves obtained for the different combinations listed in Table 21 using electrons data sets.



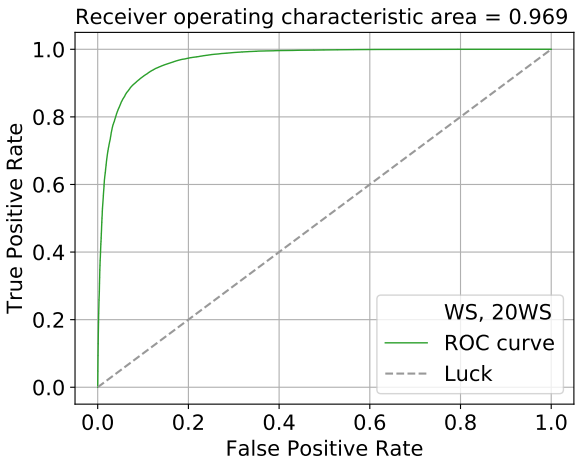
(a) SAMME-BDT, MS, 20SF, WMCW



(b) XG-BDT, MS, N20SF, WMCW

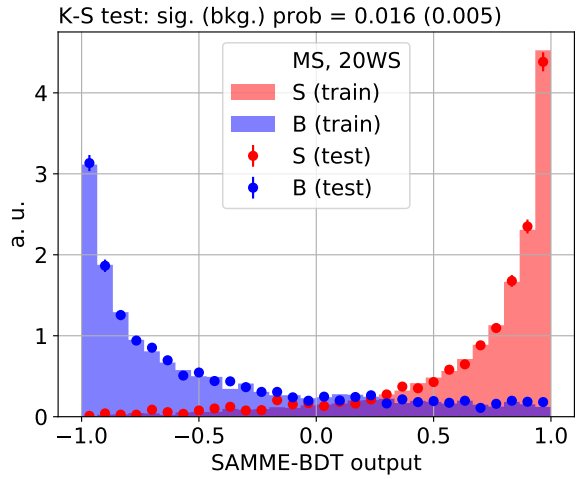


(c) XG-BDT, MS, 20SF, NMCW

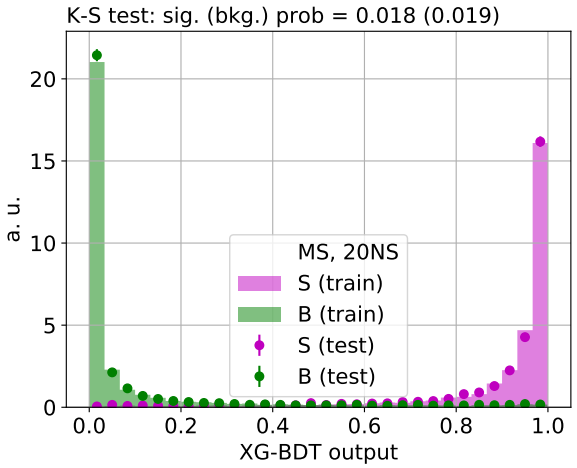


(d) XG-BDT, WS, 20SF, WMCW

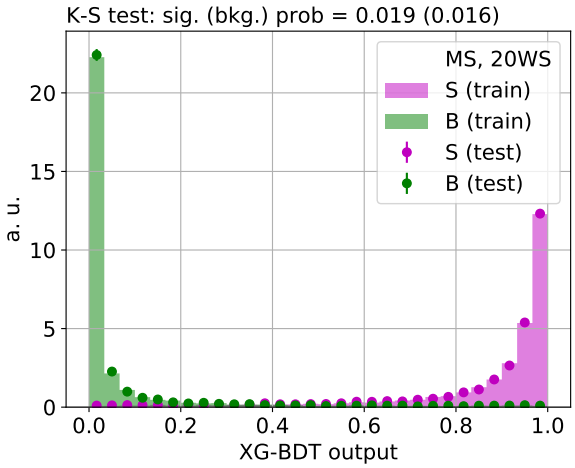
Figure 51: Comparison of ROC curves obtained for the different combinations listed in Table 21 using muons data sets.



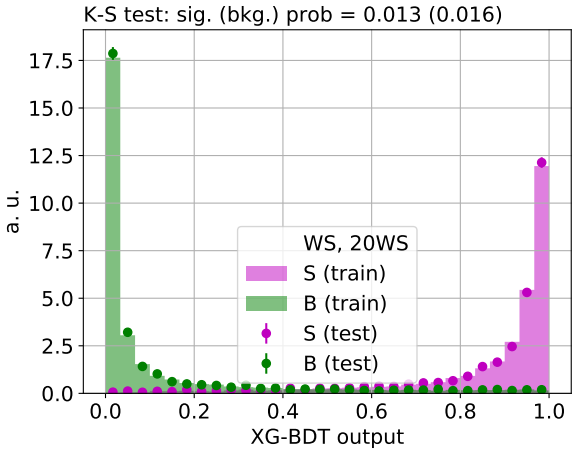
(a) SAMME-BDT, MS, 20SF, WMCW



(b) XG-BDT, MS, N20SF, WMCW

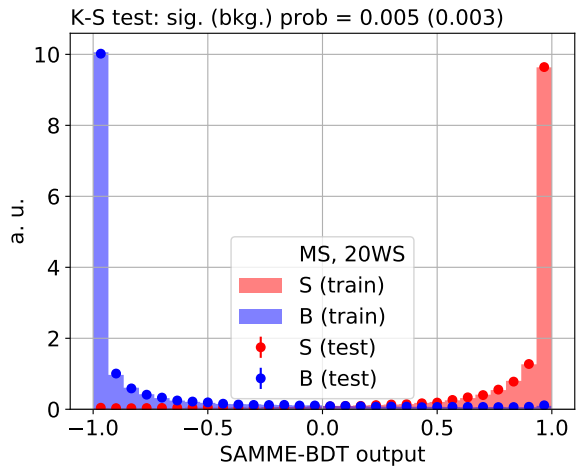


(c) XG-BDT, MS, 20SF, NMCW

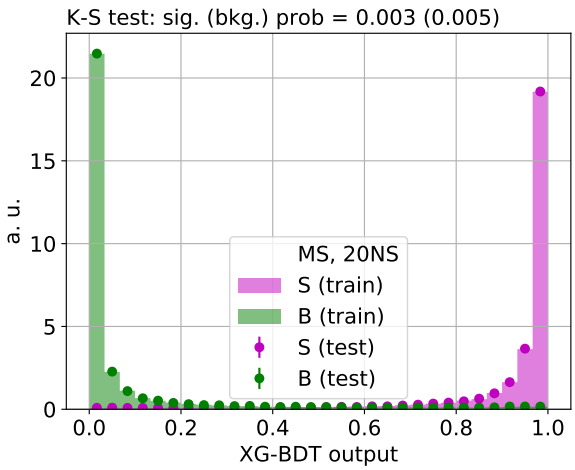


(d) XG-BDT, WS, 20SF, WMCW

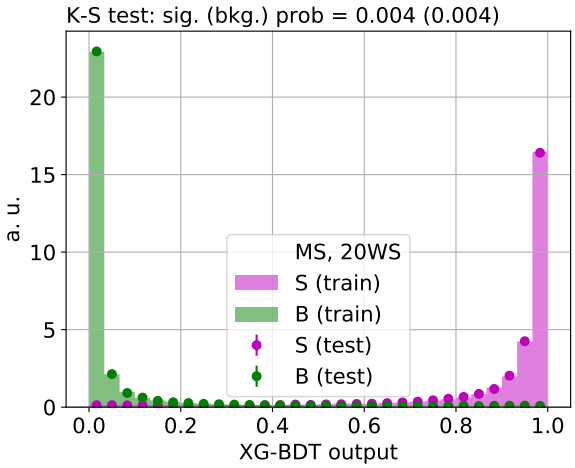
Figure 52: Comparison of overtraining obtained for the different combinations listed in Table 21 using electrons data sets.



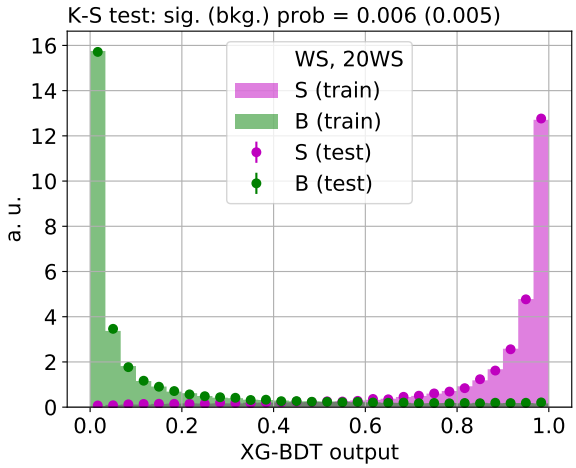
(a) SAMME-BDT, MS, 20SF, WMCW



(b) XG-BDT, MS, N20SF, WMCW



(c) XG-BDT, MS, 20SF, NMCW



(d) XG-BDT, WS, 20SF, WMCW

Figure 53: Comparison of overtraining obtained for the different combinations listed in Table 21 using muons data sets.